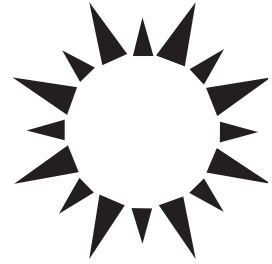


Ecological Footprints and Energy

MATHIS WACKERNAGEL and CHAD MONFREDA

Redefining Progress

Oakland, California, United States



-
1. Ecological Overshoot
 2. Conceptual Predecessors of the Ecological Footprint
 3. Component- and Compound-Based Approaches
 4. Method of National Ecological Footprint Accounting
 5. Bioproductive Areas of the Ecological Footprint
 6. Fossil Energy
 7. Nuclear Energy
 8. Renewable Energy
 9. Energy Footprint of Preindustrial Society
 10. Energy Footprint of Industrial Society

Glossary

biocapacity The potential productivity of the biologically productive space within a specified country, region, or territory.

biologically productive space Areas of land and water capable of supporting photosynthesis at sufficient rates to provide economically useful concentrations of biomass. Marginal and unproductive regions, such as deserts, tundra, and the deep oceans, are excluded. The global biologically productive space totals 11.4 billion ha.

ecological deficit The amount by which the ecological footprint of a population (e.g., a country or region) exceeds the biocapacity of the space available to that population.

ecological footprint A measure of how much biocapacity a population, organization, or process requires to produce its resources and absorb its waste using prevailing technology.

ecological overshoot Ecological deficit met through the overexploitation of resources or accumulation of waste. William Catton defines the term as “growth beyond an area’s carrying capacity, leading to crash.”

embodied energy The energy used during a product’s entire life cycle for manufacture, transport, operation, and/or disposal.

global hectare One hectare of biologically productive space with world-average productivity.

productivity A measurement of the potential rate of biological production of a given area. Various indicators

can be used to weight productivity, including agricultural suitability, potential net primary production, and useful biomass extraction.

The ecological footprint is a measurement that compares rates of human resource consumption and waste generation with the biosphere’s rates of resource regeneration and waste assimilation, expressed in terms of the area necessary to maintain these flows. Ecological footprints represent the biologically productive space required to produce the resources and absorb the wastes of a given population, organization, or product using prevailing management and technology.

1. ECOLOGICAL OVERSHOOT

The nature of ecological limits may be one of the most influential misconceptions of the sustainability debate. A former executive director of the United Nations (UN) Population Fund stated that “many environmentalists think [that the carrying capacity of Earth] is 4 billion, maximum. But now we have 6 billion people.” The continued growth of the human population and resource consumption, however, is only an apparent contradiction. In reality, ecological limits can be exceeded for a period of time because nature reacts with inertia. More precisely, natural capital can be harvested faster than it regenerates, thereby depleting the natural capital stock. Ecologists call this state overshoot.

The concept of overshoot reconciles the apparently contradictory assertions that standards of living are waxing while ecological capacity is waning. Nevertheless, uncertainties remain regarding the possible impacts of prolonged overshoot. Some fear that unsustainable harvests will trigger the sudden collapse of ecosystems, leaving fisheries and other dynamic resource stocks irreversibly damaged.

Others hope that the contraction of ecosystem services will be slower and more forgiving.

The delayed effects of overshoot and difficulty of valuing ecosystem services prevent markets and uninformed policy from forming corrective feedback mechanisms. Without adequate feedback mechanisms, such as systematic accounting that compares human demands on nature with nature's capacity to regenerate, overshoot can go unnoticed, thereby undermining humanity's future. Industrial economies in particular can exceed "sustainable limits" by masking resource depletion and decreasing energy returns on investment with technological advances, inexpensive energy, and easy access to distant resources. Particularly on the waste side, feedback is weak: CO₂ from fossil fuel burning leaves tailpipes with ease, independent of the increasing CO₂ concentration in the atmosphere.

No national government or UN agency operates a comprehensive accounting system to document the extent to which human demands fit within the capacity of existing ecosystems. The ecological footprint is one of the few tools that attempts integrated resource accounting.

The purpose of ecological footprints is not only to illustrate the possibility of overshoot but also to offer a robust tool to demonstrate its occurrence. As a resource accounting tool, it can help determine who is contributing, and how much, to the overall impact of humanity. Most important, it can evaluate potential strategies for avoiding overshoot.

2. CONCEPTUAL PREDECESSORS OF THE ECOLOGICAL FOOTPRINT

Biophysical assessments of human dependence on nature date back thousands of years to tales about the relationship between people and land. Records on the concept of carrying capacity trace at least as far back as Plato's "Laws," Book V, which states that a "suitable total for the number of citizens cannot be fixed without considering the land and the neighboring states. The land must be extensive enough to support a given number of people in modest comfort, and not a foot more is needed."

The first scholarly book in English on sustainable resource use may have been John Evelyn's 1664 "Sylva: A Discourse of Forest Trees and the Propagation of Timber." In North America, George Perkins Marsh's widely read 1864 study, "Man and Nature," was influential in increasing the awareness

of nature's limited capacity to provide for human demands. Formal ecological accounting dates to 1758, when François Quesnay published "Tableau Economique," which captures the relationship between land productivity and wealth creation. Intellectual groundwork for more recent studies was laid in the 1960s and 1970s, particularly through George Borgstrom's concept of "ghost acreage," Howard Odum's energy analyses, and Jay Forrester's model on world resource dynamics presented in Donella Meadows *et al.*'s "Limits to Growth" report to the Club of Rome in 1972.

Natural capital accounts focused on regional and global limits are increasingly prominent in the political debate. Peter Vitousek's 1986 landmark study, "Human Appropriation of the Products of Photosynthesis," compiled consumption data to determine the human economy's draw on terrestrial net primary productivity. Based on 1980s data, the study estimates that humanity co-opts 40% of the photosynthetic materials produced by terrestrial ecosystems, excluding the photosynthesis that assimilates fossil fuel emissions and other anthropogenic waste. Fischer-Kowalski and Hüttler used material flow analysis to advance measurements of societal metabolism, a macroindicator of the environmental performance of societies. The World Resources Institute's biennial "World Resources" report and the United Nations Environment Programme's (UNEP) "Global Environmental Outlook" provide detailed documentation of pressures facing global ecosystems but do not contain aggregated impact assessments.

3. COMPONENT- AND COMPOUND-BASED APPROACHES

There are two basic ecological footprinting methods: component- and compound-based. The first calculates the footprint of individual goods or organizations. This bottom-up, component-based method first identifies all relevant items and then assesses the ecological footprint of each item using life-cycle data. The overall accuracy of the final result depends on the completeness of the component list as well as the reliability of the life-cycle assessment (LCA) of each identified component. This approach provides a high level of detail but has limitations due to LCAs' frequent inaccuracies, incompleteness, and poorly defined boundaries. The component-based method also has difficulty with data gaps and indirect expenditures outside of an organization's inventories, such as public infrastructure and military.

The compound-based method uses a top-down approach to aggregate economic and environmental statistics into national ecological footprints. Compound footprinting, which underlies the national results presented in this article, calculates the ecological footprint using aggregate national data. This information is more complete than subnational sources and LCAs and captures both direct and indirect consumption. For instance, to calculate the paper footprint of a country, information about the total amount consumed is typically available and sufficient for the task. In contrast to the component method, there is no need to know which portions of the overall paper consumption were used for which purposes, aspects that are poorly documented in statistical data collections.

4. METHOD OF NATIONAL ECOLOGICAL FOOTPRINT ACCOUNTING

National ecological footprint accounts consist of two measurements. The footprint aggregates the total area a given population, organization, or process requires to produce food, fiber, and timber; provide space for infrastructure; and sustain energy consumption. The biocapacity aggregates the total area available to supply these demands.

The ecological footprint focuses on six potentially renewable demands: cropland, pasture, forests, built-up land, fisheries, and energy. Activities in these components are deemed sustainable if rates of resource use and waste generation do not exceed a defined limit that the biosphere can support without degrading the resource stock. By including a core set of potentially sustainable activities, the ecological footprint defines minimum conditions for sustainability.

The results underestimate human impact and overestimate the available biocapacity by counting each footprint component only once if multiple demands come from the same area; choosing the more conservative footprint estimates and more optimistic biocapacity estimates when in doubt; and including agricultural practices as if they would not cause long-term damage to soil productivity. The accounts also exclude activities for which there is insufficient data or understanding, such as acid deposition. Activities that systematically erode nature's capacity to regenerate are also excluded, including aquifer depletion, soil erosion, and the release of chlorofluorocarbons.

In order to provide a quantitative answer to the research question of how much regenerative capacity is required to maintain a given resource flow, ecological footprint accounts use a methodology grounded on six assumptions:

1. It is possible to use annual national statistics to track resource consumption and waste generation for most countries.

2. Resource flows can be measured in terms of the bioproductive area necessary for their regeneration and the assimilation of their waste. (Resource and waste flows that cannot be measured are excluded from the assessment.)

3. Bioproductive areas of different productivity can be expressed in a common unit of standardized usable biological productivity. Usable refers to the portion of biomass used by humans, reflecting the anthropocentric assumptions of the footprint measurement.

4. The sum of mutually exclusive areas needed to maintain resource flows expressed in a common unit represents aggregate demand; the sum of mutually exclusive bioproductive areas expressed in a common unit represents aggregate supply.

5. Human demand (footprint) and nature's supply (biocapacity) are directly comparable.

6. Area demand can exceed area supply, meaning that activities can stress natural capital beyond its regenerative capacity. For example, the products from a forest harvested at twice its regeneration rate have a footprint twice the size of the forest. A footprint greater than biocapacity indicates ecological deficit. Ecological deficits are compensated in two ways: Either the deficit is balanced through imports (ecological trade deficit) or the deficit is met through overuse of domestic resources, leading to natural capital depletion (ecological overshoot).

Cropland, forests, pastures, and fisheries vary in biological productivity, or their capacity to provide ecological goods and services through photosynthesis. One hectare of arid rangeland, for example, has less capacity to recycle nutrients, produce food, and support diverse land-use patterns than one hectare of temperate forest. The ecological footprint therefore normalizes each bioproductive area—cropland, pasture, forests, built-up land, and fisheries—into common units of “global hectares” (gha).

Current accounts weight productivity according to agricultural suitability—a function of numerous factors, including temperature, precipitation, soils, and slope. The Food and Agriculture Organization and International Institute for Applied Systems

TABLE I

Global Biocapacity^a

Area	Equivalence factor (gha/ha)	Global area (billion ha)	Biocapacity (billion gha)
Cropland	2.1	1.5	3.2
Pasture	0.5	3.5	1.6
Forest	1.3	3.8	5.2
Built-up land	2.2	0.3	0.6
Fisheries	0.4	2.3	0.8
Total	1.0	11.4	11.4

^aThe relative productivities of major global bioproductive areas, expressed in actual area (hectares) and productivity-weighted area (global-hectares). Equivalence factors are based on the suitability index of Global Agro-Ecological Zones 2000 (FAO/IIASA). Data are rounded.

Analysis have created a spatially explicit distribution of these variables in the “suitability index” of Global Agro-Ecological Zones 2000. Recent ecological footprint accounts use this index to translate unweighted hectares into global hectares (Table I). Other possibilities for weighting productivity include potential primary production and the rate of economically useful biomass production.

5. BIOPRODUCTIVE AREAS OF THE ECOLOGICAL FOOTPRINT

Earth has a surface area of 51.0 billion ha, of which 36.6 billion ha is ocean and inland waters and 14.4 billion is land. Ecological footprint accounts estimate that less than one-fourth of Earth’s surface—9.1 billion ha of land and 2.3 billion ha of water—provides sufficient concentrations of economically useful biomass to be considered biologically productive. The 11.4 billion ha of bioproductive area is differentiated into five components: cropland, pastures, forests, built-up land, and fisheries. The remaining 39.6 billion ha is marginally productive or unproductive for human use because it consists of the deep oceans, is covered by ice, or lacks fertile soils and fresh water.

Cropland occupies approximately 1.5 billion ha of the most productive land worldwide. Current accounts include 74 crops and 15 secondary products and report details on how food crops are used (waste, seeds, feed, food, and processing), traded, and stored. These numbers document the area dedicated to cultivation and omit impacts from

unsustainable agricultural practices, such as long-term damage from topsoil erosion, salinization, and contamination of aquifers with agrochemicals.

Pastures, which cover 3.5 billion ha, supply a substantial part, perhaps half, of the metabolic needs of livestock worldwide. Differing definitions of what constitutes pasture, poorly documented temporal and spatial variations in forage production and quality, and difficulties in separating metabolic needs supplied through concentrate feeds, crop residues, and other nonpasture sources make the pasture footprint difficult to determine. Healthy pastures are a critical component of food security, and even a preliminary assessment of their global capacity to support livestock would improve the current dearth of data.

Forests provide lumber, pulp, and fuelwood and include both natural forests and plantations. The Food and Agriculture Organization’s “Forest Resource Assessment 2000” places the total area at 3.8 billion ha. Estimates of timber productivity are derived from the UN’s Economic Commission for Europe’s “Temperate and Boreal Forest Resource Assessment 2000” and data from the Intergovernmental Panel on Climate Change. These sources also provide information on plantation type, coverage, timber yield, and areas of protected and economically inaccessible forest.

Fisheries are concentrated on a relatively narrow portion of the ocean located on continental shelves. Excluding inaccessible or unproductive waters, these include an estimated 2.0 billion ha. Although a mere fraction of the ocean’s 36.3 billion ha, this 2.0 billion ha provides 95% of the marine fish catch. Inland waters comprise an additional 0.3 billion ha.

Higher trophic-level fish—species high on the food chain—consume a proportionally greater portion of the ocean’s primary productivity than lower trophic-level fish. It is estimated that each level up in the food chain translates into a 10-fold increase in demand for photosynthetic production per unit of biomass. Thus, 1 ton of cod at trophic level 4 has a footprint 10 times larger than 1 ton of sardines at trophic level 3. (Trophic level 1 corresponds to photosynthetic species at the base of the food chain—the primary producers.)

Built-up land for housing, transportation, industrial production, and capturing hydroelectric power occupies approximately 0.3 billion ha. Built-up land is poorly documented at a global level, partially due to different definitions used in various studies and the difficulty of satellites to view smaller objects and linear developments such as roads. However, some

countries, such as those in the European Union, have more reliable data sets. Current accounts assume that built-up land replaces cropland because most human settlements favor fertile regions well suited for agriculture. Hence, urban expansion encroaches typically on cropland.

6. FOSSIL ENERGY

The footprint of fossil fuels can be assessed by a number of approaches. The question of how much regenerative capacity is required to maintain the throughput of fossil fuel through the human economy can be addressed from three perspectives: absorbing the waste, maintaining the energy supply, and maintaining the resource. While locally significant, the area required for extraction, solid waste disposal, and power production is comparatively small at 3.2 ha/MW. In contrast, the waste assimilation method places the footprint for carbon emissions from the operation, construction, and maintenance of a typical coal-fired facility at 1900 gha/MW and the footprint of a similar-sized natural gas plant at 1050 gha/MW.

6.1 Waste Assimilation Method

The most common method for determining the ecological footprint of fossil fuels is based on the assumption that the primary limitation for fossil fuel use is the biosphere's ability to absorb CO₂ waste emissions. Following the dictum of the 1992 UN Framework Convention on Climate Change that calls for CO₂ stabilization in the atmosphere, the waste assimilation approach estimates how much larger the biosphere would need to be to provide the absorptive services to remove excess carbon emissions. Most accounts calculate this area by estimating the additional forest area required to fully sequester CO₂ emitted from the burning of fossil fuels.

Of the average annual 6.3 Gt of anthropogenic carbon emitted from fossil fuel combustion and cement production during the 1990s, the ocean absorbed an estimated 1.7 Gt and the land another 1.4 Gt. This imbalance has led to an increase in the atmospheric concentration of CO₂ from a preindustrial level of 280 parts per million (ppm) in 1750 to 370 ppm at the end of the 20th century. Given that carbon has a much shorter residence time in the terrestrial sink than in the oceanic sink, reflects prior biomass removal, and is susceptible to re-release from land-use and climate changes, current accounts

only deduct the emissions sequestered in the ocean from the footprint.

Although possibilities for carbon sequestration include the fertilization of oceans and the pumping of carbon emissions underground, the reversion to forests from human-altered landscapes—cropland, pastures, and settlements—is used to calculate the footprint because it best reflects the biosphere's sequestration potential without human management. Reforestation measures have been implemented on a small scale and actively debated as a legitimate source of carbon credits in international climate negotiations. However, this approach does not suggest that the problem of increasing CO₂ concentration in the atmosphere can be solved through sequestration alone. Rather, it shows that the biosphere does not provide sufficient biological capacity for sequestration, and that it is highly unlikely that the CO₂ concentration in the atmosphere can be stabilized without reducing carbon emissions.

According to this method, the ecological footprint of fossil fuels equals the area that would be needed to assimilate carbon emissions through reforestation. Although sequestration rates vary with the temporal and spatial scales involved, and global average rates cannot be determined with great precision, sequestration would decrease significantly after 30–50 years as replanted forests mature. Although some studies have found that mature forests have retained their sink capacity, others have found that they lose their capacity or even become net carbon sources. Using the waste assimilation approach, the ecological footprint of fossil fuels provides a conservative estimate of the bioproductive area needed to assimilate carbon emissions. Assuming a generous sequestration rate of 0.95 tons C/ha/year and deducting the oceanic sink, a reforested area of more than 6 billion gha would be needed to balance current emissions. The total bioproductive land mass amounts to only 9.1 billion ha, equivalent to 10.6 billion gha.

6.2 Renewable Substitution Method

This approach builds on the idea of maintaining the energy stock. This means that depleted fossil fuel would be compensated with equivalent amounts of energy stocks. The most common long-term energy stock the biosphere accumulates is in wood. Hence, this method for calculating the ecological footprint of fossil fuels, and nonrenewable fuels in general, estimates the area required to supply their energy equivalent in fuelwood.

The ecological footprint of fossil fuels thus becomes the area needed to renew the energy equivalent of fossil fuel combustion. Unlike large-scale hydropower or geothermal energy, the energy stored in fuelwood is available in unmanaged systems and has been the mainstay of most pre-industrial societies (although many have managed fuelwood stocks). The ecological footprint expressed in fuelwood compares societies' nonrenewable energy demands with the biosphere's unmanaged renewable supply. On average, forests produce approximately 40 GJ of fuelwood/gha/year. Harvesting young, quickly growing trees may increase this capacity to 60 GJ/gha/year. Much of the energy captured by forests through net primary production is lost to decomposition and herbivory, and a portion of the remaining biomass energy, or net ecosystem productivity, is in foliage, roots, or other sources less suitable as fuel.

Variations of this technique describe the area needed to derive the energy equivalent of fossil carriers not in fuelwood but in cultivated biomass such as switchgrass, ethanol, or methanol derived from agriculture or forestry. Converting biomass fuels into liquid and gaseous mediums for use in prevailing infrastructure would net less energy than the total captured in unconverted biomass.

Andrew Ferguson of Optimum Population Trust promotes a slightly different approach by estimating an appropriate mix of renewable energy sources—photovoltaic arrays, offshore and onshore wind turbines, biomass-based ethanol, etc. The composition of the mix reflects the limitations of wind power due to its irregular availability and the cost barriers of photovoltaic arrays. He estimates that this mix could generate the current energy demand at approximately 95 GJ/ha/year. This would most likely occupy above-average-quality land, indicating an energy yield less than 95 GJ/gha/year.

6.3 Fossil Carrier Regeneration

This method treats fossil fuels like a renewable resource. The waste assimilation and renewable substitution approaches give the areas needed to balance emissions and energy stocks, not the area required to renew fossil fuels per se. This approach considers fossil fuels as a very slowly renewable resource limited by sedimentation and accumulation rates, and it defines their use as sustainable only when the fossil fuel stocks in the lithosphere are neither diminished nor transferred to the atmosphere (Table II). This method provides an assessment of the

TABLE II

Regeneration Rates of Fossil Energy^a

Fossil energy carriers	Yield (MJ/m ² /year)
Coal	0.05
Fuels	0.08
Natural gas	0.01

^a Annual accumulation rates of fossil energy carriers based on estimated rates of carbon sedimentation. The data are presented in unweighted hectares. The ecological footprint typically weighs hectares by their productivity. From Ströglehner (2003).

area needed to continue fossil fuel consumption indefinitely. With a global footprint of 100 gha/capita—50 times greater than global biocapacity and 100 times greater than the footprint from carbon sequestration and fuelwood substitution—it also demonstrates that the continued disparity between rates of fossil fuel consumption and creation will lead to the stock's depletion.

7. NUCLEAR ENERGY

As with fossil fuels, the area occupied through extraction, production, and waste disposal of nuclear energy is relatively minor, amounting to 0.4 ha/MW. The embodied energy of nuclear infrastructure is small relative to the energy output because of the high energy density of nuclear fuel. By itself, nuclear fuel makes relatively few demands on biological productivity when contained, but intentional and accidental releases of radioactive materials have compromised human and environmental health. Failures of nuclear power plants can appropriate large bioproductive areas by making them unsuitable for human use for extended periods of time. The meltdown of Chernobyl has completely removed a 170,000-ha “zone of alienation” from economic turnover and restricted activities on hundreds of thousands of additional hectares since the 1986 accident and possibly for thousands of years into the future. Fallout contaminated larger areas extending from nearby Ukraine, Belarus, and southwestern Russia to distant Poland, Norway, and Sweden. The actual biocapacity occupied by nuclear energy is uncertain and speculative. Some studies treat the nuclear footprint as negligible or subsume these demands within national built-up areas and fossil fuel expenditures. Most nuclear footprints use the footprint of the fossil fuel energy equivalent.

8. RENEWABLE ENERGY

Some studies have examined the ecological footprint of renewable energy sources but more work is needed (Table III). This inattention is partly due to the fact that renewable sources provide an insignificant fraction of all commercial energy. Moreover, rapidly developing technologies make generalizations about conversion efficiencies, embodied energies, and life spans difficult. Although the energy embodied within the fossil fuel infrastructure is miniscule when compared with the energy expended in the fuels' extraction, transportation, and combustion, the energy embodied in most renewables' infrastructure dominates an essentially free energy source.

8.1 Hydropower

Hydropower is the largest renewable source of commercial energy. In 1995, it provided 7% of all commercial energy and 20% of electricity globally. The ecological footprint of hydropower is based on the area inundated by the reservoir created by the

TABLE III

Energy Yield of Renewable Sources^a

Renewable sources	Yield (MJ/m ² /year)
Electricity, photovoltaic	450
Electricity, wind	900
Electricity, water power	100
Solar heating	1200
Wood extensive	7
Wood intensive	190
Straw as by-product	4
Miscanthus	25
Rape oil	6
Methyl ester of rapeseed	5
Organic ethanol from sugarcane	9
Organic ethanol from sugar beet	8
Organic ethanol from wood	5
Organic ethanol from wheat	5
Organic ethanol from maize	2
Organic methanol from wood	12
Organic biogas, digester gas, and landfill gas	5

^aThe data are presented in unadjusted hectares that do not reflect differences in biological productivity. The ecological footprint weighs hectares by their productivity. These data may differ from other estimates depending on technology or other factors. For instance, sources indicate slightly higher ethanol yields in Brazil of 12 MJ/m²/year. From Stöglehner (2003).

dam, the energy embodied in the dam, and the biological productivity lost because of the dam, primarily through depleted fish stocks. Sergio Pacca and Arpad Horvath assessed the carbon emissions released in the construction, maintenance, and operation of various electricity facilities. Basing the hydropower analysis on the Glen Canyon Dam, which generated 5.55 TWh of electricity in 1999, they determined these processes would release 25,000 tons of CO₂ per year averaged over a 20-year life cycle. The associated carbon footprint is 10.5 gha/MW.

The ratio of energy output to inundated area and the productivity of inundated land are highly variable. Energy output fluctuates with power demand and reservoir levels, although typical outputs for hydropower facilities in the United States are 40% of maximum capacity. A random survey of 24 hydropower dams in the Pacific Northwest of the United States found that the inundated areas produce an average of 3200 GJ/ha/year (10 ha/MW), varying widely between 15 GJ/ha/year (lower-course dams) and 11,000 GJ/ha/year (upper-course dams in mountain areas). A similar survey of 28 dams in 17 countries suggests an annual production of 900 GJ/ha/year (35 ha/MW), ranging from a minimum of 12 GJ/ha/year for Ghana's 833-MW Akosombo Dam to a maximum of 15,300 GJ/ha/year for Chile's 500-MW Pehuenche Dam. These values are highly variable and underestimate energy output per hectare by including the original river area as part of the inundated area. Hydroelectric plants located far from populated centers have long-distance transmission inefficiencies that increase the footprint of useful electricity to reach homes and commercial facilities.

8.2 Wind Power

The most attractive sites for wind power are often offshore or in remote regions, potentially limiting demands for biocapacity. Moreover, the actual area occupied by turbines and other infrastructure is only 5% of the total area of wind farms, leaving the remainder available for pasture, fisheries, or other dual uses. At an average output of 8200 GJ/ha/year, the built-up footprint of a wind power plant on average-quality pasture is 1.8 gha/MW.

The energy embodied in the extensive infrastructure needed to capture a diffuse and intermittent energy source forms a large part of the footprint of wind power. Pacca and Horvath estimate that the CO₂ embodied in the construction and materials of a wind farm capable of generating 5.55 TWh/year

would total 800,000 tons over 20 years—approximately 10 times greater than the embodied emissions of a similarly sized coal-powered facility (excluding operational emissions). Assuming a 20-year life span, the carbon sequestration footprint is 16.9 gha/MW, one order of magnitude greater than the built-up area of the wind farm. If wind or another renewable carrier supplied the construction energy, the footprint of embodied energy would be greatly reduced.

8.3 Photovoltaic

Like other renewable energy technologies, photovoltaic (PV) arrays contend with low power densities. Nevertheless, annual energy output can be impressive. Assuming an annual power density of 15 W/m² typical in commercial systems and allowing additional space for shading and access within large-scale installations, PV systems occupy 10 ha/MW. One hectare of PV panels can deliver the thermal energy contained in 100 tons of oil annually. Without including the excess capacity required in a working infrastructure, PV arrays covering a square 160 km on each side could supply the electricity demand of the United States. With a population of 280 million people, this amounts to 0.01 ha/capita. The global area amounts to .002 ha/capita, covering a square 310 km on each side. The complete replacement of fossil fuels with a hydrogen fuel cell-based economy may require an additional area manyfold larger. PV installations atop buildings or in arid regions can mitigate competition for bioproductive areas, an advantage shared with wind power. Also, although potential PV sites may be more widespread than those for wind power, the embodied energy is also greater. Pacca and Horvath estimate the CO₂ emissions embodied in a PV system producing 5.55 TWh/year to be 10 million tons over 20 years, or 211 gha/MW.

9. ENERGY FOOTPRINT OF PREINDUSTRIAL SOCIETY

Historically, renewable energy captured chemically through photosynthesis drove nearly every aspect of society. Renewable kinetic sources—wind and water—held much less global prominence, although they expanded rapidly in the 18th and 19th centuries, when wind mills spurred the colonization of the American frontier and waterwheels became increasingly important in industrialization. In most

areas, however, fuelwood supplied the majority of heat energy for small-scale activities, such as domestic heating and cooking, and larger scale enterprises such as metal smelting.

In many areas, fuelwood limitations were one factor restricting the size of preindustrial cities. According to Vaclav Smil, 12th-century Paris, then a large city of 100,000 people, may have required 1.5 tons of fuelwood per person annually, amounting to 3 PJ/year. Even if half of this demand was met through wood grown on cropland or other nonforest sources, as in some low-income countries today, the fuelwood footprint would have approached 0.4 gha/capita. The forest area required for energy demands alone equaled 50–200 times the urban area. Overland transport of fuelwood was prohibitively expensive at distances of more than 30 km, and cities with river access drew resources from great distances. Preindustrial Beijing, for example, imported huge rafts of wood from as far as Szechwan province more than 1000 km away. Fuelwood limits halted economic activities in less fortunate cities, where villages and industry competed for energy sources. According to one estimate, in 17th-century Europe fuelwood shortages halted production every other year in a typical ironworks furnace, which in 1 year of operation consumed 4000 ha of forest. Assuming a typical European growth rate of 5 tons/ha/year, this would equal a footprint of approximately 10,000 gha. Fuelwood demands increased severalfold during industrialization. By the mid-19th century, the United States fuelwood demand was 5 tons/capita, with a possible fuelwood footprint of 2.5 gha/capita—a national total of 60 million gha.

Crop residues and other cultivated biomass have played a significant but limited role in providing heat since their value as food and fertilizer introduced competing demands. David Pimentel estimates that currently fuelwood supplies half of all the world's biomass fuel by weight, whereas crop residues supply one-third. The higher energy density of fuelwood, however, means that the actual contribution of crop residues to biomass energy is smaller. More important, the chemical energy of cultivated biomass fed people and livestock, the predominant source of mechanical energy for transportation and labor. Powerful draft horses, for example, provided the labor of 10 men but could reach peak performance only when well nourished. Such an animal would eat the daily equivalent of 5 kg of oats, matching the food needs of 6 people. Nevertheless, the mechanical energy available in draft animals retained its advantage well into industrialization, and in the

United States, the demand for horses reached its height in the second decade of the 20th century. At an average yield of just over 1100 kg of oats/ha/year, a working horse may have had a cropland footprint of 3.5 gha. Some estimates place horse feed as constituting 20% of cultivated cropland in the United States at that time—amounting to approximately 60 million gha or 0.55 gha/capita.

The ecological footprint of energy obtained through fuelwood and agriculture faced local limits. Transgression of these constraints leads to an ecological deficit, a state in which ecological demands exceed local capacity. An ecological deficit signifies either imports of ecological capacity from abroad or ecological overshoot of local resources. Although ecological imports may contribute to overshoot abroad, local overshoot has immediate domestic implications. In the case of renaissance England, an escalation in charcoal demands for metal smelting led to widespread deforestation. Abundant coal supplies yielded a ready charcoal replacement and weaned England from fuelwood by the late 17th century, ushering in an early industrialization and the dawn of the first fossil fuel-dependent society.

10. ENERGY FOOTPRINT OF INDUSTRIAL SOCIETY

Tapping fossil fuel reservoirs accumulated over hundreds of millions of years exceeds local rates of fossil fuel formation, leading to local overshoot. The widespread transition to fossil fuels, however, has resulted in global overshoot and a qualitative shift in humanity's relation to the biosphere (Table IV; Fig. 1). Despite radical local and regional ecological changes from preindustrial societies, the human economy only surpassed global limits when fossil fuel use exceeded ecologically sustainable rates and induced the expansion and intensification of extractive industries such as agriculture and forestry. Although local overshoot portends immediate domestic resource constraints, global overshoot from fossil fuels externalizes costs, diffusing the feedback needed to adjust behavior.

Renewable energy sources bounded the growth of preindustrial economies, but the switch to fossil fuels curtails photosynthetic limitations. Energy production from biomass equaled that from coal and oil at the end of the 19th century. Fifty years later, however, biomass energy amounted to only one-fourth of global energy production, and its contribu-

TABLE IV

Global Ecological Footprint and Biocapacity, 1999

Area	Footprint (gha/cap)	Biocapacity (gha/cap)
Cropland	0.5	0.5
Pasture	0.1	0.3
Forest	0.3	0.9
Fisheries	0.1	0.1
Built-up land	0.1	0.1
Fossil fuels and nuclear energy	1.2	0.0
Total	2.3	1.9

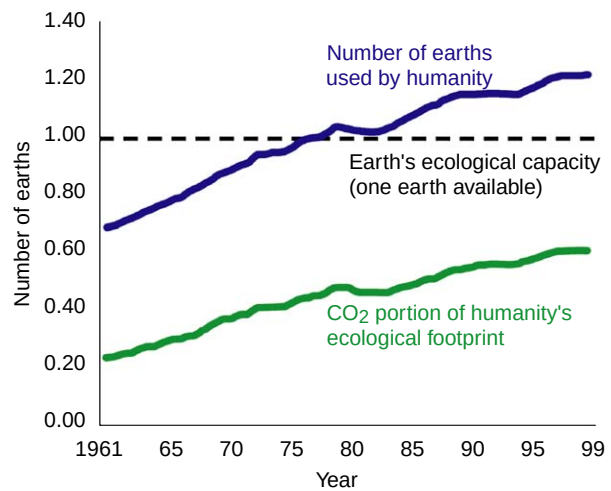


FIGURE 1 Global ecological footprint versus biocapacity, 1961–1999. Human demand exceeded the biosphere's supply during this time period, most clearly manifest in increasing CO₂ concentrations and overexploited fisheries.

tion to useful work was far smaller. Meanwhile, oil has increased in global prominence to become the largest primary energy source (Table V). The magnitude of this transformation has been immense. At more than 320 EJ/year, the global production of fossil fuels, for example, dwarfs the energy captured in food production by an order of magnitude.

With the advent of industrial civilization, non-renewable sources became the prime economic drivers, supplanting not only fuelwood for heat but also livestock for transportation and labor. Fertile areas freed by the shift from biomass energy to fossil energy, decreasing transportation costs, global markets, and the diffusion of burdens from fossil fuel combustion helped spark an enormous expansion,

TABLE V

Global Ecological Footprint of Energy, 1999

Energy source	Footprint (gha/cap)
Oil	0.5
Coal	0.4
Natural gas	0.2
Nuclear	0.1
Fuelwood	0.05
Hydro	0.002
Total	1.2

Data are rounded.

beginning with the human population in the mid-19th century and soon followed by the global economy. Today, the footprint of timber products in high-income countries is 0.7 gha/capita. Mechanized forestry boosted labor productivity with modern mills, harvesters, and chainsaws—Canadian timber extraction, for example, was 3.1 gha per Canadian in 1999. In contrast, the biocapacity of global forests approaches 0.9 gha/capita, of which only 0.4 gha/capita is exploitable. The fossil fuel transition also enables industrial fishing fleets to harvest previously inaccessible stocks, raising the footprint of the global marine catch to its estimated biocapacity of 0.1 gha/capita. The footprint of marine fish consumption averages at least triple this rate in high-income countries.

The mechanization of agriculture relieved the 20% of cropland used to feed work horses in the early 20th-century United States. The freed area became a source of meat and milk production, and in 1999 meat- and milk-producing livestock consumed 60% of U.S. maize. At the same time, the intensification of agriculture, along with new breeding techniques, initiated dramatic increases in crop yields and the human population. A substantial portion of the food energy consumed by people relies on fertilizers and agricultural chemicals synthesized through fossil fuels.

Fossil fuels have created a transitory carrying capacity of significant proportions. The ecological footprint accounts make this “phantom planet” visible. As the coiner of this term, William Catton, points out, this extra capacity gives humanity the illusion of living on a larger planet and encourages human expansion beyond the regenerative capacity of the biosphere. Consider the following comparison: 0.7 TW of human food produced today occupies more

than 40% of the biosphere’s regenerative capacity; in comparison, fossil fuel consumption exceeds 10 TW. From this perspective, the fossil fuel provides humanity with a temporary income that corresponds to the capacity of more than five biospheres ($0.4\% \text{ biospheres} \times 10 \text{ TW} / 0.7 \text{ TW} = 5.7 \text{ biospheres}$). Therefore, some biologists liken humanity’s aggregate behavior to that of species who come in contact with a large but finite food stock. Examples are deer introduced to an island or yeast cells placed in a cup full of sugar. Such an abundance of food typically leads to population growth beyond the habitat’s carrying capacity, resulting in a population crash. So far, the growth curves of the fossil fuel economy have resembled the yeast cells in a finite cup of sugar. The difference is that people have choice. Like financial budgeting, ecological footprint accounts help society choose the most desirable allocation of the limited budget nature provides. The alternative is to continue liquidating natural capital.

SEE ALSO THE FOLLOWING ARTICLES

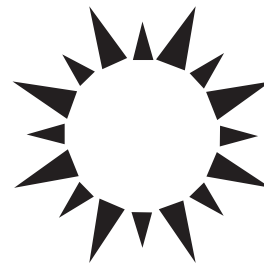
Aquaculture and Energy Use • Complex Systems and Energy • Earth’s Energy Balance • Easter Island: Resource Depletion and Collapse • Ecological Risk Assessment Applied to Energy Development • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Modeling Energy Supply and Demand: A Comparison of Approaches • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

- Costanza, R. (2000). The dynamics of the ecological footprint concept. *Ecol. Econ.* 32(3), 341–346.
- Ferguson, A. (1999). The logical foundations of ecological footprints. *Environ. Resource Econ.* 1, 149–156.
- Ferguson, A. (2002). The assumptions underlying eco-footprinting. *Population Environ.* 23, 303–313.
- Hoffert, M., Caldeira, K., Benford, G., Criswell, D., Green, C., Herzog, H., Jain, A., Kheshgi, H., Lackner, K., Lewis, J., Lightfoot, H., Manheimer, W., Mankins, J., Mauel, M., Perkins, L., Schlesinger, M., Volk, T., and Wigley, T. (2002). Advanced technology paths to global climate stability: Energy for a greenhouse planet. *Science* 298, 981–987.
- Loh, J. (ed.) (2002). “Living Planet Report 2002.” Worldwide Fund for Nature. Gland, Switzerland.
- Moffatt, I. (2000). Ecological footprints and sustainable development. *Ecol. Econ.* 32(3), 359–362.
- Pacca, S., and Horvath, A. (2002). Greenhouse gas emissions from building and operating electric power plants in the

- Upper Colorado River Basin. *Environ. Sci. Technol.* **36**, 3194–3200.
- Rees, W. (2000). Eco-footprint analysis: Merits and brickbats. *Ecol. Econ.* **32**, 371–374.
- Simmons, C., Lewis, K., and Moore, J. (2000). Two feet–two approaches: A component-based model of ecological footprinting. *Ecol. Econ.* **32**(3), 375–380.
- Stöglehner, G. (2003). Ecological footprint—A tool for assessing sustainable energy supplies. *J. Cleaner Production* **11**, 267–277.
- Wackernagel, M., Monfreda, C., and Deumling, D. (2002a). The ecological footprints of nations: How much nature do they use? How much nature do they have, Sustainability Issue Brief. Redefining Progress, Oakland, CA.
- Wackernagel, M., Schulz, N., Deumling, D., Callejas Linares, A., Jenkins, M., Kapos, V., Monfreda, C., Loh, J., Myers, N., Norgaard, R., and Randers, J. (2002b). Tracking the ecological overshoot of the human economy. *Proc. Natl. Acad. Sci.* **99**(14), 9266–9271.

Ecological Risk Assessment Applied to Energy Development



CHARLES R. HARMAN

AMEC Earth & Environmental
Somerset, New Jersey, United States

WILLIAM R. ALSOP and PAUL D. ANDERSON

AMEC Earth & Environmental
Westford, Massachusetts, United States

-
1. Components of an Ecological Risk Assessment
 2. Application to Specific Energy-Development Activities
 3. Conclusions

Glossary

analysis The second phase of the ecological risk assessment. It consists of the characterization of exposure, which is the measurement of the temporal and spatial distribution of a stressor, and characterization of effects, which is the identification and quantification of adverse effects elicited by the stressor.

assessment endpoint An explicit written expression of an environmental value that is to be protected in an ecological risk assessment. The environmental value consists of an entity (e.g., fish-eating birds) and a property of that entity (e.g., abundance). For example, sustaining the population of fish-eating birds in the vicinity of a water body.

conceptual site model A written description or visual representation of predicted relationships regarding how a stressor may affect the ecological components (ranging from an individual organism to a specific population, communities of populations, etc.) of the natural environment.

ecological risk assessment A process that evaluates the likelihood that adverse ecological effects may occur or are occurring as a result of the exposure of terrestrial, avian, and aquatic receptors to one or more stressors. As currently practiced in the United States, the process consists of three steps: problem formulation, analysis, and risk characterization.

measurement endpoint A concentration-dependent measurable response (typically an adverse effect such as lowered reproduction) to a stressor that is used to evaluate potential adverse effects to the environmental

value specified in the assessment endpoint. For example, a comparison of the number of kingfisher (a fish-eating bird) young produced per nest at an impacted water body with the number of kingfisher young produced per nest at a nonimpacted water body.

problem formulation The first step in the ecological risk assessment, which includes a preliminary characterization of exposure; examination of available data and data needs, policy, and regulatory issues; and an evaluation of site-specific factors to define the feasibility, scope, and objectives of the ecological risk assessment. The outcome of the problem-formulation step is the selection of assessment and measurement endpoints and the development of a conceptual site model.

receptor A biotic component (from an individual organism to a population, a species to a community, etc.) of the ecosystem that is, or may be, adversely affected by a stressor.

risk characterization The third and final phase of the ecological risk assessment. It uses the results of the exposure and effects analyses to evaluate the likelihood of adverse ecological effects associated with exposure to a stressor.

stressor Any physical, chemical, or biological entity that can induce an adverse effect on a receptor.

Since the mid-1970s, public concern regarding the effects of chemicals released into the environment has increased, leading state and federal regulatory agencies to develop and adopt procedures for assessing both potential and actual ecological impacts that occur as the result of these releases. These effects are evaluated through a process called an ecological risk assessment (ERA). The requirement for ERAs was brought about by the Comprehensive Environmental Response Compensation and Liability Act (also

known as Superfund), Resource Conservation and Recovery Act, and other environmental statutes. However, it was not until the late 1980s when the Science Advisory Board, the research arm of the U.S. Environmental Protection Agency, recommended that the agency place as much emphasis on evaluating ecological risks as on human health risks. Efforts to then standardize the ERA process began.

The existence of a well-established process to evaluate potential human health risks has had an important effect on the ERA process. In human health risk assessment, humans are the receptor of concern (unlike in ecological risk assessment, in which many receptors can be exposed and can potentially be at risk). Society also places a great deal of value on each individual human and on the prevention of very subtle adverse effects. In human health risk assessment, for example, an increase in mortality of 1% caused by a chemical release would be unacceptable. In ecological risk assessment, an increase in mortality of more than 1% may be acceptable if it occurs in a common and abundant species. In order to ensure that unacceptable levels of risk do not occur in humans, human health risk assessment uses numerous conservative assumptions. A conservative assumption is one that leads to overprotection rather than underprotection. Because society generally places more value on humans than on ecological receptors, and also places a unique value on different groups of ecological receptors, practitioners and users of ecological risk assessment need to be cognizant of the degree of conservatism contained in an ERA, that it is appropriate, and that it is consistent with societal norms. There is a tendency to allow the conservatism found in human health risk assessments to creep into ecological risk assessments and for ecological relevance to be lost.

With regard to the development of energy, ERAs are typically used to evaluate the effects of chemicals released into the environment, although the process has been developed to deal with “stressors” in general. The development of energy presents special challenges to evaluating impacts to environmental quality because exploration, development, refining, distribution, and use of energy from different types of energy systems have the potential for releasing a variety of chemicals into the environment over scales ranging from the small and local to regional and global levels. This article outlines concerns and issues to be considered during the preparation of ERAs for energy systems with examples from associated activities. This article focuses on the potential effects of chemicals released during energy development and

use and not on other stressors, such as nuclear radiation, heat, or the habitat disturbance that may occur during mining or oil drilling. A brief discussion of other stressors (e.g., physical effects to habitats from hydropower) associated with energy development is presented.

1. COMPONENTS OF AN ECOLOGICAL RISK ASSESSMENT

Ecological risk assessments (ERAs) typically consist of three steps: problem formulation, analysis, and risk characterization. Problem formulation provides the overall rationale and justification for the subsequent steps in the ERA. Problem formulation summarizes available information and uses it to develop clearly defined goals for the ERA as well as unambiguous criteria for evaluating potential risks. Endpoints that are ambiguous or difficult to define and measure (such as “ecosystem health”) will only result in disagreement about what was intended (e.g., protection of individual receptors versus protection of a population of receptors versus protection of certain ecosystem processes such as providing a food source). The results of the problem-formulation step are a conceptual site model (CSM) and the selection of assessment and measurement endpoints. The CSM considers the dominant physical, chemical, and biological characteristics of a system; identifies relevant receptors; and identifies potentially complete pathways by which relevant receptors can be exposed to stressors. It is common for a CSM to identify 5–10 receptors to evaluate in an ERA. Typically, 1 receptor is evaluated in each assessment endpoint. For example, for a release of chemicals to a water body, one assessment endpoint may be the protection of fish-eating birds relying on that water body as a food source. The selected receptor for this endpoint might be a kingfisher (a fish-eating bird species) that has been observed feeding at the water body.

The CSM provides the basis for the selection of assessment endpoints and the measurement endpoints used to evaluate them. Site-specific assessment endpoints should be identified that address stressor-specific potential adverse effects to individual organisms (in the case of threatened or endangered species) or to local populations or communities of plants and animals. Typically, several assessment endpoints are selected to evaluate potential ecological effects based on the potentially complete pathways between the source of the chemicals and the receptors that might

be affected. For the example of a chemical release to a water body (discussed previously), an additional assessment endpoint might be the continued survival of the largemouth bass population. A third assessment endpoint would be the evaluation of the potential adverse effects to a terrestrial mammal such as a mink or raccoon that uses the water body for food and drinking water.

Measurement endpoints represent observations, measurements, or analyses that evaluate an assessment endpoint. One to several measurement endpoints are used to evaluate each assessment endpoint. For the kingfisher described previously, one set of measurement endpoints might focus on whether the aquatic community in the pond can still provide enough food to sustain the kingfisher population living near the pond. Measurement endpoints may include direct comparison of water body chemical concentrations to ambient water quality criteria; laboratory bioassays of site water with standard aquatic organisms; and field surveys of aquatic communities comparing the site to a reference area that has the same features as the site water body, except for the presence of site-related chemicals.

For measurement (and assessment) endpoints evaluating ecological communities, the distinction between structure and function of the endpoints is important. For the example of a chemical release to surface water, historically many ecological risk assessments have included the structure of the benthic community (e.g., diversity) as an assessment endpoint. It may be more appropriate to focus on the functional aspects of the benthic community (e.g., its function as a prey base for higher trophic-level predators). A change in the structure (i.e., the diversity of organisms present) of a lower trophic-level community that serves as a prey base for higher trophic-level predators may be less important than a change in the biomass of prey produced (i.e., the weight of prey available to be eaten by predators) by the community.

A second important distinction is between chemicals that can be directly toxic but do not bioaccumulate and chemicals that are directly toxic and also bioaccumulate. For chemicals that can bioaccumulate (e.g., PCBs and mercury), assessment endpoints generally include higher trophic-level predators that may accumulate these chemicals from the food chain. For these compounds and receptors, endpoints evaluating indirect toxicity through food chain exposures may be more relevant than endpoints evaluating direct toxicity. Conversely, for chemicals with low potential for bioaccumulation, selection of

assessment endpoints may concentrate on lower trophic-level receptors that may have higher direct exposures to the chemicals. For these compounds and receptors, assessment endpoints evaluating direct toxicity to fish or benthic invertebrates (e.g., reduced growth, reduced reproduction, or even mortality) may be more relevant than those evaluating indirect toxicity from food chain exposures.

Analysis is the second step in the ERA process. In this step, exposures of receptors to stressors are evaluated and the relationship between stressor levels and ecological effects is documented. Exposure characterization describes potential or actual contact or cooccurrence of stressors with receptors.

In large and complex ERAs, exposures of ecological receptors to chemicals may vary over time and space depending on the characteristics of the selected receptors. Migratory species will spend less time being exposed to site-specific chemical concentrations than resident species. For species with large home ranges that forage over large areas that do not contain site-related chemicals, exposures will be lower than those for species with smaller home ranges that may be exposed to site-related chemicals for their entire lifetimes. Exposures may be measured directly by sampling receptors or the organisms that comprise their diet, but more commonly exposures are estimated using a series of standard equations. The equations require information about the concentration of chemicals in food, foraging areas and habits, food consumption rates, body weight, water consumption rates, and incidental media-specific ingestion rates (e.g., the amount of soil ingested by robins feeding on worms). This information is obtained from sources ranging from literature summary values to species-specific literature values and results from experimental studies conducted in the field.

Ecological effects characterization identifies the effects (e.g., mortality and reduced reproductive success) elicited by a stressor, the relationship between these effects and the assessment endpoints (e.g., sustaining the population of fish-eating birds in the vicinity of a water body), and provides an evaluation of how these effects change with varying stressor levels. The characterization begins by evaluating available effects data to specify (i) the effects that are elicited by the stressor, (ii) the plausibility that effects may occur or are occurring as a result of exposure to the stressor consistent with the conceptual site model, and (iii) the links between measurable ecological effects and assessment endpoints when the latter cannot be directly measured.

Effects data may consist of chemical-specific literature values from laboratory studies on the selected receptor species, results of chemical-specific studies on taxonomically related species, and results of studies on structurally similar chemicals. Site-specific field studies may also be evaluated as part of the effects characterization, depending on the selected assessment endpoints (e.g., sustaining the population of fish-eating birds in the vicinity of a water body) and measurement endpoints (e.g., field studies of reproductive success). Data typically collected from laboratory studies include effects ranging from the molecular level (e.g., enzyme induction) to the organ level (e.g., liver or kidney toxicity) and the organism level (e.g., mortality). Data typically collected from field studies include population-level effects (e.g., reduced numbers or reduced reproduction).

The selection of the appropriate study to use in the ecological effects characterization requires linking the study to the selected assessment endpoint. If the assessment endpoint includes a threatened or endangered species, then the effects characterization may focus on effects to an individual. However, if the assessment endpoint is sustaining the population of fish-eating birds in the vicinity of a water body and one measurement endpoint is sustaining the food supply for the fish-eating birds, then the effects characterization may focus on population-level effects to prey fish species.

The last step of the ERA process is risk characterization. This step combines the results of the analysis phase to develop estimates of the potential risks posed to the ecological receptors included in the assessment endpoints identified in the problem-formulation step. Risk characterization evaluates whether the doses measured or estimated in the exposure assessment are higher than the doses (identified in the effects assessment) that cause adverse effects and whether the estimated exposures are likely to cause adverse effects to the selected assessment endpoint. Because ERAs generally contain conservative assumptions, it is possible that an estimated exposure exceeds the allowable exposure but that an adverse effect is not expected to the assessment endpoint. Risk estimates developed from measurement endpoints are evaluated in the context of the significance of any adverse effects and lines of evidence supporting their likelihood of occurrence. For the assessment endpoint of a threatened or endangered species, this may include an evaluation of whether individual-level effects are likely to occur.

For assessment endpoints that are being evaluated using multiple measurement endpoints, a weight of evidence approach is typically used. In this approach, the weight conferred on each measurement endpoint, the magnitude of the response to each endpoint, and the concurrence among measurement endpoints are used to evaluate whether significant ecological effects are likely to occur. The weight of each measurement endpoint is a function of the degree of association between the measurement and assessment endpoints, the study design, and the quality of the data supporting the measurement endpoint. In the second component, strong unambiguous responses provide greater weight than weak responses. Finally, when a number of different measurement endpoints indicate similar responses, this provides additional weight. For example, one measurement endpoint [e.g., comparison of the number of kingfisher (a fish-eating bird) young produced per nest at an impacted water body with the number of kingfisher young produced per nest at a nonimpacted water body] may indicate no adverse effects to the assessment endpoint (e.g., sustaining the population of fish-eating birds), but an additional assessment endpoint including an evaluation of the potential for adverse effects to the aquatic community that serves as the food source for the kingfisher may indicate the potential for an adverse effect. A weight of evidence approach is used to evaluate whether the estimated effects to the food supply are sufficient to cause adverse effects to the selected assessment endpoint. Considerations in this evaluation also include the presence of alternative sources of food and the likelihood of their use by the kingfishers.

2. APPLICATION TO SPECIFIC ENERGY-DEVELOPMENT ACTIVITIES

The development of energy, although providing a benefit to society, can adversely affect the environment. ERAs can be used to assess the magnitude or severity of the perturbation and in determining cleanup levels in circumstances in which a cleanup is required. As previously mentioned, the focus of this article is on the evaluation of potential ecological effects associated with chemical releases from energy development.

The ERA process can be used to evaluate other nonchemical stressors, such as physical disturbance, radiation, and thermal changes in an environment. For example, nuclear power-generating stations have

the potential to release radioactive materials into the environment. As a stressor, radionuclides can be examined via the same ERA process as that used for chemical stressors. This process includes identification of appropriate potential receptors and endpoints, determination of levels of radionuclides in the environment, analysis of the levels of existing radiation versus benchmarks levels for organisms of concern or the endpoints that are being assessed, followed by the characterization of the risks to populations of various organisms in the release area. As with other energy-development activities, the critical step in the process is a thorough and appropriate problem-formulation step, including the development of a conceptual site model that emphasizes the circumstances of the release area and takes into consideration the movement and ultimate effects of radiation on natural ecosystems.

Although the ERA process evaluates the potential for chemicals to affect biological organisms, it is important to understand the impact that physical disturbance and habitat alteration may have on the ecosystem versus the release of development-related chemicals. Hydroelectric power, which releases few chemicals to the environment, may block the movement of fish up and downstream; may reduce large fluctuations in hydrologic regimes that may hinder plant and animal species that are competitively more adapted to living in flood-prone areas; or may change nesting areas, resulting in the modification of faunal communities in an area. Nuclear and fossil fuel power plants may impinge or entrain large quantities of fish and aquatic life through water intake for cooling purposes. These facilities may discharge heated water into the environment that may kill fish, block fish migrations, or result in the encroachment of nuisance species. Mining for coal can result in the stripping of mountaintops and the resulting loss of habitat used by terrestrial wildlife. All these are “nonchemical” stressors with the potential to pose ecological risk that can be evaluated using ERAs.

The point is that an ERA is a tool used to evaluate the effects of stressors on the environment. Its results must be placed in a realistic context of the breadth of stressors that result from a given energy-development activity. As discussed previously, and discussed throughout much of this section, the problem formulation stage is the critical first step in identifying what potential issues must be evaluated in the ERA and how they relate to other stressors in the environment. The following sections discuss issues related to the use of ERAs for specific energy-development activities.

2.1 Ecological Risk Assessment Applied to Oil and Gas Activities

Because energy development is dominated by the use of oil and gas as fuel sources (approximately 40% of the world's energy supply is from crude oil), this article focuses on the ecological effects associated with the various stages of energy development from oil and gas. Oil- and gas-related activities have the potential to adversely affect ecological receptors through the release of a variety of chemicals to the environment. Glen W. Suter outlines a general framework for conducting ecological risk assessments at sites at which petroleum hydrocarbons are the dominant chemicals that may potentially pose an ecological concern.

2.1.1 *Petroleum Hydrocarbons in the Environment*

In 1989, the U.S. Environmental Protection Agency (EPA) reported that 20.9 million barrels of production waste and 361 million barrels of drilling waste were generated from approximately 835,000 oil and gas wells nationwide. Additionally, they reported that on average approximately 9000 oil spills occurred each year, releasing approximately 11 million barrels of oil into the environment. Ecological impacts associated with these releases ranged from short-term effects such as fish kills to long-term impacts such as chronic changes in benthic macro-invertebrate community structure (e.g., the number and diversity of invertebrates living in sediments).

The ecological impacts produced by these releases occurred as a result of physical actions, such as smothering and fouling, and the direct toxicity of the various chemicals in the petroleum mixtures. Key organic compounds in such releases include polycyclic aromatic hydrocarbons [e.g., phenanthrene and benzo(*a*)pyrene] and volatile organic compounds (e.g., benzene). Although the potential effects from inorganics (e.g., metals) are generally not of concern where petroleum mixtures are spilled in the environment, metals such as arsenic, barium, cadmium, chromium, copper, lead, mercury, and selenium can be released at levels that are toxic to terrestrial and aquatic receptors during oil and gas production activities.

Petroleum is a mixture of hydrocarbons that can be classified into two broad categories: aliphatic hydrocarbons, which can be further divided into alkanes, alkenes, and cycloalkanes, and the aromatic hydrocarbons. The percentage of each group present varies with both the source of crude oil and the

refined product. The compounds within these two chemical groups exhibit a range of toxicity, with some being extremely toxic, whereas others are relatively benign. In general, the toxicity of hydrocarbons is related to their chemical structure and their affinity for water, with more soluble compounds being relatively less toxic. Although this is a general relationship and can be affected by chemical structures, compounds that are more soluble in water are less soluble in fat. As the solubility in fat or lipids increases, solubility in water decreases and the ability of the hydrocarbon to cross the cell membrane and produce biological damage increases. This relationship also suggests that as lipid solubility increases, the potential for the chemical to be passed up through the food chain (from an original food source to a lower level predator and then to ever higher levels of predators) increases.

Understanding these toxicological properties is important because of the general manner in which many regulatory agencies evaluate the presence of petroleum at sites. As opposed to measuring the concentration of each chemical in a hydrocarbon mixture, many states regulate and evaluate petroleum sites based on the quantification of total petroleum hydrocarbons (TPHs) in surface water, sediments, and soil. The EPA defines TPH as “a measure of the concentration or mass of petroleum hydrocarbon constituents present in a given amount of air, soil, or water” and further notes that

the term total is a misnomer, in that few, if any, of the procedures for quantifying hydrocarbons are capable of measuring all fractions of petroleum hydrocarbons present in the sample. Volatile hydrocarbons are usually lost in the process and not quantified. Additionally, some nonpetroleum hydrocarbons may be included in the analysis.

To the extent that the hydrocarbon extraction efficiency is not identical for each method, the same sample analyzed by different TPH methods will produce different TPH concentrations. TPH analytical procedures use a variety of techniques, such as gas chromatography (GC), infrared spectroscopy (IR), gravimetric techniques, and immunoassays. Because of different technical processes, each technique detects a different range of hydrocarbons. GC techniques normally measure the broadest range of hydrocarbons, whereas IR techniques measure most hydrocarbons with the exception of volatiles and compounds with high numbers of carbon atoms. Gravimetric methods measure anything that is extractable except volatiles, and immunoassays measure

only aromatic hydrocarbons, such as polycyclic aromatic hydrocarbons and benzene, toluene, ethylbenzene, and xylenes compounds. Interference with the analytical procedure of some of the methods can result from the presence of natural oils, humic substances, and peat moss. TPH measurements have been made of grass (14,000 mg/kg of TPH), pine needles (16,000 mg/kg of TPH), and oak leaves (18,000 mg/kg of TPH). Assessing risk based only on TPH is difficult because the analytical procedures identify all hydrocarbons and not specific ones, regardless of the source. As such, TPH analysis does not adequately characterize a petroleum mixture and provides little basis for characterizing the fate, transport, and toxicological properties of the mixture.

Under these conditions, using TPH leads to overestimates of ecological risk if, for example, the allowable TPH concentration was based on a “toxic” petroleum mixture and the petroleum being assessed was from a relatively “inert” source. Conversely, ecological risk may be underestimated if the reverse were true—the TPH in the field was from a petroleum source that was more toxic than that used to establish the allowable TPH level. The more appropriate manner for characterizing site conditions related to petroleum releases is the use of analytical procedures that evaluate the semivolatile and volatile organic compounds using standard EPA protocols. This is supported by guidance documents from Oak Ridge National Laboratory that state that TPH analysis provides little basis for performing an ERA because it supplies limited information about the composition and, therefore, the properties that determine the potential fate and toxicity of the material. However, TPH may be useful for determining the extent of contamination or the locations of the greatest contamination.

In addition to the shortcomings in adequately characterizing differences in toxicity in different petroleum hydrocarbons, TPH analyses also fail to consider the processes that cause petroleum hydrocarbons to partition into the environment. Mixtures of petroleum hydrocarbons change once released into the environment because weathering, the process by which natural chemical and physical processes alter the composition of the mixture, changes the bioavailability of the remaining hydrocarbons and hence their toxicity. Weathering is primarily driven by dissolution in water, biodegradation, and volatilization.

Because weathering affects each class of hydrocarbons differently, the toxicity of a particular mixture changes as it weathers. The solubility and volatility of all compounds generally decrease with

an increase in molecular weight. Aromatic hydrocarbons, those hydrocarbons with one or more benzene rings, are more soluble in water, whereas low-molecular-weight aliphatic hydrocarbons, characterized by a straight chain of carbon atoms, are more volatile. Low-molecular-weight compounds, which are more soluble and volatile, tend to migrate from contaminated soil at a much faster rate than high-molecular-weight compounds. Residual compounds left in highly weathered hydrocarbons tend to be the higher molecular-weight compounds that are less mobile and generally less toxic.

2.1.2 Oil and Gas Production

Potential ecological impacts from releases of chemicals during oil production activities are generally small in scale, although they can produce severe effects to locally affected communities. Compounds of ecological concern can occur in drilling mud and fluids, which when released into the environment can produce fish kills, smother benthic invertebrate communities, and destroy vegetation. Such compounds include arsenic, chromium, copper, iron, and other metals as well as oil, grease, chlorides, nitrates, and other substances. These substances tend to be directly toxic to fish, wildlife, and other biological receptors but generally do not bioaccumulate at successively higher trophic levels of the food chain.

As a result of the small scale of individual production activities at a single well, ERAs may overestimate risks to populations if they used conservative assumptions and focused on individuals rather than having the problem-formulation phase consider the issue of scale in the context of appropriate receptors and endpoints. In examining a single drilling site or release via the standard three-step ERA process, consideration must be given to applying realistic exposure parameters to the typical receptors found in a given area. Because the focus of the ERA must be on evaluating the potential for ecological impacts to local populations, care must be taken to ensure that the inference of impacts to local organisms is not considered as a significant impact to the population. A drill rig located on a $\frac{1}{2}$ -acre site in a 1000-acre tract that supports a population of the meadow vole *Microtus pennsylvanicus* might only affect 4–6 animals based solely on the territory that the animals occupy. Out of the greater population of animals, in which an individual female may bear as many as 150 young per year, this would place into question the ecological relevance of a risk assessment that is not focused in the problem formulation on population issues. However, if the 1000-acre tract

were covered with a large number of drilling sites, the cumulative effect of many sites, all releasing approximately the same quantity of compounds, could result in ecological impacts to populations of animals and plants.

2.1.3 Petroleum Refining, Storage, and Processing

From an ecological standpoint, evaluating the ecological impacts associated with residual petroleum releases at major oil-production facilities can be difficult because of the complexity of the sites, the large number of chemicals involved, and the ongoing production practices. An examination of many of the large refineries or terminals shows that these facilities are often located adjacent to sensitive environments, such as coastal waterways, rivers, and harbors. In New Jersey, the largest refineries can be found at locations along the Delaware River in the south and adjacent to Raritan Bay, Arthur Kill, and Newark Bay in the New York/New Jersey Harbor area. These locations present ample opportunity for the discharge of petroleum and petroleum-related products into productive ecosystems containing receptors that will be in close contact with the released materials.

Part of the complexity in dealing with such facilities is their size, the diversity of habitats and the variety of ecological receptors that can be exposed, and the age of the facilities. Refineries can be very large, in excess of 150 acres, and contain freshwater and coastal wetlands, ponds, and forests. Although production activities within refineries and terminals dominate land use, most refineries also contain land that is undisturbed, fallow, or covered with storage tank fields, landfills, and waste management units. Many of the areas supporting seemingly natural habitats may have historically been used for storage or production. At several refineries in New Jersey, forests and shrublands have grown on parts of the property that previously contained storage tanks that were removed from service and dismantled. These areas of terrestrial vegetation support a number of wildlife species, including white-tailed deer (*Odocoileus virginianus*) and avian predators such as red-tailed hawks (*Buteo jamaicensis*). However, these habitats have developed in areas created by fill and often contain residual levels of petroleum hydrocarbons. From a gross standpoint, the fact that these areas have revegetated substantiates the limited effects of residual hydrocarbons in the environment. Nevertheless, ERAs are generally required in such circumstances to evaluate any potential risks associated with residual levels.

In preparing an ERA for a refinery or large production facility, great care must be given to the problem-formulation step. As opposed to ERAs conducted under production or drilling scenarios, refinery ERAs must consider three complex issues. First, refineries often contain hundreds of existing or former discharge sources. These sources can include crude oil or basic petroleum products as well as more refined petrochemical products that may potentially be more toxic than crude oil. Second, the pathways associated with a refinery ERA often include both aquatic and terrestrial environments, with numerous assessment endpoints in each of these environments that may need to be evaluated. Typical assessment endpoints include the maintenance of viable terrestrial plant communities; maintenance of viable aquatic and wetland plant communities; maintenance of aquatic wildlife, fish, and benthic invertebrate communities; and maintenance of both resident and migratory terrestrial wildlife populations. Third, developing a conceptual site model for a refinery or terminal is often complicated by the age and location of these facilities. In industrial sections of the northeastern United States, some refineries began operations in the late 1800s or early 1900s, whereas general industrialization extends back to the late 1700s. In these settings, it is often difficult to distinguish impacts associated with refinery operations from those that occurred as a result of general urbanization and to isolate site-related ecological impacts from impacts that may have occurred as a result of neighboring facilities that may also be releasing constituents of concern to the environment.

In settings in which refineries have residual concentrations of chemicals but still support viable populations of terrestrial and aquatic flora and fauna, comparison to background chemical data can be an important indicator of site-related effects. In this case, it is essential to collect background chemical data from locations similar in habitat but still subject to the same level of historic disturbance as the refinery. Refineries are typically located adjacent to, or close to, major harbors, with more than 60 petroleum facilities located in the Newark Bay estuary alone. Between 1982 and 1991, a total of 1453 accidental incidents released more than 18 million gallons of hazardous materials and petroleum products into Newark Bay. Thus, identification of background locations can be a contentious issue in industrialized settings because chemical concentrations can be quite high—high enough that they would require cleanup if the site were in a pristine setting.

The EPA has issued guidance on use of background data in ERAs. Although the agency recognizes that sources other than the facility being investigated may release chemicals into the environment, the expectation is that the effect of background concentrations will be included in risk estimates, even if background is the primary source of risk, and be qualitatively discussed in the risk characterization section of the ERA. The importance of the background concentrations of chemicals of concern is applicable in the risk management phase following the completion of the risk assessment. The assumption is that even if the background chemical concentrations were anthropogenically produced, taking action to address their presence at a particular site would be neither practical nor cost-effective if the background levels are widespread. Furthermore, remediated sites, which receive a constant stream of off-site contamination, will only be recontaminated from upstream sources. Particularly in aquatic sediments, a remediated area can be recontaminated by the movement of contaminated sediments from upstream areas through the action of currents or by the release of pollutants from upstream facilities.

2.1.4 Distribution

Distribution of oil and gas through pipelines can cause large-scale effects that are not related to chemical releases (e.g., alteration of many miles of habitat during construction) and relatively local effects from releases of oil during a pipeline failure or releases of other chemicals associated with transport of oil and gas. The most well-known of these was the release of PCBs at gas pipeline compressor stations. Unlike most petroleum constituents, PCBs have the potential to bioaccumulate. Consequently, assessments of ecological risks associated with PCB releases required food web modeling to predict ecological risks to upper trophic-level predators.

The other common mode of oil distribution is via ships. The release of oil in aquatic ecosystems has two primary modes of impact. One is physical fouling or impairment of biological organisms, such as birds, fish, and plants. The other is the toxic effect of the chemicals in the oil or refined product or any residual compounds. Physical fouling can be the most destructive action, but its effects are short term and mitigated by the dilution and degradation of the released oil. The toxic effects are often longer term and are assessed through the ERA process.

2.1.5 Retail Service Stations

Typically, individual gasoline service stations do not pose a significant threat to ecological resources. Service stations are small in relation to the surrounding environment and most are located in developed, urban areas or along paved roadways. These areas generally do not support large communities of native plants or animals that can be exposed to petroleum-related products nor are they generally located near sensitive habitats. Additionally, the major contamination sources at service stations are leaking underground storage tanks used to store gasoline and other products. Because the tanks are located in the ground, exposure pathways to ecological receptors in the surface environment are generally not complete. Without a complete exposure pathway, there can be no risk. However, exposure can occur if a leaking tank contaminates groundwater, which in turn is discharged to a surface water body or a wetland. In such instances, chemicals to which ecological receptors may be exposed include volatile organic compounds such as benzene, toluene, xylene, ethylbenzene, and methyl *tert*-butyl ether (MTBE) and, depending on the age of the release, inorganic constituents such as lead.

In evaluating the ecological risks associated with groundwater contamination originating from a service station located in the Pine Barrens Management Area of southern New Jersey, environmental sampling indicated that ethylbenzene, xylenes, MTBE, and lead were being released over a 50-ft stretch of a small stream flowing through a mixed hardwood forested wetland. The problem formulation indicated that the release had the potential to affect invertebrates within the stream sediments, fish living in the stream, and wildlife that either used the stream as a water source or fed on plants growing adjacent to the stream. Investigation of the release found maximum concentrations in sediments of 34 µg/kg of xylenes, 50 µg/kg of MTBE, and 24 mg/kg of lead. The only chemicals identified in surface water were MTBE (0.5 µg/liter) and lead (26.5 µg/liter). To characterize potential risk, chemical concentrations in surface water and sediments were compared to values for specific chemicals that are meant to be protective of the environment under most conditions. Slight exceedances existed, so a site-specific biological survey of benthic invertebrates in sediments was conducted. Benthic invertebrate surveys conducted at reference locations and site-sampling locations found no difference between the two areas in the kinds of species present, indicating that the aquatic community in the affected stream

was similar in all respects to the aquatic community in unaffected streams. The ERA concluded that no adverse ecological impacts were associated with the discharge of contaminated groundwater. Although remedial actions to address groundwater contamination and potential human health risks were implemented, no actions were necessary within the stream or wetlands to protect ecological resources.

2.2 Ecological Risk Assessments Associated with Coal Mining

Whereas ecological risk issues complicated by industrial development and background contamination dominate some aspects of the petroleum industry, coal mining is generally an activity that occurs in more natural settings. The obvious ecological impact associated with coal mining is the destruction of habitat. However, chemicals can be released into the environment from the development of ore piles, mine tailings, and spoil heaps.

The chemical release of greatest ecological concern is acid mine drainage (AMD). In certain circumstances, surface water and shallow groundwater will react with rocks containing sulfur-bearing minerals to create sulfuric acid. This acidified water then leaches metals such as aluminum, iron, and manganese from rocks that come in contact with the water. Aluminum concentrations of 57 mg/liter have been measured in AMD from mining in Appalachia. The EPA chronic water quality criterion for aluminum is 87 µg/liter under alkaline conditions (the potential for aluminum toxicity increases with increasing acidity), which is approximately 1000 times lower than the aluminum concentration in AMD. The EPA chronic water quality criteria for acidity (pH) ranges between 6.5 and 9 standard units, whereas pH in AMD from Appalachia has been reported to be as low as 2.7. AMD has caused the complete elimination of fish from some lakes and streams in the Appalachians or the presence of only a few acid-tolerant fish species in other waters.

2.3 Ecological Risk Assessment Procedures Applied to Power Generation

Power generation through the burning of fossil fuels can lead to a variety of chemical-related impacts to the environment. Also, direct habitat loss can occur as a result of facility construction, tank storage, or coal pile maintenance. Chemical impacts can occur as the result of catastrophic releases of oil or waste materials

such as caustic soda, the generation of AMD (from coal piles), or the improper storage or mishandling of waste sludges and fly ash wastes. Fly ash wastes from coal-fired power plants contain metals, such as aluminum, arsenic, cadmium, chromium, copper, mercury, and selenium, that can be discharged at concentrations sufficient to significantly impact benthic invertebrate populations and fish communities as well as terrestrial plant and animal communities.

The most wide-ranging ecological issue associated with power plants is the release of chemicals through stacks into the atmosphere. Conceptual site models for air-dispersion ERAs must take into account the numerous potentially complete exposure pathways that can spread hundreds of miles from the site and affect many receptors. In such instances, ERAs must focus on select indicator organisms that are used as the most likely impacted organisms as well as sensitive receptors that could be most adversely impacted by the stack emissions. Particularly important issues related to air emissions, mercury, and acid rain are discussed in detail in the following sections.

2.3.1 Potential Ecological Impacts from Mercury Emissions

The EPA completed an assessment of the potential impacts of mercury emissions, part of which consisted of an ERA. It estimated that approximately 87% of the estimated 144 Mg (158 tons) of mercury emitted annually into the atmosphere by anthropogenic sources in the United States is from combustion point sources, 10% is from manufacturing point sources, 2% is from area sources, and 1% is from miscellaneous sources. Four specific source categories account for approximately 80% of the total anthropogenic emissions: coal-fired utility boilers (33%), municipal waste combustion (19%), commercial/industrial boilers (18%), and medical waste incinerators (10%). These sources represent high-temperature waste combustion or fossil fuel processes. For each of these operations, mercury is present as a trace chemical in the fuel or feedstock. Because of its relatively low boiling point, mercury is volatilized during high-temperature operations and discharged to the atmosphere with the exhaust gas.

Mercury is emitted from stacks primarily in its elemental or inorganic (combined with other inorganics, such as sulfide or chloride) form. Mercury is transported and deposited onto soils and surface waters, where some is converted to organic mercury (e.g., methylmercury). Inorganic and elemental mercury are directly toxic through ingestion and inhalation, but organic mercury tends to bioaccumulate

and toxicity is expected through the food chain in higher trophic-level organisms, not in directly exposed lower trophic-level organisms. That is, because mercury accumulates through the food chain, exposures experienced by higher trophic-level organisms to bioaccumulated levels of mercury can cause direct toxicity, whereas exposures experienced by lower trophic-level organisms to nonbioaccumulated levels may not be sufficient to cause toxicity.

The EPA's risk assessment evaluated the fate and transport of mercury emissions and used estimated atmospheric mercury concentrations and deposition rates as inputs in a model to derive concentrations of mercury in soils and surface waters over the drainage basis for an entire water body. These soil and water concentrations were used to estimate mercury (as methylmercury) concentrations in the associated biota and fish, which humans and other animals were assumed to consume. Methylmercury intake (on a per body weight basis) was estimated for six wildlife species: kingfisher, river otter, loon, mink, osprey, and bald eagle. Methylmercury exposures for the most exposed wildlife species (the kingfisher) were estimated to be up to two orders of magnitude higher than human exposures from contaminated freshwater fish. For wildlife, the importance of site-specific effects of mercury exposure is anticipated to be greater than that for humans in the general population because fish-eating wildlife obtain their fish from a much more limited geographic area than do people, and the diet of some wildlife species is composed almost entirely of fish.

The broad ecosystem effects of mercury are not completely understood. No applicable studies of the effects of mercury on intact ecosystems were found. Consequently, characterization of risk for nonhuman species did not attempt to quantify effects of mercury on ecosystems, communities, or species diversity. The characterization focused on quantities of mercury that may adversely affect the health of sensitive subpopulations of wildlife species and on the collocation of these populations with areas of elevated mercury exposures from anthropogenic emissions of mercury.

The results of the EPA's assessment suggest that ecosystems most at risk from airborne releases of mercury exhibit one or more of the following characteristics:

1. They are located in areas where atmospheric deposition of mercury is high.
2. They include surface waters already impacted by acid deposition.

3. They possess characteristics other than low pH that result in high levels of bioaccumulation.
4. They include sensitive species.

2.3.2 Ecological Impacts Associated with Acid Rain

Acid rain is the generic term for the wet deposition of substances that make the rain more acidic than the normal pH 5.6 of carbon dioxide in equilibrium with distilled water. Acidic deposition includes both wet and dry deposition of material more acidic than normal based on the concentrations of sulfates and oxides of nitrogen. In 1997, the major sources of sulfur dioxide emissions were electric utilities (60%), industrial combustion (17%), and industrial processes (8%). Transportation sources (including cars, trucks, and nonroad vehicles such as construction equipment) accounted for more than 50% of all nitrogen oxide emissions. Other major sources of nitrogen oxides include electric utilities (26%) and industrial combustion (14%). According to regional studies conducted by the EPA since 1990, agricultural activities, particularly manure handling, constitute the largest source of ammonia emissions. Motor vehicles and industrial processes also contribute to ammonia emissions.

Although regulatory controls in the 1970s and 1990s resulted in decreased sulfur dioxide emissions, levels remained high compared to background conditions. From 1995 to 1997, wet deposition of sulfate in the northeast was approximately 20% lower than levels in the preceding 3 years with implementation of the 1990 Clean Air Act Amendments. Controls on nitrogen oxides and ammonia were not fully addressed during the same period; consequently, emissions of these compounds have remained largely unchanged.

The 1990 National Acid Precipitation Assessment Program report to Congress concluded that there was insubstantial evidence that acid deposition had caused the decline of trees, other than red spruce, growing at high elevations. Research has shown that acid deposition contributes to the decline of red spruce trees throughout the eastern United States and sugar maple trees in central and western Pennsylvania. Reported symptoms of tree decline include poor crown condition, reduced tree growth, and unusually high levels of tree mortality.

Acid deposition degrades water quality by lowering pH levels (i.e., increasing acidity), decreasing acid-neutralizing capacity, and increasing aluminum concentrations. Although sulfate concentrations in lakes and streams have decreased during the past 20 years, they remain high compared to background

conditions. Moreover, improvement in the acidity of lakes and streams in the northeast has been limited. For example, many Adirondack region lakes are still acidic or subject to short-term pulses in acidity associated with precipitation events.

3. CONCLUSIONS

The ERA process consists of three steps: problem formulation (including developing a CSM and selecting assessment and measurement endpoints), analysis (exposure and effects), and risk characterization. Problem formulation (including selection of the ecological values to be protected) is the most important step because it identifies the ecological values that need to be protected and evaluated. Although scientific considerations play a role in identifying these values, value judgments (subjectivity or professional judgment) have an equal if not more important role, for example, in selecting which species to use as an indicator, whether to examine individual or population effects, or whether to consider structure or function of a plant or benthic invertebrate community. Problem formulation also involves the evaluation of complete versus incomplete exposure pathways. The analysis phase includes the evaluation of potential exposures of receptors to chemicals and the evaluation of the potential toxicity of these chemicals to the receptors.

The last step of the ERA process is risk characterization, which evaluates whether the doses measured or estimated in the exposure assessment are higher than the doses (identified in the effects assessment) that cause adverse effects and whether the estimated exposures are likely to cause adverse effects to the selected assessment endpoint. Because ERAs generally contain conservative assumptions, it is important to evaluate the contribution of background concentrations.

In the development of energy and energy sources, wherever release of chemicals to the environment occurs, an ERA may be needed to evaluate the potential for ecological impacts associated with that release. ERAs must be tailored to meet the technical requirements and challenges of the energy-development industry.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Climate Change and Public Health: Emerging Infectious Diseases •

Crude Oil Spills, Environmental Impact of • Ecological Footprints and Energy • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Hazardous Waste from Fossil Fuels • Nuclear Power: Risk Analysis • Risk Analysis Applied to Energy Systems

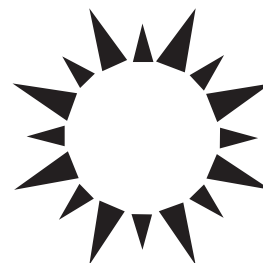
Further Reading

- Suter, G. W. II. (1993). "Ecological Risk Assessment." Lewis, Boca Raton, FL.
- Suter, G. W. II. (1997). "A framework for assessing ecological risks of petroleum-derived materials in soils, Oak Ridge National Laboratory Environmental Sciences Division Report ORNL/TM-13408." Oak Ridge National Laboratory, Oak Ridge, TN.
- U.S. Environmental Protection Agency (1989). "The nature and extent of ecological risks at Superfund sites and RCRA facilities, Office of Policy Planning and Evaluation EPA-230-03-89-043." U.S. Environmental Protection Agency, Washington, DC.
- U.S. Environmental Protection Agency (1989). "Summary of ecological risks, assessment methods, and risk management decisions in Superfund and RCRA facilities, Office of Policy Planning and Evaluation EPA-230-03-89-046." U.S. Environmental Protection Agency, Washington, DC.
- U.S. Environmental Protection Agency (1992). "Framework for ecological risk assessment, Office of Research and Development/Risk Assessment Forum EPA/630/R-92/001." U.S. Environmental Protection Agency, Washington, DC.
- U.S. Environmental Protection Agency (1993). "Wildlife Exposure Factors Handbook, EPA/600/R-31/187a." U.S. Environmental Protection Agency, Office of Research and Development, Washington, DC.
- U.S. Environmental Protection Agency (1997). "Ecological risk assessment guidance for Superfund: Process for designing and conducting ecological risk assessments, Interim final, EPA 540/R-97/006." U.S. Environmental Protection Agency, Washington, DC.
- U.S. Environmental Protection Agency (1998). "Guidelines for ecological risk assessment, 63FR26845, EPA/630/R-95/002F." U.S. Environmental Protection Agency, Washington, DC.
- U.S. Environmental Protection Agency (1999). "Issuance of final guidance: Ecological risk assessment and risk management principles for Superfund sites, Office of Solid Waste and Emergency Response Directive 9285.7-28 P." U.S. Environmental Protection Agency, Washington, DC.

Economic Geography of Energy

BARRY D. SOLOMON

Michigan Technological University
Houghton, Michigan, United States



-
1. Eight Basic Problems
 2. Fossil Fuels
 3. Nuclear Power
 4. Renewable Energy
 5. International Comparisons of Energy Consumption

Glossary

energy reserves The portion of the total resource base that has been located with certainty that can be economically recovered and used with present technology. This quantity is not fixed.

energy resources The portion of the total resource base that has been estimated with at least a small degree of certainty, and which might be recovered in the future.

location theory Spatial theory developed to solve optimization problems for selecting a preferred zone or location for a facility, which involves trading off transportation costs with production and processing costs, production balances, and market delivery levels.

Organization of Petroleum Exporting Countries (OPEC) Members include Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela.

resource/production ratio The ratio of total oil or gas resources to the current production level; provides a rough estimate of the scarcity of fossil fuels in a country, region, or the world.

The economic geography of energy involves quantitative and qualitative study of the spatial patterns of energy production, transportation, and use at a variety of geographic scales, and analysis of the determinants of those patterns. It is a surprisingly diverse subject area, encompassing everything from locational and transportation models, to economic impact studies of energy development and use, to international, national, regional, and site-specific problems associated with particular energy sectors. Site-specific sector problems include transportation constraints asso-

ciated with natural gas in various countries, the growing scarcity of oil and gas, emergency planning at nuclear power plants and radioactive waste disposal, environmental impacts, and factors influencing the international rate of diffusion of wind power technology. Given this diversity of subject matter, the discussion here provides a broad overview of the economic geography of energy.

1. EIGHT BASIC PROBLEMS

Eight broad research problems are representative of the rich range of energy subjects that are investigated in the field of economic geography: supply and demand location, fossil fuel scarcity, environmental impacts, nuclear proliferation, renewable energy development, electric utility deregulation, geopolitics, and technological changes and energy efficiency.

1.1 Separation between Supply and Demand Centers

From prehistory until the Industrial Revolution, most energy sources used by humans were localized (i.e., available within 5–10 miles of end users). Energy sources included draft animals, human slaves, and renewable sources such as biomass (wood and wastes), water mills, and wind power. Following the onset of the Industrial Revolution, with advancements in transportation technology and increased rates of deforestation in many regions, societies increasingly relied on long-distance shipments of coal, wood, and eventually oil, natural gas, and electric power. With increased separation between the location of energy resources and demand centers came a need to investigate the optimal transportation modes, production rates, inventories, and scheduling of energy production and delivery systems. Industrial location theory and mathematical programming techniques

have been used by geographers to analyze the trade-off between energy development and production costs, and shipment costs to market centers to explain the siting pattern of energy facilities. A major factor determining the economic value of an energy source or carrier to its consumers is whether it can be economically stored or controlled for later use, which in turn is a function of the demand at a particular location. This interplays with other determinants of the desirability of a specific energy source or carrier, such as pecuniary cost, convenience of access and use, energy efficiency, steadiness of flow and renewability, safety, quality and other environmental properties, and end-use technology.

1.2 The Growing Scarcity of Fossil Fuels

Policy concerns have been raised about the scarcity of fossil fuels, especially oil and natural gas, since the 1970s. The important distinction between oil and gas reserves and resources leads to different prognoses in different parts of the world (see the Glossary). A useful indicator of scarcity is the resource/production ratio, which can be instructive to calculate for various nations. For instance, Indonesia had 7.4 billion barrels of oil resources in the ground in 2000 and produced 515 million barrels in 2001, leading to a ratio of 14.4, an indicator of the number of years of production left before its supply is exhausted, *ceteris paribus*. The same ratio for the United States in 2001 was 62.1. These ratios will not change with shifts in production technology or in the availability and cost of alternative energy sources.

1.3 Environmental Impact

The production and use of all energy sources results in undesirable environmental effects, which vary based on the health of the existing ecosystem, the size and health of the human population, energy production and consumption technology, and chemical properties of the energy source or conversion device. A shorthand equation for the environmental impact of energy production and use is $I = P \cdot A \cdot T$, where I is environmental impact, P is the size of the human population, A is the affluence of the population (e.g., per capita income), and T is technology (e.g., energy efficiency, emission rate of air and water pollution). Such impacts have been operationalized by geographers in case studies around the world. In some cases, regional econometric and input-output models have been used in these studies.

1.4 Nuclear Proliferation

Geographers researching the development of nuclear power have shifted emphasis from commercialization, cost, risk, public acceptance, and power plant siting in the 1950s through the early 1980s to reactor decommissioning and radioactive waste disposal since then. With nuclear power development on hold in most countries, attention has also been given to nuclear weapons facilities and weapons proliferation in an increasingly dangerous world. With the distinction, shaky at best, between civil and military nuclear programs made by the United States and United Kingdom under the Nonproliferation Treaty of 1970, nuclear programs or policies of nations with unofficial nuclear weapons programs (i.e., South Africa, Israel, Argentina, Brazil, Taiwan, and perhaps Iran, among others) will come under increased scrutiny. This is in addition to the eight or more nations with official nuclear weapons (and power) programs by 2003: the United States, Russia (and perhaps other former Soviet republics), United Kingdom, France, China, India, Pakistan, and North Korea.

1.5 Development of Renewable Energy Sources

Amid growing concern in the international scientific community concerning the scarcity of fossil fuels, greenhouse gas emissions, and global climate change, analytical interest in renewable energy development has grown. Because renewable energy sources are widely yet unevenly available, geographic analysis of development potential and policies is important. Biomass energies and hydroelectric power are the traditional sources of renewable energy; development policy has focused on wind, solar, geothermal, and hydrogen power technologies. Wind energy development in the United States has been slow but steady since the 1990s, whereas Germany became the world leader in installed capacity in 1997. Other nations with significant programs in renewable energy development include Spain (wind), Italy (geothermal), Denmark (wind), Iceland (hydrogen and geothermal), Israel (solar), Japan (solar and geothermal), Philippines (geothermal), India (wind and solar), Indonesia (geothermal), Brazil (biomass and hydrogen), Mexico (geothermal), and South Africa (solar).

1.6 Electric Utility Deregulation

Electric utility deregulation offers the great promise of market forces leading to lower electric rates, lower

air pollution, greater energy (and economic) efficiency, and perhaps greater use of renewable energy sources. Ideally, deregulation involves the restructuring of a previously monopolized or nationalized electric utility into separate generation, transmission, distribution, and marketing companies, and allowing wholesale and retail choice of generation company or power marketer. Deregulation has occurred to varying degrees since 1989 in the United Kingdom, Norway, Australia, New Zealand, Chile, Argentina, and about 20 states in the United States. There have been promising results in a few countries and in some U.S. states in some respects, especially lower rates and lower air pollution. In most cases, competitive markets have yet to be realized and lower rates can be attributed to other causes, such as previously planned amortization or retirement of expensive power plants, unexpected surplus in natural gas, rate caps, etc. In addition, deregulation has had only a slight beneficial effect on the use of renewable electricity sources. The promise of electric utility deregulation is thus unfulfilled and deserves further study.

1.7 Role of Geopolitics

Geopolitical considerations have played a major role in many energy policy decisions, e.g., in domestic debates over gasoline taxes, pipeline construction, radioactive waste disposal, and acid rain control legislation in the United States, and in petroleum-related violence in Nigeria. The most prominent role for geopolitics in energy policy has probably involved international discussions on controlling greenhouse gas emissions, and in oil markets. In the cases of the Kyoto Protocol of 1997 and the 1992 Framework Convention on Climate Change, nations carefully considered their national economic interests, domestic politics, and international trade during the negotiations. European countries, with the lowest rates of population and economic growth along with strong domestic environmental lobbies, have pursued a greater rate of greenhouse gas reduction. The United States, in contrast, has been stubbornly cautious and backed out of the treaty in 2001 (arguing it is not in its economic best interests), and the oil-rich nations of the Middle East have been least supportive of any emissions controls. In the case of oil markets, with the United States now dependent on imports for over half its supply, energy policy and trade strategy have played major roles in the pursuit of new oil discoveries in Alaska and in warfare in Kuwait, Iraq, and perhaps Afghanistan.

1.8 Technological Change and Energy Efficiency

Although technological change (usually involving an improvement in energy efficiency) is not inherently a geographic process, it does not occur uniformly over geographic space. Consequently, mathematical modeling and behavioral and innovation diffusion studies by geographers have improved our understanding of the pervasive phenomenon of technological change in energy markets. Similarly, creative government policies in Germany since 1990 have led to a faster rate of adoption of renewable energy sources, compared to the United States. Other studies have helped to explain why Japan, Italy, Germany, and Scandinavian countries have lower levels of energy consumption per capita or per unit of gross domestic product (GDP) than is the case in the United States and Canada. The answer to this question is a complex mix of price, government incentives and standards, economic structure, education, and life-style factors.

2. FOSSIL FUELS

2.1 Oil

The economic geography of the world oil industry has been studied in detail since the 1963 seminal study by Peter Odell. This study provided an analysis of the factors that contributed to the spatial distribution of the production, refining, transport, distribution, and consumption of oil. One interesting comparison is that although the share of proved crude oil reserves located in the Middle East at the time of the study was 60%, the share has risen only slightly since then, to about 65% by 2002. The top five nations of the world in terms of crude oil reserves are all in the Middle East, with the United States a distant eighth (Table I). Given these patterns, it is not surprising that the Organization of Petroleum Exporting Countries (OPEC) accounted for 39% of the world's crude oil production in 2002, with 26% coming out of the Persian Gulf. Saudi Arabia was the leading producer at 8.4 million barrels/day, followed by Russia and the United States. Before the rise of OPEC in the 1960s, the majority of global production was controlled by multinational oil companies. Although OPEC's production share peaked at 55% in 1973 and steadily declined following the OPEC oil embargo of that year to a low of 29% in 1985, it has gradually risen again since then to the current 39%

TABLE 1

Proved Reserves of Fossil Fuels in Selected Countries in 2002^a

Nation	Oil (10 ⁹ barrels)	Gas (10 ¹² m ³)	Coal (10 ⁶ short tons)
Saudi Arabia	261.8	6.36	0
Iraq	112.5	3.11	0
United Arab Emirates	97.8	6.01	0
Kuwait	96.5	1.49	0
Iran	89.7	23.00	1885
Venezuela	77.8	4.19	479
Russia	60.0	47.57	157,010
United States	30.4	5.19	249,994
Libya	29.5	1.31	0
Mexico	12.6	0.25	1211
China	18.3	1.51	114,500
Qatar	15.2	14.40	0
India	5.4	0.76	84,396
Australia	3.5	2.55	82,090
Germany	0.3	0.34	66,000

^aData from the British Petroleum Company (2003) and the Energy Information Administration (2003).

level. Oil prices have been remarkably stable since 1974, except for the period from 1980 to 1986 following the Iranian Revolution and during the Iran–Iraq War.

The global patterns in crude oil refining, consumption, and trade illustrate how integrated the oil markets have become in recent decades. In the case

of refining, the United States is by far the global leader. It refined more than 15 million barrels/day in 2001, accounting for 21.6% of the world total. Europe as a whole (not counting the European portion of the former Soviet Union) was not far behind at 21.2%, though capacity is split widely among Germany, Italy, France, the United Kingdom, Netherlands, Spain, and others. The Middle East accounted for 8% of the global refining throughput in 2001, led by Saudi Arabia and Iran. South America and Central America were next at 7.9%, led by Brazil and Venezuela. The next largest refiner was the former Soviet Union at 7% (mostly Russia), though its capacity has fallen by about 30% since 1991. Other leading oil refiners include China and Japan. Global patterns of oil consumption are more geographically concentrated than is refining, with the United States, Europe (except for the former Soviet Union), Russia, Japan, and China accounting for 64% of the total.

The global trade pattern in crude oil and refined petroleum products is shown in Fig. 1. The large production coming out of the Middle East, which accounts for more than 40% of the total trade flow, is marketed to Japan, the Asia Pacific region, Europe, and the United States. The United States made a concerted effort after the 1973 OPEC embargo to be less dependent on the OPEC supply, but it actually increased its foreign crude oil and refined product import dependence over this period, going from 36% in 1973 to over 60% in 2002. The former Soviet Union is also a major oil trader, with most of its exports destined for refineries in western Europe.

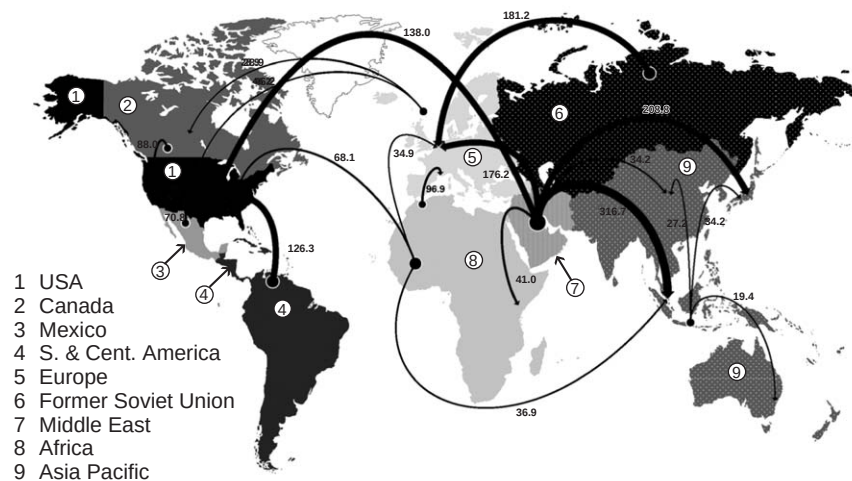


FIGURE 1 Major trade movements of oil (million tonnes). From the British Petroleum Company (2002), with permission.

Other significant oil shipment routes include Venezuela to the United States, Algeria and Libya to southern Europe, and Nigeria to the United States and the Asia Pacific region (Fig. 1).

Synthetic forms of oil supplies also can be developed. Although the vast technical potentials of tar sands and shale oil have been known for decades, these resources are beset by economic, environmental, and political problems and the industries are only in their infancy. Tar sands reserves, which are more properly known as bitumen or petroleum with very high viscosity, low resource quality, and low energy content, are concentrated in northern Alberta, Canada, where total production through surface and *in situ* mining amounted to just 30 million tonnes in 1999; an additional 3.3 million tonnes was produced in Venezuela. Canada has 979 million tonnes of proved recoverable reserves of synthetic crude oil or raw bitumen; Venezuela has 373 million tonnes. This compares with an additional 4231 million tonnes of additional reserves of foreseeable economic interest in the United States (half of the total is located in Utah and Alaska). The tar sands reserves of the United States and Canada are slightly larger than their oil reserves, whereas the tar sands reserves are much smaller than oil reserves in Venezuela.

The ultimate development potential of oil shales, if development proves to be economically and environmentally sound, is much greater than that of tar sands. The term “shale oil” is commonly used, although it is a misnomer because the organic material is chiefly kerogen and the hard rock it is usually found in is marl. This alternative oil resource can also be mined by conventional or *in situ* methods. The United States has by far the largest quantity of proved recoverable reserves of shale oil, at 60,000–80,000 million tonnes of oil equivalent. A small commercial facility was founded in Autun, France in 1839. Other nations that have produced shale oil include Canada, Scotland, Russia, China, Brazil (195 thousand tonnes in 1999), Estonia (151 thousand tonnes in 1999), and Australia (5 thousand tonnes in 1999, though production has grown since then).

2.2 Natural Gas

The spatial patterns of the world natural gas industry are somewhat different than those for oil, because gas has not been explored for and developed with the same vigor, is much less convenient to store, and is much more expensive to ship because of its low energy density (especially internationally). Geographic studies of this industry have concentrated

on its production potential and have modeled its expanded development, transport via pipeline, and trade, especially in North America, Russia, and the rest of Europe.

The gas reserve base is geographically concentrated, with Russia, Iran, and Qatar accounting for 55% of the world total in 2002 (Table I). Other leading holders of gas reserves are Saudi Arabia, United Arab Emirates, the United States, Algeria, and Venezuela. The gas reserve base does not include coal bed methane, the reserve of which may be equal to or greater than that of conventional world gas in Canada and Russia alone. The United States and Russia are the only countries that account for more than 10% of the share of global gas production, each at about 22%. United States production totaled 555 billion m³ in 2001; Russia produced 542 billion m³. Canada was a distant third, at 172 billion m³/yr. Although Russia used to dominate global output, it has alternated places with the United States as global leader since 1997.

Gas production figures, of course, reflect a nation's consumption level and involvement in international trade. The United States, the world's leading gas consumer, accounts for more than one-fourth of global consumption and is thus a slight importer, mostly from Canada (Fig. 2). Russia, conversely, only uses 15.5% of the world's gas output, and is a net exporter, to Europe and Turkey. European gas consumption, led by the United Kingdom, accounts for nearly 20% of global output. The total volume of liquefied natural gas (LNG) in world trade was 143 billion m³ in 2001, or one-third the level of movement by pipeline in gaseous form. The leading shippers of LNG in recent years have been Indonesia, Algeria, Malaysia, and Qatar. About half of this LNG is sold to Japan, with another seventh of the total going to South Korea.

2.3 Coal

Economic geography studies of the coal industry began in the 1930s. The early work focused on the location of coal seams, local economic development, and transportation to markets in the United States and Europe. Geographic research since the 1970s has been more analytical and has often used mathematical programming models, because such studies have emphasized optimal coal development and shipment patterns, power plant siting, the response of the industry to technological change, growing environmental constraints, and electric utility deregulation. In the United States, e.g., coal mining has gradually

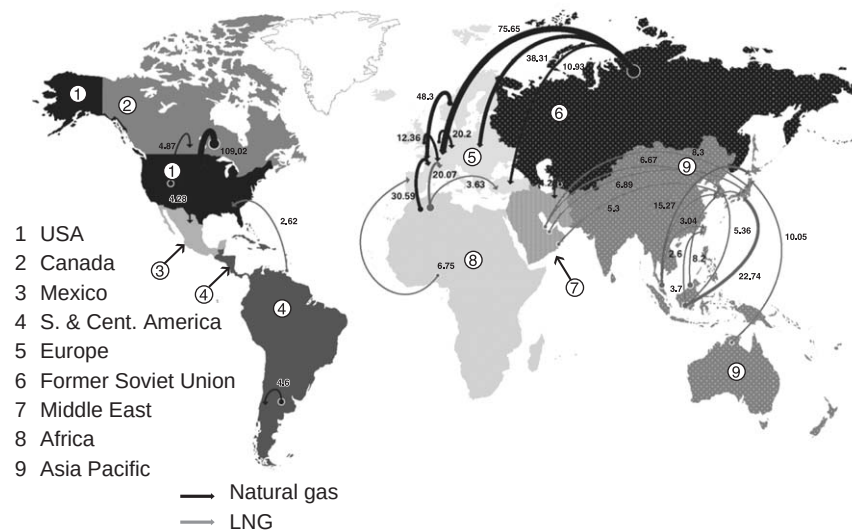


FIGURE 2 Major trade movements of natural gas (billion cubic meters). From the British Petroleum Company (2002), with permission.

shifted westward, because the economic advantages of that region's extensive near-surface deposits and lower sulfur content have increasingly outweighed the greater distances from electric power plants and other markets. European coal mining in the United Kingdom, Germany, Poland, and Russia, in contrast, has felt the greater brunt of acid rain and global warming restrictions, along with declining economics.

The coal reserve base is even more geographically concentrated than that of natural gas, with the United States, Russia, China, India, Australia, and Germany accounting for 77% of the world total in 2002 (Table I). Currently, only the United States, India, and Australia seem intent on using most of this resource base. Although Russian coal production has fallen by about half since the collapse of the former Soviet Union in 1991, the status of China is more uncertain. China led the world in coal production at 1545 million short tons in 1996, which fell to 1314 million short tons in 2000 as part of national policy to control air pollution and improve production efficiency (Germany implemented similar policy), but grew again in 2001 by more than 9%. Regarding global climate change, given the conflicting and unsettled policies of the United States, China, India, and Australia, the disposition of the huge coal resource base remains highly uncertain. The United States and China alone account for almost half of the world's coal consumption.

3. NUCLEAR POWER

The development of civil nuclear power was a stepchild of the military program of the United States to end World War II and of U.S. President Dwight D. Eisenhower's Atoms for Peace Speech to the United Nations General Assembly in 1953. The first commercial plants were commissioned in the United Kingdom, United States, and Russia in the late 1950s, and these nations actively developed this industry in the 1960s. Following the OPEC oil embargo in 1973, France and Japan developed major nuclear programs as well as strategies for energy autarky.

Significant political opposition to nuclear power plants and technology began to surface in the middle to late 1970s, first in the United States and later in several western European nations. Public opposition to nuclear power even spread to Taiwan in the 1990s. The opposition focused on the risk of accidents and radiological contamination, a prospect made worse by reactor sitings close to human population centers in New York, Boston, Philadelphia, Chicago, and coastal United Kingdom. In the aftermath of the accident at the Three Mile Island Unit 2 reactor near Harrisburg, Pennsylvania on March 28, 1979, it was clear that evacuation plans and procedures were inadequate in most nations (though they were improved thereafter).

The United States continued its global lead in nuclear power generation, with 104 nuclear reactor

units and 98.5 gigawatts (electrical; GWe) of installed capacity in 2003 (Table II). France leads large countries in reliance on nuclear power, which accounted for three-fourths of its electricity generation in 2002. Although the commitment to nuclear power and the location of the world's 441 nuclear reactors as of September 2003 (with 362 GWe of generating capacity) was widely distributed, most facilities outside of North America were located in Europe. Of the top 12 nuclear power users, 8 are European nations, with Japan and South Korea also making this list (Table II). The commitment to new reactor sitings and development, however, has stalled. Since 1995, opening of new nuclear power plants occurred only in France, the Czech Republic, Brazil, and especially in Asia (though a few others countries, such as the United States, have continued to expand their output as the capacity factor of many nuclear reactors has increased). There has been a *de facto* moratorium on new nuclear power in most of the world since 1980, with phaseout decisions pronounced in Sweden, Germany, Switzerland, Spain, Belgium, and the Netherlands.

In addition to heightened concern about nuclear proliferation and reactor safety in the 21st century, the nuclear power industry has been stifled by the twin Achilles' heels of reactor decommissioning and radioactive waste disposal. Although disposal sites have been in operation for shorter-lived, low- and medium-level wastes in several countries for decades,

most attention has been given to the development of long-term, deep-geologic-disposal sites for high-level wastes (spent fuel rods and vitrified reprocessing wastes). Most progress on this front has been made in the United States and Scandinavia. The Finnish Parliament in May 2001 approved a plan in principle for deep-geologic disposal, though the repository would not be built until the 2010s and would not open until 2020. The disposal site is expected to be near the Olkiluoto nuclear plant in the town of Eurajoki on the west coast (a low-level waste site already exists there). In Sweden, repository siting is at an advanced stage and has been narrowed down to the Oskarshamn and Forsmark nuclear plant sites. A low- and medium-level waste site has existed near Forsmark since 1988 in offshore granite bedrock, 60 m below the Baltic Sea. In the United States, the Waste Isolation Pilot Plant in southeastern New Mexico opened on March 26, 1999 after a delay of 25 years; it accepts military transuranic and mixed hazardous wastes in bedded salt caverns 650 m underground. A high-level waste site in volcanic tuffs at Yucca Mountain, Nevada, 100 miles northwest of Las Vegas (also opposed and delayed), was approved by the U.S. Senate on July 9, 2002 and could open by 2010. Even so, scientific problems continue to be raised at the U.S. waste sites.

4. RENEWABLE ENERGY

The major renewable sources of energy include hydroelectricity, biomass, wind, solar, and geothermal. Renewable energy sources account for over one-fifth of total world energy use. Geothermal energy is categorized by federal statute in the United States as renewable, but strictly speaking, it is not, because it can be depleted at specific sites. All renewable energy resources can be used to generate electricity. In addition, biomass, solar, and geothermal energies are used for heating and cooking, and biomass is also used for transportation (as ethanol or methanol).

4.1 Hydroelectricity

Hydroelectric power is the largest and cheapest renewable electricity source in the world, although biomass easily surpasses it when other energy applications are considered. This is not to say that hydro power is without adverse environmental impacts, however. Large-scale dams are well known to lead to significant ecological damages, including loss of biodiversity, localized deforestation, soil

TABLE II

Top 12 Nuclear Power Producers in 2003^a

Nation	Nuclear reactors	Capacity (GWe) ^b	Share of electricity (%)
United States	104	98.5	20
France	59	63.3	78
Japan	53	44.1	39
Germany	19	21.1	30
Russia	30	20.8	16
South Korea	18	14.9	39
United Kingdom	27	12.1	22
Canada	16	11.3	13
Ukraine	13	11.2	46
Sweden	11	9.5	46
Spain	9	7.4	26
Belgium	7	5.7	58

^aData from the World Nuclear Energy Association (2003); available on the Internet at <http://www.uic.com.au/reactors.htm>.

^bGWe, gigawatts (electrical).

erosion, and disruption of free-flowing rivers, as well as displacement of large numbers of humans in many cases. Even so, hydroelectric power releases no direct greenhouse gas emissions and is a significant source of power in many parts of the world.

The United States, China, Canada, and Brazil are the leading producers of hydroelectricity (Table III). Brazil and Canada are the most dependent on hydro power, which accounted for 89 and 61%, of the electricity used in these two countries, respectively, in 2001, although persistent drought and rationing led to power cutbacks in Brazil in that year. Overall, Brazil relies on renewable energy for more than half of its total energy supply. Yet some small nations are even more dependent on renewable sources of electricity. Norway and Iceland, for instance, rely on renewable energy for 99% of their electricity consumption. In recent years, there have been few new large-scale hydroelectric power projects in developed nations because the best dam sites have generally already been tapped. A major exception to this pattern is the 2000-megawatt (MW) Lower Churchill Project on Gull Island in Newfoundland,

Canada. In contrast, there have been several highly publicized dam removals and downstream ecological restoration projects in the northern United States. These trends, plus occasional drought, reduced the reliance of the United States on hydro power for electricity to 6% in 2001, down from a peak of 356 million MW-hr in 1997.

Large hydro projects can be found in many developing countries, such as China, India, Malaysia, Vietnam, Brazil, and Azerbaijan. China is building the controversial 18,200-MW Three Gorges Dam Project (the largest in the world) on the Yangtze River; the project is expected to reach full capacity by 2009. More than 1 million people will have to be relocated before the dam's reservoirs can be fully flooded. India, with 150-GWe of hydroelectric potential, has begun to focus on expanding this renewable source of energy, but not without difficulty in obtaining financing and in addressing protests from environmental and human rights groups. Most attention has focused on the Narmada Valley Development Project, which is planned to include up to 30 large dams in addition to numerous smaller ones.

TABLE III

Power Capacity From Hydro, Wind and Geothermal in Selected Countries, 2003^a

Nation	Gigawatts (electrical)		
	Hydroelectric	Wind	Geothermal
United States	98.9	5.3	2.8
China	70.0	0.4	— ^b
Canada	67.2	0.2	0
Brazil	59.0	— ^b	0
Russia	43.4	— ^b	— ^b
Norway	27.4	— ^b	0
India	24.5	1.7	0
Japan	21.5	0.3	0.5
Italy	13.1	0.7	0.8
Spain	11.5	5.0	0
Mexico	9.6	0.3	0.9
Germany	3.0	12.8	0
Philippines	2.3	— ^b	1.9
Denmark	— ^b	2.8	0

^aData from the Energy Information Administration (EIA) (2003); the EIA, Renewable Energy Annual 2002, DOE/EIA-0603(2002), Washington, D.C., November 2002; the International Energy Agency, Wind Energy Annual Report 2003, Paris; and the International Geothermal Association, 2002, available on the Internet at <http://iga.igg.cnr.it/electricitygeneration.php>.

^bLess than 100 MW.

4.2 Wind and Geothermal

Wind power has been the fastest growing energy source in the world since it became more cost-effective in the 1990s, though it has a long way to go to catch up with hydro power. Leading the way is Germany, supported by its 1990 Electricity Feed Law that requires electric utilities to pay 80–90% of the retail rate for wind power purchases. Germany passed the United States as world leader in wind power development in 1997, and now has almost 13 GWe in capacity (Table III). Other leading wind power producers include Spain, Denmark, and India, though Italy, the United Kingdom, Ireland, and several other European nations are rapidly expanding their wind power programs, including at many offshore sites. Denmark stands out as being most dependent on wind power, which accounts for 20% of that nation's electricity supply.

The two leading developers of geothermal energy are the United States and the Philippines, with most of their facilities being committed to electric power generation and much less to direct use (Table III). In the case of the Philippines, geothermal energy accounted for 21.5% of the nation's electricity supply in 2000. Other leading producers include Mexico, Italy, Indonesia, Japan, and New Zealand. Among these nations, the United States, Mexico,

Indonesia, and the Philippines have the largest expansion plans for geothermal energy facilities.

4.3 Biomass and Solar

Biomass energy in its various forms and uses (firewood, charcoal, crop residues, alcohol fuels, municipal and animal wastes, electricity) accounts for one-seventh of the world's primary energy supplies and over 90% in some developing nations. The United States is the leading electric power producer, with 10.1 GWe of capacity from biomass fuels (wood and wastes). Over half of this is cogenerated, mostly from the paper industry. This makes biomass the next largest source of renewable electricity in the United States after hydro power. The United States also has a significant (though subsidized) ethanol industry, which produced over 2 billion gallons in 2002, mostly from corn, with a refining capacity of over 3 billion gallons. Brazil has an even larger ethanol fuel program from sugarcane; production, which was subsidized, was between 3 and 3.5 billion gallons in recent years. Wood is a common source of biomass energy in Scandinavia, China, Pakistan, India, and sub-Saharan Africa. Unfortunately, much of the biomass energy data are incomplete or unreliable, rendering international comparison difficult. Consequently, most geographers working on this subject conduct case studies, especially of wood fuel use and sustainable development.

Solar energy has many applications. Most attention has been given to low-cost solar cookers, water and

space heaters, and photovoltaic cells, though California also 384 MW of solar power plant capacity. The price of photovoltaic modules declined a hundredfold from 1972 to 1992. The module price had fallen further to \$3.40/W and \$2.46/W by 2001 for low- and high-volume purchases, which makes them cost-effective in limited applications. Japan and Germany lead the market for photovoltaics, with over 400 and 250 MW, respectively. Japan is also the world's leading user of household solar thermal collectors. Israel is even more dependent on solar energy. Over 80% of Israeli households have solar water heaters (more than 1.3 million installations), and solar energy accounted for 3% of its total energy use in 2001.

5. INTERNATIONAL COMPARISONS OF ENERGY CONSUMPTION

International comparisons of energy consumption (per capita, per unit of GDP, etc.) are often used as indicators of energy efficiency and gluttony. On this basis, the United States, Canada, Russia, Japan, Australia, Saudi Arabia, and much of western Europe are the "haves," whereas China, India, Pakistan, Bangladesh, Indonesia, the Philippines, Vietnam, Brazil, and most of Africa are the "have nots" (Fig. 3). In some cases these differences in energy use per capita are more than 5:1. Such comparisons, although interesting, are flawed. Aggregate indicators of energy use per capita mask the real determinants of energy use. Decomposition of these patterns has



FIGURE 3 Energy consumption per capita by nation (tonnes of oil equivalent). From the British Petroleum Company (2002), with permission.

revealed the varying roles played by technological efficiency, structural economic change, and economic growth in determining energy consumption. Nonetheless, whatever the future energy trends and geographic patterns are in the world, such large inequities in energy use will not be ignored.

SEE ALSO THE FOLLOWING ARTICLES

Economics of Energy Demand • Economics of Energy Supply • Geographic Thought, History of Energy in • Global Energy Use: Status and Trends • Global Material Cycles and Energy • Natural Gas Resources, Global Distribution of • Oil and Natural Gas Liquids: Global Magnitude and Distribution

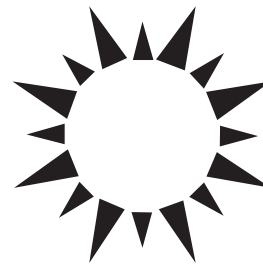
Further Reading

- British Petroleum Company. (2003). "BP Statistical Review of World Energy." British Petroleum Corporate Communication Services, London.
- Cleveland, C. J., and Kaufmann, R. K. (2003). Oil supply and oil politics: Deja Vu all over again. *Energy Policy* 31, 485–489.
- DiFrancesco, R. J., and Anderson, W. P. (1999). Developing Canada's Arctic oil resources. *Environ. Planning A* 31, 459–476.
- Elmes, G. A., and Harris, T. M. (1996). Industrial restructuring and the changing geography of the United States coal-energy system, 1972–1990. *Ann. Assoc. Am. Geogr.* 86(No. 3), 507–529.
- Energy Information Administration. (2003). International Energy Annual 2000. DOE/EIA-0219 (2001). U.S. Department of Energy, Washington, D.C.
- Kuby, M. (1996). "Population growth, energy use, and pollution: Understanding the driving forces of global change." Commission on College Geography 2. Association of American Geographers, Washington, D.C.
- Lakshmanan, T. R. (1983). A multiregional model of the economy, environment, and energy demand in the United States. *Econ. Geogr.* 59, 296–320.
- Luten, D. B. (1971). The economic geography of energy. *Sci. Am.* 225(No. 3), 109–117.
- Miller, E. W. (2002). Renewable energy's place in the global energy market. In "Renewable Energy: Trends and Prospects" (S. K. Majumdar, E. W. Miller, and A. I. Panah, eds.), pp. 1–18. Pennsylvania Academy of Science, Easton, Pennsylvania.
- Odell, P. R. (1963). "An Economic Geography of Oil." Praeger, New York.
- Schipper, L., Unander, F., Murtishaw, S., and Ting, M. (2001). Indicators of energy use and carbon emissions: explaining the energy economy link. *Annu. Rev. Energy Environ.* 26, 49–81.
- Smil, V. (1994). "Energy in World History." Westview Press, Boulder, Colorado.
- Smil, V. (2001). Geography of energy use. In "Macmillan Encyclopedia of Energy" (J. Zumerchik, ed.), Vol. 2, pp. 567–572. Macmillan Reference, New York.
- World Energy Council. (2003). WEC Survey of Energy Resources 2001. London. Available on the Internet at <http://www.worldenergy.org/wec-geis/publications/reports/ser/>. Last accessed January 2, 2003.

Economic Growth and Energy

DAVID I. STERN

Rensselaer Polytechnic Institute
Troy, New York, United States



-
1. Introduction
 2. Theory of Production and Growth
 3. Factors Affecting Linkage between Energy and Growth
 4. Empirical Testing
 5. Environmental Implications

Glossary

capital There are two meanings in economics. One is the monetary value of assets owned by an individual firm, etc. or the amount that is available to invest. The second meaning, used herein, refers to a human-made input into the production process, including things such as buildings and machinery, and infrastructure (telecommunications systems and the like). Capital in these forms is also described as reproducible capital, manufactured capital, or human-made capital. Some writers use the term “capital” more widely, to include human capital and/or environmental or natural capital. Human capital refers to expertise or know-how embodied in people through processes of education and training. The term “natural capital” has been used to refer to the stock of environmental resources.

demand elasticity The percentage response in the quantity of a good demanded for a 1% increase in its price.

diminishing returns If use of an input to production is increased, holding the amounts of the other inputs constant, the law of diminishing returns states that eventually the marginal product of an input will decline, though not necessarily becoming negative.

ecological economics A transdisciplinary field including ideas from economics, natural sciences, and other social sciences, in which scholars and practitioners are interested in the broad question of the use of resources by society and in understanding the nature, behavior, and dynamics of the economy–environment system. Many, but not all, ecological economists are highly critical of the mainstream (neoclassical) school of economics.

econometrics The specialized branch of mathematical statistics applied to the analysis of economic data.

economic growth An increase in economy-wide economic production, usually measured by an increase in gross domestic product (GDP); also, the process of the economy growing over time.

elasticity of substitution A measure of the responsiveness of the ratio of the quantities of two productive inputs to a change in the relative prices of the quantities. For example, if the price of energy increases 1% relative to the price of capital and the amount of energy used per dollar of capital decreases by 1%, then the elasticity of substitution is 1. If the decrease in energy use is less, then the elasticity of substitution is less than 1, and vice versa.

embodied energy Used directly in the production of goods and services, energy is also used indirectly through its involvement in producing the other inputs to production. The sum of the two uses is called the embodied energy of the good or service.

externality A negative externality, or external effect, occurs when a production or consumption activity has unintended damaging effects, such as pollution, on other firms or individuals, and no compensation is paid by the producer or consumer to the affected parties. A positive externality occurs when activities have beneficial effects for others who do not pay the generator of the effect, such as freely available research results.

institutions Commonly accepted rules, practices, etc. that govern human social and economic behavior, such as laws, customs, markets, etc.

marginal product If use of an input to production is increased, holding the amounts of the other inputs constant, the increase in output produced is called the marginal product of that input.

neoclassical economics The mainstream school or paradigm of economics characterized by mathematical models based on optimization by individual consumers and producers. A paradigm that arose in the late 19th century and emerged to its dominant position after the World War II; other paradigms include Keynesian, Marxist, Institutional, and neo-Ricardian, and post-Keynesian models.

nonrenewable resource A natural resource such as petroleum, which has zero natural growth over ordinary human timescales; though it may be replenished over geological time. Renewable resources such as forests

grow or regenerate at rates that are relatively fast on human timescales.

production function A mathematical representation of the quantitative relationship between productive inputs and the level of output of some good or service, expressing the maximum quantity of output that can be obtained from any particular amounts of inputs.

productivity An index of economic output relative to input. Increases in productivity are often seen as being due to improvements in technology but may also be due to other factors. The most general productivity index is known as total factor productivity. Labor productivity is just output per worker hour and is the measure usually discussed in the financial media.

public goods The consumption of a public good by one person, which does not reduce the amount available for others and once provided to some people, other people cannot be excluded from consuming them.

substitution Shifting the mix of inputs used to produce a given quantity of output. Substitution occurs among already known production methods and does not imply that innovations are introduced.

sustainability As used herein, nondeclining individual consumption or utility over time; many other definitions exist.

technological change The invention and introduction of new methods of production as well as new products; the main focus in economics and herein is on the former.

utility The level of satisfaction (or happiness) of an individual. In a more aggregate context, we refer to the social welfare of a group of people, which is an indicator related to the utilities of the individuals in society.

Resource and ecological economists have criticized the theory that energy plays a minor role in economic growth on a number of grounds, especially the implications of thermodynamics for economic production and the long-term prospects of the economy. Although a fully worked out alternative model of the growth process does not seem to exist, extensive empirical work has examined the role of energy in the growth process. The principal finding is that energy used per unit of economic output has declined, but that this is to a large extent due to a shift in energy use from direct use of fossil fuels such as coal to the use of higher quality fuels, and especially electricity. When this shift in the composition of final energy use is accounted for, energy use and the level of economic activity are found to be tightly coupled. When these and other trends are taken into account, the prospects for further large reductions in the energy intensity of economic activity seem limited. These findings have important implications for environmental quality and economic and environmental policy.

1. INTRODUCTION

This article surveys the relation between energy and economic growth, and more generally the role of energy in economic production. Business and financial economists pay significant attention to the impact of oil and other energy prices on economic activity, but the mainstream theory of economic growth pays little or no attention to the role of energy or other natural resources in promoting or enabling economic growth. The extensive discussions concerning the “productivity slowdown” following the 1970s oil crises were an exception.

This article is structured to cover these key points in a systematic fashion. In the first section there is a review of the background theory of production and growth from different points of view—those based in economics and those based in the natural sciences. The starting premise is that gaining an understanding of the role of energy in economic growth cannot be achieved without first understanding the role of energy in production. The scientific basis of the role of energy in production, and hence also in the increasing scale of production involved in economic growth, is considered first. However, institutional phenomena also affect how this role plays out and therefore the economic view of growth and production and of the potential role of energy is necessarily more complex than just this scientific understanding. The mainstream theory of economic growth is, therefore, reviewed next. The limitations of its consideration of energy and other resource issues have been the subject of strong criticism grounded in the biophysical theory of the role of energy. A review of these alternative viewpoints completes this first section of the article.

The mainstream economics concept of the production function is next used to examine the key factors that could reduce or strengthen the linkage between energy use and economic activity over time. This production theory is very general and is less subject to criticism than are the specific models of economic growth. These key factors are (1) substitution between energy and other inputs within an existing technology, (2) technological change, (3) shifts in the composition of the energy input, and (4) shifts in the composition of economic output. Each of these themes will be discussed in detail.

Numerous ideas and views exist about the potential linkages between energy and economic growth. Choice between these theories has to be on the basis of both inherent plausibility and consistency and, perhaps more crucially, empirical

evidence. Therefore, studies that investigate the strength of the linkage between energy and growth are reviewed. To be useful, such studies must not be grounded in a single theory, potential mechanism, or school of thought. Therefore, the studies reviewed here do not specify structural linkages between energy and output. Because correlation and regression analysis does not imply causality from one variable to another, most of these studies employ the econometric notions of Granger causality and cointegration to test the presence of and direction of causality between the variables.

Finally, the implications of the theory for environmental quality are examined and the recent empirical literature on that topic is discussed.

2. THEORY OF PRODUCTION AND GROWTH

2.1 Energy in Production: Physical Theory and Economic Models

Reproducibility is a key concept in the economics of production. Some inputs to production are nonreproducible, whereas others can be manufactured at a cost within the economic production system. Primary factors of production are inputs that exist at the beginning of the period under consideration and are not directly used up in production (though they can be degraded and can be added to), whereas intermediate inputs are created during the production period under consideration and are used up entirely in production. Mainstream economists usually think of capital, labor, and land as the primary factors of production, and goods such as fuels and materials as the intermediate inputs. The prices paid for all the different inputs are seen as eventually being payments to the owners of the primary inputs for the services provided directly or embodied in the produced intermediate inputs. In the theory of growth, this approach has led to a focus on the primary inputs, in particular on capital and land, and a much lesser and somewhat indirect treatment of the role of energy in the growth process. The primary energy inputs are stock resources such as oil deposits. But these are not given an explicit role in the standard growth theories, which focus on labor and capital. However, capital, labor, and, in the longer term, even natural resources are reproducible factors of production, whereas energy is a nonreproducible factor of production, although, of course, energy vectors (fuels) are reproducible factors. Therefore,

natural scientists and some ecological economists have placed a very heavy emphasis on the role of energy and its availability in the economic production and growth processes.

The first law of thermodynamics (the conservation law) implies the mass-balance principle. In order to obtain a given material output, greater or equal quantities of matter must enter the production process as inputs, with the residual as a pollutant or waste product. Therefore, there are minimum material input requirements for any production process producing material outputs. The second law of thermodynamics (the efficiency law) implies that a minimum quantity of energy is required to carry out the transformation of matter. All production involves the transformation or movement of matter in some way, and all such transformations require energy. Therefore, there must be limits to the substitution of other factors of production for energy. Energy is also an essential factor of production. Though all economic processes require energy, some service activities may not require the direct processing of materials. However, this is only true at the micro level. At the macro level (economy-wide level), all economic processes require the indirect use of materials, either in the maintenance of labor or in the production of capital goods.

Some aspects of organized matter—that is, information—might also be considered to be nonreproducible inputs. Several analysts argue that information is a fundamentally nonreproducible factor of production in the same way as energy. Energy is necessary to extract information from the environment, but energy cannot be made active use of without information and possibly accumulated knowledge. Unlike energy, information and knowledge cannot be easily quantified. But these latter factors of production must be incorporated into machines, workers, and materials in order to be made useful. This provides a biophysical justification for treating capital, labor, etc. as factors of production. Though capital and labor are easier to measure than information and knowledge, their measurement is, still, very imperfect compared to that of energy.

In the mainstream neoclassical economics approach (discussed later), the quantity of energy available to the economy in any period is endogenous, though restricted by biophysical constraints (such as the pressure in oil reservoirs) and economic constraints (such as the amount of installed extraction, refining, and generating capacity), and the possible speeds and efficiencies with which these processes can proceed. Nevertheless, this analytical approach leads

to a downplaying of the role of energy as a driver of economic growth and production. Some alternative, biophysical models of the economy propose that energy is the only primary factor of production. This could be understood as there being a given stock of energy that is degraded (but, due to the law of the conservation of energy, not used up) in the process of providing services to the economy. But this means that the available energy in each period needs to be exogenously determined. In some biophysical models, geological constraints fix the rate of energy extraction. Capital and labor are treated as flows of capital consumption and labor services rather than as stocks. These flows are computed in terms of the embodied energy use associated with them and the entire value added in the economy is regarded as the rent accruing to the energy used in the economy. The actual distribution of the surplus depends on the relative bargaining power of the different social classes and foreign suppliers of fuel. Therefore, the owners of labor, capital, and land appropriate the energy surplus. The production process in the economy as a whole can be represented as an input–output model as originally developed by Nobel Laureate Wassily Leontief and adapted in the energy context by Bruce Hannon and others. The input–output model specifies the required quantities of each input needed to produce each output, with each output potentially an input to another process. It is also possible to trace all final production back to its use of a single primary factor of production, in this case energy.

These ecological economists argue that the energy used to produce intermediate resource inputs such as fuels increases as the quality of resources (such as oil reservoirs) declines. This is like negative productivity growth or technological change. Thus, changing resource quality is represented by changes in the embodied energy of the resources rather than by changes in the input–output coefficients. If resource stocks were explicitly represented, energy would no longer be the only primary factor of production. The neo-Ricardian models developed by Charles Perrings and Martin O'Connor, like all other neo-Ricardian models, have a fixed-proportions technology in terms of capital stocks instead of the flows in the Leontief model. They do not distinguish between primary and intermediate factors of production. Yet that approach can still take biophysical constraints such as mass balance and energy conservation into account.

If the economy can be represented as an input–output model where there is no substitution between factors of production, the embodied knowledge in factors of production can be ignored. An accurate

accounting for all energy used to support final production is important. But in the real world the contribution of knowledge to production cannot be assumed to be proportional to its embodied energy. Though thermodynamics places constraints on substitution, the actual degree of substitutability among capital stocks embodying knowledge and energy is an empirical question. Neither the Leontief nor the neo-Ricardian model allows substitution between inputs. The neoclassical production models that are considered next, however, do.

2.2 The Mainstream Theory of Growth

As already explained, there is an inbuilt bias in mainstream production and growth theory to downplay the role of resources in the economy, though there is nothing inherent in economics that restricts the potential role of resources in the economy. The basic model of economic growth is the Nobel prize-winning work by Robert Solow that does not include resources at all. This model subsequently was extended with nonrenewable resources, renewable resources, and some waste assimilation services. These extended models, however, have only been applied in the context of debates about environmental sustainability, not in standard macroeconomic applications.

2.2.1 The Basic Growth Model

Economic growth models examine the evolution of a hypothetical economy over time as the quantities and/or the qualities of various inputs into the production process and the methods of using those inputs change. Here, the simplest model based on the work of Solow is described. In this model, a constant-sized labor force using manufactured capital produces output in terms of gross domestic product (GDP). The model assumes that output increases at a decreasing rate as the amount of capital employed rises. The uppermost curve in Fig. 1 shows this relationship between output (Y) and capital (K). This is termed diminishing returns to capital. It is assumed that a constant proportion, s , of the output is saved and invested in the capital stock. A constant proportion of the existing capital stock depreciates (becomes productively useless) in each period of time. The capital stock is in equilibrium (and so unchanging in size) when saving equals depreciation. This is also shown in Fig. 1. The dynamics implied by Fig. 1 are very simple. To the left of K^* , where capital per worker is scarce, capital investment generates a relatively large increase in future income, and so will

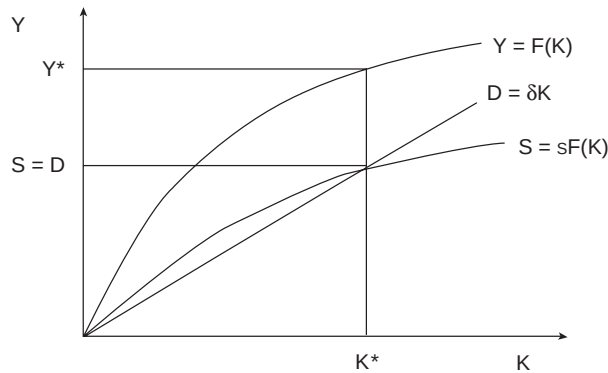


FIGURE 1 The neoclassical growth model.

offer a high rate of return. Moreover, it is clear from the relative positions of the S and D curves to the left of K^* that the addition to the capital stock (S) is greater than depreciation (D) and so capital rises. When the capital stock has reached K^* , it will be at a stationary, or equilibrium, state. Additions to capital due to saving are exactly offset by reductions in capital from depreciation.

This simple economy must sooner or later reach a stationary state in which there is no net (additional) investment and economic growth must eventually halt. In a transition process, while a country is moving toward this stationary state, growth can and will occur. An underdeveloped economy, with a small capital stock per worker, can achieve fast growth while it is building up its capital stock. But all economies will eventually settle into a zero growth equilibrium if the savings rate remains constant. No country can grow in perpetuity merely by accumulating capital, due to the diminishing returns to capital accumulation. If the savings rate is increased, growth will occur for a while until a new equilibrium is reached. Of course, the higher the savings the rate, the lower the current standard of living of the population.

If the labor force grows at a fixed rate over time, the total capital stock and the total quantity of output will rise, but capital per worker and output per worker will remain constant once an economy has developed to its equilibrium. The only necessary adjustment to Fig. 1 is that all units are now measured in per capita terms.

According to neoclassical growth theory, the only cause of continuing economic growth is technological progress. As the level of technological knowledge rises, the functional relationship between productive inputs and output changes. Greater quantities or better qualities of output can be produced from the

same quantity of inputs. In the simple model being examined, technological progress continually shifts the output function upward, and so raises the equilibrium per capita capital stock and output levels. Intuitively, increases in the state of technological knowledge raise the rate of return to capital, thereby offsetting the diminishing returns to capital that would otherwise apply a brake to growth.

2.2.2 Endogenous Technological Change

The simple model just described does not explain how improvements in technology come about. They are just assumed to happen exogenously, so that this model is said to have exogenous technological change. More recent models attempt to endogenize technological change, explaining technological progress within the growth model as the outcome of decisions taken by firms and individuals. In endogenous growth models, the relationship between capital and output can be written in the form $Y = AK$. Capital, K , is defined more broadly than in the neoclassical model. It is a composite of manufactured and knowledge-based capital. Endogenous growth theorists have been able to show that, under reasonable assumptions, the A term in the preceding expression is a constant, and so growth can continue indefinitely as capital is accumulated.

The key point is that technological knowledge can be thought of as a form of capital. It is accumulated through research and development (R&D) and other knowledge-creating processes. Technological knowledge has two special properties. First, it is a public good: the stock of this form of capital is not depleted with use. This is important because it implies that the knowledge stock can be stored over time, even when it is being used. Second, it generates positive externalities in production: although the firm doing R&D obtains benefits from the knowledge acquired, others benefit too—the benefits that the firm accrues when it learns and innovates are only partly appropriated by itself. There are beneficial spillovers to the economy from the R&D process so that the social benefits of innovation exceed the private benefits to the original innovator. These externalities create momentum in the growth process. As firms install new capital, this tends to be associated with process and product innovations. The incentive to devote resources to innovation comes from the prospect of temporary monopoly profits for successful innovations. The growth of K thus means the growth of a composite stock of capital and disembodied technological knowledge. Therefore, output is able to rise as a constant proportion (A) of the

composite capital stock, and is not subject to the diminishing returns shown in Fig. 1.

So, in an endogenous growth model, the economy can sustain a constant growth rate in which the diminishing returns to manufactured capital are exactly offset by the technological growth external effect just described. The growth rate is permanently influenced by the savings rate; a higher savings rate increases the economy's growth rate, not merely its equilibrium level of income.

2.2.3 Growth Models with Natural Resources

The growth models examined so far do not include any natural resources. All natural resources exist in finite stocks or flows, though some, such as sunlight or deuterium, are available in very large quantities. Some environmental resources are nonrenewable, and many renewable resources are potentially exhaustible. Finiteness and exhaustibility of resources make the notion of indefinite economic growth problematic. Even maintaining the current level of output indefinitely—so-called sustainable development—may be difficult.

When there is more than one input (both capital and natural resources), there are many alternative paths that economic growth can take. The path taken is determined by the institutional arrangements that are assumed to exist. Analysts have looked at both optimal growth models that attempt to either maximize the sum of discounted social welfare over some relevant time horizon (often an infinite horizon) or achieve sustainability (nondeclining social welfare) and models intended to represent real economies with perfectly competitive markets or other arrangements.

The neoclassical literature on growth and resources centers on what conditions permit continuing growth, or at least nondeclining consumption or utility. Technical and institutional conditions determine whether such sustainability is possible. Technical conditions refer to things such as the mix of renewable and nonrenewable resources, the initial endowments of capital and natural resources, and the ease of substitution among inputs. The institutional setting includes things such as market structure (competition versus central planning), the system of property rights (private versus common property), and the system of values regarding the welfare of future generations.

In a 1974 paper published in a special issue of *Review of Economic Studies*, Solow showed that sustainability was achievable in a model with a finite and nonrenewable natural resource with no extrac-

tion costs and nondepreciating capital, which was produced using capital and the natural resource. However, the same model economy under competition results in exhaustion of the resource and consumption and social welfare eventually falling to zero. In another paper in the same special journal issue, Dasgupta and Heal showed that with any constant discount rate the so-called optimal growth path also leads to eventual depletion of the natural resource and the collapse of the economy. A common interpretation of this theory is that substitution and technical change can effectively decouple economic growth from resources and environmental services. Depleted resources or degraded environmental services can be replaced by more abundant substitutes, or by “equivalent” forms of human-made capital (people, machines, factories, etc.). But this is a misinterpretation. Neoclassical economists are primarily interested in what institutional arrangements, and not what technical arrangements, will lead to sustainability, so that they typically assume *a priori* that sustainability is technically feasible, and then investigate what institutional arrangements might lead to sustainability if it is technically feasible. However, there is a tendency among mainstream economists to assume that sustainability is technically feasible unless proved otherwise.

The elasticity of substitution (σ) between capital (factories, machines, etc.) and inputs from the environment (natural resources, waste assimilation, ecosystem services) is a critical technical term that indicates by how much one of the inputs must be increased to maintain the same level of production, when the use of the other input is reduced. A large value of σ implies that this substitution is easy, and vice versa. Figure 2 shows the different combinations of two inputs that can produce a given level of output for different values of σ . Different levels of output were chosen for the three values of σ to make the figure clearer.

The marginal product is the additional contribution to production of using one more unit of an input, holding the levels of the other inputs constant. A unitary elasticity of substitution ($\sigma = 1$), referred to as “perfect substitutability,” means that as the ratio of the two inputs is changed by a given percentage, holding output constant; the ratio of their marginal products changes by the same percentage (in the opposite direction). This relation is shown by the curve (known as an isoquant) in Fig. 2, which is asymptotic to both axes. As resource use decreases toward zero, production can be maintained by increasing capital use toward infinity. Additionally,

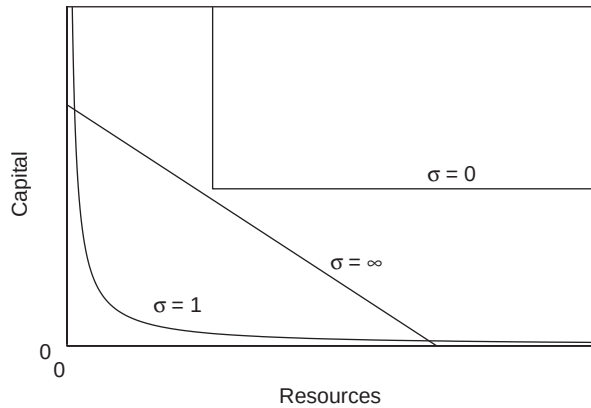


FIGURE 2 The elasticity of substitution.

the total cost of production is constant along the isoquant. Perfect substitutability does not mean that resources and capital are equivalently useful—that case is termed “infinite substitutability” ($\sigma = \infty$). In the latter case, producers see no difference between the two inputs and use the cheapest one. Whenever the elasticity of substitution is greater than 1, the isoquants cross the axes and inputs are nonessential for production. The figure also illustrates the case where no substitution is possible ($\sigma = 0$). In this case, the two inputs must be used in a fixed ratio. As is discussed later, perfect substitutability is an unrealistic assumption from a biophysical perspective, at least if it is assumed to apply at all possible ratios of capital and resources. Demand elasticities for energy, which in theory are related to the elasticity of substitution, also indicate that the elasticities of substitution between energy and other inputs and among different fuels may be between 0 and 1.

Economists such as Solow explicitly dispose of cases where σ for nonrenewable resources and capital is greater or less than unity. In the former case, substitution possibilities are large and therefore the possibility of nonsustainability is not an issue. In the latter case, sustainability is not feasible if an economy uses only nonrenewable resources. Of course, where there are renewable resources, sustainability is technically feasible, at least in the absence of population growth. Neoclassical economists argue that the class of growth models that includes resources can account for mass balance and thermodynamic constraints with the “essentiality condition.” If σ is greater than 1, then resources are “nonessential.” If σ is less than or equal to 1, then resources are “essential.” Essential in this case means that, given positive nonresource inputs, output is

only zero when the resource input is zero, and strictly positive otherwise. The Cobb–Douglas production function, a form frequently used in growth models, has the essentiality condition. Economists argue that this at least accounts for the fact that some amount of energy and materials is required to produce goods and services. But when the elasticity of substitution is unity, this “essential” amount can be infinitesimal if sufficient manufactured capital is applied. Economists also note that resources and capital are interdependent in the neoclassical models in that some positive quantity of resources is required to produce capital assets. Thus, the capital stock cannot be increased without depleting the resource stock.

Substitution that is technically possible will not occur unless society invests in sufficient capital over time to replace the depleted natural resources and ecosystem services. How much investment does take place depends on the institutional setting of the economy. For example, in an economy in which sustainability is just technically feasible ($\sigma = 1$) and there are only nonrenewable resources, sustainability will not occur in either a competitive or centrally planned economy where the decision rule is the maximization of the discounted flow of utility of future generations using a constant and positive discount rate. Consumption per capita will eventually decline to zero after an initial period of economic growth because resources and ecosystem services are depleted faster than capital can be accumulated to replace them. However, if the utility of individuals is given equal weight without regard to when they happen to live, and the aim is to maximize the sum of utilities over time, then growth in consumption can occur indefinitely. This is equivalent to maximizing net present value with a zero discount rate. The Hartwick rule is an investment rule that can, in theory, achieve sustainability if sustainability is technically feasible. A constant level of consumption can be achieved by reinvesting resource rents in other forms of capital, which, in turn, can substitute for resources.

The other key factor permitting growth in the face of a limited resource base is technological change. A technological improvement is defined as a gain in total factor productivity, which implies that output increases while a weighted sum of the inputs to production is held constant. Therefore, growth would be possible in the face of finite resources. Studies that examine the roles of resources in growth models with endogenous technological change have so far been less general in their assumptions, compared to research using the exogenous technological change

assumption, and do not yet provide necessary or sufficient conditions for the achievement of sustainability. For example, Sjak Smulders and Michiel de Nooij assume in their model that energy use per capita continues to grow over time and that conservation policies consist of either a reduction in the growth rate or a one-time cut in the level of energy use. This set of assumptions is very different from those in the models just discussed. The results depend on whether the growth rate of energy use is reduced or the level of energy use is cut and on the value of the elasticity of substitution. In these models, energy scarcity or conservation policies can induce technological change. Though this can mitigate the effects of scarcity, non-energy-related research and development may be crowded out. This means that attempts to conserve energy can depress the long-run economic growth rate, even when they would not in a model with exogenous technological change. Integrating these ideas into research on sustainability policy will clearly be a future area of active research.

2.3 Critique and Alternative Views

Many ecological economists have a “preanalytic vision” of the economic process that is fundamentally different from that presented in neoclassical economics. Mainstream growth theory focuses on institutional limits to growth. Ecological economists tend instead to focus on the material basis of the economy. The criticism of growth theory focuses on the limits to substitution and the limits to technological progress as ways of mitigating the scarcity of resources. Substitution of manufactured capital for resources and technological change could potentially get more production out of a limited resource input and circumvent the limited capacity of natural environments to absorb the impacts of energy and resource use by reducing those impacts. But if these two processes are limited, then limited resources or excessive environmental impacts may restrict growth (empirical evidence on these questions is reviewed in Sections 4 and 5).

2.3.1 Limits to Substitution

There is more than one type of substitution between inputs and, therefore, there is more than one reason why substitution may be limited. There can be substitution within a category of similar production inputs (for example, between different fuels) and between different categories of inputs (for example, between energy and machines). There is also a distinction to be made between substitution at the

micro level (for example, in a single engineering process or in a single firm) and at the macro level (in the economy as a whole). Additionally, some types of substitution that are possible in a single country are not possible globally.

The first type of substitution (within category), and, in particular, the substitution of renewable for nonrenewable resources, is undoubtedly important. The long-run pattern of energy use in industrial economies has been dominated by substitutions: from wood to coal, to oil, to natural gas, and to primary electricity. It is possible that the elasticity of substitution for within-category types of substitution exceeds unity. This would imply that some particular inputs are nonessential.

Ecological economists emphasize the importance of limits to the other type of substitution, and, in particular, the substitution of manufactured capital for natural capital. Natural capital is needed both for energy capture and for absorbing the impacts of energy and resource use—the sink function. Even if smaller amounts of energy were needed for production, all productive activity, being a work process transforming the materials using energy, will disrupt the natural environment. Often one form of environmental disruption (for example, pollution) is replaced by another form of environmental disruption (for example, hydroelectric dams). Potential reasons for limited substitutability are discussed in the following section.

2.3.1.1 Thermodynamic Limits to Substitution

Thermodynamic limits to substitution are easily identified for individual processes by an energy–materials analysis that defines both the fundamental limitations of transforming materials into different thermodynamic states and the use of energy to achieve that transformation. These types of analyses have shown where technological improvements exhibit strong diminishing returns due to thermodynamic limits, and where there is substantial room for improvements in the efficiency of energy and material use. For example, the thermal efficiency of power plants has been relatively constant for many years, reflecting the fact that power plant design is approaching the thermodynamic limit.

2.3.1.2 Complementarity Limits Substitution

Production is a work process that uses energy to transform materials into goods and services. Nicholas Georgescu-Roegen’s fund-flow model describes production as a transformation process in which a flow of materials, energy, and information is transformed

by two agents of transformation: human labor and manufactured capital. The flow of energy, materials, and services from natural capital is what is being transformed, while manufactured capital effects the transformation. Thus, some ecological economists argue that, for example, adding to the stock of pulp mills does not produce an increase in pulp unless there also is the wood fiber to feed them. The latter is essential an argument about material balance.

Mainstream economists think about this question differently. First, they argue that though additional capital cannot conjure wood fibers out of a vacuum, more capital can be used with each amount of wood fibers to produce more sophisticated and valuable products from them and that this is the relevant substitution between capital and resources. In the energy industries, more capital can extract more oil from a petroleum reservoir and extract more useful work downstream in cleaner ways, only subject to thermodynamic limits. Even thermodynamic limits apply only to production of physical product. There is no limit in this view to the potential value of product created through sophisticated manipulation using larger amounts of capital.

2.3.1.3 Physical Interdependence and Macroeconomic Limits to Substitution The construction, operation, and maintenance of tools, machines, and factories require a flow of materials, i.e., energy from natural capital. Similarly, the humans that direct manufactured capital consume energy and materials (i.e., food and water). Thus, producing more of the “substitute,” i.e., manufactured capital, requires more of the thing for which it is supposed to substitute. Ecological economists argue that production functions used in growth models do not account for this interdependence, and thus assume a degree of substitutability that does not exist. But both environmental and ecological economics have erred by not distinguishing among the micro- and macro-applications of production functions. Substitution is fundamentally more constrained at the macro level of analysis than at the micro level. For example, home insulation directly substitutes for heating fuel within the household sector. But interdependence means that insulation requires fuel to manufacture, so, for the economy as a whole, the net substitution of insulation for fuel is less than that indicated by an analysis of the household sector, in isolation from the rest of the economy.

In Fig. 3, the curve $E=f(M)$ is a neoclassical isoquant for a constant level of output, where E is

energy and M is materials. The indirect energy costs of materials are represented by $g(M)$. For simplicity, the diagram unrealistically assumes that no materials are required in the extraction or capture of energy. Addition of direct and indirect energy costs results in the “net” isoquant $E=h(M)$. Generalizing for material costs to energy extraction appears to indicate that there are eventually decreasing returns to all factors at the macro level, and therefore the socially efficient region of the aggregate production function does not include areas with extreme factor ratios.

At a global level, a country such as Kuwait or Nauru can deplete its natural resources and invest in manufactured capital offshore through the financial markets. But this route to substituting manufactured capital for natural capital is clearly not possible for the world as a whole.

2.3.1.4 Critical Natural Capital Limits Substitution Ecological economists have also argued that at the macro level some forms of “natural capital” are not replaceable by produced capital, at least beyond certain minimum stock sizes. These stocks may provide life-support services to the economy or represent pools of irreplaceable genetic information or “biodiversity.” The limited substitutability argument has also been extended to incorporate non-linear dynamics and irreversible changes. The fear is that excessive substitution of human-made capitals for natural capital will cause the system to approach a threshold beyond which natural systems will lose resilience and suffer catastrophic collapse. Though we cannot demonstrate these forms of nonsubstitutability from basic physical laws, they may be just as important as thermodynamics in constraining actual production functions. In the energy context, this argument is most relevant regarding the sink func-

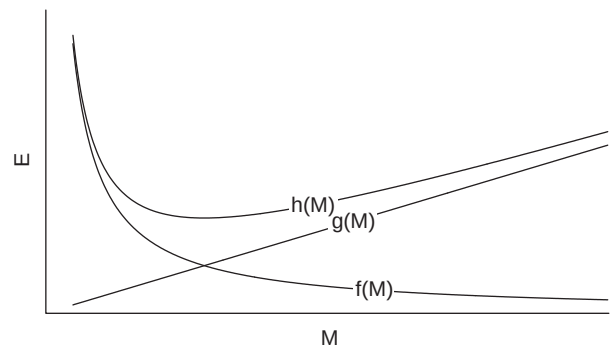


FIGURE 3 Macro-level limits to substitution.

tion of the environment. Using more and more of the environment as a sink for pollution means less and less of it is available for other life-support uses. Compared to current fossil fuel technologies, alternative energy sources may require larger areas of the environment for energy capture and may also generate wastes, etc. in the production of the energy capture and transmission capitals.

2.3.2 Limits to Technological Change

Even if substitution possibilities are limited, sustainability is possible if technological change is natural capital augmenting and unlimited in scope. The arguments for technological change as a solution would be more convincing if technological change were really something different from substitution. The neoclassical approach assumes that an infinite number of efficient techniques coexist at any one point in time. Substitution occurs among these techniques. Changes in technology occur when new, more efficient techniques are developed. However, in a sense, these new techniques represent the substitution of knowledge for the other factors of production. The knowledge is embodied in improved capital goods and more skilled workers and managers, all of which require energy, materials, and ecosystem services to produce and maintain. Thus, however sophisticated the workers and machinery become, there are still thermodynamic restrictions on the extent to which energy and material flows can be reduced.

Another question is whether technology will follow the “right” direction. If natural resources are not priced correctly due to market failure—a common and pervasive phenomenon that is the main topic of study of mainstream environmental economics—then there will be insufficient incentives to develop technologies that reduce resource and energy use. Instead, technological change would result in more resource use, not less.

3. FACTORS AFFECTING LINKAGE BETWEEN ENERGY AND GROWTH

There has been extensive debate concerning the trend in energy efficiency in the developed economies, especially since the two oil price shocks of the 1970s. Taking the example of the U.S. economy, energy consumption hardly changed in the period 1973 to 1991 (Fig. 4). This was despite a significant increase in GDP. These facts are indisputable. The reasons for

the break in the trend have been the subject of argument. It is commonly asserted that there has been a decoupling of economic output and resources, which implies that the limits to growth are no longer as restricting as in the past.

The discussion here starts from the neoclassical perspective of the production function to examine the factors that could reduce or strengthen the linkage between energy use and economic activity over time. A general production function can be represented as follows:

$$(Q_1, \dots, Q_m)' = f(A, X_1, \dots, X_n, E_1, \dots, E_p), \quad (1)$$

where the Q_i are various outputs (such as manufactured goods and services), the X_j are various inputs (such as capital, labor, etc.), the E_k are different energy inputs (such as coal, oil, etc.), and A is the state of technology as defined by the total factor productivity indicator. The relationship between energy and an aggregate of output such as gross domestic product can then be affected by substitution between energy and other inputs, technological change (a change in A), shifts in the composition of the energy input, and shifts in the composition of output. Also, shifts in the mix of the other inputs—for example, to a more capital-intensive economy from a more labor-intensive economy—can affect the relationship between energy and output, but this issue has not been extensively discussed in the literature and so will not be pursued further here. It is also possible for the input variables to affect total factor productivity, though in models that invoke exogenous technological change, this is assumed not to occur.

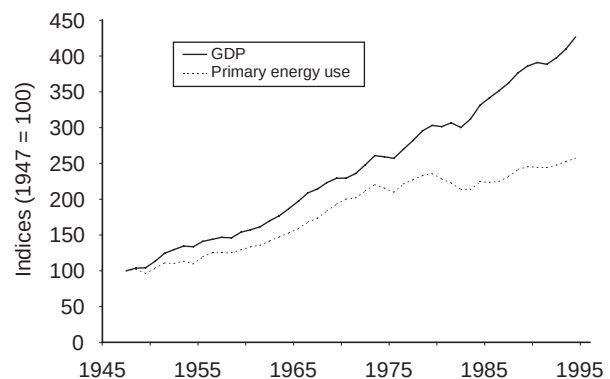


FIGURE 4 United States gross domestic product (GDP) and total primary energy use. GDP is in constant dollars, i.e., adjusted for inflation. Both variables are indexed to 100 in 1947.

3.1 Energy and Capital: Substitution and Complementarity

Econometric studies employing the translog and other functional forms have come to varying conclusions regarding whether capital and energy are complements or substitutes. These studies all estimate elasticities at the industry level. It seems that capital and energy are at best weak substitutes and possibly are complements. The degree of complementarity likely varies across industries, the level of aggregation considered, and the time frame considered.

There are few studies that look at macroeconomic substitution possibilities. In a 1991 paper, Robert Kaufmann and Irisita Azary-Lee demonstrated the importance of accounting for the physical interdependency between manufactured and natural capital. They used a standard production function to account for the indirect energy used elsewhere in the economy to produce the capital substituted for fuel in the U.S. forest products sector. They found that from 1958 to 1984, the indirect energy costs of capital offset a significant fraction of the direct fuel savings. In some years, the indirect energy costs of capital are greater than the direct fuel savings. The results of Kaufmann and Azary-Lee's analysis are consistent with the arguments made previously that scale is critical in assessing substitution possibilities. In this case, the assessment of substitution at one scale (the individual sector) overestimates the energy savings at a larger scale (the entire economy).

3.2 Innovation and Energy Efficiency

Substitution between different techniques of production using different ratios of inputs occurs due to changes in the relative prices of the inputs. Changes in the energy/GDP ratio that are not related to changes in the relative price of energy are called changes in the autonomous energy efficiency index. These could be due to any of the determinants of the relationship between energy and output listed at the beginning of this section, not just technological change. Actually, when there is endogenous technological change, changes in prices induce technological changes. In reality, technological change is at least in part endogenous. As a result, an increase in energy prices does tend to accelerate the development of energy-saving technologies. Periods of decreasing energy prices may result in technological development tending toward more intense use of energy.

Estimates of the trend in autonomous energy efficiency or the related energy augmentation index are mixed. This is likely because the direction of change has not been constant and varies across different sectors of the economy. Jorgenson and Wilcoxon estimated that autonomous energy efficiency in the United States is declining. Jorgenson also found that technological change tended to result in an increase in the share of production costs allocated to energy use—so-called biased technological change. Berndt *et al.* used a model in which the energy augmentation index is assumed to change at a constant rate. They estimate that in the U.S. manufacturing industry between 1965 and 1987, the energy augmentation index was increasing between 1.75 and 13.09% per annum, depending on the assumptions made. Judson *et al.* estimate separate environmental Kuznets curve relations for energy consumption in each of a number of energy-consuming sectors for a large panel of data using spline regression. The sectors are industry and construction, transportation, households and others, energy sector, nonenergy uses, and total apparent consumption, as well as households and agriculture, which are subsets of households and others. Judson *et al.* estimate time effects that show rising energy consumption over time in the household and other sectors but flat to declining time effects in industry and construction. Technical innovations tend to introduce more energy-using appliances to households and energy-saving techniques to industry.

Figure 5 presents the author's estimate of the autonomous energy efficiency index using the Kalman filter, an advanced time series technique that can extract unobserved stochastic series from time series data given a theoretical model that defines the trend of interest. This approach is more sophisticated than that of the studies cited previously, which assume that the index follows a simple linear or quadratic time path. Relative to an overall upward trend, the index shows large fluctuations. Until the mid-1960s, autonomous energy efficiency is increasing and then it starts a sharp decline. However, the results show that the first oil shock in 1973 does not disrupt the overall downward trend in energy efficiency. Only after the second oil shock does the trend reverse and energy efficiency increase.

The Khazzoom–Brookes postulate argues that energy-saving innovations can end up causing even more energy to be used, because the money saved is spent on other goods and services that require energy in their production. Energy services are demanded by the producer or consumer and are produced using

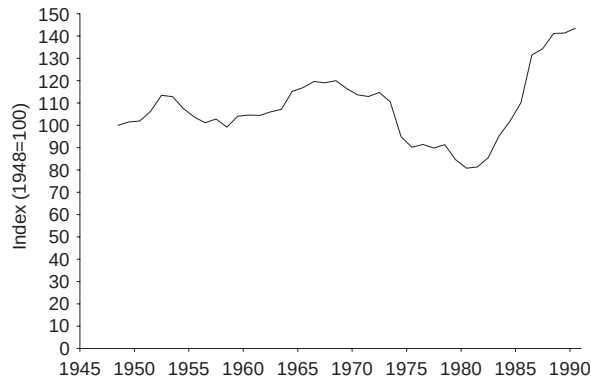


FIGURE 5 United States autonomous energy efficiency index.

energy. An innovation that reduces the amount of energy required to produce a unit of energy services lowers the effective price of energy services. This results in an increase in demand for energy services and therefore for energy. The lower price of energy also results in an income effect that increases demand for all goods in the economy and therefore for the energy required to produce them. There may also be adjustments in capital stocks that result in an even further increased long-run demand response for energy. This adjustment in capital stocks is termed a “macroeconomic feedback.” In a 1997 paper, Richard Howarth argued persuasively that the rebound effect is less than the initial innovation-induced reduction in energy use, so improvements in energy efficiency do, in fact, reduce total energy demand.

3.3 Energy Quality and Shifts in Composition of Energy Input

Energy quality is the relative economic usefulness per heat equivalent unit of different fuels and electricity. One way of measuring energy quality is the marginal product of the fuel, which is the marginal increase in the quantity of a good or service produced by the use of one additional heat unit of fuel. These services also include services received directly from energy by consumers. Some fuels can be used for a larger number of activities and/or for more valuable activities. For example, coal cannot be used directly to power a computer whereas electricity can. The marginal product of a fuel is determined in part by a complex set of attributes unique to each fuel: physical scarcity, capacity to do useful work, energy density, cleanliness, amenability to storage, safety, flexibility of use, cost of conversion, and so on. But

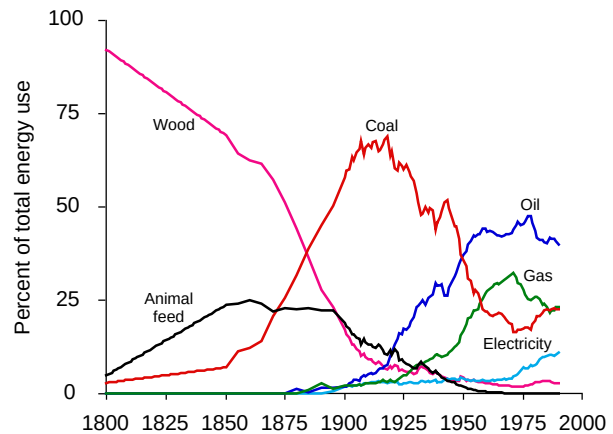


FIGURE 6 Composition of U.S. energy input.

also the marginal product is not uniquely fixed by these attributes. Rather the energy vector’s marginal product varies according to what activities it is used in, how much and what form of capital, labor, and materials it is used in conjunction with, and how much energy is used in each application. Therefore, energy qualities are not fixed over time. However, it is generally believed that electricity is the highest quality type of energy, followed by natural gas, oil, coal, and wood and other biofuels in descending order of quality. This is supported by the typical prices of these fuels per unit of energy. According to economic theory, the price paid for a fuel should be proportional to its marginal product.

Samuel Schurr was among the first to recognize the economic importance of energy quality. Noting that the composition of energy use has changed significantly over time (Fig. 6), Schurr argued that the general shift to higher quality fuels reduces the amount of energy required to produce a dollar’s worth of GDP. If this is ignored, apparent total factor productivity (TFP) growth is greater than is really the case. Researchers such as Cutler Cleveland, Robert Kaufmann, and the Office of Technology Assessment, have presented analyses that explain much of the decline in U.S. energy intensity in terms of structural shifts in the economy and shifts from lower quality fuels to higher quality fuels. In a recent paper, Kaufmann estimated a vector autoregressive model of the energy/GDP ratio, household energy expenditures, fuel mix variables, and energy price variables for the United States. He found that shifting away from coal use and in particular shifting toward the use of oil reduced energy intensity. This shift away from coal contributed to declining energy intensity over the entire 1929–1999 time period.

Figure 7 illustrates the increase in the second half of the 20th century in U.S. GDP and a quality-adjusted index of energy use computed by the author. The index accounts for differences in the productivity of different fuels by weighting them by their prices. There is clearly less evidence of decoupling of energy use and GDP in these data than in those in Fig. 4. If decoupling is mainly due to the shift to higher quality fuels, then there appear to be limits to that substitution. In particular, exhaustion of low-cost oil supplies could mean that economies have to revert to lower quality fuels such as coal.

3.4 Shifts in the Composition of Output

Typically, over the course of economic development, the output mix changes. In the earlier phases of development there is a shift away from agriculture towards heavy industry, and in the later stages of development there is a shift from the more resource-intensive extractive and heavy industrial sectors toward services and lighter manufacturing. Different industries have different energy intensities. It is often argued that this will result in an increase in energy used per unit of output in the early stages of economic development and a reduction in energy used per unit output in the later stages of economic development. Pollution and environmental disruption would be expected to follow a similar path.

Service industries still need large energy and resource inputs. The product being sold may be

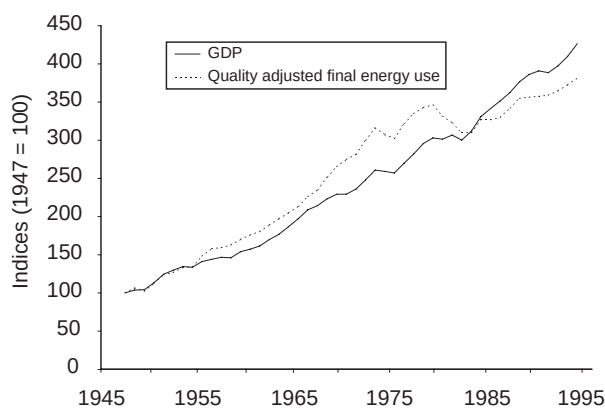


FIGURE 7 United States gross domestic product (GDP) and quality-adjusted final energy use GDP is in constant dollars, i.e., adjusted for inflation. Energy use is a Divisia index (the accumulation of weighted sums of growth rates) of the principal final energy use categories (oil, natural gas, coal, electricity, biofuels, etc.). The different fuels are weighted according to their average prices.

immaterial but the office towers, shopping malls, warehouses, rental apartment complexes, etc. where the activity is conducted are very material intensive and energy is used in their functioning as well as in their construction and maintenance. Other service industries such as transport clearly heavily draw on resources and energy. Furthermore, consumers use large amounts of energy and resources in commuting to work, to shop, etc. Therefore, a complete decoupling of energy and growth as a result of shifting to the service sector seems unlikely. When the indirect energy use embodied in manufactured products and services is taken into account, the U.S. service and household sectors are not much less energy intensive than are the other sectors of the economy, and there is little evidence that the shift in output mix that has occurred in the past few decades has significantly lowered the energy/GDP ratio. Rather, changes in the mix of energy used are primarily responsible. Furthermore, on a global scale, there may be limits to the extent to which developing countries can replicate the structural shift that has occurred in the developed economies and, additionally, it is not clear that the developed world can continue to shift in that direction indefinitely.

4. EMPIRICAL TESTING

Here the focus is on the empirical evidence for the tightness of coupling between energy use and economic output. Ordinary linear regression or correlation methods cannot be used to establish a casual relation among variables. In particular, it is well known that when two or more totally unrelated variables are trending over time they will appear to be correlated simply because of the shared directionality. Even after removing any trends by appropriate means, the correlations among variables could be due to causality between them or due to their relations with other variables not included in the analysis. Two methods for testing for causality among time series variables are Granger causality tests and cointegration analysis.

Granger causality tests whether one variable in a linear relation can be meaningfully described as a dependent variable and the other variable as an independent variable, whether the relation is bidirectional, or whether no functional relation exists at all. This is usually done by testing whether lagged values of one of the variables significantly add to the explanatory power of a model that already includes

lagged values of the dependent variable, and perhaps also lagged values of other variables.

Whereas Granger causality can be applied to both stationary and integrated time series (time series that follow a random walk), cointegration applies only to linear models of integrated time series. The irregular trend in integrated series is known as a stochastic trend, as opposed to a simple linear deterministic time trend. Time series of GDP and energy use are integrated. Cointegration analysis aims to uncover causal relations among variables by determining if the stochastic trends in a group of variables are shared by the series. If these trends are shared, either one variable causes the other or they are both driven by a third variable. It can also be used to test if there are residual stochastic trends that are not shared by any other variables. This may be an indication that important variables have been omitted from the regression model or that the variable with the residual trend does not have long-run interactions with the other variables. The presence of cointegration can also be interpreted as the presence of a long-run equilibrium relationship between the variables in question.

Both cointegration and Granger causality tests are usually carried out within the context of vector autoregression models. These models consist of group of regression equations in which each dependent variable is regressed on lagged values of itself and of all the other variables in the system.

Many analysts have used Granger causality tests or the related test developed by Christopher Sims to test whether energy use causes economic growth or whether energy use is determined by the level of output in the context of a bivariate vector autoregression. The results have been generally inconclusive. When nominally significant results were obtained, the majority indicated that causality runs from output to energy use. In a 1993 study, the author tested for Granger causality in a multivariate setting using a vector autoregression (VAR) model of GDP, energy use, capital, and labor inputs in the United States over a the post-World War II period. He also used a quality-adjusted index of energy input in place of gross energy use. The multivariate methodology is important because changes in energy use are frequently countered by the substitution of other factors of production, resulting in an insignificant overall impact on output. When both of these innovations are employed, energy is found to “Granger-cause” GDP. These results are supported by other studies that found that changes in oil prices Granger-cause changes in GDP and unemployment

in VAR models, whereas oil prices are exogenous to the system.

Yu and Jin were the first to test whether energy and output cointegrate. They found that no such relationship exists between energy use and either employment or an index of industrial production in the United States. However, the lack of a long-run equilibrium relationship between gross energy use and output alone does not necessarily imply that there is no relation between the variables. Few analysts believe that capital, labor, and technical change play no significant role in determining output. If these variables are integrated, then there will be no cointegration between energy and output, no matter whether there is a relationship between the latter two variables. Also, decreasing energy intensity (due to increased energy efficiency), shifts in the composition of the energy input, and structural change in the economy mean that energy and output will drift apart. Similar comments apply to the bivariate energy-employment relationship. Further, using total energy use in the economy as a whole but measuring output as industrial output alone may compound the insensitivity of the test.

It would seem that if a multivariate approach helps in uncovering the Granger causality relations between energy and GDP, a multivariate approach should be used to investigate the cointegration relations among the variables. The author investigated the time series properties of GDP, quality weighted energy, labor, and capital series, estimating a dynamic cointegration model using the Johansen methodology. The cointegration analysis showed that energy is significant in explaining GDP. It also showed that there is cointegration in a relationship including GDP, capital, labor, and energy. Other analysts have found that energy, GDP, and energy prices cointegrate and that when all three variables are included there is mutual causation between energy and GDP. The inconclusive results of the earlier tests of Granger causality are probably due to the omission of necessary variables—either the quantities of other inputs (and quality adjustment of the energy input) or energy prices.

5. ENVIRONMENTAL IMPLICATIONS

5.1 Theory

The argument has been made here that there is a very strong link between energy use and both the levels of

economic activity and economic growth. What are the implications of this linkage for the quality of the environment? The focus here is now on immediate environmental impacts (issues of sustainability were discussed in Section 2).

Energy use has a variety of impacts. Energy extraction and processing always involve some forms of environmental disruption, including both geomorphological and ecological disruption as well as pollution. Energy use involves both pollution and other impacts, such as noise from transport, and land-use impacts, such as the construction of roads, etc. As all human activities require energy use; in fact, all human impacts on the environment can be seen as the consequences of energy use. Energy use is sometimes seen as a proxy for the environmental impact of human activity in general. Creating order in the economic system always implies creating disorder in nature, though this disorder could be in the sun or space rather than on Earth. The factors that reduce the total amount of energy needed to produce a dollar's worth of GDP, discussed in Section 3, therefore, also act to reduce the environmental impact of economic growth in exactly the same way as they reduce energy consumption. However, not all impacts of energy use are equally harmful to the environment and human health.

A shift from lower to higher quality energy sources not only reduces the total energy required to produce a unit of GDP, but also may reduce the environmental impact of the remaining energy use. An obvious example would be a shift from coal use to natural gas use. Natural gas is cleaner burning and produces less carbon dioxide per unit of energy derived. However, we need to be careful here. Nuclear-generated electricity is a higher quality fuel compared to coal, at least as measured by current prices, but its long-term environmental impacts are not necessarily lower. Incorporating the cost of currently unpriced environmental externalities into the prices of fuels would, though, raise the apparent quality of those fuels. This is not a contradiction, because a higher market price would mean that those fuels would only be used in more productive activities.

The environmental impact of energy use may also change over time due to technological innovation that reduces the emissions of various pollutants or other environmental impacts associated with each energy source. This is in addition to general energy-conserving technological changes that reduce the energy requirements of production, already discussed. Therefore, despite the strong connections

between energy use and economic growth, there are several pathways through which the environmental impact of growth can be reduced. Again, however, if there are limits to substitution and technological change, then the potential reduction in the environmental intensity of economic production is eventually limited. Innovations that reduce one type of emission (for example, flue gas desulfurization) often produce a different type of waste that must be disposed of (in this case, gypsum that is possibly contaminated with heavy metals) as well as other disruptions required to implement the technology (in this example, limestone mining).

5.2 Empirical Evidence: The Environmental Kuznets Curve

In the 1970s, the debate on the relation between growth and the environment focused on the report, *The Limits to Growth*, by the Club of Rome. The common wisdom was that economic growth meant greater environmental impacts and the main way to environmental gains was reduced population and consumption. Economists and others argued that substitution and innovation could reduce environmental impacts (as was described in Section 2). But these were minority views. The mainstream view was that environment and economy were conflicting goals. By the late 1980s, however, with the emergence of the idea of sustainable development, the conventional wisdom shifted to one of "too poor to be green." Less developed countries lacked the resources for environmental protection and growth, and development was needed to provide for environmental protection. This idea was embodied in the empirical models that became known as environmental Kuznets curves (EKC). The hypothesis states that there is an inverted U-shape relation between various indicators of environmental degradation and income per capita, with pollution or other forms of degradation rising in the early stages of economic development and falling in the later stages. The EKC is named for the economist Kuznets, who hypothesized that the relationship between a measure of inequality in the distribution of income and the level of income takes the form of an inverted U-shape curve. However, Kuznets had no part in developing the EKC concept.

The idea has become one of the "stylized facts" of environmental and resource economics. This is despite considerable criticism on both theoretical and empirical grounds. The EKC has been inter-

preted by many as indicating that no effort should be made to adopt environmental policies in developing countries, i.e., when those countries become rich, the current environmental problems will be addressed by policy changes adopted at that later time. As a corollary, it is implied that little in the way of environmental clean-up activity is being conducted in developing countries. These views are challenged by recent evidence that, in fact, pollution problems are being addressed and remedied in developing economies. In addition to the data and case studies provided by Susmita Dasgupta *et al.*, the author shows that for sulfur (widely believed to show the inverted U-shape relation between emission and income per capita), emissions in fact rise with increasing income at all levels of income, but that there are strong time effects reducing emissions in all countries across all income levels. In the author's opinion, this new evidence supersedes the debate about whether some pollutants show an inverted U-shaped curve and others—for example, carbon dioxide and “new toxics”—show a monotonic relationship. All pollutants show a monotonic relation with income, but over time pollution has been reduced at all income levels, *ceteris paribus*. Similarly, the debate about whether the downward sloping portion of the EKC is an illusion resulting from the movement of polluting industries to off-shore locations is also now moot. This phenomenon might lower the income elasticity of pollution in developed economies relative to developing economies, but it does not seem sufficient to make it negative. The true form of the emissions–income relationship is a mix of two of the scenarios proposed by Dasgupta *et al.*, illustrated in Fig. 8. The overall shape is that of their “new toxics” EKC—a monotonic increase of emissions in income. But over time this curve shifts down. This is analogous to their “revised EKC” scenario, which is intended to indicate that over time the conventional EKC curve shifts down.

Although the environmental Kuznets curve is clearly not a “stylized fact” and is unlikely to be a useful model, this does not mean that it is not possible to reduce emissions of sulfur and other pollutants. The time effects from an EKC estimated in first differences and from an econometric emissions decomposition model both show that considerable time-related reductions in sulfur emissions have been achieved in countries at many different levels of income. Dasgupta *et al.* provide data and case studies that illustrate the progress already made in developing countries to reduce pollution. In addition, the

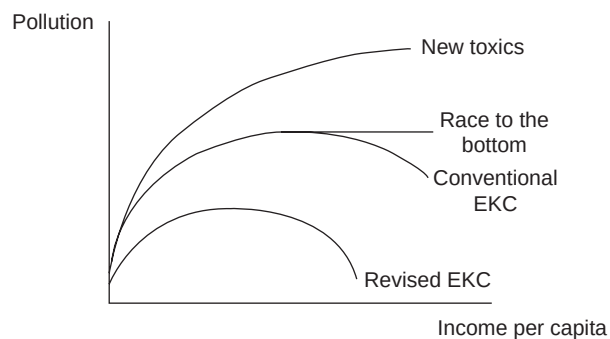


FIGURE 8 Environmental Kuznets curve (EKC): different scenarios. From Dasgupta *et al.* (2002).

income elasticity of emissions is likely to be less than 1—but not negative in wealthy countries, as suggested by the EKC hypothesis. In slower growing economies, emissions-reducing technological change can overcome the scale effect of rising income per capita on emissions. As a result, substantial reductions in sulfur emissions per capita have been observed in many Organization for Economic Cooperation and Development (OECD) countries in the past few decades. In faster growing middle income economies, the effects of rising income overwhelm the contribution of technological in reducing emissions.

This picture relates very closely to the theoretical model described in the previous Section 5.1. Overall, there is a strong link between rising energy use, economic growth, and pollution. However, the linkages between these three can be mitigated by a number of factors, including shifting to higher quality fuels and technological change aimed both at general increases in economic productivity and, specifically, at reducing pollution. However, given the skepticism regarding the potential for unlimited substitution or technological progress, there may be limits to the extent that these linkages can continue to be loosened in the future.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Economic Growth, Liberalization, and the Environment • Economics of Energy Demand • Economics of Energy Supply • Energy Ladder in Developing Nations • Environmental Kuznets Curve • Population Growth and Energy • Sustainable Development:

Basic Concepts and Application to Energy • Thermodynamics and Economics, Overview

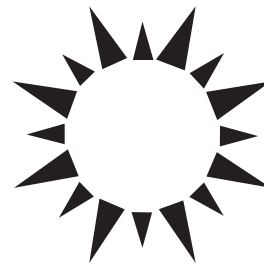
Further Reading

- Berndt, E. R., Kolstad, C., and Lee, J.-K. (1993). Measuring the energy efficiency and productivity impacts of embodied technical change. *Energy J.* **14**, 33–55.
- Brookes, L. (1990). The greenhouse effect: The fallacies in the energy efficiency solution. *Energy Policy* **18**, 199–201.
- Cleveland, C. J., Costanza, R., Hall, C. A. S., and Kaufmann, R. K. (1984). Energy and the U.S. economy: A biophysical perspective. *Science* **225**, 890–897.
- Dasgupta, P., Heal, G. (1974). The optimal depletion of exhaustible resources. Review of economic studies. Symp. Econ. Exhaustible Resources, pp. 3–28.
- Dasgupta, S., Laplante, B., Wang, H., and Wheeler, D. (2002). Confronting the environmental Kuznets curve. *J. Econ. Perspect.* **16**, 147–168.
- Georgescu-Roegen, N. (1976). “Energy and Economic Myths.” Pergamon, New York.
- Hannon, B. (1973). The structure of the ecosystem. *J. Theoret. Biol.* **41**, 535–546.
- Hartwick, J. M. (1977). Intergenerational equity and the investing of rents from exhaustible resources. *Am. Econ. Rev.* **66**, 972–974.
- Howarth, R. B. (1997). Energy efficiency and economic growth. *Contemp. Econ. Policy* **25**, 1–9.
- Jorgenson, D. W. (1984). The role of energy in productivity growth. *Energy J.* **5**(3), 11–26.
- Jorgenson, D. W., and Wilcoxon, P. J. (1993). Reducing U.S. carbon emissions: An econometric general equilibrium assessment. *Resour. Energy Econ.* **15**, 7–25.
- Judson, R. A., Schmalensee, R., and Stoker, T. M. (1999). Economic development and the structure of demand for commercial energy. *Energy J.* **20**(2), 29–57.
- Kaufmann, R. K. (2003). The mechanisms for autonomous energy efficiency increases: A cointegration analysis of the US energy/GDP ratio. *Energy J.* (in press).
- Kaufmann, R. K., Azary-Lee, I. G. (1991). A biophysical analysis of substitution: Does substitution save energy in the U.S. forest products industry? Proc. Ecol. Econ.: Implications Forest Manage. Practice, 1991.
- Khazzoom, D. J. (1980). Economic implications of mandated efficiency standards for household appliances. *Energy J.* **1**(4), 21–39.
- O'Connor, M. P. (1993). Entropic irreversibility and uncontrolled technological change in the economy and environment. *J. Evolut. Econ.* **34**, 285–315.
- Perrings, C. A. (1987). “Economy and Environment: A Theoretical Essay on the Interdependence of Economic and Environmental Systems.” Cambridge University Press, Cambridge.
- Schurr, S., and Netschert, B. (1960). “Energy and the American Economy, 1850–1975.” Johns Hopkins University Press, Baltimore.
- Sims, C. A. (1972). Money, income and causality. *Am. Econ. Rev.* **62**, 540–552.
- Smulders, S., and de Nooij, M. (2003). The impact of energy conservation on technology and economic growth. *Resour. Energy Econ.* **25**, 59–79.
- Solow, R. M. (1956). A contribution to the theory of economic growth. *Q. J. Econ.* **70**, 65–94.
- Solow, R. M. (1974). Intergenerational equity and exhaustible resources. Review of economic studies. Symp. Econ. Exhaustible Resources, pp. 29–46.
- Solow, R. M. (1997). Reply: Georgescu-Roegen versus Solow/Stiglitz. *Ecol. Econ.* **22**, 267–268.
- Stern, D. I. (1993). Energy use and economic growth in the USA. A multivariate approach. *Energy Econ.* **15**, 137–150.
- Stern, D. I. (2000). A multivariate cointegration analysis of the role of energy in the U.S. macroeconomy. *Energy Econ.* **22**, 267–283.
- Stern, D. I. (2002). Explaining changes in global sulfur emissions: An econometric decomposition approach. *Ecol. Econ.* **42**, 201–220.
- Stern, D. I., and Common, M. S. (2001). Is there an environmental Kuznets curve for sulfur? *J. Environ. Econ. Manage.* **41**, 162–178.
- U.S. Congress, Office of Technology Assessment (1990). “Energy Use and the U.S. Economy.” OTA-BP-E-57. U.S. Government Printing Office, Washington, D.C.
- Yu, E. S. H., and Jin, J. C. (1992). Cointegration tests of energy consumption, income, and employment. *Resour. Energy* **14**, 259–266.

Economic Growth, Liberalization, and the Environment

SJAK SMULDERS

Tilburg University
Tilburg, The Netherlands



-
1. Interactions between the Economy and the Environment
 2. Trends in Economic Growth and Resource Use
 3. Explaining the Link between Growth and Resource Use
 4. International Trade and the Environment
 5. Environmental Regulation and Productivity

Glossary

comparative advantage A country has a comparative advantage in a particular good if the ratio of costs to produce the good, relative to the cost to produce another good, is lower than the same cost ratio in other countries. Comparative advantage is the basis for trade: Countries gain if they export the good they produce relatively most efficient.

economic growth Steady increases in gross domestic product.

energy intensity Energy supply divided by gross domestic product.

externality A situation in which market transactions affect persons who are not involved in the transaction directly (i.e., not through market prices).

gross domestic product (GDP) The sum of all output produced by economic activity within a country. Gross national product includes net income from abroad (e.g., rent and profits).

general equilibrium A situation in which all markets in the economy are in equilibrium.

market equilibrium A situation characterized by a certain volume of trade and a price of a certain good such that no market party wants to change behavior: No seller of the good wants to increase or decrease supply, and no consumer wants to reduce or increase demand.

Organization for Economic Cooperation and Development (OECD) The North American and Western European countries, Greece, Turkey, and Mexico.

real price Relative price of a commodity in terms of another basket of commodities. GDP is often measured in real prices (also labeled “at constant prices”) to

eliminate the effects of inflation (where inflation is the upward trend common in all prices, leading to a downward trend in the purchasing power of money).
total factor productivity (also multifactor productivity) Outputs per unit of input. Changes in total factor productivity measure technological change by accounting for how much of the increase in production is due to other factors besides changes in inputs.

Economic growth is the steady increase in output produced by economic activity within a country (gross domestic product). A central question is whether economic growth is compatible with a nondeteriorating environment and whether the environmental consequences of economic growth have a major impact on welfare levels derived from growing output of marketed goods. International trade has grown even faster than national income. It has been questioned whether increased trade has been bad for the environment and whether it has caused rich countries to relocate their polluting industries to developing countries. The following is another important question: How large are the costs associated with environmental regulation in terms of competitiveness, productivity, and employment? This article deals with these questions by reviewing empirical evidence and economic theory on resource use and economic growth.

1. INTERACTIONS BETWEEN THE ECONOMY AND THE ENVIRONMENT

1.1 Economic Growth and Resource Use

Statistics on national production levels and indicators of environmental pressure have been collected during

the past few decades (and reconstructed for the more distant past) to document the link between economic growth and the environment. The theory of economic growth and economic theories of natural resources provide explanations and interpretations of the trends as well as methods to assess future developments. The basic approach in most statistical and theoretical analyses is the decomposition of aggregate production of a nation into, on the one hand, the current state of technology and, on the other hand, the inputs into the national production process. The main inputs are labor (hours worked), capital (produced means of production that can be used over a certain time period), and natural resources.

Natural resources comprise energy inputs and material inputs. An important distinction is between renewable and nonrenewable resources. Renewable resources are biotic populations, flora, and fauna, which have the potential to grow by means of natural reproduction. For example, the theory of renewable resources studies fishery and forestry. Nonrenewable resources are minerals of which physical supplies do not grow on a relevant timescale for humans. Examples are metals (e.g., aluminum, copper, and gold) and fossil fuels (e.g., oil, coal, and natural gas).

The basic characteristic of natural resources is that they cannot be produced, but they serve as an essential input in production. A higher rate of resource usage implies faster depletion of the available stock of the resource. In this sense, resource inputs differ in a fundamental way from other inputs used in production, such as labor and capital: Labor is not used up and man-made capital can be accumulated. Another important difference is that the available resource stock may directly affect welfare as an amenity (nonuse value). Resource economics studies the trade-off between extraction of the resource for current consumption and conservation for later dates. Environmental quality can also be considered as an input in the aggregate production process. Pollution is an inevitable by-product of production. Pollution can be seen as a degradation of environmental resources, such as clean air and soil.

1.2 Externalities, Markets, and Incentives

Economic growth is the result of growth in inputs and increases in the productivity of the inputs—that is, changes in technology. Economics further explains why inputs and technology change over time. Inputs such as materials, labor, and capital are traded in factor markets. Prices match demand and supply and provide consumers and suppliers with information

and incentives. Population growth, for example, increases the potential working force and the supply of labor inputs in the economy increases. This fuels growth if wages adjust so that the growing working force is employed in production. With more labor in production, capital can be operated at a higher productivity level. This increase in the return to capital fuels investment in new capital so that growth in capital inputs also fuels output growth. Similarly, the demand for materials and energy increases. If these are available in limited supply, their prices increase, which provides incentives to develop technologies that save on their use. Thus, although growing scarcity of inputs may impose a drag on growth, it may also induce technological change that fuels economic growth.

Markets provide a powerful coordination mechanism. In sectors of the economy in which the demand for resources such as materials or labor is high, prices increase and this attracts resources to these sectors, away from other sectors. Thus, resources tend to be allocated to those sectors in which they are in highest demand. The high demand reflects a high willingness to pay for the resources and implies that their value is considered to be high. This explains why the market mechanism in principle can enhance efficiency and welfare in the economy.

The market can only coordinate social preferences (demand for certain goods or services) and (technical) possibilities and resource endowments in a welfare-maximizing way if prices reflect these preferences and possibilities. This requires (i) that the owners of production factors (the inputs such as labor, capital, and materials) can charge a price that reflects their production cost and (ii) that consumers can express their willingness to pay. For many environmental goods, these preconditions are not met. In the absence of environmental regulation, there are no markets for clean air or climate. Pollution entails a cost to society but not a cost to the polluting producers. Waste can be disposed freely but imposes a cost on society. When economic activity (i.e., market transactions) imposes costs or benefits on persons who are not involved in the transaction directly, externalities are said to be present. Externalities cause markets to function imperfectly and welfare to be impaired. This is why environmental policies may improve welfare. These policies have to correct environmental externalities, for example, by making firms pay for the pollution costs they impose on society.

We can now more precisely define the cost of pollution for society. From an economic perspective,

pollution (and, similarly, environmental degradation or resource depletion) is costly to the degree that it causes externalities. Because the value of environmental quality is not reflected in the prices in the economy, the economic decisions that are based on these prices give rise to more pollution than society finds optimal.

The ideal (optimal) environmental regulation imposes measures to eliminate all external costs of pollution; that is, it makes all parties in society internalize all social costs of pollution. In such a situation, welfare is maximal, but pollution is not necessarily (and very unlikely) at the technical minimum. Although depletion of natural resources and polluting consequences of production are by definition bad for the environment, they are not necessarily bad for welfare. They provide the inputs for production of valuable things. As long as the benefits from increases in production outweigh the costs of environmental degradation, economic growth enhances welfare.

The following sections examine how economic growth, international trade, and environmental regulation affect the environment and welfare. To examine the impact on aggregate welfare, changes in the physical state of the environment have to be made comparable to changes in economic production, which requires a valuation method to place values on environmental improvements. Since this is a difficult and subjective task, I also discuss the physical effects, by which I mean the impact on pollution, pollution concentration, and resource stocks measured in physical terms. This is the conservationist perspective, to be contrasted with the economic approach that deals with welfare effects.

2. TRENDS IN ECONOMIC GROWTH AND RESOURCE USE

2.1 Trends in Economic Growth

From 1820 to 1995, total world production [gross domestic product (GDP)] grew on average at an annual rate of 2.2%, which amounts to a more than 40-fold increase over a mere 175 years. This rate of change is unprecedented: In the 1800 years before this period, world GDP increased less than 7-fold. Per capita economic growth has also been enormous during the past century. With the current world population of approximately 5.5 billion, five times as large as in 1820, per capita GDP is now almost eight times higher than it was 175 years ago.

Differences among countries are major. In 1988, output per worker in the United States was more

than 30 times higher than that in sub-Saharan Africa. Developing countries often have per capita income levels that are less than 10% of U.S. levels (e.g., in 1988, India, 9%; China, 6%). Per capita income in Western Europe is on average 30% lower than that in North America, but income gaps are smaller than they were four decades ago. Typically, among the richer countries, there is convergence in income levels, whereas the gap between the poorest and richest countries grows.

2.2 Trends in Energy and Materials

Economic growth directly translates into growing demand for energy resources and materials since they are necessary inputs in production. However, in the process of growth, some energy resources or materials may be substituted for others so that not all of them are used at an increasing rate. Technological change has allowed many sectors of the economy or even entire economies to produce at a lower overall energy intensity.

Primary energy supply in the Organization for Economic Cooperation and Development (OECD) countries was approximately 50% higher in 1999 than in 1971, but GDP increased faster. Indeed, from 1960 to 2000, energy use was decoupled from economic growth, at least in the European Union, Japan, and the United States (at annual rates of 0.4, 0.2, and 1.4%, respectively). The oil shortages in the 1970s induced energy conservation programs. Also, the growing relative importance of services relative to manufacturing caused energy efficiency to increase. In recent years, growth in energy use has accelerated due to increased transport and electricity use in service sectors. Trends in materials use vary considerably for different materials. Iron and steel have grown slower than GDP, paper use has grown in line with GDP, but plastics and aluminum have grown faster than GDP.

The concern is that growing energy and materials use results in scarcity of resources and imposes a threat to future growth. Stocks of virgin ores decline. The stock of materials embedded in products grows, leaving less available for recycling. If growth fuels demand for resources and supply becomes smaller, prices can be expected to rise. However, most price indicators do not reflect growing scarcity. The prices of most commodities have declined since 1960. Here, prices are measured in real terms (i.e., relative to the prices of a representative basket of produced goods). Most economists have concluded that resource scarcity will not limit growth, at least not during

the coming decades and not if we consider scarcity of those resources that are traded in well-developed global markets, such as most minerals and energy resources.

Declining prices may reflect technological progress, which allows mining at lower cost or exploration of new resource deposits. According to most economic theories, this is likely to be a temporary phenomenon, and in the long term prices of nonrenewable resources are expected to increase. This does not necessarily imply that future growth will no longer be fuelled by lower costs of energy and materials, since technological change, as well as exploration and development of new economically useful resources, may overcome scarcity limits.

The growing use of certain energy resources and materials has severe impacts on the natural environment. In general, economists agree that free market forces do not serve very well society's demand for environmental amenities or conservation of natural resources for which markets are illegal, incomplete, or do not exist (e.g., water and air quality, rare species, biodiversity, and ecosystem services). Emissions of CO₂ from fossil fuel combustion pose major threats to air pollution and climate. Despite the decoupling of energy use from economic growth in the developed economies and despite the decline in the carbon content per unit of energy used, total CO₂ emissions have risen in almost all countries in the past few decades. In large developing countries such as India, China, and Brazil, absolute CO₂ emissions have increased very rapidly. In contrast to the OECD experience, their CO₂ emissions per unit of GDP have risen at fast rates (up to 2.6% percent per year in India from 1971 to 1998).

Other major threats to the environment come from the use of hazardous industrial waste, emissions, and toxic chemicals that are used in the production process and discarded as some form of industrial waste. As a result of environmental regulation, emissions from production activities (industry) have actually declined. However, waste and emissions from consumption activities are growing. Municipal waste and hazardous waste have increased in recent decades in developed countries at a rate faster than those of the population and GDP.

2.3 Cross-Country Differences in Resource Intensities

Differences in energy used per unit of GDP are large. In North America, per capita energy use is more than

twice that in Japan, France, or the United Kingdom and approximately 1.5 times that in Germany and Australia, mainly due to differences in per capita income and in energy used per unit of output. Differences in the structure of production (in terms of the relative importance of industrial production sectors) are hardly responsible for the differences in per capita energy use.

3. EXPLAINING THE LINK BETWEEN GROWTH AND RESOURCE USE

3.1 The Environment–Income Relationship

Economists have tried to identify underlying mechanisms in resource use and growth trends. A central question is how environmental pressure changes with economic growth. One hypothesis is the so-called environmental Kuznets curve (EKC) hypothesis, according to which environmental pressure or resource use first increases with GDP but, after a certain threshold income level, declines with per capita GDP. The hypothesis has been tested for several resources, energy sources, and emissions. The evidence is mixed: The EKC shows up, but only for some pollutants and not in all countries.

In the typical approach, pollution in the major cities (or other measurement stations) of a set of countries is compared and related to the level of income in these cities. If pollution rises with the level of income but also declines with the square of the level of income, the EKC holds.

The EKC is found for water pollutants and for air pollutants, such as SO₂, suspended particulate matter (dark matter as well as heavy particles), NO_x, and CO. These pollutants cause health problems close to the place where they are emitted (i.e., these are local pollutants), whereas their impact occurs simultaneous with emissions (i.e., these are short-lived or flow pollutants). The EKC is generally rejected for municipal waste, CO₂, and aggregate energy consumption: These variables monotonically increase with the level of income in a cross section of countries. Since it is the accumulation of municipal waste in landfills, rather than the annual flow, that imposes environmental pressure, it is a stock pollutant. CO₂ is partly a local flow pollutant, but its main impact is through concentration of CO₂ in the atmosphere on the global climate (i.e., it is a global pollutant as well as a stock pollutant). Other evidence on the main global stock pollutants and environmental indicators is only

indirect. For example, wilderness land is monotonically declining with the level of income, which is a crude measure of the negative relation between income and biodiversity.

3.2 The Role of Environmental Regulation

The common pattern is that local pollutants and flow pollutants are more likely to first rise and then fall with income than global pollutants or stock pollutants. The main driving force behind the lower pollution levels in richer countries is policies to reduce pollution. In richer countries, citizens are willing to pay more for improvements in environmental quality and succeed in convincing their local and national governments to impose the necessary regulation to reduce urgent pollution problems. Since reducing local flow pollutants has the most visible effects for the constituencies of the local authorities, regulation is mainly directed at these rather than global or stock pollutants. The latter affect other regions or future generations, respectively, but not so much the authorities' electorate. In this interpretation, environmental improvement only occurs when environmental regulation becomes more stringent in richer economies. That is, environmental problems are not solved automatically by economic growth. Growth will be accompanied by reductions in pollution only if a number of conditions are fulfilled.

First, when becoming richer, people must care more about environmental quality. There is no strong evidence in favor of this claim. For example, in developing countries, environmental quality and availability of natural resources are often much more crucial for households than in developed countries. However, within developed countries, willingness to pay for improvements in health increases with income. As a result, richer citizens may be more in favor of costly measures to reduce pollutants with direct health impacts.

Second, only if the costs of reducing pollution do not increase too much with national income will citizens vote for pollution reduction. Technological change is important in this respect. On the one hand, cleaner production processes have been developed, mitigating the cost of pollution reduction. On the other hand, some new production processes and the production of new goods require more energy and resource use. For example, the rapid spread of computers and the growing demand for transport make the economies heavily dependent on electricity

and fossil fuels so that reductions in energy use have a major cost in terms of loss of production. This may explain why electricity and fossil fuel use do not decline with income, despite concern about the environment.

Third, actual environmental policies need to reflect societies' preferences for environmental quality. Indeed, in countries with a relatively high degree of political and civil freedom, environmental problems decline more rapidly with income. In countries with high indicators of corruption, natural resource depletion is faster.

Pollutants differ in terms of their (health) impact, the degree to which they are local and short-lived, and the cost of reduction. As a result, when comparing emissions of different pollutants in a particular country, some may increase and others may decrease with income. That is, the Kuznets curves, if they exist at all for the pollutants at hand, overlap. It implies that although environmental quality improves due to the reduction in one pollutant, aggregate measures of environmental quality may deteriorate because this pollutant is replaced by other pollutants. This may be a problem particularly since stock pollutants replace short-lived pollutants, which amounts to a shift of the environmental burden to future generations. Similarly, Kuznets curves of different countries overlap. While rich countries reduce emissions of a certain pollutant, poor countries may simultaneously increase emissions in their growth process. Thus, pollution is shifted from rich to poor countries. As a result, the existence of an EKC for a certain pollutant does not guarantee that worldwide environmental quality improves over time. This phenomenon is particularly important because the poor countries, which tend to increase emissions with production, account for a much larger share of the world economy than rich ones and some of them are growing at above-average rates (notably China).

4. INTERNATIONAL TRADE AND THE ENVIRONMENT

4.1 The Pollution Haven Hypothesis

The evidence on the EKC indicates that national environmental policies differ between high-income and low-income countries, and that as a result of income growth pollution in the Northern hemisphere may be reduced over time at the cost of increasing pollution in the Southern hemisphere. There have

been concerns that this process is reinforced by international trade of goods and by foreign direct investment between North and South. More stringent environmental policies in the North force pollution-intensive industries located in the North to raise their prices relative to the prices charged by firms in the South. Consumers in the North may then profitably import more goods from the South and firms in the North may profitably relocate their plants to the South. This claim on the effects of differences in stringency of environmental regulation is called the pollution haven hypothesis.

Empirical evidence for the pollution haven hypothesis is not very strong. The hypothesis has been tested in various (mainly indirect) ways. Typically, in a first step polluting industries are identified based on their emissions per unit (dollar) of output or pollution control costs as a fraction of value added. At the top of these rankings are industries such as iron and steel, nonferrous metals, mining, pulp and paper, industrial chemicals, petroleum refineries, rubber, and leather products. In a second step, the performance of these sectors in different countries is related to the country's environmental stringency, controlling as much as possible for other country differences. Stringency is either measured as an index based on law and regulation or a measure of pollution control expenditures.

Within the United States, polluting industries tend to export smaller volumes and build fewer new plants, and there is less investment in states with more stringent environmental regulation than in states with more lenient environmental regulation. Among different countries, the effects of differences in environmental regulations are more difficult to trace. Overall, industrial countries have reduced their share in world trade of goods from polluting industries relative to developing countries during the past few decades. However, there is no evidence that this trend is mainly due to differences in environmental regulation or due to international trade. The pattern of growth in these countries may be caused by a shift to pollution-intensive industries (as a movement along the upward sloping part of the EKC). Moreover, industrial countries remain the largest exporters in the most polluting industries. Hence, we must reject the claim that internationally most pollution is created in countries with lax environmental standards. The main explanation for this finding is that other factors besides environmental regulation are much more decisive in determining production cost or location advantages. The share of compliance costs for environmental

regulation is small relative to the cost savings that can be gained from the availability of skilled workers able to operate advanced technologies, access to large consumer markets, and political stability. Polluting industries sometimes have high transport costs, so proximity to consumer markets is important. Polluting industries are also typically capital-intensive and operated by skilled workers. These factors make developed countries a more attractive location for polluting industries. Hence, differences between North and South in availability (i.e., endowments) of capital and skilled workers offset the differences in environmental regulation.

4.2 Is Trade Beneficial?

For given differences in endowments of capital and labor, relatively lax environmental regulation in some countries, particularly poorer ones, is likely to shift some of the dirty industries to the South. The question is whether this is always undesirable. According to a standard economic argument, there may be mutual benefits from relocation of pollution-intensive industries between different countries. For example, countries with low income but enough space to keep pollution away from their populations may be willing to accept pollution if this can greatly boost income. High-income countries with high population density care less about increases in income if they are obtained at the cost of pollution problems. In this case, both types of countries benefit if pollution is relocated from high-income to low-income countries.

Although lax environmental policies in the South seem to burden citizens in the South with pollution havens, workers in the North may be concerned about the stringent environmental policies in their countries. The North loses jobs in pollution-intensive industries and export performance in these sectors declines when these industries move to the South. However, at the same time, trade with the South allows the North to improve performance in clean industries so that again trade creates mutual benefits for North and South (although these benefits may be obtained only after some time-consuming, and therefore costly, adjustments in the economy). Indeed, also if the North did not engage in trade with the South, environmental regulation would result in job losses in polluting industries and shifts to clean industries, but it would forego the benefits from importing and consuming goods that can be produced at lower costs in other countries and the opportunity to export goods that can be produced at a comparative advantage in the North.

In general, differences in environmental regulation may reflect differences in social costs to deal with pollution and form a true basis for mutually beneficial trade. However, if national environmental regulation ignores the average citizen's preference for pollution and fails to reflect the social costs of pollution, trade is not always mutually beneficial. This is why large multinational corporations have been criticized when they negotiate concessions with elite governments and exploit their monopoly power. However, evidence on unbalanced benefits and costs from the activities of such firms is not available.

Other situations in which trade does not necessarily lead to mutual gains occur when pollution is transboundary—that is, when pollution from one country harms citizens of other countries. Since national governments are not likely to take into account these pollution spillover effects, they tend to set pollution regulation too lax. For all countries to gain from trade, environmental regulation in each country should reflect the cost of pollution not just to its own citizens but also to the world community as a whole. Thus, if SO₂ emissions in the United Kingdom hardly affect British citizens but cause forests in Scandinavia to die, the United Kingdom may be tempted to impose lax regulation of SO₂ emissions, thus creating a comparative advantage in SO₂-intensive goods and boosting export of these goods. This trade would harm Scandinavia.

A final situation of harmful international trade may arise when countries have strategic reasons to set overly lax or overly stringent environmental policies. Some environmentalists fear that governments set environmental policies overly lax. Governments may have an incentive to accept environmental standards that are lower than is socially desirable in order to attract foreign direct investment or give domestic industry a cost advantage over foreign firms, thus enabling them to make higher profits (which has been called ecological dumping). However, if many governments try to protect their polluting industries in this way, they may all fail in a “race to the bottom”: The lax standard in one country is offset by the response in the other country to also lower environmental regulation. Only if firms' location choices and cost levels are sufficiently responsive to differences in environmental standards can national governments use environmental policies as a strategic instrument against other countries. In the past, the lack of strong evidence for pollution havens has been interpreted as evidence against significant firm responses to environmental stringency and the fear of strategic use of environmental policy seemed ungrounded. Recently,

however, evidence of firm responses has become stronger so that strategic environmental policy may still be reason for concern.

4.3 Is Trade Good for the Environment?

Although there is no hard evidence on the magnitude of the net benefits (or costs) of trade in polluting goods, and there is no strong evidence regarding major effects of environmental regulation on trade and foreign direct investment, there are detailed analyses of the effects that an increase in world trade has on pollution in different countries. Trade has significant effects on pollution, but often more trade implies a cleaner environment.

World trade has increased enormously during the past few decades. Between 1973 and 1998, world merchandise exports and global foreign direct investment increased at annual rates of 9 and 14%, respectively. This expansion of trade is not responsible for increases in world pollution. The share of dirty products in world trade has declined. Comparing open economies (countries with high shares of exports and imports in GDP) to closed economies (countries with little foreign trade), statistical analysis shows that open economies have less pollution per unit of GDP (other things being equal). Open economies tend to adopt clean technologies more rapidly.

Trade often has a beneficial impact on the environment because of the interaction between comparative advantages and adjustment of environmental regulation. High-income countries tend to have higher environmental standards. However, high-income countries can also produce pollution-intensive (which are often skilled-labor or capital-intensive) products at a relatively low cost due to access to advanced technology and large endowments of skilled labor and capital. This implies that high-income countries have a comparative advantage in producing pollution-intensive goods. When they liberalize trade, they tend to export more of these goods. Thus, world trade becomes greener since trade shifts production of polluting goods to countries in which environmental standards are higher. Moreover, international trade increases income, which leads to a policy response since richer countries tend to impose more stringent environmental regulation. Thus, countries with lower incomes tighten their environmental regulation. Multinational firms adopt environmental practices overseas that are (partly) determined by environmental regulations in their home countries under pressure of consumer actions and for fear of loss of reputation.

Although evidence on an aggregate level suggests that international trade is benign to the environment, in certain areas trade has a serious adverse impact on the environment. Certain types of pollution or natural resource use remain unregulated, and it is in these areas that there are concerns. One example is transport. The growth in world trade has been accompanied by increases in transport movements, which are heavily energy intensive. The pollution costs (externalities) associated with transport are not fully reflected in regulation so that welfare losses and excessive pollution may stem from the increase in transport. Since the costs of transport cross borders, appropriate policies require international coordination, which is often difficult to establish. Second, transport involves the problem of invasive species: The accidental or intentional introduction of harmful nonindigenous species of plants and animals can damage environmental resources. The cost of screening and quarantines is often prohibitive. Third, the concentration of agriculture in certain locations without appropriate regulation puts stress on biological diversity with potential loss of soil quality in larger regions in the long term. It destroys the habitat of species and may lead to extinction of species and loss of biodiversity.

5. ENVIRONMENTAL REGULATION AND PRODUCTIVITY

5.1 The Cost of Environmental Regulation

Environmental regulation aims at improving environmental quality but may come at the cost of jobs, productivity, or other undesirable economic effects. To justify environmental regulation on economic grounds, it should be shown in a cost-benefit analysis whether the benefits are large enough to incur the costs. In some cases, the costs are found to be small or even negative; that is, there are benefits to be reaped not only in terms of environmental improvement but also in other fields. Estimates of the total costs of environmental regulation vary enormously between studies.

The traditional argument is that environmental regulation imposes costs on firms as well as governments and consumers. Governments incur costs of administration of the regulation. Firms are burdened with costs because they have to install and operate pollution control equipment and intensify monitoring, and they face higher administrative burdens.

Second, under certain types of regulation, they may face pollution charges, expenditures on buying pollution permits or quotas, or fines insofar as they emit more than the quotas that are allocated to them. Third, depending on the degree that they pass on costs through higher prices, demand for their output will decline and they will suffer from lower profits.

The total burden on polluting firms depends heavily on firm characteristics and the employed technology. Estimates of direct costs of complying with environmental regulation vary from less than 1% of GDP in 1972 to more than 2.5% at the end of the 1990s in the United States and slightly lower in Western Europe. Petroleum and coal industries faced pollution control capital expenditures up to one-fourth of total capital expenditures, whereas for rubber and plastics production industries the number was only 2%.

More crucial from an economic perspective is that the burden of regulation depends on the design of the regulation, particularly the flexibility firms are given to choose their own way to comply with the regulation. At the one extreme, the regulator prescribes the technology the firms have to use or the pollution control actions they have to take, leaving little flexibility for the firms. At the other extreme, the regulator allows firms a specified emissions allowance and leaves it to the firms' discretion to determine by what means to reduce emissions as well as whether to trade pollution allowances with other firms (system of marketable pollution permits). Compared to the former type of regulation (the command-and-control approach), the latter type (the market-based approach) minimizes the costs of regulation. First, it allows firms to choose the cheapest way to reduce emissions. Profit-maximizing firms will have an incentive to do so. Second, it gives incentives to firms that can reduce emissions at lowest cost to do most of the reduction and sell their pollution allowances to other firms. Thus, total pollution control costs are minimized. Alternative market-based instruments are charges (taxes) proportional to emissions.

Market-based instruments have been applied less frequently than command-and-control instruments, but their importance has increased. Marketable permits have been successful with respect to reductions of lead content in fuel, air pollutants, and SO₂ emissions in particular. Estimates of the cost savings potential of changing from command-and-control to market-based instruments range from 50 to 90%.

Sometimes, regulated firms have zero cost options to comply with more stringent environmental

regulation, or even make profits. Case studies have documented situations in which the reductions of materials or toxic chemical use for environmental reasons saved enough of the firm's expenditures on inputs to easily repay the investment cost of adjusting the production process. Even without environmental regulation, the investment would have been justified since it allowed for a real efficiency improvement.

It is unlikely that these situations arise systematically since managers of firms are supposed to search for opportunities to improve efficiency and exploit them even without intervention by environmental regulators. Thus, when firms maximize profits and regulation constrains their choice of technologies, the smaller choice menu must lead to fewer opportunities to operate at certain profit levels and regulation is costly. This conclusion is unlikely to change if we acknowledge the limits to profit maximization because of informational and managerial problems. No doubt, inefficiencies in firms' production processes and organization remain pervasive since managers' span of control is insufficient to assess and eliminate all of them. Because of the small share of environment-related costs, managers may systematically ignore environment-related inefficiencies. Then, environmental regulation may make managers more focused on cost savings potentials related to pollutants and materials use. However, it may come at a cost by turning their attention from other inefficiencies so that on balance it is unlikely that environmental regulation can systematically (i.e., economywide) improve overall efficiency.

Opportunities for cost savings have been explored by engineering or "bottom-up" studies. These studies start from an emissions reduction target for the economy as a whole and then identify all the technological measures by which this target can be achieved. For each of them, the researcher estimates the cost incurred by firms and adds these up to an economywide total. Often, the numbers are surprisingly low, suggesting that technological opportunities abound to improve the environment at low cost. The problem with these studies is that many hidden costs are ignored. Adoption of new technologies will not only change firms' emissions but also change product characteristics and quality, employment structure, and input demand. It requires adjustments beyond the simple installation of pollution control equipment. The entire production and marketing process may be affected.

The alternative method for assessing pollution control costs, the "top-down" approach, involves a statistical (and econometric) analysis of the relation-

ship between production costs and environmental regulation and between emissions and regulation across industries. The statistical results are used to make a projection of how costs are affected by a hypothetical change in emissions, which corresponds to the policy target. This requires building a stylized model of the entire economy, which involves the incorporation of many assumptions and additional statistical results regarding how markets work and react to policy changes. Top-down studies suggest much higher costs of environmental policy than do bottom-up studies. Hence, adjustment costs are substantial. Drawbacks of this method are that it necessarily relies on historical data on firms' reactions to regulation and it cannot incorporate the details of available new technologies to meet the environmental target. The results are also quite sensitive to the modeling structure and assumptions made.

Econometric techniques have also been used to estimate the effects of environmental regulation on employment. One would expect to find a negative correlation between stringency of environmental regulation and employment changes in pollution-intensive industries. Most studies find minor effects when relating environmental regulation across U.S. states to employment. Often, the correlation between employment in dirty industries and environmental regulation is positive. The main explanation is that we should not expect a causal relation from environmental stringency to employment but rather view both employment and regulation as being determined simultaneously by economic and political factors. States with a strong economy, industries with high productivity, and high income can afford—and may actually prefer—more stringent regulations than poorer states; at the same time, the former may experience strong employment growth as a result of the favorable economic conditions.

5.2 General Equilibrium Effects

To assess the overall costs of environmental regulation, one must measure not only how it directly imposes costs on a regulator and regulated firms but also how it affects other areas of the government's budget and how it affects customers and suppliers of regulated firms. Almost all activities in the economy are linked to each other: A consumer buys goods from firms and receives his or her income from (other) firms and pays taxes to governments, which subsidize firms and provide infrastructural goods for firms. Hence, regulating certain firms may affect the entire economy through these linkages. Such indirect

effects are labeled general equilibrium effects because they reflect the effects after all markets in the economy have reacted and adjusted such that a new equilibrium in the economy as a whole results.

The linkages between economic activities may sometimes mitigate the costs of environmental regulation and sometimes reinforce them. When regulated firms reduce their polluting activities, they buy less from their suppliers, which also experience a loss in sales. However, other firms may benefit. Regulated firms may need to buy specialized equipment to control pollution, so firms selling this equipment benefit. When regulated firms pass on their higher costs through higher prices, their customers switch to other producers. Hence, firms likely to benefit from regulation are those that supply goods that customers view as a suitable substitute for the regulated firms' goods.

The overall cost of environmental regulation is the sum of all gains and losses. The indirect gains are not likely to fully offset the direct costs for one obvious reason: Regulation aims at reducing pollution, and this requires reductions in polluting inputs. If the economy as a whole has to cut back on inputs, total output is likely to decline, too. This effect may be reinforced over time through effects on investment. Investment in new capital goods fuels growth in the economy. The return to investment is likely to be smaller if new capital goods have to be run with less energy or polluting inputs so that the incentive to invest is crowded out through more stringent environmental regulation. Thus, growth may be impaired by environmental regulation.

The economywide impact of environmental regulations can only be estimated, rather than directly measured, since we cannot observe how the economy would have evolved in the absence of the regulation. Estimates therefore heavily depend on the model that is used for the estimation. Based on standard macroeconomic or growth-theoretic models, estimates of the total cost are on the same order of magnitude as the direct pollution control costs, approximately 2% of GDP. Some researchers have noted that the surge in regulatory costs coincides with the growth slowdown in the 1970s. However, all calculations show that only a very small portion of the slowdown can be attributed to environmental regulation.

5.3 In Search of "Double Dividends"

Despite the fact that environmental regulation amounts to restricting inputs in the economy and is

therefore likely to reduce output, counterarguments have been advanced that environmental regulation may bring about large gains. The key argument is that environmental policy shifts economic activity away from polluting sectors to other sectors in which the contribution of economic activity to welfare is larger. Hence, although regulation forces the economy to reduce polluting inputs, it may increase total output by improving the average efficiency at which inputs are employed.

5.3.1 Elimination of Inefficient Subsidy and Tax Programs

Efficiency gains can arise from environmental regulation if efficiency differs between sectors, particularly if efficiency in polluting sectors is lower than in clean sectors. Normally, differences in efficiency are eliminated by market forces, but (nonenvironmental) regulation may distort these market forces and create sectors in which resources are inefficiently used. Important examples are agriculture and energy. In the OECD, agricultural output is heavily subsidized so that high-cost/low-productivity agricultural activities in these countries remain viable. This support creates inefficiencies, and the tax revenue could be better spent from an efficiency standpoint. The widespread use of energy subsidies to firms also creates support for low-productivity, energy-intensive production. If these subsidies are abolished and replaced by environmental regulation that reflects, for example, the adverse impacts of pesticide on soil and water quality or those of energy use on world climate, efficiency gains can be large by shifting economic activity away from inefficient farming and energy-intensive industries.

5.3.2 Environmental Tax Reform

Gains from environmental regulations have also been claimed to be feasible when they can be combined with reductions in certain taxes. On the one hand, governments have relied on taxes on labor and capital to fund their budgets. These taxes may reduce the incentive to participate in the labor market or to invest in capital so that the taxes distort markets and create inefficiently low levels of employment and investment. On the other hand, environmental taxes (charges on emissions and auctioning off of pollution permits) not only give incentive to address environmental problems but also raise revenue.

Indeed, the first important insight from the theory of public finance theory is that it is better to "recycle" the revenues from environmental taxes through reductions in distortionary taxes than to use them

to fund the general government budget. However, the second insight is that replacing distortionary taxes on labor and capital by revenue-raising environmental taxes will reduce the efficiency of the tax system and increase the burden of taxation. That is, the improvement in the environment (a first dividend) is not likely to be offset by a cheaper way for the government to raise taxes (which would give a second, nonenvironmental, dividend). The reason is that a change from taxing labor and capital to taxing pollution narrows the base on which the taxes can be levied. Capital and labor income each constitute a much larger fraction of total income than the share of pollution in total cost. As noted previously, environmental taxation reduces inputs in production and forces firms to spend more on pollution control. This reduces output and income in the economy, the tax base becomes smaller, and the government raises less tax revenue from nonenvironmental taxes. The revenues from environmental taxes are insufficient to make up for the loss of revenues from nonenvironmental taxes.

In practice, the replacement of nonenvironmental taxes by revenue-raising environmental taxes may yield a double dividend if the existing tax system is highly inefficient in other, nonenvironmental dimensions (e.g., if capital is excessively taxed relative to labor). In this case, the tax system can be improved even without the introduction of environmental taxes, but vested interests or distributional concerns may prevent these tax reforms. Results from economic simulation models suggest that gains from the combination of environmental taxation and tax reform may yield substantial benefits in practice but not for all groups in society. The major challenge is to find ways to compensate losers from tax reform without providing the wrong incentives. For example, workers and capital owners in polluting industry are likely to be negatively affected by environmental tax reform, but compensating them for losses should not reduce their incentive to invest funds and to find alternative employment outside the polluting industry.

5.3.3 Environmental Quality as a Source of Productivity

Environmental regulation has short-term costs but long-term gains. These gains are primarily in terms of a cleaner environment; they take time to materialize since the recovery of environmental quality is time-consuming. However, as a feedback effect, a cleaner environment may improve productivity in the economy, which may offset the economic costs. The most obvious example is the effect on productivity in

sectors that heavily rely on environmental resources. Policies to improve water and air quality of certain regions may boost the attractiveness of these regions for tourism or residential development. Hence, profits and productivity of tourism and construction may increase. Policies aimed at marine resources may also boost fish populations, thus improving productivity of the fishery. Forest management may give rise to both environmental benefits and improved yields for forestry.

In general, productivity of sectors or the economy as a whole is measured by multifactor productivity, which is a measure of the amount of outputs per unit of inputs. The factor productivity of conventional inputs, such as labor and capital, may increase if environmental quality improves. Capital equipment and residential buildings may depreciate more slowly if there is less air pollution. The productivity of labor may benefit from health improvements that are induced by better environmental quality. Less working days are lost if health improves, which increases output per unit of input. Health care costs decrease, alleviating private and government budgets. Substantial costs can be avoided if climate change, such as an increase in sea level (which directly affects agriculture, urban areas in coastal regions, and dike maintenance costs) and the loss of harvest in agriculture due to draughts or changes in precipitation, is mitigated.

5.4 Confronting Costs and Benefits

Most empirical research suggests that improvements in environmental quality require costly regulation, despite the various general equilibrium effects on efficiency and productivity that may mitigate the direct costs of regulation. The focus on costs of environmental regulation may distract attention from its benefits: These costs are incurred to achieve environmental benefits. It is worthwhile to incur the costs if the benefits outweigh the costs in a social cost–benefit analysis.

The social benefits from environmental improvements are often major, but they are less visible than the costs of environmental regulation. Measuring the benefits requires putting a value on improvements in environmental quality, health, and environmental amenities. When questioned in surveys, individuals indicate that they are willing to give up large amounts of money and income to pay for these improvements. We can also infer individuals' valuation of environmental amenities from the amount of money individuals spend on traveling to locations with

environmental amenities or the premium people are willing to pay on houses located near them. The data from these studies can be used to construct an aggregate indicator of the value of environmental improvements, which can then be compared to the cost estimates to undertake a social cost-benefit analysis.

The problem that the value of environmental amenities is unobservable (that there are no market prices) can thus be solved. However, perceptions may still be biased. Since neither market transactions nor market prices are associated with environmental goods, the current system of national accounts does not include the value of environmental goods; it only measures the value of marketed goods. As a result, the environmental benefits that environmental policies create are not part of national income, and environmental improvements do not count as economic progress. Attempts have been made to adjust national income data so as to include imputed values of environmental improvements and degradation. These “green GDP” data would more appropriately measure well-being and progress than do current GDP data. As long as these adjustments are not standard, there seems to be a trade-off between environment and growth. Once we take into account the value of environment for well-being, however, there is a close positive correlation between environment and economic progress.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Economic Growth and Energy •

Energy Ladder in Developing Nations • Equity and Distribution in Energy Policy • Global Energy Use: Status and Trends • Market-Based Instruments, Overview • Market Failures in Energy Markets • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use

Further Reading

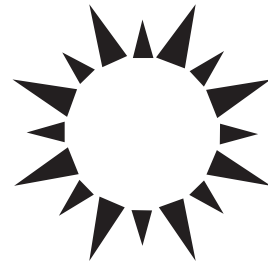
- Bovenberg, A. L., and Goulder, L. H. (2002). Environmental taxation and regulation. In “Handbook of Public Economics” (A. Auerbach and M. S. Feldstein, Eds.), Vol. 3. North-Holland, Amsterdam.
- Jaffe, A. B., Peterson, S. R., Portney, P. R., and Stavins, R. N. (1995). Environmental regulation and the competitiveness of U.S. manufacturing: What does the evidence tell us? *J. Econ. Literature* 33, 132–163.
- Krautkraemer, J. (1998). Nonrenewable resource scarcity. *J. Econ. Literature* 36, 2065–2107.
- Lieb, Chr. M. (2003). “The environmental Kuznets curve—A survey of the empirical evidence and of possible causes, Study No. 391.” Interdisciplinary Institute of Environmental Economics, University of Heidelberg, Heidelberg, Germany. [Available at <http://www.uni-heidelberg.de>].
- Schulze, G., and Ursprung, H. (eds.). (2001). “International Environmental Economics; A Survey of the Issues.” Oxford Univ. Press, Oxford, UK.
- Simpson, D., and Toman, M. (eds.). (2004). “Scarcity and Growth in the New Millennium.” Resources for the Future, Washington, DC.
- Smulders, S. (2000). Economic growth and environmental quality. In “Principles of Environmental Economics” (H. Folmer and L. Gabel, Eds.), pp. 602–664. Elgar, Cheltenham, UK.

Economics of Energy Demand

KENNETH B. MEDLOCK III

Rice University

Houston, Texas, United States



-
1. The Demand for Energy: An Introduction
 2. Modeling Energy Demand: An Analytical Framework
 3. Modeling Energy Demand: Empirical Analysis

Glossary

asymmetry In the present context, this refers to the asymmetric effect of changes in income or price on energy demand. Specifically, changes in energy demand caused by a Δ increase in income or price will not necessarily be of equal magnitude to changes caused by a Δ decrease in income or price (see also imperfect price-reversibility).

dematerialization The reduction of raw material (energy and material) intensity of economic activities, measured as the ratio of material (or energy) consumption in physical terms to gross domestic product in deflated constant terms.

economic structure The structure of production and consumption in an economy. Specifically, the share of goods and services produced and consumed by various economic sectors.

energy security The concept of maintaining a stable supply of energy at a “reasonable” price to avoid the macro-economic costs associated with interruption of energy supply or increases in energy price.

imperfect price-reversibility In the present context, this refers to the asymmetric effect of changes in price on energy demand. Specifically, changes in energy demand caused by a Δ increase in price, though opposite in impact, will not necessarily be of equal magnitude to changes caused by a Δ decrease in price.

income elasticity of energy demand The percentage change in energy demand caused by a 1% change in income, holding all other factors constant.

price elasticity of energy demand The percentage change in energy demand caused by a 1% change in price, holding all other factors constant.

primary energy requirements The quantity of energy consumed by the end user (total final consumption) plus the losses incurred during the conversion, transmission, and distribution.

rental price of capital The opportunity cost of purchasing/owning capital.

total final consumption The total quantity of energy consumed at end use.

user cost of capital The rental price plus the cost of operating and maintaining capital.

The demand for energy is a *derived* demand. Energy’s value is assessed not by its intrinsic appeal, but rather by its ability to provide a set of desired services. When combined with energy-using capital equipment, energy provides the work necessary to produce goods and services in both industry and in the household. As such, energy consumption, whether in the firm or household, is the result of a set of simultaneous decisions. Energy consumers must determine the quantity and type of capital equipment to purchase, where type is differentiated by technological characteristics such as efficiency and fuel input, and they must also determine a rate of capital utilization. Thus, the quantity of energy consumed is a function of capital utilization, as well as the engineering characteristics of capital equipment.

1. THE DEMAND FOR ENERGY: AN INTRODUCTION

To the extent that technologies make it possible, consumers of energy can choose between different sources of energy, and, in fact, energy use by source varies greatly across countries. In industrialized countries, the majority of energy consumed is referred to as commercial energy, which is differentiated from traditional forms of energy. The distinction between commercial and traditional energy can be made as an economic one. Specifically, commercial energy (for example, oil, natural gas, coal, and electricity) is traded in broader, larger-scale markets, whereas, traditional energy (for example,

waste and other biomass) is usually only traded, if at all, in smaller local markets. Typically, only in less developed countries will there be large shares of total energy consumed in traditional forms, since the proliferation of stocks of energy-using capital equipment, an occurrence associated with rising incomes, begets a substitution to commercial energy.

On a global scale, the 20th century was dominated by the use of fossil fuels. According to the U.S. Department of Energy, in the year 2000 global commercial energy consumption consisted largely of petroleum (39%), followed by coal (24%), and natural gas (23%), with hydro (6%) and nuclear (7%) splitting the majority of the remainder. However, this breakdown has not been constant over time either globally or within nations. In the United States, for example, wood was the primary source of energy consumption prior to industrialization (see Fig. 1). At the onset of industrialization, however, fossil fuels became the dominant energy source, and a number of factors, such as relative prices and environmental considerations, have dictated the composition of energy use since.

1.1 Energy Demand: A Context for Discussion

Energy is crucial to the economic progress and social development of nations. The development of new technologies and emergence of new production processes during the 19th and 20th centuries enabled an increase in productivity through the substitution of mechanical power for human and animal power. This, in turn, led to unprecedented economic growth and improved living standards, as well as, of course, large increases in energy consumption. For example, the consumption of coal and crude oil products for the locomotive, and later the automobile, significantly reduced the cost of transportation of people and goods, thereby creating growth opportunities for businesses and entrepreneurs. In addition, electrification has played a pivotal role in social development and welfare by making possible greater access to information via radio and television, providing a cleaner, more efficient means of storing and preparing food, and controlling enclosed environments to provide warmth and cooling. In the industrialized

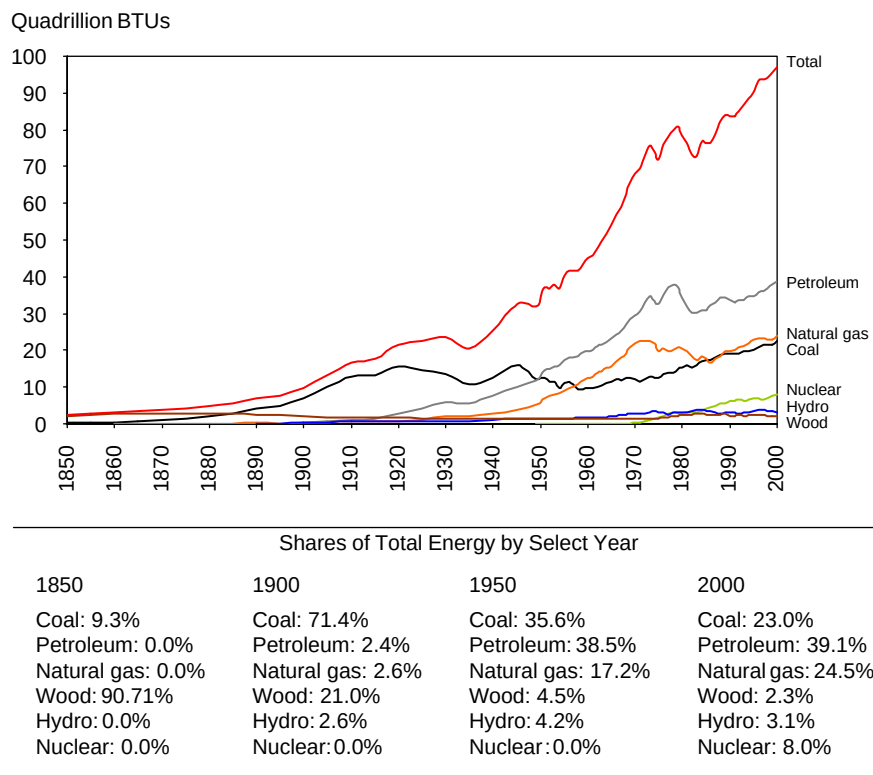


FIGURE 1 Energy demand in the United States increased fairly consistently from 1850–2000, but the source of that energy continually changed. Initially, wood was the primary form of energy use, but, driven by structural and technological change, coal, and later petroleum and natural gas, took the majority share of energy demand as the U.S. economy expanded. From the U.S. Energy Information Administration.

society of the 21st century, energy is necessary lest modern economic activity would cease and modern standards of living would diminish. Moreover, in developing nations, a lack of availability of energy supplies will prohibit further economic and social development.

Due to the importance of energy to economic and social progress, it is difficult to separate a discussion of energy demand from energy policy. In fact, a reoccurring theme in energy economics in general has been that of energy security. According to Bohi, Toman, and Walls, "Energy security refers to loss of economic welfare that may occur as a result of a change in the price or availability of energy" (p. 1), thus suggesting a link between energy and economy. The energy-economy link is supported by an abundance of empirical research demonstrating a strong correlation between increases in oil prices and decreases in macroeconomic performance, for both the United States and many other oil-importing industrialized countries. In fact, almost every recession in the post-World War II era in the United States, as well as many other energy-importing nations, has been preceded by a spike to the price of energy (see, for example, Hamilton; Ferderer; and Mork, Mysen, and Olsen). Because of this, governments, especially in those countries that are dependent on imported energy resources, have been concerned with maintaining security of supply at a reasonable price.

Projections of energy demand are necessary to develop national energy policies aimed at minimizing exposure to international price fluctuations and maximizing energy availability. On a less global scale, projections of energy demand are also important to the development of corporate business plans and the establishment of operational parameters for public utilities that must develop strategies to fulfill public service obligations. Therefore, the modeling and projection of energy demand is very important in the formation of policy in general, at both the national and local levels as well as in the private sector. Sound energy policy requires an adequate understanding of the economics of energy.

The economics of energy are generally concerned with the manner in which both nonrenewable and renewable energy resources are allocated over time and across space. A resource is considered nonrenewable when the sum, over time, of the services it provides is finite, or at least not replaceable in a reasonable time frame. Energy economists have been heretofore concerned primarily with the allocation of nonrenewable resources, largely because global en-

ergy consumption has consisted primarily of the consumption of fossil fuels. Dating back to Hotelling's classic paper, there have been numerous theoretical studies of optimal depletion rates and associated pricing rules given different sets of assumptions, as well as empirical studies designed to test theory. The well-known r -percent rule, as well as an apparent disregard for the impact of still underappreciated factors such as technological change in the extractive industries, the development of lower cost alternatives, and expansion of the resource base through exploration activities, influenced predictions by many scholars and industry analysts in the late 1970s and early 1980s that oil prices would top \$100 per barrel by 2000. This projection never occurred.

Also of considerable influence to these erroneous projections was an apparent extrapolation of demand trends without properly considering the impact of allowing for, among other things, structural change in economies as they develop and the effect of increased energy prices on the composition and characteristics of capital stocks. In general, as economies develop they move from agricultural to industrial to service economies. This structural transformation has important implications for the manner in which energy demand grows relative to income. Specifically, as less energy intensive activities make up a larger proportion of total economic activity the income elasticity of demand will decline. In short, the income elasticity of nations is not constant.

The characteristics of energy-using capital stocks are also an important determinant of energy demand that cannot be ignored when forecasting. When considering an entire economy, the type of capital employed by firms and in households is largely dependent on economic structure, but certain characteristics of capital, such as embodied technology, are sensitive to price. Therefore, price-induced technological change will alter the energy requirements for a given level of output, allowing for an expansion of economic activity without an appreciable increase in energy demand. To illustrate, consider the United States transportation sector. According to the Energy Information Administration, the average fuel rate of a passenger automobile increased from about 14.3 miles per gallon in 1978 to about 20.2 miles per gallon in 1990, an average annual increase of 2.9%. During the same time period, transportation fuel consumption increased at only 0.7% per year, despite increases in both motor vehicle stocks and motor vehicle utilization (miles driven per vehicle). Clearly, efficiency gains offset increased utilization and growing vehicle stocks thus

limiting the increase in fuel consumption. Apart from allowing output to increase for a given level of fuel consumption, an increase in the energy efficiency of capital stocks also leads to asymmetry in the response of energy demand to changes in price, as noted in Gately and Huntington. Specifically, if certain capital investments aimed at increasing energy efficiency are made in response to a price increase, they will not be undone when price falls. Thus, energy demand tends to respond differently to price increases than to price decreases.

1.2 Elasticity of Energy Demand

The experiences of the 1970s and 1980s led to an explosion in the number of energy demand studies, a trend that has been somewhat revitalized by the emergence of concerns about the emissions of greenhouse gases from the combustion of fossil fuels. A large majority of these studies have resulted in the estimation of, or made use of previously estimated, income and price elasticities of energy demand. Income and price elasticities of energy demand are of the utmost importance to forecasting energy demand. Moreover, the manner in which these relationships may or may not evolve in response to future economic growth, changes in economic structure, and various policy initiatives is also important. Income and price elasticities are, in general, the rules of thumb by which broader energy policy is directed. There is, however, no real consensus as to an appropriate value of income or price elasticity. Much of the disagreement in empirical research is certainly due to different model specifications, but it is also due to some confounding issues that are difficult to isolate, such as the effects of changes in technology, changes in capital stock utilization, capital stock turnover, shifts in economic structure, and changes in energy policy.

The income elasticity of energy demand is defined as the percentage change in energy demand given a 1% change in income holding all else constant, or

$$\varepsilon_y = \frac{\% \Delta e}{\% \Delta y} = \frac{de}{dy} \cdot \frac{y}{e},$$

where e denotes energy demand and y denotes income. This measure provides an indication of how demand will change as income changes. The income elasticity of energy demand had long been reported as being greater than or near to 1. Research, however, has indicated that this may be greatly overstated for industrialized countries, as there is increasing evidence that energy intensity is inversely

related to economic development, being greatly influenced by changes in economic structure and the state of technology.

The price elasticity of energy demand is defined as the percentage change in energy demand given a 1% change in price holding all else constant, or

$$\varepsilon_p = \frac{\% \Delta e}{\% \Delta p} = \frac{de}{dp} \cdot \frac{p}{e},$$

where p denotes the price of energy. This measure gauges the influence of energy price on energy demand and is often used as a barometer to assess the impact of various conservationist policies, such as emissions abatement programs or energy tax/subsidy schemes. For example, one can approximate the reductions in emissions achieved by taxing energy if the impact of that tax on the quantity and composition of energy consumption is known. The price elasticity can also be used to determine the impact of energy price disruptions on energy demand and, hence, any subsequent loss in macroeconomic output. However, the long- and short-run effects of energy prices can be difficult to disentangle. Energy demand is predicated on ownership of an energy-using capital good, the owner's decision as to the energy efficiency of the capital good, which is made when the capital is purchased, and utilization of the capital good. Once a capital good has been purchased, the capital utilization decision will dictate the responsiveness of energy demand to price in the short run, assuming the short run implies that capital and technology are fixed so that the consumer cannot "trade up" to higher energy efficiency. In the long run, however, the consumer is free to turn over capital equipment in favor of higher energy efficiency. Thus, the long-run response to price involves a decision on both capital utilization rates and energy efficiency.

There is an important consideration that bears mention here with regard to calculating income and price elasticity of energy demand. When estimating an income elasticity of energy demand, for example, it is not sufficient to simply divide the percentage change in energy demand by the percentage change in income. Nor it is not sufficient to simply use linear regression analysis to estimate an income elasticity of energy demand if income is the only explanatory variable. Both of these approaches completely ignore the influence that other variables may have on energy demand, which can lead to erroneous, statistically biased results. Nonetheless these are common mistakes often made by well-meaning energy analysts.

To illustrate, recall that income elasticity of energy demand is the percentage change in energy demand given a 1% change in income, *holding all else constant*. If we take energy demand to be a function of income, price, and possibly some other set of variables, then, in general, the total derivative of energy demand is given as

$$de = \frac{\partial e}{\partial y} dy + \frac{\partial e}{\partial p} dp + \dots$$

If we hold all factors but income constant, the preceding equation reduces to

$$de = \frac{\partial e}{\partial y} dy.$$

We can then multiply both sides by

$$\frac{1}{dy} \cdot \frac{y}{e}$$

to yield

$$\frac{de}{dy} \cdot \frac{y}{e} = \frac{\partial e}{\partial y} \cdot \frac{y}{e}.$$

The left-hand side of this equation is the income elasticity of energy demand, as defined earlier, but we only arrive at this result when all variables except income are held constant. It is a simple matter to show that not holding all other variables constant will result in a remainder term on the right-hand side of the preceding equation (i.e., the term in brackets):

$$\frac{de}{dy} \cdot \frac{y}{e} = \frac{\partial e}{\partial y} \cdot \frac{y}{e} + \left\{ \frac{\partial e}{\partial p} \cdot dp \cdot \frac{1}{dy} \cdot \frac{y}{e} + \dots \right\}.$$

If, therefore, we seek an estimate of income elasticity, we cannot ignore the other relevant variables as it introduces a bias similar to the term in brackets. For example, if energy consumption has been increasing at 2% per year for 10 years, and income has been increasing at 3% per year for the same time period, a naïve approximation of income elasticity of energy demand would be 0.67 (= 2%/3%). However, if we consider that price may have been changing during the past 10 years, then our naïve estimate is incorrect. Specifically, if price had been falling, then, given a downward sloping demand curve, an income elasticity of 0.67 is an overestimate and can lead to serious problems when forecasting future demand. Therefore, when modeling energy demand, it is important to recognize that many variables influence energy use, and those variables must be carefully accounted for. This is precisely why a sound

theoretical foundation is necessary when analyzing energy demand.

1.3 A Note on Energy Accounting

Before discussing the aspects of energy demand modeling about which the researcher must be conscious, it is important to highlight exactly what is referred to when discussing energy demand. When modeling energy demand, the researcher is usually interested in the quantity of energy required to perform a specified activity. For example, operating a drill-press requires electricity, driving an automobile requires gasoline or diesel fuel, and heating a household requires natural gas, heating oil, or electricity. Simply accounting for the energy consumed at end use, however, does not fully account for all the energy consumed to make the activities listed above possible. Specifically, raw energy supplies (such as coal, crude oil, and wet natural gas), while they can be directly consumed, are generally converted into products (such as gasoline, heating oil, propane, ethane, dry natural gas, and electricity) that are then consumed in various end uses. The energy conversion industries, such as refining, processing, and electricity production, use primary energy inputs to produce usable products. Energy balance tables are a means of accounting for the process of energy conversion and eventual consumption. Figure 2 illustrates the general flow that energy balance tables capture. Total primary energy requirements (TPER) are the energy needed to produce the energy necessary for total final consumption (TFC). TPER is greater than TFC due to the energy expended in the conversion, transmission, and distribution activities.

2. MODELING ENERGY DEMAND: AN ANALYTICAL FRAMEWORK

As stated earlier, the demand for energy is derived from the demand for goods and services provided by energy-using capital equipment. In residences and commercial establishments, energy is consumed, in conjunction with some energy-using durable good, to provide services such as space heating and cooling, lighting, cooking, and cleaning. Similarly, in industry, energy is used to produce steam and power automated processes, and in transportation, energy is required to make use of the services provided by automobiles, airplanes, trains, and so on. Therefore,

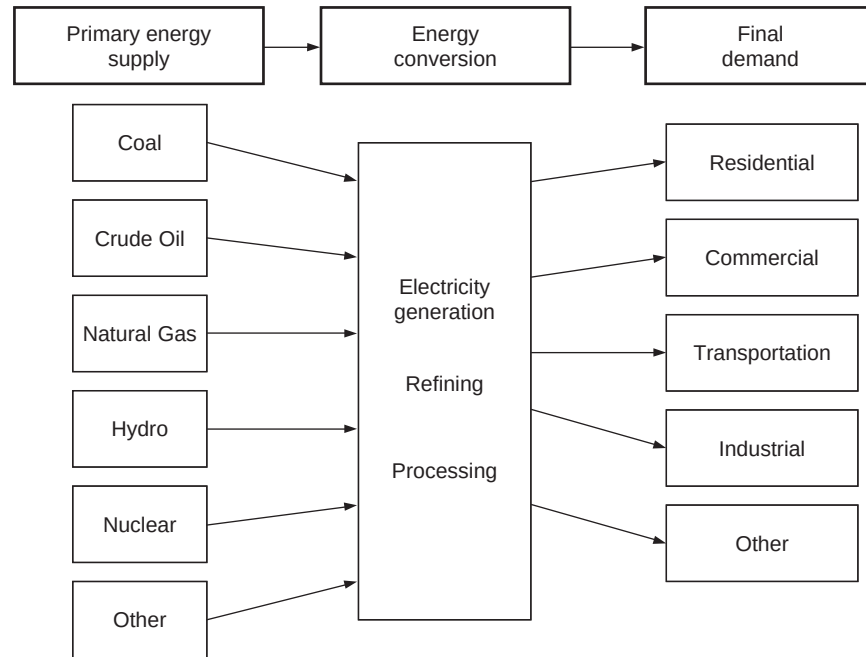


FIGURE 2 Primary energy supply is the energy required to meet final demand. Some energy is expended in the energy conversion industries in order to make usable energy products, so primary energy supply is greater than final demand.

implicit in the decision to consume energy is an underlying set of decisions that each energy consumer faces. Consumers of energy must decide to purchase the capital equipment that provides the desired service (heating, cooling, transportation, and manufacturing), they must choose the characteristics of the capital equipment (energy efficiency and type of energy input), and they must choose the rate at which the capital will be utilized. These decisions are conditional on, among other things, the energy cost of capital utilization. It is this particular aspect of demand modeling that is sometimes difficult to capture. In the short run, capital stocks are assumed fixed, so energy price fluctuations will influence the utilization rate of capital. In the long run, the energy consumer can replace capital equipment thereby improving energy efficiency. In empirical studies, this can pose difficulties due to limited time series data (both in length and breadth), thereby confounding the differentiation between the long- and short-run impacts of energy price movements. It is therefore important to develop a theoretical framework that is capable of addressing these concerns in order to lend guidance to assessing and conducting empirical analysis.

The quantity of energy consumed in the production of goods and services is dependent on the technical nature of energy-using capital equipment. The relationship between energy consumption and

capital equipment can be given as

$$\text{energy} \equiv \left(\frac{\text{capital utilization rate}}{\text{capital equipment efficiency}} \right) \times \text{capital stock}, \quad (1)$$

where efficiency is an equipment-specific technological parameter, defined to be service output per unit energy. It is important to note that this relationship is, in general, equipment specific, and there is one such relationship for each type of capital equipment. This is important because it is an oversimplification to aggregate capital stocks and energy if one wishes to analyze the relationship between capital and energy. In the case of motor vehicles, efficiency is miles per gallon, and in the case of central air conditioners, efficiency is British thermal units (BTUs) per watt-hour. As efficiency increases, the energy required per unit of service declines. The capital utilization rate denotes the service rendered by capital equipment, such as miles driven per vehicle, thus energy use is positively related to capital utilization. Equation (1) is an identity because its units simplify to energy $\equiv$ energy.

Notice that the right-hand-side of Eq. (1) contains the three choice variables involved in any decision to consume energy. Because capital stocks and technology are key determinants of energy use, static models of energy demand are limited in their ability to

capture the full response of energy use to changes in economic variables, such as income and price. Nevertheless, the static model built on sound micro-economic principles provides a reasonable starting point. As such, and because static models are widely used, we will begin with a brief description of them before moving to a dynamic framework.

2.1 Static Models

Following the energy balance structure indicated in Fig. 2, the demand for energy originates in different end-use sectors: industrial, residential, commercial, transport, and other. More generally, energy consumption can be modeled as either providing utility to households or as an input in the production process for firms. For treatment of the former problem, it is generally assumed that a consumer maximizes his or her welfare by consuming some bundle of goods and services subject to a budget constraint. Energy is assumed to be proportional to the service it provides, and therefore consumer welfare is influenced by energy consumption. (As demonstrated later, if one wishes to allow for substitution possibilities across types of energy commodities, it is a simple matter to incorporate various types of energy along with their specific costs.) The resulting demand for energy by the representative individual in this problem will be influenced by a number of factors, including but not limited to energy price, the price of other goods, and income (other variables, such as weather and durable goods stocks, can also have an influence, depending on the model specifications).

Expressing the preceding argument mathematically, we have that a representative consumer maximizes utility, $U(z, e)$, which is function of energy consumption, e , and all other consumption, z , subject to the constraint that expenditures cannot exceed income, y . As usual, we assume utility is a concave twice-differentiable function of its arguments. If the modeler wishes to examine the substitution possibilities across energy products, the energy variable can be a vector of n energy products, $e = [e_1, e_2, \dots, e_n]$. Allowing the price of good j to be represented as p_j , the consumer is assumed to

$$\max_{z, e_1, \dots, e_n} U(z, e_1, \dots, e_n)$$

subject to $y \geq p_z z + p_{e_1} e_1 + \dots + p_{e_n} e_n$.

The first-order necessary conditions for a maximum for this problem can be solved to yield demand equations for each of the energy products and for all other consumption. The resulting function for energy

demand is given as

$$e_j^* = e_j(y, p_z, p_{e_1}, \dots, p_{e_n}), \quad (2)$$

where the star denotes an optimal choice. Thus, the demand for energy commodity j is a function of income, own price, the price of all other energy commodities, and the price of all other consumption goods.

A firm also demands energy to produce goods and services. The typical firm is interested in maximizing profits, or minimizing the cost of producing a given level of output, and it must purchase energy as well as all other inputs to production. As it turns out, the firm's problem is not unlike the household's problem. The firm's demand for energy in this problem can be shown to be a function of its output, the price of energy, and the price of all other inputs.

Formalization of this problem is as follows. Consider a representative firm that minimizes costs, c , for a given level of output, $\bar{q}$. Output is a function of capital (k), materials (m), labor (l), and energy, $q = f(k, l, e_1, \dots, e_n, m)$, and costs are the sum of payments to the factors of production, $c = rk + wl + p_{e_1} e_1 + \dots + p_{e_n} e_n + p_m m$, where r are rent payments to capital, w are wages paid to labor, and p_j is the price of factor j . The firm's problem is then stated as

$$\min_{k, l, e_1, \dots, e_n, m} c$$

subject to

$$\bar{q} = f(k, l, e_1, \dots, e_n, m)$$

$$c = rk + wl + p_{e_1} e_1 + \dots + p_{e_n} e_n + p_m m.$$

The reader can verify that the solution to this problem is not very different than the solution to the utility maximization problem, other than a change of terminology. (In fact, the dual to the household's [firm's] problem is one of cost minimization [profit maximization].) The firm will choose inputs such that their marginal values in production will be equal. Solution of the first-order conditions for a maximum for this problem yields a general function for the demand for energy

$$e_j^* = e_j(\bar{q}, r, w, p_{e_1}, \dots, p_{e_n}, p_m).$$

Thus, by Eq. (3), the firm weighs the price of all possible factor inputs when choosing its energy consumption for a given level of output.

These problems are static representations of energy demand. As such, they are most appropriate in deriving short-run demand functions, when capital and technology are fixed. Comparative statics can be a useful tool in this setting for analyzing the impact of a rise in capital stocks, say, but they disregard the

intertemporal choices an energy consumer faces with regard to type of capital, utilization of capital, and efficiency of capital. A more general approach would allow for a dynamic specification of demand. A dynamic model accounts for intertemporal choice of capital stock characteristics and utilization, as well as the effects of each on energy demand. As such, a dynamic framework is capable of addressing the asymmetries that exist in the energy-income and energy-price relationships. In the static setting such explorations are left to pontification and conjecture.

2.2 Dynamic Models

Dynamic models of capital ownership and energy use allow for a more complete analysis of energy demand because, for one, they are capable of capturing factors that generate the asymmetries mentioned previously. In addition, dynamic models incorporate the intertemporal choices that a consumer (firm) must make when maximizing welfare (profits) over some time horizon. Medlock and Soligo (2002) developed a dynamic model of motor vehicle ownership and fuel use that generalizes to many other types of energy-using capital equipment. They showed that the demand for energy is determined by the consumer's wealth, the price of energy, the price of other goods, the user cost (or cost of owning and maintaining) capital, and the energy efficiency of the energy-using capital stock. Because the model they developed is dynamic, it captures the notion that the decision to consume energy is made jointly with the decision to purchase and maintain a stock of energy-using capital equipment, the latter of which incorporates the elimination of potential intertemporal arbitrage opportunities.

It is useful to consider the framework developed by Medlock and Soligo (2002). To begin, consider a modification of the household's problem. For ease of exposition, energy consumption has been denoted as a single commodity. Relaxation of this constraint allows for multiple types of capital and hence multiple types of energy consumption. Denoting time using the subscript t the consumer will maximize the discounted sum of lifetime utility,

$$\sum_{t=0}^T \beta^t U(z_t, e_t),$$

subject to the constraint that capital goods purchases (i_t), purchases of other goods (z_t), purchases of energy (e_t), and savings (s_t) in each period cannot exceed this period's income (y_t) plus the return on

last period's savings ($((1+r)s_{t-1})$). It is assumed that capital goods depreciate at a rate δ , savings earn a rate return r , the discount rate is such that $0 < \beta < 1$, and all initial conditions are given. (Note we can complicate the problem somewhat by allowing access to credit markets, but this adds little to the present context.)

The problem is thusly formulated. Consumers will

$$\max_{z,e,s} \sum_{t=0}^T \beta^t U(z_t, e_t)$$

subject to

$$p_{z,t}z_t + p_{e,t}e_t + p_{k,t}i_t + s_t \leq y_t + (1+r)s_{t-1}$$

$$e_t \equiv \frac{u_t}{\varepsilon_t} k_t$$

$$i_t = k_t - (1-\delta)k_{t-1}$$

$$z_t, u_t, k_t \geq 0 \quad t = 1, \dots, T$$

where $p_{j,t}$ is the price of good j and all initial values are given. Note the second constraint is a restatement of Eq. (1). It is through this constraint that the relationship between energy and capital is made explicit.

The marginal conditions for this consumer's problem lend considerable insight into the manner in which consumers will respond to changes in income and energy prices. Specifically, the marginal conditions yield

$$U_k \frac{u_t^*}{\varepsilon_t} = U_z \left(p_{e,t} \frac{u_t^*}{\varepsilon_t} + p_{k,t} - p_{k,t+1} \left(\frac{1-\delta}{1+r} \right) \right), \quad (4)$$

where the star denotes the optimality. Equation (4) indicates that the consumer will allocate income such that the marginal value of the energy services attained from the capital stock is equal to the marginal value of consumption of all other goods. The term in brackets in Eq. (4),

$$p_{e,t} \frac{u_t^*}{\varepsilon_t} + p_{k,t} - p_{k,t+1} \left(\frac{1-\delta}{1+r} \right),$$

is defined as the "user cost," μ_k , of capital stock, and it is this term that bears significant mention. The first term, $p_{e,t} \frac{u_t^*}{\varepsilon_t}$, indicates that the consumer chooses user cost to the extent that capital utilization is a choice variables. If the optimal choice of utilization were such that $u_t^* = 0$ user cost reduces to

$$p_{a,t} - p_{a,t+1} \left(\frac{1-\delta}{1+r} \right),$$

which, following Diewert, is the rental price ($p_{R,t}$) of owning capital equipment.

Simultaneous solution of the first-order conditions will yield a set of demand equations, one for each of the arguments in the utility function at each of the

$t+1$ time periods. By Eq. (1), once a solution is obtained for k_t^* and u_t^* , we can solve for the optimal consumption of energy. Thus, the optimal choice of energy demand is determined by the optimal choice of capital stocks and capital utilization, with efficiency given. The demand for energy is found to be a function of income, the user cost of capital, and the price of all other goods. Since user cost is a function of energy price, energy efficiency, and the rental price of capital, we can generically state the demand for energy as

$$e_t^* = e(y_t, p_{z,t}, p_{e,t}, p_{R,t}, \varepsilon_t). \quad (5)$$

Note that Eq. (5) is altered if efficiency is taken to be choice variable, although a distinction must be made between the long and short run. This is not done here for ease of exposition, as it would require the differentiation of types of possible capital purchases by the use of additional constraints. (One such example is Tishler.) The added insight gained from allowing a choice across multiple energy efficiencies is that consumers may choose to increase efficiency if energy prices rise in order to prevent the user cost of capital from rising. This type of response allows for the previously mentioned asymmetries.

We could also increase the complexity of the problem by allowing for multiple types of capital equipment (a household typically has refrigerators, air conditioners, ovens, etc.) and multiple fuel sources, which would result in, for example, natural gas prices appearing in, say, an electricity demand function. Again, however, this is not done here for ease of exposition. Such analysis is left to the reader for further exploration.

3. MODELING ENERGY DEMAND: EMPIRICAL ANALYSIS

Empirical studies often use the general form of Eq. (2) or Eq. (5) to specify a function to be estimated as an approximation of the “true” demand function. This can lead to substantial variation in published results since parameter estimates are sensitive to the model specification. This is not an unimportant fact since policy recommendations are often made on the basis of empirical research. Moreover, it highlights the importance of sound theoretical justification for model specification.

A simple and popular demand specification is the log-linear demand equation given as

$$\ln e_t^* = a + \alpha \ln y_t + \eta \ln p_{e,t}, \quad (6)$$

where e^* is the “optimal” quantity of energy demanded, y is income, p_e is energy price, and a is a constant. The subscript t denotes the time period. (Such a demand specification implies a demand function of the form $e_t^* = Ay_t^\alpha p_{e,t}^\eta$). Equation (6) is easily estimated to directly yield an income elasticity, α and a price elasticity, η . Because energy demand is not necessarily a function of only income and price, additional parameters are also sometimes included in the specified demand equation. For example, when estimating energy demand across countries, a *per capita* income measure is generally preferred because it normalizes income with respect to population. Population may also be added explicitly as an independent variable, as well as other parameters of importance such as capital stocks, weather/climate, and the prices of alternative fuels, the latter two of which are particularly important when modeling the demand for individual fuel types whose consumption varies seasonally.

An alternative specification to Eq. (6) allows energy demand to be a function of past realizations of income and price. It can be shown that this is equivalent to assuming the existence of a lag adjustment mechanism of the form

$$(\ln e_t - \ln e_{t-1}) = \gamma (\ln e_t^* - \ln e_{t-1}),$$

where $\gamma \in [0, 1]$ is the speed of adjustment. Eliminating the term $\ln e^*$ in Eq. (6) by direct substitution results in the equation

$$\ln e_t = \gamma a + \gamma \alpha \ln y_t + \gamma \eta \ln p_{e,t} + (1 - \gamma) \ln e_{t-1}, \quad (7)$$

where α is the long-run elasticity of demand with respect to income, η is the long-run elasticity of demand with respect to price, and $\gamma \alpha$ and $\gamma \eta$ are the respective short-run elasticities.

While econometric specifications such as Eqs. (6) and (7) are attractive because they are easily estimated and yield elasticity terms directly, they implicitly assume that elasticity is constant across all ranges of income and price, a point that has been challenged in the literature. Moreover, it does not adequately capture the influence of changes in the composition of capital stocks over time, a phenomenon that could alter energy required per unit output and result in asymmetric demand responses to changes in income and price. One means of controlling for these factors directly is to estimate the components of Eq. (1) in order to derive an estimate of energy demand for a specific end use. Olof and Schipper take such an approach with regard to forecasting long-term motor fuel demand in several

countries. The advantage of such an approach is that it is robust to the simultaneous set of decisions the consumer makes when determining energy consumption.

Other adjustments can and have been made to Eqs. (6) and (7). For example, one can employ dummy variables on the price and income coefficients, thereby separating price and income effects into increases and decreases, in order to capture any asymmetric response of energy demand. One can also specify a demand function that is nonlinear in all or some of the included variables. Such a specification could allow, for example, a declining income elasticity of energy demand. Still other approaches involve an entirely different approach to deriving an equation for energy demand. This class of models involves estimating the first-order conditions that would arise from a theoretical model similar to that presented earlier. Thus, rather than using general solutions such as those given in Eqs. (2) or (5), consumer utility functions or production functions are to be specified, and analytical solutions are obtained. While potentially cumbersome mathematically, this type of approach has the advantage of including all relevant variables, and the interactions therein, from theoretical framework in a simultaneous equations setting.

In sum, there are numerous specifications that can be potentially explored when analyzing energy demand. But it must be noted that the different types of alternative specifications are tailored to identify specific characteristics of the data being analyzed and are usually specific to the question posed by the researcher. While no one specification is necessarily more correct than another, some are certainly more adequate than others in capturing certain aspects of energy demand, and thus they highlight what should, and sometimes should not, be considered in empirical modeling.

Regardless of the specification chosen, a strong understanding of the economic forces at work in individual markets is crucial to empirical analysis. For instance, when modeling the near-term demand for natural gas, it is important to account for both weather and alternative fuel prices (heating oil for example), as well as income and own price. This is due to the fact that natural gas demand increases considerably in the winter months as heating demand increases. Therefore, for reasons discussed earlier, not accounting for weather and alternative fuel prices will lead to biased estimates of income and price elasticity.

3.1 Empirical Considerations: The Energy-Income Relationship

3.1.1 Economic Development and Dematerialization

The assumption of constant income elasticity of energy demand is inconsistent with empirical evidence that energy intensity increases then decreases during the course of economic development. In general, econometric analysis of countries across time and space yield the result that the income elasticity of energy demand is higher in lower income countries. The energy intensity of an economy resembles an inverted U-shape across increasing levels of per capita income, as discussed in 2001 by Medlock and Soligo. This arises because various sectors grow at different rates and at different times during the course of economic development. Industrialization results in large increases in commercial energy use. This is followed by smaller increases in energy demand during the post-industrial phase of development, in which the service sector grows relative to all other sectors. Figure 3 illustrates the general pattern of energy intensity that arises during the course of economic development, as well as the general relationship between energy intensity and income elasticity of energy demand. The reader can readily verify that if the income elasticity of energy demand exceeds unity, then, holding all else constant, energy intensity is rising. Similarly, if the

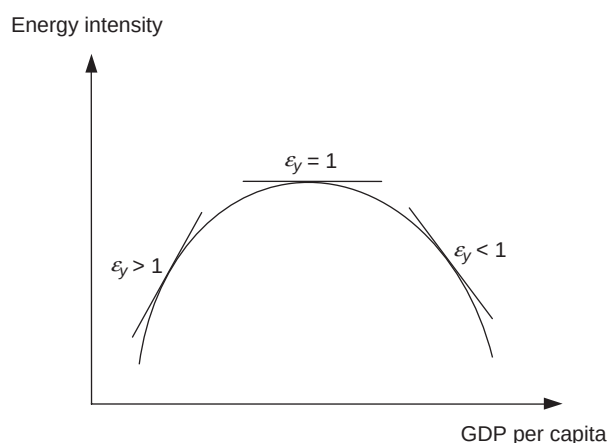


FIGURE 3 Generally, due to structural transformation and technological change, energy intensity follows an inverted U-shape during the course of economic development. This has implications for income elasticity. When income elasticity of energy demand exceeds unity, then, holding all else constant, energy intensity is rising. Similarly, if the income elasticity of energy demand is less than unity, then, holding all else constant, energy intensity is falling.

income elasticity of energy demand is less than unity, then, holding all else constant, energy intensity is falling. Note that declining energy intensity does not imply that energy demand declines, only that energy demand grows more slowly than output.

A rising then declining energy intensity is generally consistent with the theory of dematerialization, which was defined by Bernardini and Galli in 1993 as “the reduction of raw material (energy and material) intensity of economic activities, measured as the ratio of material (or energy) consumption in physical terms to gross domestic product in deflated constant terms.” There are two basic ideas that constitute the theory of dematerialization. One, intensity-of-use of energy and materials initially increases then decreases with increasing GDP in all countries. Two, the later in time a country develops the lower its maximum intensity-of-use in energy and materials will be. These concepts, as they pertain to energy, are illustrated in Fig. 4.

The idea that the intensity-of-use of energy first increases then decreases is related to the impact of changing economic structure on energy use. In general, the process of industrialization results in large increases in commercial energy consumption, which causes energy intensity to initially increase. As

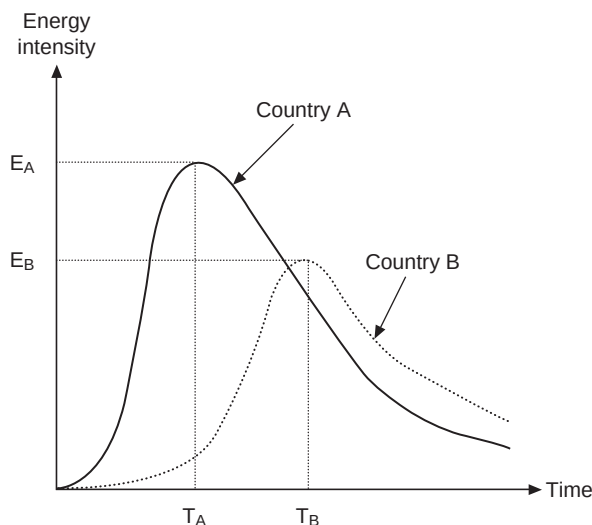


FIGURE 4 Consider two countries, country A and country B. Assume that gross domestic product per capita is positively correlated with time, and allow country A to develop more rapidly than country B. *Ceteris paribus*, the energy intensity of country A, reaches its peak (T_A , E_A) before country B (T_B , E_B). Moreover, country B benefits from the production processes and technologies developed by country A. Therefore, the peak energy intensity in country B is lower than the peak energy intensity in country A. Or, otherwise stated, $E_A > E_B$ because $T_A < T_B$.

consumer wealth rises, however, the structure of consumption changes increasing the demand for light manufactures and services. As these new demands are met, the structure of a nation's capital stock will change to reflect its output. Growth in the service sector relative to other sectors serves to reduce energy intensity as the energy required per unit of output from capital employed in the service sector is generally lower than that in the industrial sector.

The decline in intensity-of-use of energy and materials can also be hastened by the introduction of newer energy resources and materials with price and/or quality advantages, and also by the introduction of technologies that increase the energy efficiency of capital equipment. Industrialization in the West saw massive increases in the use of coal, as it was the primary fuel of choice for meeting energy requirements. However, increasing wealth generated new demands, and better, more suitable alternatives, namely oil and natural gas, eventually took the place of coal. For example, growth in private and small-scale transportation could be more easily facilitated with petroleum products than with coal, since the latter is more cumbersome and less efficient to meet that end. As environmental pressures mount and the associated costs are internalized, it is probable that ever cleaner alternative energy sources will become economically competitive, and therefore it is likely the intensity-of-use of other fuel sources will continue to rise over time as the intensity-of-use of fossil fuels continues to decline.

The second of the ideas central to the theory of dematerialization, which comes to bear because the diffusion of newer technologies and the economic/environmental lessons learned in the industrialized world could contribute to lowering the maximum energy intensity of developing nations, is difficult to test in a casual manner. While one can examine cross-country data to make comparisons, it is necessary to control for factors specific to individual countries that could influence energy intensity. Such research has not to date been completed. Nonetheless, the implication of the time dependency of maximum energy intensity is substantial. If developing countries, in general, will never reach the maximum intensity of energy use that was reached in industrialized countries in the 20th century, then expectations should be gauged accordingly. Long-term forecasting, therefore, should anticipate the possibility of these developments, particularly when assessing the impact of future economic development on the demand for energy.

3.1.2 Asymmetry in the Energy-Income Relationship

When considering the energy-income relationship, we must be mindful of the fact that the manner in which income influences energy demand is not symmetric. In other words, an increase in macroeconomic output will increase energy demand, but a decline in output will not necessarily have an equal and opposite effect. The reason for this is related to the arguments underpinning dematerialization, and possible explanations can be tied to Eq. (1). Specifically, different sectors grow at different rates and at different times during the course of economic development. As mentioned earlier, it is an oversimplification to assume Eq. (1) holds at the aggregate level. Doing so ignores the fact that capital stock characteristics are tied to specific capital stocks, and the proliferation of types of capital is very much dependent on economic structure. For example, the capital employed in industrial and manufacturing facilities is differentiated from the capital employed in households by the service they provide and the engineering characteristics they possess. Moreover, these items are more or less prevalent given the level of economic and social development, as well as the structure of production and consumption, within a nation.

Asymmetry in the energy-income relationship arises largely because not all expansions (recessions) are created equal. Some sectors of an economy may expand (contract) more rapidly than others during certain periods of time. Since different economic sectors are generally characterized by different types of capital, it is possible that some periods of macroeconomic expansion (contraction) will result in the expansion (contraction) of a sector in which energy intensity is very high relative to other sectors. Therefore, growth (reduction) in energy use could be larger in those time periods.

To illustrate, consider a two-sector economy (denoted as sectors A and B). Total energy consumption is given as the sum of energy use in each of the two sectors, $e = e_A + e_B$. Sector A has an energy intensity measured as

$$\frac{e_A}{y_A},$$

and sector B has energy intensity measured as

$$\frac{e_B}{y_B},$$

where y_j is the output of sector j and total output is given as $y = y_A + y_B$. Assume that sector A is more

energy intensive than sector B, or $\frac{e_A}{y_A} > \frac{e_B}{y_B}$. Note that total energy intensity can be written as

$$\begin{aligned} \frac{e}{y} &= \frac{e_A + e_B}{y} = \frac{e_A}{y_A} \cdot \frac{y_A}{y} + \frac{e_B}{y_B} \cdot \frac{y_B}{y} \\ &= \frac{e_A}{y_A} \cdot \theta_A + \frac{e_B}{y_B} \cdot \theta_B, \end{aligned}$$

where θ_j is the sector j share of total output.

It follows for a given rate of total economic growth, if sector A grows faster than sector B, the effect on total energy demand will be larger than if sector B grows faster than sector A. To see this, hold energy intensity within each sector constant. Taking the partial derivative of the preceding equation with respect to the output shares yields

$$\frac{\partial \left(\frac{e}{y} \right)}{\partial \theta_A} = \frac{e_A}{y_A} \quad \text{and} \quad \frac{\partial \left(\frac{e}{y} \right)}{\partial \theta_B} = \frac{e_B}{y_B}.$$

But we know, by assumption, that

$$\frac{e_A}{y_A} > \frac{e_B}{y_B},$$

so growth in sector A has a larger impact on total energy intensity than growth in sector B.

It should be noted that reality is not as simple as this example. Specifically, technological innovations do cause sector-specific energy intensities to change over time (although, as the example illustrates, they need not do so for total energy intensity to change over time). However, if we assume that the time period is reasonably short, our assumption of constant sector-specific energy intensity is not a gross approximation. Allowing for a dynamic sector-specific energy intensity would, however, indicate just how complicated the long-term energy-income relationship can become.

3.2 Energy Demand and Imperfect Price Reversibility

It is well documented that energy prices have a significant negative impact on both energy demand and macroeconomic performance. However, this relationship is not symmetric. In other words, much as is the case with the energy-income relationship, rising energy prices will reduce growth in macroeconomic output and energy demand, but falling prices will not have an equal and opposite effect. This imperfect price reversibility arises due to the characteristics of the energy-consuming capital stock. When energy prices rise, there is incentive for the holders of energy-using capital stocks to take steps to increase efficiency. This could come in the

form of either retrofitting existing capital or replacing it with newer capital. In either case, these steps require capital investment that will not be reversed when energy prices fall. Likewise, improvements in technologies (most of which are embodied in newer capital) induced by higher energy prices will not be reversed when prices fall.

Imperfect price reversibility follows from examination of Eq. (1). In the short run, when energy prices increase, the user cost of capital also increases (see Eq. (4)) since capital and technology are fixed. Thus, the energy consumer will re-optimize his or her portfolio by reducing energy use where possible. Since capital and technology are fixed in the short run, this can only be done by reducing capital utilization rates (u_t). Not only will the decline in capital utilization result in a decline in energy demand, it will also result in a reduction in economic activity.

In the long run, allowing capital and technology to vary will, in general, still result in a decrease in energy demand, but it can complicate the macro-economic response to energy price increases. As technologies change and capital stocks are turned over, older stocks are replaced with stocks in which newer, more energy-efficient technologies are embodied. The improved efficiency can generate asymmetry in the response of demand to price by preventing

energy use from rising appreciably when energy prices decline. In addition, better efficiency can reduce the impact on energy demand of future price increases because, by Eq. (4), increasing efficiency lowers the user cost of energy-using capital. Thus, a Δ price decrease will not have an equal and opposite effect as a Δ price increase. Similarly, a Δ price increase in the future will not necessarily have the same impact as a Δ price increase in the present. To emphasize the point, we can again consider the U.S. experience in the consumption of motor fuels to illustrate the dynamics at work in the long and short run. As indicated in Fig. 5, average fuel efficiency of passenger vehicles in the United States was virtually unchanged from 1960 to 1978. For the majority of this period, oil prices were relatively stable, and vehicle utilization (measured as miles per vehicle) grew steadily. However, the oil price shocks of the 1970s and early 1980s were followed by an immediate decrease in vehicle utilization. During the same period, average vehicle efficiency was relatively stable. From 1979 to 1992, however, average vehicle efficiency increased at close to 5% per year, prompted by the energy efficiency policies adopted following the oil price shocks of the 1970s. The increases in efficiency in the early 1980s offset increasing prices, as 1981 saw vehicle utilization, fuel efficiency, and the real price of gasoline all increase.

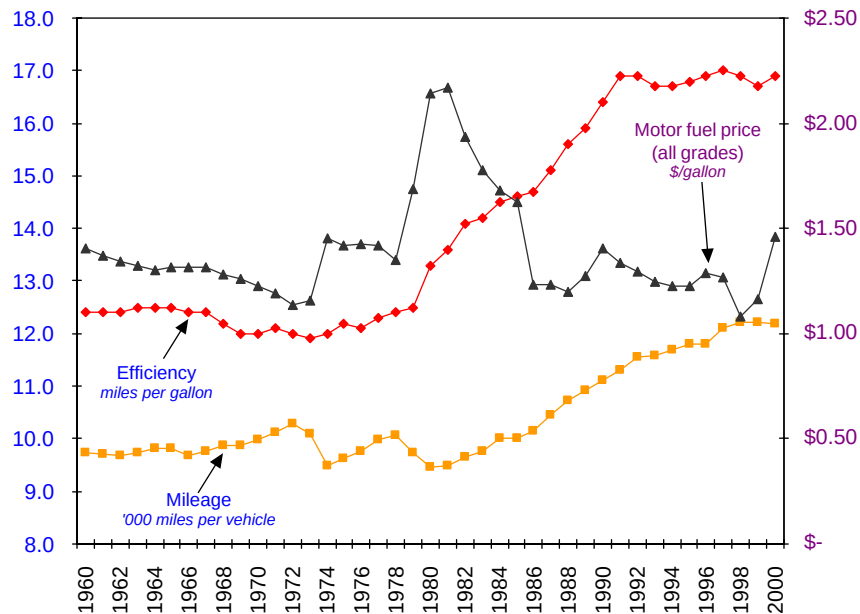


FIGURE 5 Energy prices for transportation increased dramatically from 1974 through 1982. In response, motor vehicle utilization declined initially, but a subsequent, albeit later, increase in efficiency of the motor vehicle fleet served to offset higher fuel prices, thereby reducing user costs.

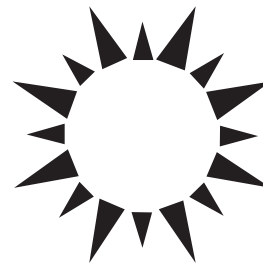
SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Climate Change: Impact on the Demand for Energy • Economic Geography of Energy • Economic Growth and Energy • Economics of Energy Supply • Energy Futures and Options • Market Failures in Energy Markets • Modeling Energy Supply and Demand: A Comparison of Approaches

Further Reading

- Balke, N. S., Brown, S. P. A., and Yücel, M. K. (2002). Oil price shocks and the U.S. economy: Where does the asymmetry originate? *The Energy J.* 23(3), 27–52.
- Bernardini, O., and Galli, R. (1993). Dematerialization: Long-term trends in the intensity of use of materials and energy. *Futures*, May, 431–448.
- Bohi, D., Toman, M., and Walls, M. (1996). “The Economics of Energy Security.” Kluwer Academic Publishers, Boston, MA.
- Diewert, W. E. (1974). Intertemporal consumer theory and the demand for durables. *Econometrica* 42, 497–516.
- Ferderer, J. P. (1996). Oil price volatility and the macroeconomy. *J. Macroeconomics*, Winter, 1–26.
- Galli, R. (1998). The relationship between energy intensity and income levels: Forecasting long term energy demand in Asian emerging countries. *Energy J.* 19(4), 85–105.
- Gately, D., and Huntington, H. (2002). The asymmetric effects of changes in price and income on energy and oil demand. *Energy J.* 23(1), 19–55.
- Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *J. Political Economy* 91(2), 228–248.
- Huntington, H. (1998). Crude oil prices and U.S. economic performance: Where does the asymmetry reside? *Energy J.* 19(4), 107–132.
- Johansson, O., and Schipper, L. (1997). Measuring the long-run fuel demand of cars: Separate estimations of vehicle stock, mean fuel intensity, and mean annual driving distance. *J. Transport Economics and Policy* 31(3), 277–292.
- Judson, R., Schmalensee, R., and Stoker, T. (1999). Economic development and the structure of the demand for commercial energy. *Energy J.* 20(2), 29–58.
- Medlock, K. B., and Soligo, R. (2001). Economic development and end-use energy demand. *Energy J.* 22(2), 77–105.
- Medlock, K. B., and Soligo, R. (2002). Car ownership and economic development with forecasts to 2015. *J. Transport Economics and Policy* 36(2), 163–188.
- Mork, K., Mysen, H. T., and Olsen, O. (1994). Macroeconomic responses to oil price increases and decreases in seven OECD countries. *Energy J.* 15, 15–38.
- Tishler, A. (1983). The demand for cars and gasoline: A simultaneous approach. *European Economic Rev.* 10, 271–287.

Economics of Energy Efficiency



ADAM B. JAFFE

Brandeis University and National Bureau of Economic Research
Waltham, Massachusetts, United States

RICHARD G. NEWELL

Resources for the Future
Washington, D.C., United States

ROBERT N. STAVINS

Harvard University
Cambridge, Massachusetts, United States

-
1. Overview
 2. The Process of Technological Change
 3. Understanding the Energy Efficiency Gap
 4. Market Failures
 5. Non-Market-Failure Explanations of the Energy Efficiency Gap
 6. Discounting and Implicit Discount Rates
 7. Summary

Glossary

externality An economically significant effect of an activity, the consequences of which are borne (at least in part) by parties other than the party who engages in the activity, and which are not accounted for through trade.

diffusion The gradual adoption of new process or product innovations by firms and individuals.

innovation The initial market introduction or commercialization of new process or product inventions.

invention The development and creation of a prototype new idea, process, or piece of equipment.

market barriers Disincentives to the diffusion and/or use of a good, such as high costs or prices, which may or may not represent market failures.

market failures The failure of private markets to provide certain goods at all or at the most desirable level, typically arising from a situation in which multiple parties benefit from a good without decreasing one another's benefits, and in which those who have paid for the good cannot prevent others from benefiting from it.

technological change The process of invention, innovation, and diffusion whereby greater and/or higher quality outputs can be produced using fewer inputs.

This article reviews economic concepts relevant to decision making about energy efficiency investments and related policy choices. We describe economic perspectives on the process of energy-saving technological change, the distinction between market failures and market barriers in energy-using product markets, and the important role that discounting plays in this area.

1. OVERVIEW

Energy efficiency is defined here to mean energy services provided per unit of energy input (for example, gallons of water heated to a specified temperature per British thermal units of natural gas input). Within this framework, energy efficiency is conceived primarily at the disaggregated, product level, rather than at a more aggregated sectoral level. As with virtually all economic problems, the economics of energy efficiency is at its heart a question of balancing of costs and benefits. For the individual energy user, this involves weighing the higher initial cost of purchasing energy-efficient products against the expected benefits of future cost savings when operating the products, among other considerations. Due to this difference in the timing of energy efficiency costs and benefits, issues related to discounting feature prominently in analyses of energy-efficient technology adoption. For suppliers of energy-using products, decisions regarding energy-efficient innovations likewise depend on the expected profits from such technology development. Profits from innovation depend in turn on the expected

demand for energy-efficient technologies and the degree to which firms can appropriate the value created by their innovations.

On both the demand and supply side of this market for energy-efficient technology, potential market imperfections can lead to underinvestment in energy efficiency. This raises the possibility that corrective government policies could provide economic gains if they are practicable and provide net benefits after inclusion of all public and private implementation costs. The degree to which such opportunities exist in practice is the subject of significant debate. Finally, environmental pollution associated with energy production—particularly carbon dioxide emissions from fossil fuels—represents an additional reason why private markets might underprovide energy efficiency if energy users do not face the cost of any resultant environmental harm.

2. THE PROCESS OF TECHNOLOGICAL CHANGE

Many readers may be unfamiliar with the way economists typically view the process of technological change, thus it is useful to first establish this common understanding. Furthermore, to understand the potential for public policy to affect energy efficiency, it is also necessary to understand the process through which technology evolves: invention, innovation, diffusion, and product use. Invention involves the development of a new idea, process, or piece of equipment. This activity typically takes place inside the laboratory. The second step is the process of technology innovation, in which new processes or products are brought to market. Another way of describing this stage is commercialization. The third step is diffusion, the gradual adoption of new processes or products by firms and individuals, which then also decide how intensively to use new products or processes.

Tying this all together, it would be possible, for example, to think of a fundamentally new kind of automobile engine being invented. This might be an alternative to the internal combustion engine, such as a system dependent on energy-efficient fuel cells. The innovation step would be the work carried out by automobile manufacturers or others to commercialize this new engine—that is, bring it to market, then offer it for sale. The diffusion process, then, would reflect the purchase by firms and individuals of automobiles with this new engine. Finally, the degree of use of these new automobiles will be of great

significance to demand for particular types of energy. The reason it is important to distinguish carefully among these different conceptual steps—invention, innovation, diffusion, and use—is that public policies can be designed to affect various stages and will have specific and differential effects.

2.1 Technology Adoption and Diffusion

Beginning at the end of the technological change process, research has consistently shown that diffusion of new, economically superior technologies is never instantaneous. The S-shaped diffusion path shown in Fig. 1 has typically been used to describe the progress of new technologies making their way into the marketplace. The figure portrays how a new technology is adopted at first gradually and then with increasing rapidity, until at some point its saturation in the economy is reached. The explanation for this typical path of diffusion that has most relevance for energy conservation investments is related to differences in the characteristics of adopters and potential adopters. This includes differences in the type and vintage of their existing equipment, other elements of the cost structure (such as access to and cost of labor, material, and energy), and their access to technical information. Such heterogeneity leads to differences in the expected returns to adoption and, as a result, only potential adopters for whom it is especially profitable will adopt at first. Over time, however, more and more will find it profitable as the cost of the technology decreases, quality improves, information about the technology becomes more widely available, and existing equipment stocks depreciate. The longevity of much energy-using equipment reinforces the importance of taking a longer term view toward energy efficiency improvements—on the order of decades (see Table I).

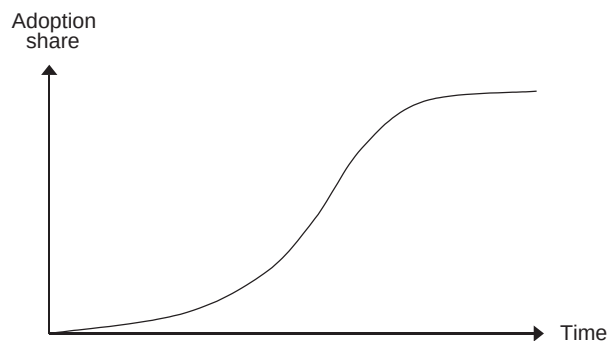


FIGURE 1 The gradual S-shaped path of technology diffusion.

Several studies have explored the effect of energy prices and technology adoption costs on energy efficiency investments. From a policy perspective, the effect of higher energy prices can be interpreted as suggesting what the likely effects of taxes on energy use (or carbon dioxide) would be and the effects of changes in adoption costs can be interpreted as indicating what the effects of technology adoption subsidies would be. As suggested by the economic cost–benefit paradigm, these studies have found that higher energy prices increase and adoption costs decrease the extent of adoption of energy-efficient technology (for example, building insulation, more efficient home appliances, or more efficient industrial motors). An additional interesting finding in this line of research is that the adoption of these technologies is more sensitive to the cost of the equipment than it is to the expected cost of energy. This implies that a policy of subsidizing the purchase of new efficient equipment may be more effective than a policy of taxing resource use, for policies that should in theory create the same magnitude of economic incentive.

There are at least three possible explanations for this divergence. One possibility is a behavioral bias that causes purchasers to focus more on up-front cost than they do on the lifetime operating costs of an investment. An alternative view is that purchasers

focus equally on both, but uncertainty about future energy prices or whether they will face these costs (because they could move, for example) makes them give less weight to energy prices than they do to capital cost, which is known. A final interpretation might be that consumers have reasonably accurate expectations about future energy prices, and their decisions reflect those expectations, but the proxies for these expectations that are used by researchers are flawed, causing their measured effect to be smaller than their true effect. For example, studies often use current realized energy prices as a proxy for expected future energy prices. Current prices fluctuate more than expected future prices, however, leading to a downward bias in the coefficient on the energy price proxy relative to the true relationship with expected prices.

Although empirical evidence indicates that subsidies may be more effective than comparable taxes in encouraging technology diffusion, it is important to recognize some disadvantages of such subsidy approaches. First, unlike energy prices, adoption subsidies do not provide incentives to reduce utilization. Second, technology subsidies and tax credits can require large public expenditures per unit of effect, because consumers who would have purchased the product even in the absence of the subsidy still receive it.

2.2 Technology Innovation and Invention

Now it is possible to move back in the process of technological change from diffusion to innovation. In the energy efficiency area, it is helpful to think of the innovation process as affecting improvements in the attributes or characteristics of products. In Fig. 2, this process is represented as the shifting inward over time of a curve representing the trade-offs between different product characteristics for the range of products available on the market. On one axis is the cost of the product, and on the other axis is the energy flow (use; that is, the energy intensity) associated with a product. The downward slope of the curves indicates the trade-off between equipment cost of energy efficiency. Innovation means an inward shift of the curve—greater energy efficiency at the same cost, or lower cost for a given energy efficiency.

As with technology diffusion, studies have shown that increases in the price of energy have induced technological innovations in energy efficiency of commercialized products, such as household appliances (e.g., air conditioners, water heaters), automobiles, tractors, and jet aircraft. Moving back even further in the process of technological change to

TABLE 1

Technology Diffusion and the Rate of Capital Stock Turnover^a

Type of asset	Typical service life (years)
Household appliances	8–12
Automobiles	10–20
Industrial equipment/machinery	10–70
Aircraft	30–40
Electricity generators	50–70
Commercial/industrial buildings	40–80
Residential buildings	60–100

^aTechnology diffusion is closely related to the concept of “capital stock turnover,” which describes the rate at which old equipment is replaced and augmented by new. New equipment can be purchased either to replace worn out and obsolete units or as a first-time purchase. A primary driver of replacement purchases for durable energy-using products is the useful lifetime of the product. The rate of economic growth is also important, especially for first-time durable-goods purchases; the rate of home construction is particularly relevant for residential equipment. The typical lifetimes for a range of energy-using assets illustrate that the appropriate time frame for thinking about the diffusion of many energy-intensive goods is on the order of decades.

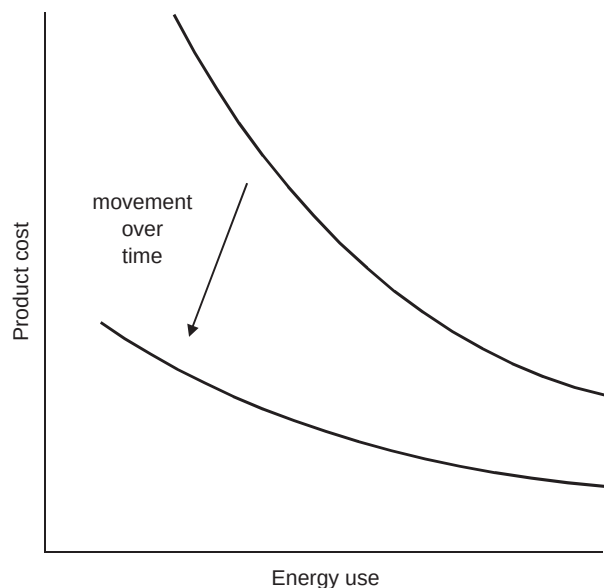


FIGURE 2 Innovation in product characteristics.

examine invention, other studies have analyzed U.S. patent data, finding that the rate of energy-conserving device patent applications (e.g., for waste heat, heat pumps, solar energy, and fuel cells) is significantly and positively associated with the price of energy.

3. UNDERSTANDING THE ENERGY EFFICIENCY GAP

Renewed attention is now being given by policy-makers to energy efficiency, primarily due to concerns about global climate change associated with carbon dioxide emissions from fossil fuels. Much attention has been placed on the role that technological improvements in energy efficiency can play in reducing carbon emissions and in lowering the cost of those reductions. In such policy contexts, it is frequently asserted that there exists an “energy efficiency gap” between current energy use, on the one hand, and optimal energy use, on the other hand. From this perspective, the energy efficiency gap discussed earlier can now be thought of as a debate mainly about the gradual diffusion of energy-saving technologies that seem to be cost-effective.

Clearly, an estimate of the magnitude of this gap will be determined mainly by how optimal behavior is defined. Several major conceptual issues surround different approaches to defining an optimal energy-use scenario, with implications for guidance regarding public policy for energy and energy technologies. The standard economic approach defines “good” public policy to be that which maximizes the appropriately

weighted sum of the values of goods and services enjoyed by society throughout time (including intangibles such as the environment). From this perspective, energy efficiency is not a goal in itself, but only a means to the end of overall economically efficient (and equitable) resource allocation. To the extent that energy generation and/or use creates environmental problems—one of the primary reasons why energy use is a public policy concern in the first place—such effects can in principle be incorporated in analyses by placing appropriate values on the environmental and nonenvironmental benefits and costs associated with energy generation and use.

The crux of the debate surrounding the energy efficiency gap lies in differing interpretations of what has been called the paradox of gradual diffusion of apparently cost-effective energy-efficient technologies. Why are compact fluorescent light bulbs, improved thermal insulation materials, and energy-efficient appliances not more widely used? Many studies have demonstrated that there exist such technologies (and processes)—ones that simple net-present-value calculations show to be cost-effective at current prices and market interest rates, but which enjoy only limited market success. The phrase market barriers has been used to refer to any factors that may account for this apparent anomaly. Others have used the phrase market barriers even more broadly to include, for example, low energy prices that are a disincentive to the adoption of more energy-efficient technologies. Differing views about the nature of these barriers lead to fundamentally different views about optimal energy use and the role of government policy.

As already noted, the diffusion of economically superior technologies is typically gradual. Awareness of this empirical reality should make the existence of the energy efficiency gap much less perplexing, but it does not answer the question of whether the optimal rate of diffusion is greater than the observed rate. To some degree, responses to the observation that new technologies always diffuse slowly depend on subjective perspectives. Some are inclined to note that if technology diffusion is typically not optimal, then there is no basis for presuming that the market is working efficiently, and hence no reason to eschew government intervention designed to improve it. On the other hand, others are more inclined to maintain that the government cannot possibly worry about trying to optimize the diffusion rate for all technologies, and that the government should hence stay out altogether, even if the no-intervention result is not optimal.

As will be described further here, at a theoretical level there are several reasons why technology diffusion will not, in general, occur at an economically efficient rate. But, if the purpose of assessing the efficiency gap is to identify desirable government policy interventions, then it is necessary to know whether the market barriers that cause slow diffusion can be mitigated by government intervention such that overall resource allocation is improved. In this context, the appropriate definition of the optimal level of energy efficiency is that which is consistent with efficient overall resource use, including efficient use of government resources.

Those market barriers that might justify a public policy intervention to overcome them are referred to in economics parlance as market failures. Normative economics teaches that in the presence of market failures, unfettered markets may not operate to produce outcomes that are socially optimal. When analysts speak of no-cost climate policies based on energy efficiency enhancement, they are often implicitly or explicitly assuming the presence of market failures in energy efficiency. By implication, there are some market barriers—such as “low” energy prices, “high” technology costs, and “high” discount rates—that are not necessarily market failures. In such cases, the existence of an energy efficiency gap does not, in and of itself, call for policy responses. On the other hand, there are also some market failures that do not relate to problems in individual decision-making per se, but that might still provide justification for policy interventions. In other words, even if the energy efficiency paradox is resolved in the sense that people’s decisions are found to be consistent with the costs they face, there could be other reasons—particularly environmental pollution from energy use—why the resulting behavior would deviate from the social optimum. These classes of market failures and non-market-failure explanations of the energy efficiency gap are considered next.

4. MARKET FAILURES

In this section, the focus is on several sources of potential market failure that may affect energy-conserving technology adoption rates. Three of these relate to the availability of information. There are also market failure issues that do not help explain nonadoption at current technology costs and energy prices, but which are still relevant to policy debates about the energy efficiency gap. These include environmental externalities, broader innovation and

adoption externalities, and issues related to energy supply pricing and national security.

4.1 Inadequate Information

First, information has important public good attributes: once created it can be used by many people at little or no additional cost. It may be difficult or impossible for an individual or firm that invests in information creation to prevent others who do not pay for the information from using it. It is well known that such public goods will tend to be underprovided by ordinary market activity. Second, if the act of adopting a new technology is inherently a source of useful information for others, then the act of adoption creates a positive externality by providing information to others for which the adopter is unlikely to be compensated. It takes time for the many potential users to learn of the new technology, try it, adapt it to their circumstances, and become convinced of its superiority. An important mechanism in this learning process is the observation of the adoption of the new technology by others. If a neighbor or competitor tries something new, and others see that it works, it becomes much safer and easier for the observers to try it. Hence the adopter of a new technology creates a positive externality for others, in the form of the generation of information about the existence, characteristics, and successfulness of the new technology. This phenomenon is often called learning by using. This (positive) externality is another form of market failure.

Incomplete information can also foster principal-agent problems, as when a builder or landlord chooses the level of investment in energy efficiency in a building, but the energy bills are paid by a later purchaser or a tenant. If the purchaser has incomplete information about the magnitude of the resulting energy savings, the builder or landlord may not be able to recover the cost of such investments, and hence might not undertake them. This is another potential form of market failure. A home-builder, acting as agent for the ultimate home-buyer, may have incentives to take actions that are different from what the principal would prefer. The extent to which this market failure is significant in practice is not well established.

4.2 Environmental Externalities

Economic analysis of environmental policy is based on the idea that the potentially harmful consequences of economic activities on the environment constitute

an externality. An externality is an economically significant effect of an activity, the consequences of which are borne (at least in part) by a party or parties other than the party who controls the externality-producing activity. An electricity generator that pollutes the air imposes a cost on society. The firm that owns the generator has an economic incentive to use only as much fuel as it can productively employ, because this input is costly to the firm. In jargon, the cost to society of having some of its fuel used up is internalized by the firm, because it has to pay for that input. But the firm does not (in the absence of appropriate environmental policy intervention) have an economic incentive to minimize the external costs of pollution associated with that fuel.

Although the details and refinements are important, all environmental policies, at their core, are designed to deal with this externality problem, either by internalizing environmental costs so that polluters will make efficient decisions regarding their consumption of environmental inputs, or else by imposing from the outside a level of environmental pollution that policymakers believe to be more efficient than that otherwise chosen by firms. If environmental externalities are not fully addressed by environmental policy—for political or other reasons—the resulting level of energy efficiency will likely be too low. But this may not be true if the government is already intervening to change the energy efficiency of products available in the market place (or is otherwise intervening to control the pollution associated with energy use). For example, Corporate Average Fuel Economy (CAFE) standards result in the production and sale of cars that are more efficient than consumers would otherwise demand. Depending on the actual magnitude of environmental externalities, it is possible that the efficiency of automobiles is greater than would be socially optimal.

4.3 Broader Innovation and Adoption Externalities

In addition to the externality associated with pollution, innovation and diffusion are both characterized by externalities, as well as other market failures. In the case of pollution, the problem is that a polluter imposes costs on others, and hence has an inadequate incentive, from a social perspective, to reduce those costs. With respect to technology, the problem is the reverse. A firm that develops or implements a new technology typically creates benefits for others, and hence has an inadequate

incentive to increase those benefits by investing in technology. Pollution is a negative externality, and so the invisible hand allows too much of it. Technology creates positive externalities, and so the invisible hand produces too little of it.

With respect to innovation, the positive externality derives from the public-good nature of new knowledge. Whereas patents and other institutions try to protect firms' investments in innovation, such protection is inherently imperfect. A successful innovator will capture some rewards, but those rewards will always be only a fraction—and sometimes a very small fraction—of the overall benefits to society of the innovation. Hence innovation creates positive externalities in the form of knowledge spillovers for other firms, and spillovers of value for the users of the new technology.

In addition, production costs tend to fall as manufacturers gain production experience in a process commonly called learning by doing. If this learning spills over to benefit other manufacturers, it can represent an additional externality. In any event, the existence of these innovation and adoption externalities suggests a rationale for public support of research and development in general, not for energy-efficient technology in particular. An argument particularly supporting energy-efficient technologies in this regard would require stating that spillovers related to these technologies were somehow greater than for technologies more generally in the economy.

4.4 Market Failures in Energy Supply

4.4.1 Average-Cost Pricing

Actual energy prices, particularly for electricity, may differ from marginal social cost because of subsidies and because of pricing based on average rather than marginal cost. During the 1980s, it was widely perceived that the incremental costs of increasing electricity supplies were significantly greater than the average costs of existing electrical capacity. Because the prices consumers pay are typically based on historical average costs, it was frequently suggested that consumers faced inadequate incentives to conserve electricity. Each kilowatt of capacity that did not need to be built saved society much more than it saved the persons whose conservation decisions reduced the need for such new capacity. If true, this would be one reason why public policy should seek to promote greater energy efficiency than private individuals choose on their own.

Although this argument remains conceptually valid in part, several things have changed since the 1980s that weaken the argument considerably. First, the deregulation of electricity production for about 40% of the U.S. population has meant a movement toward competitive electricity prices; this trend will continue. Second, there is widespread excess capacity in the electricity industry in most of the United States, although there is substantial regional variation. Thus, it is simply no longer true that the incremental cost of capacity is well above the price paid by most consumers. Indeed, given the availability of wholesale bulk power at prices closer to variable cost than to total cost, it could be argued that the social value of reducing electrical utility loads is actually less than the prices paid by most consumers. Furthermore, this situation is likely to prevail for many years, depending on the growth of electricity demand. Thus, regulatory distortions in the utility industry no longer provide an argument for policy interventions to foster energy conservation. One caveat, however, is that at peak demand periods, the marginal cost of supply often exceeds the price. This provides an argument not for energy efficiency in general, but rather for load-shifting to off-peak periods, energy conservation during peak periods, or time-of-day pricing.

With respect to technology incorporated in new buildings, it can be argued that these buildings are likely to last long into the future, at which time new electrical capacity may once again be needed, so that consumers choosing energy-inefficient technology are imposing a social cost above their own energy costs, albeit one that is very uncertain and remote in time. With respect to retrofit investments, however, even this argument is unavailable, because these investments can simply be postponed until such time as they are needed to forestall the need to expand capacity.

4.4.2 Security Externalities

It has also been suggested that there are externalities associated with the economic and military security costs resulting from domestic U.S. dependence on oil imported from politically unstable regions. For this argument to be valid, it would have to be the case that, at the margin, these national security costs are reduced if oil consumption is marginally reduced. This seems unlikely to be the case.

5. NON-MARKET-FAILURE EXPLANATIONS OF THE ENERGY EFFICIENCY GAP

Non-market-failure explanations of the energy efficiency gap consist essentially of explaining why observed behavior is indeed optimal from the point of view of individual energy users. To be useful, such explanations must advance beyond the tautological assertion that if the observed rate of diffusion is less than the calculated optimal rate, there must be some unobserved adoption costs that would modify calculations of what is optimal.

A slightly less tautological approach would be to maintain the assumption—based on general experience—that private agents act in their own interests, unless it can be shown that specific market failures exist. Under this approach, it can quickly be concluded that no paradox exists, unless and until it can be demonstrated that the theoretical market failures already discussed are important in explaining observed behavior. This amounts to a policy prescription that markets should be “considered innocent until proved guilty.” There is no scientific basis for such a prescription. On the other hand, it is important to keep in mind that although market failure may be considered a necessary condition for government policy intervention, it is not a sufficient condition. For government policy to be desirable in economic terms, it must be the case that a market failure exists and that there exists a government policy that, in overcoming the market failure, generates benefits in excess of the costs of implementation. Analyses of past government and utility energy conservation programs demonstrate that although some such programs certainly have had positive net benefits, the benefits of others have been lower than predicted.

5.1 Irreversibility and the Option to Wait

One such non-market-failure explanation is that uncertainty about the future benefits of energy-efficient technologies, combined with the irreversible nature of the efficiency investment, makes the effective discount rate for analyzing the net present value of energy savings significantly greater than is typically used in the calculations that suggest the existence of a paradox. Simple net-present-value analysis typically does not account for changes over time in the savings that purchasers might enjoy from an extra investment in energy efficiency, which

depends on trends and uncertainties in the prices of energy and conservation technologies. When making irreversible investments that can be delayed, the presence of this uncertainty can lead to an investment hurdle rate that is larger than the discount rate used by an analyst who ignores this uncertainty.

The magnitude of this option-to-wait effect or option value depends on project-specific factors, such as the degree of energy price volatility, the degree of uncertainty in the cost of the investment, and how fast the prices of energy and conservation technologies are changing over time. The effect is magnified when energy and technology price uncertainty is increased and when energy prices are rising and technology costs are falling more quickly. Under conditions characterizing most energy conservation investments, this effect could raise the hurdle rate by up to about 10 percentage points. On the other hand, if there is no opportunity to wait, this effect can be ignored.

Note that uncertainty, in contrast to imperfect information, is not a source of market failure in and of itself. It is reasonable and appropriate for individuals to take uncertainty into account in making investment decisions, and to apply relatively high hurdle rates to irreversible investments when returns are uncertain. To the extent that consumers' true effective discount rates are high for this reason, this would not represent a market failure or a reason for policy intervention from an economic perspective.

5.2 Unaccounted for Costs and Overestimation of Energy Savings

There are typically costs of adoption that are not included in simple cost-effectiveness calculations. It is by no means costless to learn how a technological improvement fits into a home or firm or to learn about reliable suppliers. Even after basic information about a technology has been disseminated, the purchase price of a new product is only a lower bound on its adoption cost. Another type of hidden cost is the possibility that qualitative attributes of new technologies may make them less desirable than existing, less efficient technologies. An obvious example is the difference in hue between florescent and incandescent lighting and the delay time in achieving illumination with many florescent lights.

Some have argued that not only costly information acquisition but also biased estimates of likely energy savings play a role. There is evidence that analysts have substantially overestimated the energy savings that higher efficiency levels will bring, partly

because projections often are based on highly controlled studies that do not necessarily apply to actual realized savings in a particular situation. For example, studies have found that actual savings from utility-sponsored programs typically achieve 50 to 80% of predicted savings. Other studies have drawn similar conclusions based on analysis of residential energy consumption data, finding that the actual internal rate of return to energy conservation investments in insulation was substantially below typical engineering estimates for the returns from such investments. On the other hand, the bias could be offset in the opposite direction, because some studies indicate that consumers systematically overestimate energy savings associated with some types of new technologies.

5.3 Heterogeneity in Energy Users

Another possible non-market-failure explanation for the energy efficiency gap is associated with the fact that even if a given technology is cost-effective on average, it will mostly likely not be for some individuals or firms. If the relevant population is heterogeneous—with respect to variables such as the purchaser's discount rate, the investment lifetime, the price of energy, the purchase price, and other costs—even a technology that looks very good for the average user will be unattractive for a portion of the population. Heterogeneity in these and other factors leads to differences in the expected value that individual purchasers will attach to more energy-efficient products. As a result, only purchasers for whom it is especially valuable may purchase a product. Depending on how the efficiency gap is measured, such underlying heterogeneity can provide yet another non-market-failure explanation. For example, it may not make sense for someone who will only rarely use an air conditioner to spend significantly more purchasing an energy-efficient model—they simply may not have adequate opportunity to recoup their investment through energy savings. Analysis based on single estimates for the important factors previously discussed—unless they are all very conservative—will inevitably lead to an optimal level of energy efficiency that is too high for some portion of purchasers. The size of this group, and the magnitude of the resulting inefficiency should they be constrained to choose products that are not right for them, will of course depend on the extent of heterogeneity in the population and the assumptions made by the analyst.

6. DISCOUNTING AND IMPLICIT DISCOUNT RATES

Much discussion surrounding the energy efficiency gap is couched in terms of arguments about the appropriate discount rate to use in evaluating energy savings. It is useful to distinguish among a number of separate and distinct questions regarding the appropriate discount rate. First, an attempt can be made to estimate the implicit discount rate that consumers appear to be using when they make energy efficiency decisions. Second, there can be speculation about the correct discount rate for consumers to use in making such decisions. Third, the implicit discount rate (explained below) or the correct rate can be compared with other private discount rates in the economy. Fourth and finally, it can be asked whether any of these rates are equal to the social discount rate that ought to be applied in evaluating future energy savings when making public policy decisions.

The observation that consumers have high implicit discount rates when they make energy efficiency decisions is actually neither more nor less than a restatement of the existence of the energy efficiency gap. To estimate implicit discount rates, studies examine decisions actually made and calculate the discount rate that makes those decisions privately optimal, given estimates of the costs and future energy savings of the investments and given the assumption that there are no important market failures impeding the adoption of efficient technologies. There is substantial empirical evidence of implicit discount rates for energy-conservation investment decisions in the range of 20 to over 100% for products such as air conditioners, space heating, and refrigerators. The energy efficiency gap can thus be restated as the observation that these implicit discount rates appear to be much higher than other interest rates in the economy. To observe that implicit discount rates are high, however, says nothing about the reason people make the decisions they make. One possibility is that people are applying normal discount rates in the context of significant market failures; another possibility is that people actually utilize high discount rates in evaluating future energy savings. Note that “actually” does not mean “literally.” It is not necessary that people know how to perform net-present-value calculations; rather, they make decisions balancing costs today and costs tomorrow, with those trade-offs incorporating their true discount rate. The truth is likely some combination of the two. Thus, high implicit discount rates, on

their own, are neither a market failure nor an explanation of observed behavior.

Further, if only the relevant investment decisions are observed, it is fundamentally impossible to determine whether the observed behavior results from market failures, from truly high discount rates, or perhaps from misestimation of the true costs and savings of adoption. To make that distinction would require either observing something that distinguishes the market-failure explanations from the non-market-failure ones, such as whether people with better information are more likely to purchase more efficient models, or else calculating from some basis other than the investment decisions what an appropriate discount rate for these investments would be. Thus, to investigate the energy efficiency gap by changing the discount rates people use in making investment decisions amounts to assuming the answer. If the outcome with lower discount rates is considered the optimal result, then it is implicitly assumed that all market barriers are indeed market failures. Conversely, postulating that the optimal result is the one in which consumers are assumed to discount at observed high implicit rates, then it is implicitly assumed that there are no market failures.

To make this a bit more concrete, suppose that through calculations it is found that consumers’ true discount rates for some set of energy efficiency investments are approximately 20%—higher than mortgage interest rates, but in the range of rates on personal credit cards. What would this suggest for public policy? If mandatory efficiency standards were being considered, for example, then requiring efficiency any greater than what someone with a 20% discount rate would choose would have the effect of making that consumer worse off. Thus, to the extent that high implicit discount rates correspond to truly high discount rates, rather than to market failures, there is nothing particularly wrong with those high rates, and they do not correspond to any efficiency gap that ought to be addressed by public policy. Instead, the question of whether relevant market failures exist is again the issue.

Finally, there is the issue of the social discount rate. Several decades of academic and policy debates have failed to resolve fully the question of what the social rate of discount should be, even as a conceptual matter. However, there is a strong argument that policies to increase energy efficiency can be justified only on the basis of the appropriate true private rate of discount, not some (lower) social rate. If the only reason why investment in energy-efficient technology is suboptimal is the divergence

between the private and social discount rates, then it is certainly true that a theoretical argument could be made to support public policies to increase that investment. The problem is that this same argument would also apply with equal force to all other forms of investment—plant and equipment, research, education, and so on. It may be that the government should be doing more to encourage all of these forms of investment, but that is surely an issue beyond the scope of energy policy.

Further, if the social rate of discount is really very low, then the set of available investment opportunities that should be undertaken probably exceeds the nation's annual gross national product. Obviously, the government can not undertake all of these projects. If the notion that public policy should have as its objective undertaking all investments for which rates of return exceed the social discount rate is taken seriously, then the reasonable prescription is to increase investment, beginning with the investments with the highest rates of return. But increasing investment reduces consumption, and long before investment opportunities were exhausted, consumption would be so significantly reduced that the social discount rate would rise. Increasing investment up to the point where the private and social rates of return coincide can be imagined, but it is impossible to know how much greater investment would be at that point or what the discount rate would be. Given this, a policy prescription to increase investment in energy efficiency should be based on a conclusion that the (social) rate of return to this form of investment is higher than the rate of return available on other forms of investment in the economy. Otherwise, this form of investment could well be increasing at the expense of other forms that are even more beneficial. Of course, saying that the social rate of return on energy efficiency investments exceeds the social rate of return on other investments means precisely that current investment in energy efficiency is inadequate when evaluated at the appropriate private rate of discount.

7. SUMMARY

It is possible to synthesize much of the above discussion by examining graphically the concepts behind the major, alternative notions of the energy efficiency gap. To understand the basic elements of the debate, it is helpful to distinguish first between energy efficiency and economic efficiency, as in Fig. 3. The vertical axis measures increased energy

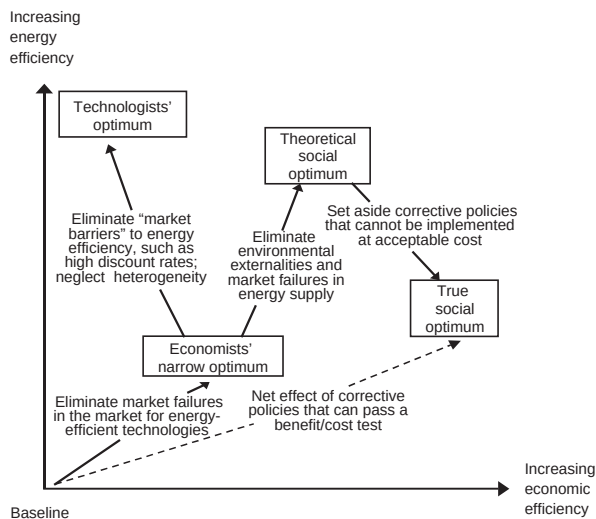


FIGURE 3 Alternative notions of the energy efficiency gap.

efficiency (decreased energy use per unit of economic activity). The horizontal axis measures increased economic efficiency (decreased overall economic cost per unit of economic activity, taking into account energy and other opportunity costs of economic goods and services). Different points in the diagram represent the possible energy-using technologies available to the economy as indicated by their energy and economic efficiency. As a concrete illustration of this distinction, consider two air conditioners that are identical except that one has higher energy efficiency and, as a result, is more costly to manufacture because high-efficiency units require more cooling coils, a larger evaporator, and a larger condenser, as well as a research and development effort. Whether it makes sense for an individual consumer to invest in more energy efficiency depends on balancing the value of energy that will be saved against the increased purchase price, which depends on the value of the additional materials and labor that were spent to manufacture the high-efficiency unit. The value to society of saving energy should also include the value of reducing any associated environmental externalities; but again this must be weighed against the costs. Starting from the baseline situation, as represented by the lower left corner of Fig. 3, the adoption of more energy-efficient technology is represented as an upward movement. But not all such movements will also enhance economic efficiency. In some cases, it is possible simultaneously to increase energy efficiency and economic efficiency. This will be the case if there are market failures that impede the most efficient allocation of society's

energy, capital, and knowledge resources in ways that also reduce energy efficiency. These are examples of what economists and others refer to as win-win or no regrets measures.

At the risk of excessive simplification, technologists can be characterized as believing that there are plentiful opportunities for low-cost or even negative-cost improvements in energy efficiency, and that realizing these opportunities will require active intervention in markets for energy-using equipment to help overcome barriers to the use of more efficient technologies. In terms of Fig. 3, the economist's notion of a narrow optimum is where market failures in the market for energy-efficient technologies have been corrected, the result being greater economic efficiency and energy efficiency. This optimum is narrow in the sense that it focuses solely on energy technology markets and does not consider possible failures in energy supply markets (such as underpriced energy due to subsidies or regulated markets) or, more important, environmental externalities associated with energy use (such as global climate change). Market failures in the choice of energy-efficient technologies could arise from a variety of sources, as discussed previously.

Eliminating broader market failures takes us to what is called the theoretical social optimum in Fig. 3. This represents both increased economic and increased energy efficiencies, compared with the economists' narrow optimum. This ignores the reality that the baseline efficiency already reflects policy measures motivated by environmental concerns. As emphasized previously, it is possible that these measures already achieve as much or more additional efficiency as would be justified by the environmental externalities of energy use. But not all market failures can be eliminated at acceptable costs. In cases in which implementation costs outweigh the gains from corrective government intervention, it will be more efficient not to attempt to overcome particular market failures. This takes us from a theoretical social optimum to what is referred to as the true social optimum in Fig. 3. Market failures have been eliminated, but only those for which elimination can pass a reasonable benefit-cost test. The result is the highest possible level of economic efficiency, but a level of energy efficiency that is intermediate compared with what would be technologically possible.

In contrast to the economist's perspective, if non-market-failure market barriers (such as high discount rates caused by uncertainty about payback) were also removed, the technologists' optimum

would be achieved. This alternative notion of an optimum is found by minimizing the total purchase and operating costs of an investment, with energy operating costs being discounted at a rate the analyst (not necessarily the purchaser) feels is appropriate. Clearly, if no distinction is made between barriers to adoption that are market failures and those that are not, and no choice is made to eliminate all barriers, a higher estimate of energy efficiency will be achieved, but not economic efficiency. An important implication of this perspective is that comparisons of an engineering ideal for a particular energy use with the average practice for existing technology are inherently misleading, because the engineering ideal does not incorporate all the real-world factors influencing energy technology decision-making.

How these alternative definitions of optimal energy efficiency and the energy efficiency gap relate to specific empirical estimates depends, of course, on the assumptions that underlie those estimates. Many published studies of the potential for energy efficiency correspond to what is labeled technologists' optimum. That is, they assume that the resolution of the energy paradox must be that the simplest calculations are correct and that a host of market failures explain observed behavior. Certainly, estimates of efficiency potential derived from forcing low discount rates into analyses of consumers' and firms' decisions do not correspond to any policy-relevant notion of optimal. Unfortunately, it is easier to explain what is wrong with existing approaches than it is to specify what is the right approach. It is clear, however, that in order to understand what is wrong with some of the existing approaches to estimating the energy efficiency gap and to begin the process of searching for the right measure, the first requirement is to disentangle market-failure and non-market-failure explanations for observed decisions regarding energy-efficiency investments.

SEE ALSO THE FOLLOWING ARTICLES

Conservation Measures for Energy, History of • Consumption, Energy, and the Environment • Discount Rates and Energy Efficiency Gap • Economics of Energy Demand • Economics of Energy Supply • Energy Efficiency and Climate Change • Energy Efficiency, Taxonomic Overview • Industrial Energy Efficiency • Obstacles to Energy Efficiency

Further Reading

- Anderson, S. T., and Newell, R. G. (2003). Information programs for technology adoption: The case of energy-efficiency audits. *Resource Energy Econ.* (in press).
- DeCanio, S. J., and Watkins, W. E. (1998). Investment in energy efficiency: Do the characteristics of firms matter? *Rev. Econ. Statist.* 80(1), 95–107.
- Hassett, K. A., and Metcalf, G. E. (1993). Energy conservation investment: Do consumers discount the future correctly? *Energy Policy* 21(6), 710–716.
- Hassett, K. A., and Metcalf, G. E. (1995). Energy tax credits and residential conservation investment: Evidence from panel data. *J. Public Econ.* 57, 201–217.
- Hausman, J. A. (1979). Individual discount rates and the purchase and utilization of energy-using durables. *Bell J. Econ.* 10, 33–54.
- Hirst, E. (1986). Actual energy savings after retrofit: Electrically heated homes in the Pacific Northwest. *Energy* 11, 299–308.
- Jaffe, A. B., and Stavins, R. N. (1994). The energy paradox and the diffusion of conservation technology. *Resource Energy Econ.* 16(2), 91–122.
- Jaffe, A. B., and Stavins, R. N. (1994). The energy-efficiency gap: What does it mean? *Energy Policy* 22(1), 804–810.
- Jaffe, A. B., Newell, R. G., and Stavins, R. N. (2003). Technological change and the environment. In “Handbook of Environmental Economics,” Vol. 1 (K.-G. Mäler and J. R. Vincent, Eds.), pp. 461–516. Elsevier, Amsterdam.
- Joskow, P. L., and Marron, D. B. (1992). What does a negawatt really cost? Evidence from utility conservation programs. *Energy J.* 13(4), 41–74.
- Levine, M. D., Koomey, J. G., McMahon, J. E., and Sanstad, A. H. (1995). Energy efficiency policy and market failures. In “Annual Review of Energy and the Environment” (J. M. Hollander, Ed.), pp. 535–555. Annual Reviews Inc., Palo Alto.
- Lind, R. C. (ed.). (1982). “Discounting for Time and Risk in Energy Policy.” Resources for the Future, Washington, D.C.
- Metcalf, G. E., and Hassett, K. A. (1997). Measuring the energy savings from home improvement investments: Evidence from monthly billing data. *Rev. Econ. Statist.* 81(3), 516–528.
- Newell, R. G., Jaffe, A. B., and Stavins, R. N. (1999). The induced innovation hypothesis and energy-saving technological change. *Q. J. Econ.*, August, 941–975.
- Sanstad, A. H., Blumstein, C., and Stoft, S. E. (1995). How high are option values in energy-efficiency investments? *Energy Policy* 23(9), 739–743.
- Sebold, F. D., and Fox, E. W. (1985). Realized savings from residential conservation activity. *Energy J.* 6(2), 73–88.
- Sutherland, R. J. (1991). Market barriers to energy-efficiency investments. *Energy J.* 12(3), 15–34.

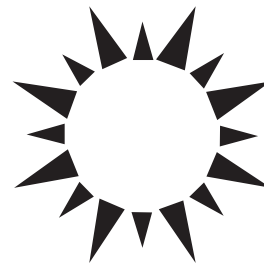
Economics of Energy Supply

JEFFREY A. KRAUTKRAEMER

Washington State University
Pullman, Washington, United States

MICHAEL A. TOMAN

Johns Hopkins University
Washington, D.C., United States



-
1. Introduction
 2. Basic Hypotheses in the Economics of Depletable Energy Supply
 3. The Simple “Cake-Eating” Model of Resource Extraction
 4. A More General Theory of Economically Efficient Depletable Resource Supply
 5. Effects of uncertainty
 6. Market Failures and Distortions
 7. Empirical Analysis of Depletable Energy Supply
 8. Concluding Remarks

Glossary

depletable resources Those natural resources for which total physical availability is fixed over any meaningful time horizon for economic decision-making; oil and coal are classic examples. The quality and other attributes of different deposits may vary.

economic scarcity The result of not having enough resources to meet all of one’s needs. Consequently, choices must be made and something of value must be foregone or given up in order to gain something else of value. In the case of depletable resources, if there is not enough of the resource to meet all needs in each time period, then there is a scarcity value to the resource stock.

Hotelling rule Named for Harold Hotelling, the intertemporal efficiency condition in a simple depletable resource model that the *in situ* value of the resource stock increases at the rate of interest. The rule does not hold in such a simple form in more complex models, such as when there are stock effects.

Hotelling valuation principle The valuation of a resource deposit on the basis of current price, marginal extraction cost, and the amount of the resource in the deposit.

***in situ* value** The incremental increase in present value that would occur with a one-unit increase in the initial resource stock. It also is referred to in the literature as the “user cost” or “scarcity rent” associated with depletion of the resource stock.

marginal cost The added economic cost of increasing an activity such as resource extraction or reserve discovery.

stock effect Usually, the increase in marginal extraction cost that occurs when a resource stock is depleted, either because it is physically more difficult to extract the resource (e.g., well pressure declines) or because lower quality or higher cost deposits are exploited.

The basic logic underlying all neoclassical economic theories of “optimal” energy supply is maximization of the present value of some stream of economic returns. The economic theory of optimal resource depletion has evolved since Hotelling’s classic 1931 article. The power of the theory to support improved empirical understanding of actual behavior is considered herein. The empirical literature indicates that this theory has so far provided only limited empirical understanding.

1. INTRODUCTION

The economic principles governing the efficient production of depletable energy resources (oil, natural gas, and coal) have been a subject of interest and controversy for over 70 years. The pathbreaking article on the subject by Hotelling in 1931 was motivated by then-current concerns about scarcity of energy and other natural resources that have a remarkably familiar ring to contemporary ears. More recently, the energy crisis of the 1970s motivated new concerns about the availability of sufficient energy resources

and the capacity of economic markets to allocate those resources effectively over space and time.

In this article, the basic logic underlying all neo-classical economic theories of optimal energy resource provision, maximization of the present value of some stream of economic returns, is first presented and discussed. Then, how the economic theory of optimal resource depletion has evolved since Hotelling's article is assessed. Considered in turn are theories of economically efficient energy resource extraction and new resource development. After presenting the basic theoretical framework, the discussion centers on extensions incorporating investment in fixed production capacity, uncertainty, and various market distortions, as well as extensions incorporating nondepletable "backstop" energy resources.

The utility of an economic theory is to be found not merely in its conceptual elegance, but also in its power to support improved empirical understanding of actual behavior. The empirical literature indicates that this work has so far provided only limited empirical understanding of observed energy supply behavior. Some key examples in which the theory and application of depletable energy supply have been developed in the literature. Some relatively recent surveys provide more detailed references (those by Cairns, Walls, Epple and Londregan, Toman and Walls, and Krautkraemer).

2. BASIC HYPOTHESES IN THE ECONOMICS OF DEPLETABLE ENERGY SUPPLY

The fundamental assumption underlying all of mainstream economic theory is that economic actors are motivated to make choices by the goal of maximizing some measure of economic returns. The textbook model of commodity supply is a simple case in point. In that model, competitive producers (producers lacking any market power, taking prices of inputs and outputs as given) seek to maximize profit, the excess of gross revenue over their opportunity cost of production, including returns to capital and entrepreneurial effort. Profit is maximized when the market price of output, a measure of incremental gain from expanding production, equals the incremental cost of output expansion taking into consideration all possible technological possibilities for cost-effectively combining various inputs.

The same assumption underlies the economics of depletable energy supply. However, in that theory, the

measure of returns assumed to motivate behavior is a present value of current and future net revenues that accrue over time from a sequence of production decisions. A present-value metric is used in depletable energy economics because (for reasons explained further later) the act of producing or developing more or less resource today has unavoidable consequences for future potential returns, owing to the interplay of economics, geology, and technology. If these future consequences were ignored by totally myopic producers, their cumulative effects would come as an unanticipated and unwelcome surprise later. Because there is every reason to assume that producers are aware of these intertemporal connections and, being aware, are motivated to account for them, it makes more sense from the standpoint of theoretical logic to suppose that producers systematically account for them in evaluating the present value of a sequence of decisions over time.

The foregoing does not imply that energy producers are omniscient or free of forecast error. That is *prima facie* not the case. Later in this article there is a discussion of how the theory of depletable energy supply has been extended to incorporate various kinds of uncertainties. Those extensions emphasize the key role that expectations play in influencing supply behavior. What is assumed in the economics of depletable energy supply, however, is that producers are capable of avoiding systematic and persistent errors over time. This assumption can and sometimes is questioned, but no known persuasive economic theory based on alternative assumptions has been formulated.

It is worth noting in this regard that the calculation of a present value involves the choice of discount rate(s) to deflate future returns relative to current returns. In characterizing producer behavior in depletable energy supply theory, the appropriate discount rates would reflect the competitive rate of return capital invested in the energy industry could earn in other uses, including adjustments for various kinds of financial risks. In practice, market-based rates of return put disproportionate weight on returns in the first few years of an activity, thus the assumption that producers are not systematically biased in their expectations of future returns is not as difficult to swallow as it might first seem.

The standard theory of depletable energy supply can be extended well beyond a simple description of a single competitive firm. In addition to uncertainty, already noted, the theory can be extended to incorporate a number of market distortions, including property rights problems, environmental spillovers, taxes, and market power. However, efforts to

describe the exercise of market power by energy producers such as the Organization of Petroleum Exporting Countries (OPEC) and its members have been especially unsuccessful. A number of political and other factors must be incorporated to treat market power properly.

3. THE SIMPLE “CAKE-EATING” MODEL OF RESOURCE EXTRACTION

In the simplest case, it is assumed that a known, finite stock of the energy resource is being allocated over some number of time periods to maximize the present value of net returns. Producers calculate returns to extraction in different periods based on price expectations (in the simple model, expectations are assumed always to be realized), and knowledge of costs based on technological and geological information. Returns across periods are linked in this model by the simple fact that any energy resource produced in some period is by definition not available for production in future periods (energy resources obviously cannot be recycled). Only a finite amount of energy resource is available in this framework, and all of it will eventually be extracted. But the finite availability of fossil fuels is not the only key feature of energy supply, even in the simplest framework. In this framework, the present value of net returns is maximized when the “equimarginal principle” is satisfied. This principle requires that the marginal net return from the extraction and sale of the energy resource be the same in every time period with positive extraction. This, in turn, requires the current value of marginal net profit to be increasing at the rate of discount applied to future returns.

This principle can be stated in simple mathematical terms. Let P_t denote the price of extracted energy resource at time t , and q_t is the quantity extracted at time t , $C(q_t)$ is the cost of extracting P_t , and δ is the discount rate. Then, under the equimarginal principle, it must be that

$$\begin{aligned} P_0 - C'(q_0) &= \frac{P_1 - C'(q_1)}{(1 + \delta)^1} = \frac{P_2 - C'(q_2)}{(1 + \delta)^2} \\ &= \frac{P_t - C'(q_t)}{(1 + \delta)^t} = \lambda \end{aligned} \quad (1)$$

where λ denotes the common present value of marginal net profit. This number can be interpreted as the present-value shadow price or *in situ* value of the resource stock, because it reflects the incremental

increase in present value that would be enjoyed if a resource owner could experience an augmentation of resource stock. It also is referred to in the literature as the “user cost” or “scarcity rent” associated with depletion of the resource stock.

At any point in time, then,

$$P_t = C'(q_t) + (1 + \delta)^t \lambda. \quad (2)$$

In any successive time periods,

$$\begin{aligned} \frac{P_{t-1} - C'(q_{t-1})}{(1 + \delta)^{t-1}} &= \frac{P_t - C'(q_t)}{(1 + \delta)^t} \quad \text{or} \\ P_t - C'(q_t) &= (1 + \delta)[P_{t-1} - C'(q_{t-1})]. \end{aligned} \quad (3)$$

These conditions are known as the Hotelling rule (see below) in honor of Hotelling’s contribution to the literature. The first efficiency condition requires that at any point in time, the marginal gain from additional extractions equals the full opportunity cost of extraction, including the present value of future losses from a diminished future resource stock as well as conventionally defined incremental production costs. The second efficiency condition requires that the rate of return to holding the asset be equal to the rate of discount. This condition can be thought of in terms of asset market equilibrium, because presumably the discount rate is the rate of return that could be earned by investing in other assets. The markets for assets will not be in equilibrium if the rate of return to holding one asset is greater than the rate of return to holding some other asset. If that were the case, investment in the higher (lower) return asset would increase (decrease). Thus, asset market equilibrium requires that the rate of return to holding a nonrenewable resource asset be the same as the rate of return to other assets in the economy as represented by the discount rate.

The user cost, or shadow price of the remaining stock, depends on future extraction costs and the future price path. These depend, in turn, on developments in resource production technology, the size of the remaining stock, and future availability of substitutes. The future price of the resource can depend on the general level of economic activity, so expectations about population and economic growth are important determinants of the scarcity rent or user cost. Technological progress can affect extraction cost, the availability of substitutes, and the remaining resource stock. For example, the development of directional drilling techniques has lowered the cost of extraction and increased the amount of recoverable oil from some deposits. Increased computing capability has enhanced the ability to use information from exploration activities

and lowered the cost of discovering and developing new oil reserves.

The model just described can be extended in several directions. Switching from considering a single competitive producer to a price-setting monopolist, it is possible to infer that the monopolist will reallocate output from earlier to later in the life of the resource base, because this generates higher prices early on, and near-term returns are worth more than discounted future returns. But because the stock of resource is fixed and is necessarily exhausted under either market structure, resources that are not produced earlier will be produced later. If, to take another scenario, the stock of the resource is fixed but uncertain, then, again, extraction early on would be less than in a situation with certainty about the resource stock, because producers would put weight on the possibility of more scarce and therefore more valuable carry-over stocks. If, on the other hand, there is a renewable backstop energy form that can be supplied if the price of depletable energy climbs high enough, the Hotelling rule will still hold, but the net price path will rise at the rate of discount until it meets the unit cost of the backstop just at the point of full depletion.

Despite its seeming intuitive appeal, however, the basic Hotelling model is built on very shaky assumptions. It assumes a known stock of a resource of homogeneous quality and that the extraction technology does not change over time. In fact, nonrenewable resource stocks are not a given: exploration for new deposits as well as the further development of existing deposits are important features of fossil fuel supply. For many nonrenewable resources, including fossil fuels, the discovery of additional deposits has exceeded consumption, so that known, or proved, reserves has actually increased over many time periods. For example, U.S. oil reserves increased from 13 billion barrels in 1930 to 33.8 billion barrels at the end of 1990, and production in that period was 124 billion barrels. More than 30 billion barrels were produced from 1991 to 2000, but reserves at the end of 2000 were 29.7 billion barrels. The increase in world oil reserves is even more dramatic—proved reserves increased from 660 billion barrels at the end of 1980 to 1009 billion barrels at the end of 1990. This increased slightly to 1046 billion barrels at the end of 2000, even though 250 billion barrels were produced from 1991 to 2000.

Nonrenewable resources, including energy resources, also vary in quality and in their cost of extraction. Moreover, cost conditions for extraction in a specific petroleum reservoir change over the

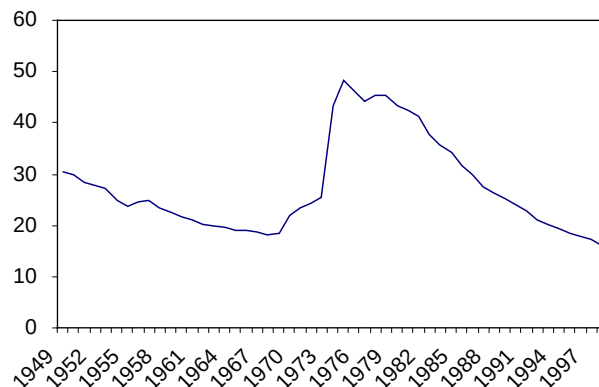


FIGURE 1 Real price of coal, 1949–1999, in chained 1996 dollars, calculated by using gross domestic product implicit price deflators. From the Energy Information Administration, United States Department of Energy, *Annual Energy Review*. Available on the Internet at <http://www.eia.doe.gov/emeu/aer/txt/tab0708.htm>.

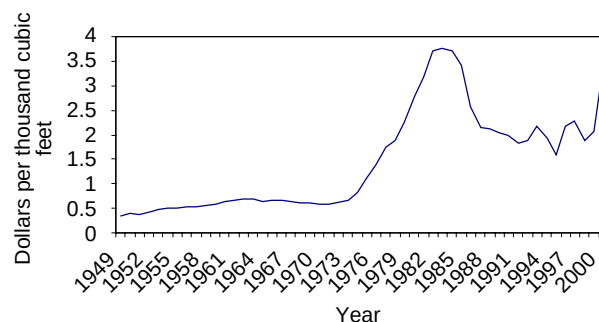


FIGURE 2 Real price of natural gas, 1949–2000, in chained 1996 dollars, calculated by using gross domestic product implicit price deflators. From the Energy Information Administration, United States Department of Energy, *Annual Energy Review*. Available on the Internet at http://www.eia.doe.gov/emeu/aer/pdf/pages/sec6_19.pdf.

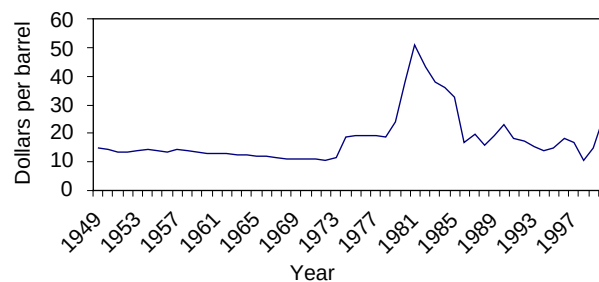


FIGURE 3 First purchase price, 1949–2000, of crude oil (domestic), in chained 1996 dollars, calculated by using gross domestic product implicit price deflators. From the Energy Information Administration, United States Department of Energy, *Annual Energy Review*. Available on the Internet at http://www.eia.doe.gov/emeu/aer/pdf/pages/sec5_39.pdf.

economic life of the reservoir: as natural drive is depleted, cost rises and yield per unit extraction effort falls, unless costly secondary and other recovery enhancement measures (e.g., injection of water or carbon dioxide) are undertaken to stimulate reservoir pressure. In short, resource stocks are neither fixed nor homogeneous, contrary to the assumptions underlying the simple Hotelling rule. It is not surprising, therefore, that the simple rule has done poorly in describing actual supply behavior (see later). We must look to a richer theory if we want to better understand observed behavior.

Empirically, the resource price should reflect both the marginal extraction cost and the scarcity rent, therefore generally it would be expected from Eq. (3) that the price of a nonrenewable resource would be increasing over time. However, there has not been a persistent increase in prices of fossil fuels (or other nonrenewable resources), but rather fluctuations over various time periods. Fossil fuel prices fell over much of the 20th century, peaked dramatically in the 1970s and early 1980s, and have generally fallen since then (see Figs. 1–3). This further illustrates the need for a richer theory to capture observed behavior.

4. A MORE GENERAL THEORY OF ECONOMICALLY EFFICIENT DEPLETABLE RESOURCE SUPPLY

A more useful general theory must take into account the stylized facts already noted—that resources are heterogeneous, with extraction costs that vary systematically over time with depletion of resource deposits—and that resources can be augmented through exploration and development. Moreover, a complete theory must account for the high degree of capital intensity of the extractive sector and the implications of this in terms of potential constraints on productive capacity.

4.1 Resource Extraction

Let us first address generalization of the Hotelling extraction model to deal with resource heterogeneity. Among the many good papers to deal with this case, those of Pindyck (in 1978) and Heal (in 1976) are particularly noteworthy. To examine this part of the problem, continue to treat reserves as fixed for the moment, and ignore capacity constraints.

The cost of extraction is assumed to be given by some function, $C(q_t, R_t)$, where R_t is the stock of

resource that has not yet been depleted. Incorporating this “stock effect” into extraction cost is a simple if somewhat imprecise way to introduce resource heterogeneity into the extraction cost relationship. The relationship is relatively straightforward if we focus on a single energy resource development: as the resource in this development declines, cost is assumed to rise (deeper pits to be dug, more waste material to be removed, weaker reservoir pressure). It is less straightforward if we are attempting to describe heterogeneity among resource deposits (as is discussed later). With cost rising as the resource is depleted, the assumption of complete physical exhaustion from extraction—which is not empirically realistic anyway—no longer makes sense in the theoretical model either. Instead, what drives extraction behavior is economic exhaustion of the resource—at some point, cost becomes so high relative to the price what buyers are willing to pay that a particular mine or field is abandoned.

With resource heterogeneity and incomplete physical exhaustion, the simple Hotelling model summarized by the efficiency conditions [Eqs. (1)–(3)] no longer applies. Instead, the equimarginal principle must be modified as in the following equation:

$$\begin{aligned} P_t &= C_{qt} + (1 + \partial)^t \lambda_t \\ &= C_{qt} + \sum_{s=t+1}^T (-C_{Rs})(1 + \partial)^{t-s}. \end{aligned} \quad (4)$$

In Eq. (4), subscripts q and R on cost C signify rates of change of cost with respect to the extraction rate and the depletion of reserves, respectively. The term λ_t is a generalized measure of user cost that reflects the rate at which future extraction cost is increased as a consequence of future depletion. This stock effect is part of the full opportunity cost of current extraction. Manipulating Eq. (4) a bit algebraically yields

$$P_t - C_{qt} = (1 + \partial)[P_{t-1} - C_{q,t-1}] + C_{Rt}. \quad (5)$$

Equation (5) shows how with increased stock providing a kind of “dividend” in terms of holding down extraction cost over time ($C_{Rt} < 0$), the net return from depletion no longer has to rise at the rate of the discount rate to induce resource owners to retain their holdings. Application of this framework to describe the depletion over time of different energy deposits is frequently done, but is somewhat more problematic. In the simplest case, where resource quality varies across deposits but is homogeneous within a deposit, the optimal extraction pattern generally requires exploiting those deposits in a

sequence from low cost to high cost. The user cost for a lower cost deposit is greater than the user cost for a higher cost deposit. At the time of transition from one deposit to the next most costly deposit, the marginal extraction cost plus user cost is the same at each deposit. This implies that the resource price rises at a rate slower than the rate of interest during the transition. Simultaneous extraction from different deposits can be optimal when marginal extraction cost at a deposit increases with the extraction rate, or if extractive capacity is fixed.

The current pattern of petroleum extraction obviously violates the rule of exploiting lower cost deposits before higher cost deposits. The cost of extraction from petroleum reserves in the Middle East is much lower than the cost of extraction from reserves in the United States. This suggests that the simple modeling device of incorporating resource inventory in the extraction cost relationship probably aggregates over and therefore blurs some important distinctions. The case of petroleum is further complicated by the actions of large oil producers in OPEC using market power to restrict their output to levels below what would occur under perfect competition. The higher-than-competitive price that results provides the incentive for higher cost producers to extract from their reserves. The higher noncompetitive price also distorts the capital investment decision, because the marginal cost of investment in capacity varies widely, from \$343 in Saudi Arabia to over \$10,000 in the United States. Efficient allocation would result in the equalization of marginal investment cost across geographical regions.

4.2 Exploration and New Resource Development

Exploration is an economic activity and, like other economic activities, marginal benefit and marginal cost play a key role in determining how much exploration occurs. Again, Pindyck published a key paper in the development of this analysis. Expenditures on exploration and development reduce current profit with an expected return of adding valuable reserves for future exploitation. The efficient level of exploration activity balances the expected marginal cost of exploration with the expected benefit of exploration.

The expected marginal benefit of exploration is not just the discovery and development of new reserves *per se*; in a model of incomplete resource exhaustion, the quantity of the resource stock is not determinative. The marginal benefit of exploration comes from the discovery of reserves for which the

cost of development and extraction is less than or equal to the cost of the reserves currently being depleted. In other words, the marginal value of new reserves equals the user cost of current resource depletion along an efficient time path of extraction and new reserve development.

One feature of the full opportunity cost of exploration activity is that it reduces the stock of exploration opportunities, so that future additions to reserves tend to become more costly. In other words, there is a user cost similar to that in Eq. (4) associated with depleting the stock of exploration opportunities, in addition to direct exploration costs. If the conditions for efficient exploration, development, and extraction are then combined, we get

$$\begin{aligned} \text{Price} &= (\text{marginal extraction cost}) \\ &\quad + (\text{extraction user cost}) \\ &= (\text{marginal extraction cost}) \\ &\quad + (\text{marginal direct cost of} \\ &\quad \text{finding new reserves}) \\ &\quad + (\text{user cost of reduced long-term} \\ &\quad \text{development prospects}). \end{aligned} \quad (6)$$

The heuristic equality in Eq. (6) can be related to intuition about opportunity cost in economics more generally. The extraction user cost is the cost of depleting an existing inventory, which can be replenished, but only (in the model) at an ever-rising cost. The sum of the marginal extraction cost and the marginal direct cost of finding new reserves gives a measure of the cost of using and replacing the inventory, but it ignores the ultimate cost of natural capital depreciation reflected in the user cost of reduced long-term development prospects.

Once we allow reserves to vary endogenously with economic (and technological) conditions, the inexorable upward movement of prices implied by Eq. (3) or even Eq. (5) no longer need hold. Depending on the degree to which current extraction costs are rising as a consequence of depletion and the cost of new reserve additions, including their user cost, it is possible in the model for new discoveries to outstrip depletion, at least for awhile, causing a U-shaped energy price path. This is a reassuring finding, because such behavior has been observed in energy and other resource markets, and a model incapable of replicating such behavior would have little use. The ability to de facto enlarge reserves through technological breakthroughs that increase recoverability and lower the costs of extraction from existing deposits, or make new deposits easier to find and access (lowering reserve

replacement costs), likewise could exert downward pressure on prices, possibly for a long period of time. The simple model of endogenous exploration and reserves makes the simplifying assumptions that reserves are found on average in order of increasing cost and can be found in infinitely divisible amounts. Given the uncertainties surrounding the exploration process, the former assumption obviously is an oversimplification. When the discovery of deposits is more random, the aggregate extraction cost function cannot be described by a reserve stock indicator alone.

If the discovery of deposits is “lumpy,” exploration and discovery can occur in “episodes,” with intervals of no exploration between those episodes. Exploration does not generally occur in strictly discrete episodes, but it does vary with the resource price, and large discoveries do put significant but temporary downward pressures on price. The theory would then predict that the *in situ* value of the resource increases between major discoveries and jumps up or down when exploration occurs, depending on whether the exploration outcome is less or more favorable than expected. The resource price path can follow a “sawtooth” pattern while still having a downward trend over some time periods, depending on the actual exploration outcomes. The likelihood that the resource price will be increasing over time increases as unexplored prospects become scarcer.

Exploration also provides new information about future exploration prospects that can lead to revised expectations about the future value of the resource stock and generate new expected time paths for the resource price and quantity extracted. For example, a firm may revise its expectations about the probability of successful exploration either upward or downward in response to information obtained from its current exploration. In this case, a variety of patterns are possible for the realized price path, including a generally downward trend. The arrival of previously unanticipated information can alter the resource price, extraction, and exploration paths so that the observed price path deviates systematically from a deterministic calculation. The observed time paths for the resource price and *in situ* resource value may represent a combination of the initial portion of many different expected price paths rather than outcomes along one fully anticipated price path. Moreover, because individual energy producers can learn something about the information possessed by others via observing their behavior (e.g., by another producer not drilling in a certain area), information has certain public good aspects. To ensure that enough information about energy resources is

collected and disseminated to society at large, it may be necessary for the government to undertake or underwrite investments in resource delineation.

4.3 Capacity Investment and Capacity Constraints

The basic nonrenewable resource model does not capture the capital intensity of extractive industries. The development and extraction of fossil fuels are capital intensive, and the timing and size of investment in extractive capacity are functions of the cost of capital and the expected price path. Once the existence of capacity constraints is incorporated, the behavior of the energy resource price along an efficient path generally will depart from the kind of behavior implied by Eq. (5) (or its extension to incorporate new reserve additions). This problem, too, has been studied by a number of authors (Lasserre, Powell and Oren, and Cairns and Lasserre).

If extraction is expected to decrease over time, either because the present value price of the resource is declining or the extraction cost increases with depletion, the extractor would seek to reduce its capital input over time if it were able to do so. However, this is not possible if capital is nonmalleable (i.e., once invested in oil fields, it cannot easily be converted into rail car investments). The energy-producing firm takes this into account when the initial investment is made, and the initial capital investment is lower than it would have been if capital were malleable. Once in place, this lower initial extractive capacity is likely to constrain the rate of extraction from the deposit, at least early in the extraction horizon. In addition, it may be relatively costly to increase extractive capacity, at least quickly. Firms may invest in advance of the need for some capacity in order to smooth out their investment costs over time, but they will also invest slowly in response to previously unanticipated price increases. Consequently, the short-run supply curve for an energy resource may be very inelastic and so variations in market demand (especially increases) may be mediated more by price changes than by quantity changes. The result is that short-term energy prices can be relatively volatile.

In the standard nonrenewable resource model, an increase (decrease) in the interest rate decreases (increases) the relative value of extraction in the future relative to current extraction, and so tilts the depletion path toward the present and away from the future. However, if a large capital investment is necessary before extraction begins, an increase in the

interest rate increases the cost of the capital investment. Thus, a higher interest rate reduces the incentive to use capital, and can lead to lower rather than greater initial extraction whether or not capital is malleable. The effect of the higher interest rate on the capital investment outweighs the effect of the higher interest rate on user cost if the initial resource stock is relatively large. A lower interest rate increases conservation if the interest rate is low (when user cost is high and capital cost is low), but decreases conservation when the interest rate is high (user cost is low and capital cost is high). A higher (lower) interest rate also can increase the capital cost of a backstop technology, increasing the maximum, or choke, price for the depletable energy resource and resulting in less (more) rapid extraction.

5. EFFECTS OF UNCERTAINTY

There is uncertainty regarding the future values of many important influences on the intertemporal supply of depletable energy resources: future price, future extraction cost, the costs and outcome of exploration and development activities, and the cost and timing of the availability of a backstop technology. The intertemporal extraction and resource development patterns and price path are affected by uncertainty in a variety of ways, depending on the source and nature of the uncertainty. From a financial perspective, the risk associated with holding reserves of a depletable energy resource can be mitigated by also owning other assets. The necessary return to holding the energy resource asset—including the rate of *in situ* appreciation plus the benefits of increased reserves in moderating costs—can be less if the return is negatively correlated with the return to other assets in a portfolio, thereby reducing overall risk.

Even with modern developments in financial markets, such as energy commodities futures that more tightly link energy and financial markets and provide added options for risk management, in practice it is not possible for energy producers to completely “insure” away fluctuations in net income. Moreover, firms will not be indifferent to risks even if their basic management posture is what economists call “risk neutral.” Work by Pindyck and others shows that these impacts depend on the curvature of incremental cost relationships, among other factors. For example, uncertainty about the growth of demand would shift extraction from the future to the present; if the degree of uncertainty increases

with time, a limiting case would be the risk of expropriation, meaning no return from future production. Uncertainty about reserves could increase or decrease the marginal value of new discoveries compared to a situation of the same average conditions with no uncertainty.

6. MARKET FAILURES AND DISTORTIONS

Assessing the effects of various market failures and distortions on depletable energy supply is complicated by the inherently intertemporal nature of these resources. Whether a market imperfection results in depletion that is more or less rapid or extensive than the socially efficient depletion rate or volume depends on how the distortions change over time and interact with dynamic changes in resource scarcity signaled by the extraction and exploration user costs. For example, a monopolist would want to restrict output to raise price and increase returns. This would mean a lower cost of extraction in the future compared to a competitive industry facing the same demand conditions and extracting more quickly. (In actuality, demand conditions would change in response to the higher monopoly prices, as energy users adapted to more energy-efficient and alternative-energy technologies.) This, in turn, would encourage more output. Nevertheless, the monopolist could restrict output over time in such a way that the cumulative depletion of the resource (and cumulative development of new resources over time) was less than under competition. This is a different prediction than the one obtained from a simple model with fixed homogeneous reserves.

Government interventions also can tilt the market’s depletion path toward the present or the future. Energy resource industries are subject to a variety of taxes in addition to the taxes paid by firms generally. One motivation for depletable resource taxation is that depletable natural resources such as fossil energy are part of the national heritage, and resource rents should accrue to the general welfare. Depletable energy resources can be subject to severance taxes per unit extracted or royalty payments as a percentage of the resource price. In general, the effect of such a tax on the intertemporal extraction pattern depends on how the present value of the tax changes over time as well as on the magnitudes of depletion effects in the cost structure. For example, the present value of a constant severance tax decreases over time and so shifts extraction from the present to the

future. However, the tax still reduces net returns in any period compared to a no-tax alternative, thereby reducing the contemporaneous incentive to extract. Although less extraction would moderate the depletion effect on cost, it is still likely that total recovery of the energy resource will fall over time as a result of the tax. In particular, reserves that were formerly marginally economic to extract become subeconomic. Moreover, the tax would reduce the return to investment in new reserves.

The extraction and consumption of energy resources are associated with a wide variety of environmental externalities. The effect of externalities on the market's depletion path can depend on the exact nature of the externality. There are two broad categories of externalities: flow externalities and stock externalities. Flow externalities represent a situation in which the environmental damage is a function of the current rate of energy production or consumption. An example relevant to energy supply would be the air pollution by-products of energy use in the process of mining. More persistent stock externalities arise when the environmental damage is a function of cumulative emissions. Examples of more persistent externalities would include atmospheric accumulation of carbon dioxide and its effect on the global climate, contamination of ground water from oil or coal extraction that is only slowly reversed by natural processes, and unremediated damage to natural landscapes through strip mining.

Flow externalities can result in a depletion rate that is too rapid or too slow relative to the socially efficient supply rate, depending on how the present value of the marginal external damage changes over time. The effect of stock externalities is similar to the impact of resource depletion on the cost of extraction: these stock effects imply a slower rate of depletion and less total depletion than the market would choose on its own.

7. EMPIRICAL ANALYSIS OF DEPLETABLE ENERGY SUPPLY

Efforts to test empirically the theoretical implications of resource depletion models as descriptions of observed energy supply behavior are hampered by a lack of data. Data for *in situ* reserve values are not readily available. Market data on the value of firms are generally available, but exclude depletable resource values, and data on extraction and development costs are usually proprietary information.

Consequently, the empirical economic analysis of resource supply and scarcity has taken a variety of less direct paths. Analysts have attempted in a number of ways to use available data to construct dynamic empirical models based on the basic theoretical tenets sketched in the preceding sections.

7.1 The Hotelling Valuation Principle for Assessing Resource Scarcity

An illustration of the methods for testing the resource scarcity implications of the depletable energy supply model involves the relationship between average reserve value and current net price using cross-section data. Miller and Upton argued that this relationship can be written as follows:

$$\frac{V_0}{S_0} = \alpha + (P_0 - C_0). \quad (7)$$

Here V_0 denotes reserve value, S_0 is the resource endowment, and $P_0 - C_0$ is the price net of extraction cost; the term α , which can be positive or negative, reflects influences such as changing cost and price trends. The model was found to be consistent with pooled, cross-section data from December 1979 to August 1981 for 39 oil- and gas-producing firms in the United States. A subsequent test of the Hotelling valuation principle using data from August 1981 to December 1983 produced a quite different result: in that analysis, an increase in the net price translated into an increase in average reserve value less than half as large.

An explanation for why the Hotelling Valuation Principle might overvalue reserves, at least in the case of oil and natural gas production, is that it affords producers greater flexibility for choosing output than they actually have. The extraction of petroleum over time is restricted by declining well pressure as the reservoir is depleted. If the rate of extraction declines at the rate a because of declining well pressure, the average reserve value is

$$\frac{V_0}{S_0} = \frac{a}{a + r - g}(P_0 - C_0), \quad (8)$$

where g is the expected rate of change in net price. This example and others in the literature provide strong empirical evidence that a simple model of finite resource depletion does not adequately explain the observed behavior of depletable energy resource prices and *in situ* reserve values. This is not terribly surprising, given the many other features of depletable resource supply already discussed—such as exploration for and discovery of new deposits, technological

change, and capital investment—that alter the implications of finite availability. It seems clear that these other factors have overshadowed finite availability of the resource as determinants of the observed dynamic behavior of nonrenewable resource prices and *in situ* values.

7.2 Data on Reserves

In the theoretical models of depletable energy supply we have discussed, measures of reserves or resources play a significant role as indicators of cost drivers, not just in terms of physical availability. Before looking at efforts to model empirically depletable energy supply, therefore, it is worth making some brief comments on the nature of available data for reserves.

The classification scheme of the United States Geological Survey defines total resources, including energy resources, as materials that have been discovered or might be discovered and used. Reserves are that portion of total resources that can be economically extracted. Undiscovered resources are classified as hypothetical, if in a known mining district, or speculative. Identified but currently noneconomic resources are categorized as paramarginal or submarginal. The physical measures of reserves often are compared to measures of the rate of use in order to determine the remaining life of the reserves.

Because reserves are defined in terms of economic recovery, whether a deposit is classified as a reserve changes with the resource price and extraction cost. In addition, because costly investment is required to “prove” reserves, there is a limited incentive to prove reserves beyond a certain point. The ratio of reserve to consumption for petroleum increased from 35 in 1972 to 45 in 1990, even though commercial energy consumption increased by more than 50% between 1971 and 1991. Physical measures of reserves thus probably have more meaning as an indicator of inventory on hand than as a more fundamental indicator of resource cost.

7.3 Models of Depletable Energy Supply Behavior

Behavioral models of depletable energy or other resource supply seek to develop empirical counterparts to theoretical efficiency conditions [such as Eqs. (4)–(6)] in order to describe how supply decisions (reserve development and extraction) respond to economic incentives (prices and cost drivers). Much of this work has focused on oil and

gas supply, though some applications to mineral resources (e.g., nickel) also can be found in the literature. Relatively little of this kind of modeling work has been done for coal, because the depletion effects central to the development of the dynamic efficiency conditions (discussed previously) seem to be of very limited relevance to coal supply (with large reserves and continued technical progress in resource exploitation).

The problem of developing empirical behavioral models based on the theory of efficient depletable energy supply is quite difficult for several reasons. One problem, noted previously, is the poor data on reserves and costs. But even with better data, it is inherently complicated to develop reliable empirical models for complex dynamic process, and resource depletion is inherently dynamic. Moreover, because behavior depends so much on expectations of unknown future influences in this dynamic framework, the empirical analyst must formulate a hypothesis about the nature of the expectations driving behavior.

Finally, the theory discussed previously applies most directly to an individual decision-making unit and perhaps even to decisions within a specific resource-bearing location. Once the analysis is extended to cover multiple extraction and investment decisions from heterogeneous deposits, some difficult aggregation problems arise. Care is needed not to omit or mix together key influences. For example, the yield from exploratory drilling likely depends on both how much the best resource prospects have been depleted by past activity and what advances in drilling technique can do to raise discoveries per unit of effort.

Basically, two types of modeling approaches are found in the literature: process optimization simulation models and econometrically estimated models. Process optimization simulation models specify and numerically solve sets of equations based on first-order efficiency conditions such as Eqs. (4)–(6). For example, the model could incorporate a specification in which the reserves discovered per foot of new drilling depend on cumulative past drilling in that province. This specification could be used along with other information to determine the efficient level of reserve development as a function of the direct cost and user cost of drilling, plus the marginal (shadow) value of additional reserves. One prominent example of such a model in the United States is the oil and gas supply component of the National Energy Modeling System (NEMS) operated by the Energy Information Administration (EIA); Deacon has provided a smaller-scale example.

Numerical parameters in such models come from a variety of sources: statistical curve fitting, engineering experience, and informed judgment, among others. Because these models are not derived from statistically fitting equations to data, it is not possible to construct classical confidence intervals and statistics for assessing the reliability of the model outputs. They are useful principally for “what if” exercises—by supplying different sets of input variables (e.g., oil prices and technical recovery parameters for drilling), the consequences of changes in these assumptions can be seen. Whether the model structure is properly specified empirically cannot be independently assessed. The simulation models also do not incorporate uncertainty and expectations formation. Implicitly in dynamic optimization models, individuals are assumed to judge the future accurately.

Econometrically estimated models also are rooted in the theoretical efficiency conditions. However, these models are parameterized using statistical methods that seek to define supply decisions as functions of causal variables. The most sophisticated of these frameworks explicitly model expectations formation using the rational expectations time series approaches pioneered in macroeconomics by Epple, Hendricks and Novales, and Walls. In brief, these methods explicitly include representations of expectations formation based on past observations of price and cost influences; assuming that producers are not systematically biased in their expectations and fairly rapidly learn about a systemic change in market or technological conditions, statistical techniques can be used to estimate the parameters of the expectations process and the underlying parameters linking supply choices to the relevant economic influences. The point is to provide consistent estimates of the basic technological and geological parameters undergirding the supply process, so that the effects of policy intervention or shifts in market conditions can be separated out and understood. However, success with this class of models has been limited, notwithstanding their conceptual appeal. The models are difficult and data intensive to construct and estimate; they are not well suited to geographical or other finer grained breakdowns of activity into constituent parts; and the results have not always been very powerful. In particular, some approaches have not coped well with what appear to be basic nonlinearities in the relationship between unit supply cost and reserves. A prospect for future work is a better blending of the engineering expertise found in energy systems modelers with the dynamic economic models discussed here.

8. CONCLUDING REMARKS

Finite availability is one defining characteristic of a depletable energy resource and generates the “Hotelling rule” that the marginal value of a nonrenewable resource stock increases at the rate of interest. However, many other factors, including exploration, capital investment, and heterogeneous reserve quality, also are important to the economics of energy depletion and development. The investigation of how these other factors affect the empirical implications of the Hotelling model has been spurred by the frequent failure of the basic Hotelling model to explain the observed dynamic behavior of nonrenewable resource prices and *in situ* values.

These other factors can affect price and depletion paths in a number of ways, particularly when considered in combination with each other. The variety of possible outcomes makes it difficult, if not impossible, to make any general predictions about the overall impact on price and extraction paths. These other factors, particularly the discovery of new deposits and technological progress that lowers the cost of extracting and processing nonrenewable resources, appear to play a relatively greater role than does finite availability in determining observed empirical outcomes. Although models that include these other factors have enriched the Hotelling model, they have not completely reconciled the economic theory of nonrenewable resources with the observed data. The distinction between the responses of supply to anticipated versus unanticipated changes in extraction cost, interest rate, reserve discoveries, availability of backstop substitutes, and other factors figures prominently as a topic for future empirical research. The empirical evidence also indicates that the discovery of new deposits and technological progress in resource extraction and discovery have significantly mitigated the impacts of finite geological availability on the relative cost and supply of depletable energy resources. Although the future holds promise with respect to the development of new substitutes for fossil energy and greater energy efficiency, improvements in extraction and exploration technologies that reduce costs and allow the economical use of lower grade reserves also are likely to continue.

SEE ALSO THE FOLLOWING ARTICLES

Depletion and Valuation of Energy Resources • Economic Geography of Energy • Economic Growth

and Energy • Economics of Energy Demand • Energy Futures and Options • Market Failures in Energy Markets • Modeling Energy Supply and Demand: A Comparison of Approaches

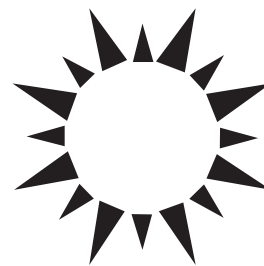
Further Reading

- Adelman, M. A. (1990). Mineral depletion with special reference to petroleum. *Rev. Econ. Statist.* 72(1), 1–10.
- Adelman, M. A. (1993). Modelling world oil supply. *Energy J.* 14(1), 1–32.
- Bohi, D. (1999). Technological improvement in petroleum exploration and development. In “Productivity in Natural Resource Industries” (R. D. Simpson, Ed.). Resources for the Future, Washington, D.C.
- British Petroleum (2001). BP Statistical Review of World Energy, June. Available on the Internet at <http://www.bp.com/> (accessed November 12, 2002).
- Brobst, D. A. (1979). Fundamental concepts for the analysis of resource availability. In “Scarcity and Growth Reconsidered” (V. K. Smith, Ed.), pp. 106–159. Johns Hopkins University Press (for Resources for the Future), Baltimore.
- Cairns, R. D. (1990). The economics of exploration for non-renewable resources. *J. Econ. Surv.* 4(4), 361–395.
- Cairns, R.D., Lasserre, P. (1990). Investment in a multiple-deposit non-renewable resource industry: Some results and remaining puzzles. *J. Environ. Econ. Manage.* 21(1), 52–66.
- Darmstadter, J. (1999). Innovation and productivity in U.S. coal mining. In “Productivity in Natural Resource Industries” (R. D. Simpson, Ed.). Resources, Washington, D.C.
- Deacon, R. T. (1993). Taxation, depletion, and welfare: a simulation study of the U.S. petroleum resource. *J. Environ. Econ. Manage.* 24(2), 159–187.
- Energy Information Administration. (2000). “The National Energy Modeling System: An Overview 2000.” U.S. Department of Energy, Washington, D.C.
- Epplé, D. (1985). The econometrics of exhaustible resource supply: a theory and an application. In “Energy Foresight and Strategy” (T. J. Sargeant, Ed.). Resources for the Future, Washington, D.C.
- Epplé, D., Londregan, J. (1993). Strategies for modeling exhaustible resources supply. In “Handbook of Natural Resource and Energy Economics,” Vol. 3 (A. V. Kneese and J. L. Sweeney, Eds.). Elsevier, Amsterdam.
- Farzin, Y. H. (1984). The effect of the discount rate on depletion of exhaustible resources. *J. Politic. Econ.* 92(5), 841–851.
- Heal, G. M. (1976). The relationship between price and extraction cost for a resource with a backstop technology. *Bell J. Econ.* 7(2), 371–378.
- Hendricks, K., and Novales, A. (1987). Estimation of dynamic investment functions in oil exploration. Stanford Hoover Institute Working Papers in Economics, pp. E87–E52. Hoover Institution, Stanford, CA.
- Hotelling, H. (1931). The economics of exhaustible resources. *J. Politic. Econ.* 39(2), 137–175.
- Krautkraemer, J. A. (1998). Nonrenewable resource scarcity. *J. Econ. Lit.* 36(4), 2065–2107.
- Lasserre, P. (1985). Capacity choice by mines. *Can. J. Econ.* 18(4), 831–842.
- Livernois, J. R., and Uhler, R. S. (1987). Extraction costs and the economics of nonrenewable resources. *J. Politic. Econ.* 95(1), 195–203.
- Miller, M. H., and Upton, C. W. (1985). A test of the hotelling valuation principle. *J. Politic. Econ.* 93(1), 1–25.
- Miller, M. H., and Upton, C. W. (1985). The pricing of oil and gas: some further results. *J. Finance* 40(3), 1009–1018.
- Pindyck, R. S. (1978). The optimal exploration and production of nonrenewable resources. *J. Politic. Econ.* 86(5), 841–861.
- Pindyck, R. S. (1980). Uncertainty and exhaustible resource markets. *J. Politic. Econ.* 88(6), 1203–1225.
- Powell, S. G., and Oren, S. S. (1989). The transition to non-depletable energy social planning and market models of capacity expansion. *Oper. Res.* 37(3), 373–383.
- Slade, M. E. (1982). Trends in natural resource commodity prices: an analysis of the time domain. *J. Environ. Econ. Manage.* 9(2), 122–137.
- Slade, M. E. (1987). Natural resources, population growth and economic well-being: Issues and evidence. In “Population Growth and Economic Development” (D. G. Johnsson and R. D. Lee, Eds.), pp. 331–372. University of Wisconsin Press, Madison.
- Swierzbinski, J. E., and Mendelsohn, R. (1989). Exploration and exhaustible resources: the microfoundations of aggregate models. *Int. Econ. Rev.* 30(1), 175–186.
- Toman, M. A., and Walls, M. (1995). Nonrenewable resource supply: theory and practice. In “Handbook of Environmental Economics” (D. W. Bromley, Ed.). Blackwell, Oxford and Cambridge.
- Walls, M. (1992). Modeling and forecasting the supply of oil and gas: a survey of existing approaches. *Resour. Energy* 14(3), 287–309.
- Walls, M. (1994). Using a “hybrid” approach to model oil and gas supply: a case study of the gulf of mexico outer continental shelf. *Land Econ.* 70(1), 1–19.
- World Resources Institute (1994). “World Resources 1994–95: People and the Environment.” Oxford University Press, Oxford.

Economic Systems and Energy, Conceptual Overview

JOHN PEET

University of Canterbury
Christchurch, Canterbury, New Zealand



-
1. Physical Foundations of Economic Activity
 2. Energy in the Economic System
 3. The Economics and Physics of Production
 4. Energy Analysis of Economic Processes
 5. Physical Accessibility and the Concept of Net Energy
 6. Energy and its Importance to Policies for Sustainability
 7. Aims of the Biophysical Systems Worldview

Glossary

direct energy The energy used in the specific process of production of a good or service.

embodied energy The energy associated, directly or indirectly, with the production of a good or service.

energy analysis A family of tools that follow economic production and consumption processes through their (physical) energy consequences.

energy carrier (also energy vector) A portable substance that can release energy when desired (e.g., liquid petroleum fuel).

indirect energy The energy used in preparation of intermediate inputs to a process.

net energy The amount of energy available from an energy transformation system after deduction of the energy expended in creating the capital equipment for mining and refining and the energy used in operating the system.

Human society is a complex system, in turn embedded in another—the natural environment. Since humans are totally dependent on the natural environment for the necessities of life, and since human economic activities strongly influence both human and ecosystem health, we need to be well informed about key aspects of the state of the human system and the natural “supersystem” within which humans exist. One such key aspect is associated with the flows of energy, both natural and human influenced, throughout the biosphere.

Economic systems depend on the availability and use of energy resources for their every function. In this context, the ability of a form of energy to “do work” in a thermodynamic sense (e.g., run a machine) is often referred to as its availability (or exergy). Highly “available” forms of energy such as electricity are better able to perform a range of important functions in the economy than, for example, fuelwood, which has a significantly lower availability. The energy in seawater is much less available, and although the overall quantity is huge, its ability to do economically useful tasks is small. The term quality is often used to describe a fuel’s economic usefulness (e.g., as gross domestic product produced per heat unit of fuel used), and although this is related to its ability to do work in a strict thermodynamic sense, it is not the same thing.

From the mainstream economic standpoint, the economy is seen as a system of production, organized according to a system of signals in the form of prices paid by consumers to producers in order to obtain a good or service. These prices are determined in a marketplace, where willing buyers and willing sellers reach an understanding of the price to be paid for the product. Anything that is scarce has a price, provided ownership can be established.

On the other hand, from the biophysical systems standpoint, an economy is seen as a social system characterized by the physical activities that take place in it. That social processes are involved in the myriad of decisions made in that society is to some extent secondary to the physical processes that are occurring. From this perspective, physics, particularly thermodynamics, provides the tools for understanding both what is happening now and what is either possible or impossible, in physical terms, in the future. Clearly, the two perspectives are very different; part of the purpose of this article is to show

how at least a partial integration of perspectives is possible.

1. PHYSICAL FOUNDATIONS OF ECONOMIC ACTIVITY

The term energy not only applies to wood, coal, oil, gas, and electricity but also has a fuller scientific meaning. This is because the laws of thermodynamics apply, without exception, to all physical processes on Earth. They are true whether applied to the actions of people and machines in an economy, all forms of life, circulation in the atmosphere and the oceans, or movements of the planets. Inputs of available energy are essential to the functioning of any economy, not only directly as fuels but also indirectly through the natural solar cycles that provide the “ecosystem services” (air, water, soil, etc.) essential for all forms of life, including the economic.

The associated transformation of available into unavailable energy resources (in accordance with the second law of thermodynamics) is an unavoidable part of the functioning of all economies. There is no process known that did not begin by using some form of high-quality raw material (available energy) and that will not end up somewhere as waste (unavailable energy). Directly or indirectly, most source energy comes from or came from the sun, and all sink energy is eventually dissipated to outer space. Heat from the earth’s interior is also an important source of energy for some economies, and it is obviously relevant in many geological processes important to human life.

Early human societies evolved in a context in which they obtained the essentials of life from their surroundings via natural mechanisms by which most of the products they used were obtained, directly or indirectly, from the sun. Needs for food, clothing, shelter, and so on were met from plants and animals by hunting and gathering, and humans traveled or migrated as necessary to ensure continued supply. Early industrial societies also used solar-derived energy resources, especially wood as fuel, often converted to charcoal for more specialized purposes (e.g., high-temperature metal smelting). By the 18th century, large-scale mining of coal, together with associated technologies for the production of iron and steel, was fueling the industrial revolution and the rapid economic and population growth that eventuated. Later techno-

logies, similarly based on prior use of high-quality energy resources, enabled exploitation of oil and gas, in turn powering further economic growth. The ability of these fuels, together with their associated technologies, to do “work” in economic production processes vastly increased the ability of humans to control and transform their physical environment.

Each stage of industrialization was built on a foundation of previous industrialization. Without exception, all required prior consumption of available energy from the environment and, implicitly, the “investment” of that energy in the creation of the artifacts (machines, processes, and buildings) needed for physical production of goods and services.

From the physics standpoint, it is also important to note that what the marketplace calls energy is actually consumer fuels (coal, oil, gas, and electricity) since they are the form in which energy is traded commercially. In the generalized scientific sense, however, energy has no substitutes. Only fuels, which can be seen as “energy carriers,” can often be substituted, one for the other (e.g., oil for coal). An energy carrier is a portable substance that can release energy when desired (e.g., petroleum fuel such as gasoline).

The interaction between energy and matter is the basis of life; therefore, it is also the basis for human life and culture. However, this is not to suggest that life is only energy. Far from it. It is nevertheless vital to recognize the key role of energy in the physical world because otherwise it is fatally easy to fall into the trap of thinking that all processes of importance in human decision making are internal to the economy. Unfortunately, this view is typical of all major economic theories. This does not leave much room in the political spectrum for other views, but it is essential to do so and thereby add important information to both.

Probably the most notable exponents of the physical basis of economics are Nicholas Georgescu-Roegen, Kenneth Boulding, and Herman Daly, economists who have made extensive contributions to the literature on physical limits to the use of resources in economic production.

The inclusion of energy in a worldview highlights the help that the biophysical systems perspective can provide in clarifying questions about the relationship between a social system and its physical environment. It also helps us see where the biophysical viewpoint contrasts with the political-economic and thus indicates areas where conventional assumptions may need deeper examination.

2. ENERGY IN THE ECONOMIC SYSTEM

Starting with primary resources such as coal, oil, and gas in the ground, raw energy is processed into “consumer” form, suitable for use in society, via power stations, cables, oil refineries, gas pipelines, and so on. The resultant fuels, such as gasoline, piped natural gas, and electricity, are then used to produce energy services—transport, warmth, cooking, lighting, and so on.

Like energy, matter also enters an economic system as raw resources from the environment. Even if temporarily accumulated within the economy (e.g., embodied in capital goods) or recycled for awhile within the economic process, it must be disposed of. This gives rise to pollution, acknowledged by the concept of externalities in economic terminology. These are also seen as feedbacks in systems terminology.

The flow of matter into and through an economic system can be characterized in terms of the high-quality available energy required to carry out the transformation processes in which the matter is involved. Thus, energy consumption can be used as a “numeraire” enabling a wide variety of goods and services to be accounted for in terms of the quantity of one important characteristic, the energy expended in the processes of transforming them into marketable products. From this viewpoint, capital accumulation and stocks of durables are seen as a temporary store for energy resources. In time, these depreciate and wear out, and they eventually end up as waste.

In physical terms, as in economic terms, energy is only of value by virtue of its ability to cause physical or chemical change or to do “work.” From the biophysical systems perspective, what is important is the amount (i.e., the physical cost) of energy used in a process and thereby embodied in a product, not its price in the marketplace. In contrast to the political-economic perception, however, energy analysis follows only the physical processes of production and not the social valuation of the product in the marketplace.

Energy is a means of following the way in which an economy works, usually via techniques of energy analysis. Embodied energy is the energy associated, directly or indirectly, with the production of a good or service. In this context, direct energy is the energy used in the specific process of production; indirect energy is used in the preparation of intermediate

inputs to that process. Although the term embodied energy is widely used, it has many different interpretations, depending on the context.

The use of energy as a complementary means of describing the workings of a social system can also be seen as an attempt to address the apparent inability of standard economics to take full account of the central position of physics in economic production.

3. THE ECONOMICS AND PHYSICS OF PRODUCTION

From a conventional economic perspective, the marketplace is where decisions are made about the uses of energy. Energy is simply a tradable commodity, like all scarce resources, the demand for which is affected by its price. Everything the economy needs to know about energy is incorporated in its price, and the freer and more competitive the market, the more accurate is the price evaluated by it.

Goods are said to be substitutes if (other things being equal) an increase in the price of one leads to a shift toward consumption of another, whose price is lower. Thus, if an increase in the price of coal causes some consumers to switch to natural gas as a heating fuel, these are substitutes. Note, however, that in the generalized scientific sense, energy has no substitutes. Only fuels, which are energy carriers, can often be substituted, one for the other.

Energy consumption has been used for decades as an indicator of the level of wealth of an economy. As a broad generalization, it appears that there has often been a strong relationship between energy consumption and gross domestic product (GDP), with the ratio $E:GDP$ commonly known as the energy intensity of GDP. There are problems with this criterion, however, because some very different countries have approximately similar energy demands, whereas other similar countries have different demands. Nevertheless, greater energy consumption was long viewed as a necessary prerequisite for economic growth.

The relationship between energy consumption and economic activity is significantly less direct than was the case in the past for a number of reasons, including substantial improvements in the technical efficiency of use of energy, especially in technologically advanced countries that previously consumed large quantities of fuel. In some cases, this increase in efficiency was the result of switching to other, higher quality fuels, which enabled a greater amount of GDP to be produced per unit of fuel used.

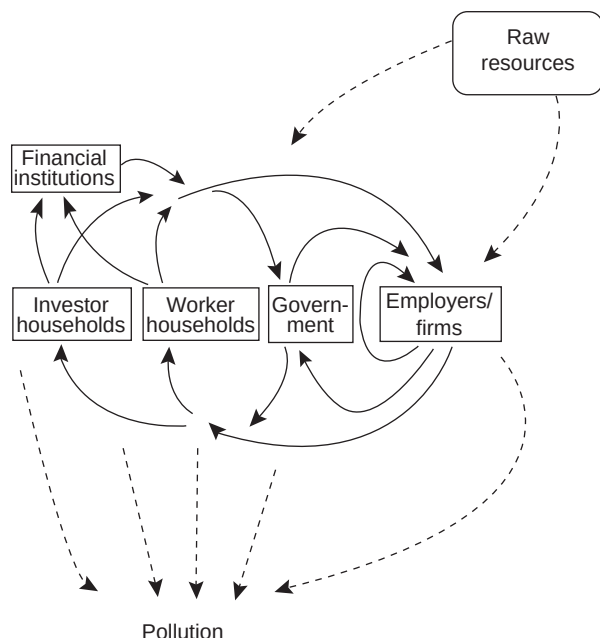


FIGURE 3 Resource requirements of economic activity.

“externalities.” In environmental economics, the solution to environmental problems becomes one of creation of markets in pollution rights and ecosystem services on the principle that damage must be paid for, which acts as a disincentive to pollute. Related tools in resource economics involve application of discounting of the value of future events such as resource depletion in order to determine the most economically efficient use of resources over time.

Therefore, there is a sense in which the mainstream economic model regards the environment as simply a subsystem of the economy—a source of resources and sink for wastes, with the primary purpose being to enable the economy to grow without limit. Figure 4 illustrates Daly’s critique of this perception. The special place of energy in the biophysical worldview is completely absent from this model, other than as just another resource.

All these attempts to incorporate the reality of environmental pollution into social policy development have involved biophysical issues interpreted via economic theories. From a scientific viewpoint, one may observe that in this approach the needs of the total system are interpreted in terms of those of one of its (dependent) subsystems (the economy) for reasons that appear to owe little to science. Nevertheless, there is little doubt that if properly applied, tools such as environmental and resource economics could give rise to more environmentally and socially benign outcomes than current outcomes via the use

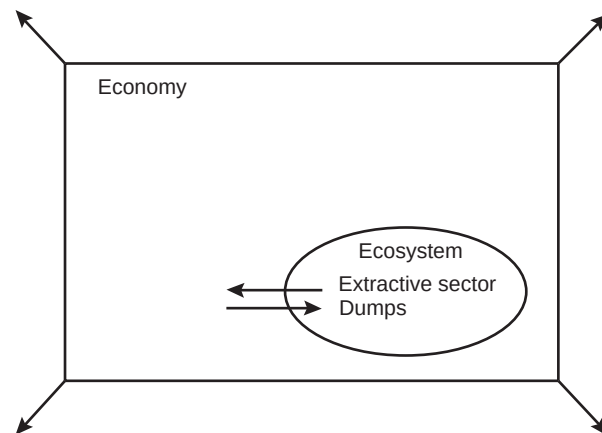


FIGURE 4 Ecosystem as subsystem of the economy. From Daly (1999), with permission.

of improved technologies for converting resources into end-use human services.

3.2 Physically Modified Economic Models of Production

The model in Fig. 3 has limitations due to the particular place energy has in the processes of production and consumption. If we separate out raw energy resources from other material resources, we obtain a different picture, despite the fact that some raw energy resources, such as coal, oil, and gas, may also be raw resources in their own right (e.g., when used for petrochemicals production). This is not a problem as long as we remember the distinction between the two functions.

Figure 5 indicates the special function of the energy sector as an energy transformation system. A flow G of raw energy resources (coal, oil or gas in the ground, water in mountain lakes, geothermal heat, solar input, and so on—often referred to as primary energy) is transformed into a flow E of consumer energy (often referred to as secondary energy)—such as delivered, refined fuels, electricity, process heat, etc.—transformed in coal mines, oil/gas wells, refineries, thermal power stations, hydro dams, boilers, and so on.

These processes always involve loss of some of the input of raw energy resources, as a direct reflection of the processes involved, such as oil drilling, pumping and refining, thermal or other electricity generation, and so on. Some of the consumer energy output of these plants is often needed to run the plants (e.g., electricity to run auxiliary machinery in a thermal power station). Note that flows of waste energy and material pollution are not included in this simplified model but are also important.

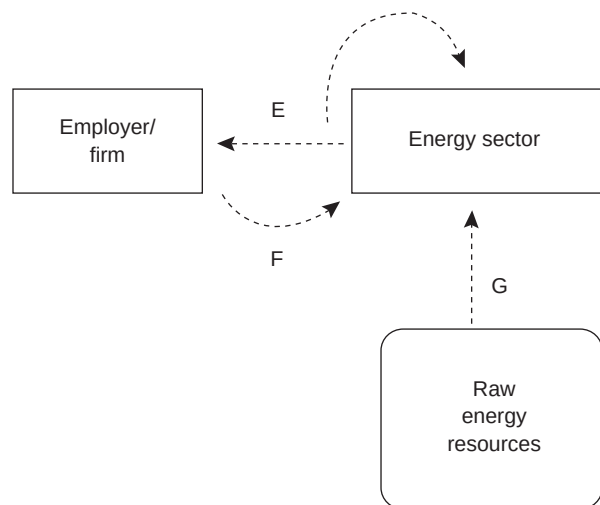


FIGURE 5 Energy transformation system.

The energy transformation system comprises large capital works constructed from steel, concrete, and other materials, in turn created using energy and material resources in the past. Figure 5 includes a feedback arrow showing the energy flow *F* embodied in the goods and services (as manufactured capital, etc.) needed from the production sector to build and maintain the energy sector.

Even consumer, or secondary, energy is not the real end product. What economic production or consumption processes actually need from energy is the services that energy provides, such as heat, light, motive power, chemical conversion, and operation of electronic equipment. This is illustrated in Nørgård's model in Fig. 6, which shows the chain of conversion of raw, primary energy through secondary energy to energy services and to the ultimate end, the energy services and other inputs to lifestyle and satisfaction of needs. In this model, the fact that energy is a means to an end is very clear. Also very clear is the importance of efficiency of conversion at each stage and the fact that each increment of satisfaction of needs requires a substantial infrastructure of capital equipment and a steady consumption of raw resources. Sooner or later, there is an associated and inevitable conversion of all the resources into waste.

Gilliland's model (Fig. 7) integrates the perspective of Fig. 5 into the economic model of Fig. 2 and shows how a significantly richer picture of economic activity is obtained if resources are separated into their material and energy components. This thermophysical model shows the economy as a pathway by which energy resources are used to run the processes

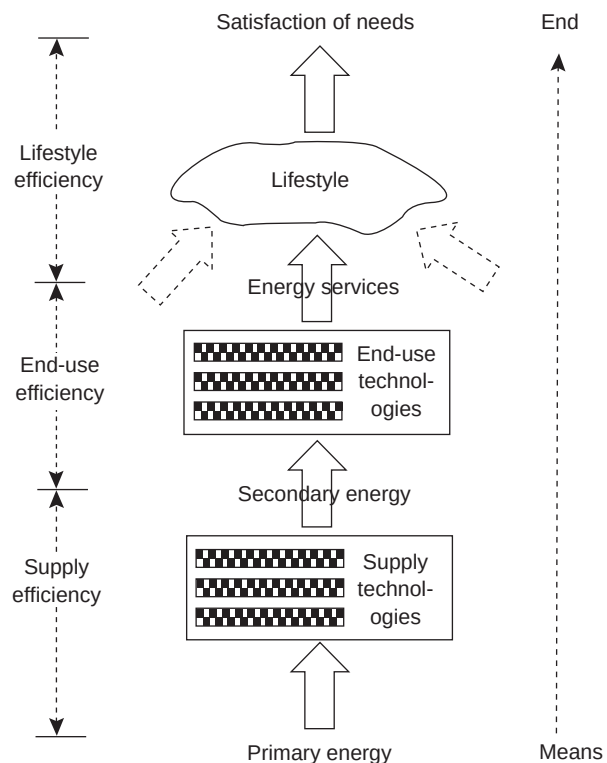


FIGURE 6 Conversion of primary energy to energy services. Courtesy of Nørgård, J. S.

that convert raw materials into the goods and services needed for everyday functioning of the economic system.

In Fig. 7, it is clear that money flows in the opposite direction as energy. In addition, although money flows are circular, energy flows are essentially linear (in one end and out the other), as is typical of all metabolic processes.

In the thermophysical context, energy use—or, more correctly, the degradation of available input energy into unavailable output energy—can be seen as representing physical “cost.” However, this is a different concept from money flow in the economic perspective, which normally represents market price. Therefore, it is not suggested that energy consumption is an adequate measure of social value in an economic context, although undoubtedly it can have significant long-term implications for prices.

Again, however, to make our approach operational, it is desirable to view the place of important component subsystems in our conceptual model of the system of society within the supersystem environment. Figure 8 shows the thermophysical economy of Fig. 7 within its total ecological environment of planet Earth. The energy concept is central to this description of the thermobiophysical “engine” that enables

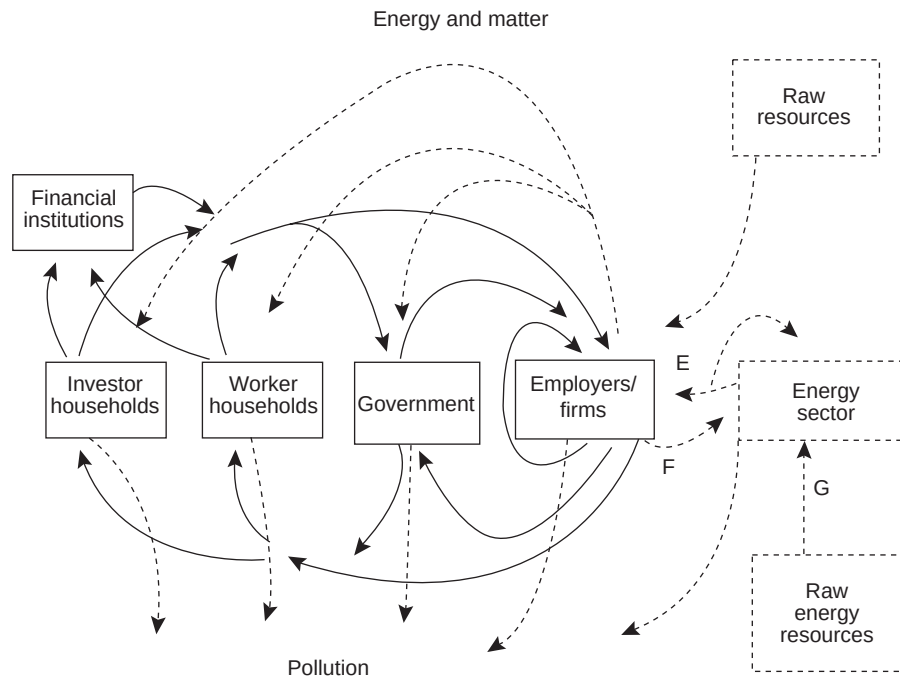


FIGURE 7 The thermophysical macroeconomic model of Gilliland.

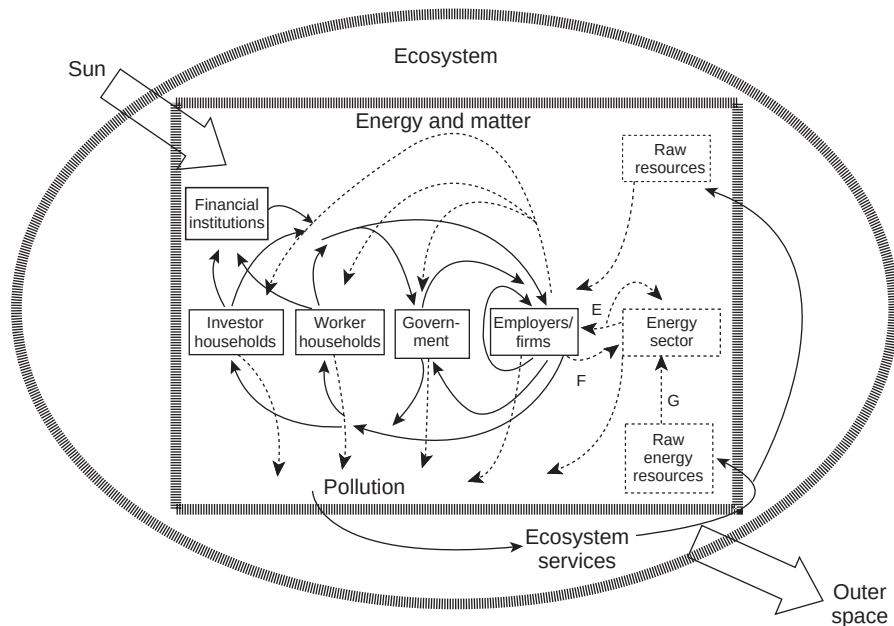


FIGURE 8 Biophysical economic model as part of planet Earth.

the global ecosystem to function, with high-quality incident solar energy driving all the natural processes on which the activities of living things depend, before eventually being rejected as waste heat into the vast sink of outer space. Within the earth's system, that energy is stored for a short time in processes such as

circulation of the atmosphere, for longer times in plants that absorb energy for photosynthesis, and for even longer times in geological deposits.

The energy in fossil fuels was originally derived from solar energy many millions of years ago, captured mainly through photosynthesis in plants

and then via the plants (and the animals that fed on the plants) whose remains were transformed over very long periods of time by geological processes into the hydrocarbon fuels coal, oil, and gas. The fact that humans currently have access to these resources (for a period that will be very short in the context of geological or even human time) does not affect the validity of Fig. 8 as a general model.

4. ENERGY ANALYSIS OF ECONOMIC PROCESSES

An important characteristic of an energy transformation sector is that it functions by absorbing capital, operating, maintenance, and other inputs from the economy, represented in Fig. 5–7 by the feedback flows F of energy embodied in those goods and services. It is only when the flow of useful, consumer energy E is greater than the sum of embodied feedback flows that there is a supply of net energy to the economy. This concept is analogous to one common in economics in that revenue or turnover alone are not enough to indicate financial success of a project; net cash flows or profit are important.

Energy analysis comprises a family of tools that follow economic production and consumption processes through their energy consequences, usually expressed in terms of the first law of thermodynamics (i.e., the heating abilities of energy). More advanced forms of energy analysis, such as exergy analysis, address the ability of energy to do work in terms of the second law of thermodynamics.

In its basic form, energy analysis involves the determination of the energy embodied in the process of delivering consumer energy or end-use energy or in making a good or service. As Fig. 7 indicates, both the losses of raw energy and the energy required to run the energy sector may be substantial. In the case of coal-fired thermal electricity generation, approximately two-thirds of the raw (primary) energy is dissipated in producing the consumer (secondary) electricity output, even before the efficiency of conversion into end-use forms is taken into account.

There are two common techniques for evaluating the embodied energy of goods and services—process analysis and input–output analysis. Process analysis involves the systematic study of the inputs to and outputs from a process, with determination of energy requirements for each input. These are then summed to give the total energy requirement of the product.

For example, a process energy analysis of running a home would first concentrate on electricity and heating fuel consumption since these are the most obvious large energy inputs. Less important direct inputs would then be assessed. Energy for agriculture, harvesting, processing, transport, merchandising, and cooking of food also needs to be evaluated, together with the energy involved in the manufacture of clothing, furniture, cleaning materials, paint, carpet, and so on from their raw materials and the energy to transport and deliver them to shops and the home. If one goes back further, one could consider the energy required to make or process the house construction materials (steel and concrete for foundations, lumber and other materials for the structure, and the energy needed to build the machines and factories where these parts were made) and, before that, the energy required to build and run the mines, forests, and so on from whence came the raw materials, and so on.

Clearly, the task is daunting, and few process analyses have been carried out beyond the first or second stage. This technique is good at giving an accurate estimate of direct energy inputs but not for evaluating the indirect inputs, such as the energy required to process the timber for the building, the lumber mill, or the steel mill or the mines from which the iron ore came to make the steel. These indirect inputs can often be very large, especially in the service-sector industries, which on the surface often appear to have relatively small energy demands but whose indirect demands may be many times higher.

The second common technique is input–output energy analysis, a modification of a standard economic tool. This employs a detailed survey of transactions in an economy. In effect, it addresses the production part of the economy, with explicit accounting for the energy inputs required to keep the economy going. By a straightforward computer-based method, all the energy inputs to the processes that preceded the output of a given good or service in the economy are summed, back to the coal mine, oil well, power station, or gas field.

In general, process analysis is appropriate for study of specific production technologies and is a technique widely used in engineering. Input–output analysis is more appropriate for examining the effects of changes in technology, resource use, etc. for a whole economy. The difference in quality, or availability to do work, of the various forms of energy is not specifically determined by either approach, although it is possible to evaluate it.

Both techniques are involved, as appropriate, in applications such as life-cycle analysis.

Incorporation of the extra information gained from a material/energy analysis inevitably raises problems of commensurability of units because activities within the economy are conventionally represented in monetary units, whereas those in ecological and industrial systems are better represented by measures of material and energy/entropy flows. The emerging field of ecological economics addresses these and related issues.

5. PHYSICAL ACCESSIBILITY AND THE CONCEPT OF NET ENERGY

There are significant differences in the ways different researchers measure the net energy performances of energy transformation systems that deliver a flow of consumer (secondary) energy E to an economy (Fig. 5) from a raw (primary) energy resource flow G using a feedback flow F . However, the net energy criterion (in whatever form) is the means of indicating the physical accessibility of a resource. It is the amount of effort (measured in energy terms) required to extract the resource, process it as required, and deliver it to the consumer. The greater the effort, the lower the accessibility of the resource. As indicated previously, production of electricity from coal via conventional thermal power stations results (in heat equivalent terms) in only approximately one-third of the energy in the mined coal being delivered to consumers in the form of electricity. However, the ability of the electricity to do work, especially running machines and electronics, makes it a more valuable energy source to an economy than the raw coal.

During the past three centuries, economies in industrialized countries have changed from an almost complete dependence on biomass-based fuels (mainly wood) to coal and then to oil and natural gas. Each transition was characterized by change to a more accessible energy source, but that change was not possible until technology developed to the point at which the improved accessibility was achievable. Thus, coal could not be supplied in large quantities until underground mining technology was developed, and open-cut coal required heavy earth-moving machines to be developed. Oil and gas were not accessible until drilling, transportation, and refining technologies were applied. All of these depended heavily on the development of systems for long-distance transportation of both raw materials and products.

In each case, the accessibility of the new resource (measured by both net energy and economic methods) was better than that of the previous one. This meant costs were lower, so substitution could proceed rapidly after the technology developed. This outcome is believed by some to be the result of factors internal to economic systems. To the extent that technological developments were required, this is partially true. The underlying physical accessibility mechanism is a separate issue, however, subject to the laws of physics and should not be excluded from the argument.

There is increasing evidence that energy is becoming more difficult to obtain and process into readily usable forms. Geological resources are not homogeneous, and humankind has exploited the easily accessible, high-quality resources first. Lower quality, less accessible resources are left for future generations in the expectation that the price mechanism will offset any shortages.

As an economy changes its sources over time from high-quality (high-availability) energy sources to low-quality (low-availability) sources, the second law of thermodynamics indicates that the efficiency with which work may be done by these resources will decline. In addition, the change from resources that are easy to mine to those that are deeper and/or more difficult to mine means that a greater amount of economic (e.g., engineering) effort will have to be expended in order to tap these resources and to supply useful net energy to the economy. Thus, the resources become less accessible. In both cases (Figs. 5–7), a greater flow of either G or F (or both) is needed to maintain a constant flow E . Over time, the energy transformation industry will have to grow and absorb a steadily increasing quantity of economic output just to maintain the system.

If there is a steady increase in the resources that need to be expended from the economy (flow F) in order to obtain a given quantity of energy E , then in the long term one can envisage a situation in which no net energy is supplied. In other words, the energy content (embodied in buildings, machinery, maintenance, and operation) of the economic effort needed to extract energy and deliver it could increase to such a level that as much embodied energy flows back into the energy supply industry as is supplied to the economy. If $F = E$, a point of futility has been reached, where much socioeconomic effort is going into a process equivalent (in energy terms) to going around in circles.

When resources are relatively inaccessible, the primary resource flow will be greater than that for

the earlier, high-accessibility case. Thus, the rate of depletion of primary resources will also increase for a given (constant) flow of net energy to the economy. Pollution and other externalities will also be greater due to the increased rejection of unavailable energy.

Of course, technical change can often act to lower energy and other costs, thereby offsetting to a greater or lesser extent the depletion effect. Improved technologies may also have an energy cost, however, especially if they involve heavy capital investment with the associated indirect energy costs of manufacture. Nevertheless, many such technologies show substantial benefits. Although very valuable in the short term, in the longer term the constraints set by the laws of physics must prevail.

Thermodynamics is concerned with the properties of systems, not of things on their own. In energy analysis, the expenditure of energy and time to do work is a linking mechanism between the natural environment system and the human socioeconomic system. This link may enable a unified, more complete analysis of the combined environment-economy system to be carried out.

5.1 Net Energy of Resources of Declining Accessibility

As an example of the type of a situation that could happen, consider a simple coal mine. Assume that the coal seam is vertical, starting at the surface. (This example is exaggerated to make a point, but it provides a good indication of the type of situation that will happen in the long term because coal that is closest to the surface will always be extracted first, leaving deeper seams until later.) Initially, only simple materials and resources are needed to mine the coal and deliver it to trucks or railway cars. As the coal face goes lower, the total work that has to be done to get the coal to the surface increases because of the greater height which it has to be lifted. The greater the height, the heavier and more complex the lifting mechanism has to be and the faster it has to run just to maintain a constant rate of delivery of coal at the surface. At this stage, a substantial proportion of the energy in the coal delivered to the surface will be needed just to manufacture, maintain, and run the machinery needed to get it out. This is quite separate from the effort expended in other activities, such as pumping water to keep the mine dry and the precautions necessary to ensure safety at the coal face, where the working conditions are becoming ever more inhospitable.

As the depth increases, the amount of net energy delivered to the surface declines because increasingly more of the mine's output is needed just to drive the mining machinery. To ensure even a constant supply of net energy E delivered to the economy at the surface, the rate of extraction of coal will have to increase. It is not difficult to appreciate that the mining rate G will accelerate in such circumstances. The faster the miners try to get the coal out and the greater the demands placed on the mine by the economy, the faster the process approaches the stage at which the whole operation is energetically pointless. Clearly, the end result is stalemate in that the mining operation does not actually do anything for the economy other than supply employment. Beyond the point at which the net energy yield is zero, the operation is in fact taking more than it supplies. This will happen irrespective of the amount of coal left in the mine. After a certain stage, therefore, further mining is pointless and the resource is economically and energetically useless.

It is possible to predict from dynamic energy analysis the maximum depth at which the work required to lift the coal exactly equals that available from the energy in the coal lifted. The lower the quality of the energy resource, the sooner this will happen. Peat or lignite will be accessible to shallower depths than bituminous coal or anthracite. The direct and indirect energy requirements of the other parts of the mining activity will ensure that in practice, the depth that is mineable will be very much shallower than that which is calculated from the energy requirements of lifting alone.

From the economic viewpoint, many people support the notion that improvements in technology will ensure that the economy will always get what it wants. Such a view is not supported by the physics. It is quite possible that lifting machinery and water pumps will be improved, but there is an absolute minimum energy requirement for lifting a tonne of coal to the surface and pumping a liter of water that is absolutely unalterable by technology, being a consequence of the second law of thermodynamics. The most that can be hoped for from technology is that it will enable mining to go a little deeper; the constraint remains. It should be mentioned, however, that few if any mines have reached the depth at which the net energy yield is zero; the point of this discussion is that the raw energy requirement of delivered consumer energy will increase, and at an accelerating rate, in the future, not that we will run out of the resource.

5.2 Dynamic Energy Systems Analysis

The example of the coal mine illustrates stages in the evolution of energy supply to an economy. To generalize the analysis, we must acknowledge that in reality the energy system and the economy are changing all the time. As described previously, the inevitability of declining resource quality means that accessibility will steadily decrease. For awhile, improvements in technology can be expected to compensate for this problem or even improve the overall situation. However, the second law of thermodynamics ensures that there are absolute limits to the improvements that are possible in the efficiency of conversion of primary energy into consumer energy and then into end-use energy. It is therefore necessary to examine the way in which the energy supply system and the economy interact over a long period of time in order to understand the nature of net energy effects.

Empirical data confirm that as an energy system expands into resources of decreasing accessibility, the primary energy intensity of economic activity does indeed increase, with the result that in some circumstances the total primary energy input to the economy can be expected to increase very rapidly for only modest increases in production of goods and services. In addition to the increased energy intensity of economic activity that results from such system growth are increases in the cost of substitute forms of energy. In an economy in which energy costs are increasing rapidly in real terms (as is likely in the later stages of life of many major natural gas fields, which is already occurring), it can be expected that the cost of plant to manufacture substitutes (e.g., synthetic natural gas from coal) will also increase.

From our understanding of the behavior of complex systems, we also must acknowledge that the future is indeterminate. Thus, we cannot make a realistic forecast of when a gas field will run out, let alone an estimate of the costs of substitute gas. The only reliable conclusion we can make is that current predictions appear to seriously underestimate the current value of gas and other high-quality scarce fuels and encourage wasteful use.

5.3 The Importance of Growth Rate

A great deal of energy-intensive capital investment goes into the development and expansion of energy sources, and under high-growth rates it is quite conceivable for energy-supply programs to be net consumers of energy. A very important factor here is

the rate of expansion; the lower the rate, the better the net energy performance.

It appears that the rate of growth of consumer energy demand E may be a critical factor in evaluating the net energy obtainable from a system because of the strong positive-feedback energy flows F required for rapid system growth. Although net energy is not the only resource that an energy technology must supply, one can conclude that it is the most important.

It is a basic tenet of neoclassical economics that substitution is always possible, which effectively makes scarcity only relative. From biophysical systems reasoning, however, substitution of a resource by another of lower accessibility is not a situation in which the two resources can be considered separately: They exist in a systemic relationship, in which one affects the other. The increased energy requirements of access to new primary energy resources suggest that the cost of delivered energy could increase at an accelerating rate in the future and that substitution is by no means automatic. The extent to which this argument applies to the transition from fossil fuels to renewable energy sources is still an open question.

5.4 Renewable Energy

Solar inputs are essential to the functioning of any economy, not only directly but also indirectly through their part in “driving” the natural cycles that provide the ecosystem services that are essential for all forms of life, including economic. Solar sources also provide small but rapidly increasing supplies of electricity (photovoltaic and wind) and refined biomass fuels (ethanol and biodiesel) to many economies. Although environmentally more benign than the use of fossil fuels, these still have to obey the thermodynamic imperatives summarized in Figs. 5–7 via energy transformation sectors. As with fossil resources, the accessibility of solar energy is not uniform, either with latitude or with the seasons. However, the fact that annual solar inputs to Earth are many orders of magnitude greater than fossil fuel inputs means that the potential for contributions to economic activity can be expected to increase enormously in the future.

A caveat is important at this point. Investment in solar energy capture for electricity, for example, requires preinvestment from fossil-fueled economic processes. In other words, solar capture technologies (wind machines, photovoltaic cells, etc.) must be manufactured from metals, glass, plastics, ceramics,

and other products of an economy in which these materials are produced predominantly by fossil-fueled industries. The transition from present-day economies to future economies that are entirely solar powered will not be easy and can be expected to take a long time. It is not clear whether present-day patterns of consumption of goods and services can be maintained in an economy that is entirely solar based.

6. ENERGY AND ITS IMPORTANCE TO POLICIES FOR SUSTAINABILITY

In the early 21st century, the issue of sustainability of economic activity has moved to center stage of policy development in many countries. For many, the policy issues involve integration of the environmental, the social, and the economic into a policy-formation nexus. Figure 9 inverts the mainstream economic model of Fig. 4, showing from the biophysical perspective Mitchell's concentric model, in which the economy's sustainability is subordinate to that of the society, whose sustainability is in turn subordinate to that of its surrounding and sustaining environment. Fig. 9 also indicates the place of branches of thermodynamics and economics in addressing the sustainability of these parts.

The fundamental question of purpose arises. In one respect, there is no doubt that by replacing fossil fuel inputs with solar inputs to the economy there would be very substantial environmental advantages. There would also likely be substantial economic advantages (in terms of increased employment and improved health due to reduced localized pollution).

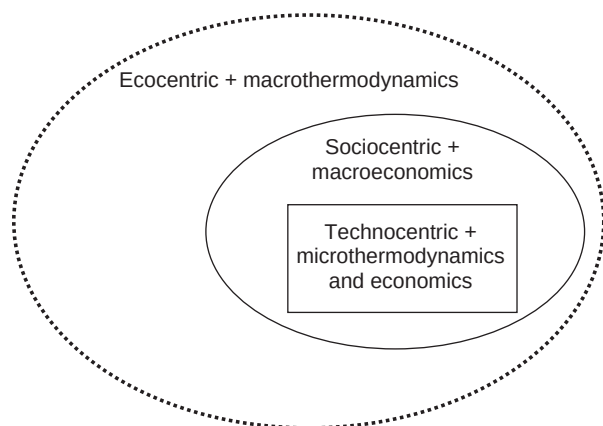


FIGURE 9 Mitchell's concentric model. From Mitchell, C. A. (2000). *Trans. Instit. Chem. Eng.* 78, B237–242, with permission.

However, there remains the question: What is it all for? If one only envisages the replacement of fossil fuels by clean solar fuels, does that mean continuing increases in the use of private cars? Does that mean 3-ton “green” SUVs spending hours in traffic jams? To take the question further, could it mean solar photovoltaic-powered or wind-powered aluminum smelters to produce green aluminum for throwaway beer or soft drink cans?

These examples raise important questions about the relationship between energy and economics. As noted previously and demonstrated in Fig. 6, the ultimate purpose of use of energy resources is to satisfy the needs of people. In other words, the key requirement is to match means with ends.

Substituting solar for fossil fuels simply addresses the requirement of minimizing adverse impacts on the natural system while doing nothing directly to better address people's basic needs. The overall issue is much more complex than the simple substitution of one primary source of energy by another, no matter how apparently benign. The ethical positions of society are inherently relevant to our discussion and need to be applied to decisions about future economic processes.

7. AIMS OF THE BIOPHYSICAL SYSTEMS WORLDVIEW

From the previous discussion, it is clear that the biophysical systems perception has little to contribute to the determination of how people are likely to react in social situations because it is based on and applicable to issues involving the physical world, not the social. With few exceptions, most physical scientists would agree that the perception is applicable to the physical constraints and long-term limits to social activities and not to the day-to-day activities of people within those limits.

The universality of the process of consumption of available energy and its rejection as unavailable energy by physical systems makes energy one logical means of studying the structure and dynamic behavior of real physical systems. It also provides one means of assisting development of a more unified understanding of the behavior of complex physical systems.

The most important attribute of the biophysical systems approach, however, is that it is an entirely different way of thinking—a different worldview. In any real system, the ruling state is metastable, reflecting physical disequilibrium. This understanding is essential to a proper understanding of economic

activity within its physical environment. This concept gives an entirely different viewpoint on development in an economy, and of its impacts on options for future development, than is gained from the standard economic perspective. Together, they help us gain a perspective that is clearer and more powerful than those obtained from either used alone.

If the insights obtained from this perspective alone were to be incorporated into resource economics and the treatment of subjects such as depletion, then the economics of sustainable alternatives such as energy demand management (energy conservation) would be seen to be much more favorable. The approach makes it clear that there are strong indications of the existence of limits to growth of economies, in any sense involving the production and consumption of material goods and services. Although the immediate aims of the biophysical systems perception are modest, the effects of acceptance of the findings would be far-reaching, not least in the field of the use of energy resources in society.

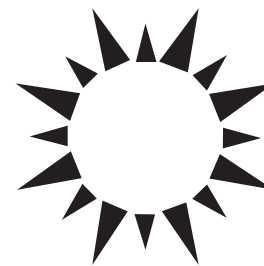
SEE ALSO THE FOLLOWING ARTICLES

Economic Growth and Energy • Economics of Energy Demand • Economics of Energy Supply • Entropy and the Economic Process • Sustainable Development: Basic Concepts and Application to Energy • Thermodynamics and Economics, Overview

Further Reading

- Daly, H. E. (1999). "Ecological Economics and the Ecology of Economics: Essays in Criticism." Elgar, Cheltenham, UK.
- Peet, J. (1992). "Energy and the Ecological Economics of Sustainability." Island Press, Washington, DC.
- Slesser, M. (1982). "Macmillan Dictionary of Energy." Macmillan, London.
- Ulgiati, S. (ed.). (1998). "Proceedings of the International Workshop: Advances in Energy Studies Energy Flows in Ecology and Economy, Porto Venere, Italy, 26/30 May 1998." Museum of Science and Scientific Information, Rome.

Economic Thought, History of Energy in



PAUL CHRISTENSEN

Hofstra University
Hempstead, New York, United States

-
1. Introduction
 2. Early Energy Ideas
 3. Early Physiological Economics: Hobbes, Petty, Boisguilbert
 4. François Quesnay: The Energetic Premises of Physiocracy
 5. Adam Smith: The Work and Powers of Nature and Labor
 6. J.-B. Say: General Theory of the Work of Nature, Labor, and Machines
 7. Malthus and Ricardo: Limits on Powers of Land
 8. Late Classical Theory: Babbage, Senior, and J. S. Mill
 9. How Energy Came Up Missing in Neoclassical Economics

Glossary

- ancient atomism** Natural bodies composed of indivisible units of matter possessing size, shape, weight, and an intrinsic source of motion.
- classical economics** Early modern theories based on the primacy of production processes.
- early idea of energy** The principle of activity, force in action, the “go of things.”
- fire** A very active and subtle substance or element used to explain motion and activity in nature and in organisms, including sensitivity in antiquity and in early modern chemistry and physiology.
- neoclassical theory** The modern exchange approach to economics based on scarcity and marginal utility.
- phlogiston** A volatile component of combustible substances in 18th century chemistry captured from sunlight by plants and released in vital processes, combustion, and reduction.
- work** A late-18th-century measure of the effect produced by machines and labor employing natural powers.

Economic theory and its history conveniently divide in two parts: an “early modern” production theory

focused on how a society uses resources provided by nature to produce physical goods and services and a “modern” exchange theory centered on the study of the efficient allocation of scarce resources to satisfy the unlimited demands of consumers. The first, pre-classical and classical economics (1650–1850), evolved in a close and ongoing relationship with early modern scientific theories of nature’s physical fecundity and organization including mechanics, matter theory, chemistry, physiology, and psychology. Mainstream neoclassical economics (1870 to the present) took its direction from the conceptual and mathematical framework of 19th-century analytical mechanics. Production is treated as an analogue of a spatial (and nondissipative) energy/output gradient, which effectively eliminated any substantive treatment of material and energy resources and physical “converters” from production theory. Neoclassical economics neglects material and energy resources and technologies and is inconsistent with their inclusion. The theory is inherently static and cannot explain physical processes, the dynamics of change in economics, or even how an economic system is coordinated.

1. INTRODUCTION

The production approach to economics that emerges in the mid-17th century took its inspiration from a broad range of sciences: Galilean mechanics, ancient ideas of active matter, Harvey’s new circulation physiology, the proto-ecological natural histories developed by the scientific academies, and psychological and political theories of sentiment and organization with roots in antiquity. For nearly two centuries, economic ideas of physical production and economic behavior drew on the natural sciences,

including physiology and medicine. This early connection between economics and natural philosophy has been little studied by economists but provides an essential key for understanding the evolution and structure of the classical approach.

In the second half of the 19th century, the focus of theoretical inquiry shifted from a production economics to a theory of market exchange. Neoclassical economics was constructed on two related ideas: the universal scarcity of resources and diminishing marginal productivity applied to resources and consumption goods alike. This basic idea, formulated in terms of simple partial derivatives, is that additions of any single input (or good), everything else held constant, yield a stream of positive albeit diminishing additions to output (or psychic benefit). But according to the conservation matter and energy, physical output cannot be increased if material and energy inputs are held constant. The central idea of neoclassical theory, the doctrine of marginal productivity, violates the laws of thermodynamics. Basing economic theory and its regulatory mechanisms on the conceptual structures and mathematics of analytical mechanics and field theory eliminated material and energetic processes (and much else including technology, organization, and psychology) from economic theory. Economics became an autonomous science isolated any real connection to nature or to society.

Our concern is to show how the material and energy ideas of early modern science were used to shape early economic theory and how these ideas were gradually lost. These early connections between science and economics have been hidden for the reason that they were considered heretical and were not generally openly avowed by the scientists and the economic thinkers who took them up to avoid censorship or religious condemnation. Contemporary readers could be assumed to know the coded language of their presentation. Modern investigators face the double difficulty of not knowing the language of the early sciences or the coded messages of radical materialism. They are also trained in economic theories that have severed any connection to nature and therefore lack any conception of the necessary interdependence that exists between material and energy processes in nature and in human economies. The discovery of these connections requires a sustained investigation of the shifting and forgotten labyrinths of early science. Thanks to the breadth of work in the history of science, an archaeology of the material and energetic foundations of early economics is now under way.

The next section surveys the main energy ideas employed in the sciences and economic thought during the first two centuries of early economic theory. Subsequent sections focus on the main economic writers and their very individual connections to the scientific ideas of their time starting with Thomas Hobbes and William Petty in the 17th century. The main developments move back and forth between Britain and France. Classical theory ends in Britain with J. S. Mill in the mid-19th century who moves away from grappling with the fundamental question of the meaning of energy for economic production. A final section briefly sketches the analytical focus that has governed neoclassical theory.

2. *EARLY ENERGY IDEAS*

The idea of energy we use is Wicken's suggestion that energy "is the go of things." We are interested in concepts that provided explanations of physical and vital motions in early modern natural philosophy and the influence of these ideas on the formation of early economic theory. Since energy comes in a material form and is not used for itself but to make or do something, we are also concerned with material transformations and physical technologies that convert energy and materials. These components of production are inextricably linked in the economic process.

Material and energy conversions are subject to physical principles: (1) matter and energy are neither created nor destroyed in ordinary processes; (2) energy is degraded in quality as it is used and becomes unavailable for use (useless energy); and (3) material processes also follow physical and chemical rules. Theories of economic production must be consistent with scientific principles including the basic complementarity that exists between materials, energy, and physical converters. If economic principles are not consistent with the sciences of nature, then economic theory must be reformulated. Economics is not an autonomous science.

The first energetic concept we isolate is the material fire, which provided the active principle of matter theory in the natural philosophies of antiquity and early modern era. Fire was a very subtle and active matter used to explain the visible fire of combustion (by a disassociation of the matter of combustible bodies by the penetration of fire atoms), the invisible flame of the vital heat in animal bodies, and sensitivity in plants and animals.

The mechanist philosophies that emerged in the wake of the victory of the new science of mechanics of Kepler and Galileo are generally assumed to have regarded matter as entirely passive. But nature required sources of motion and ideas of active matter remained central in many 17th-century philosophies. The revival of Stoic, Epicurean, and Democritean theories of matter, the close connection between a mechanical explanation of nature and ideas of active matter in antiquity, the Renaissance Aristotelianism taught in the Italian medical schools, and the growing influence of an alchemy joined to atomism all promoted ideas of active matter. Fire, the active principle of matter theory, was a basic doctrine in these theories.

Attempts to construct a unified mechanical philosophy of nature reopened the conflict between doctrines of active matter and religion. Nature, in church doctrine, was passive and God the sole source of motion in the universe. After Galileo, matter theory proved more problematic for natural philosophers than Copernicanism. Descartes canceled plans to publish *Le Monde* with its naturalistic account of the universe and moved to ground his natural philosophy on a metaphysical argument that all motion was due to God. Despite a prohibition against teaching atomism, Gassendi, a Catholic priest and astronomer protected by powerful friends, proposed a version of atomism consistent with the creation. God, according to Gassendi, could have made the world in any number of ways but chose to make it from atoms endowed with an enduring source of motion. His union of atomism with 17th-century atomistic chemistry and with Hippocratic biology provided an attractive package for the development of chemical and physiological ideas and was quickly adopted by physicians in France and by Harvey's disciples in England. His followers had to express their ideas clandestinely. Hobbes developed a entirely naturalistic account of motion but had to parry accusations of antagonists that he held a theory of self-moving matter. Newton fought materialism all his life and denied material powers to "brute matter." But by treating gravity and other active principles as powers of bodies and speculating on material ethers and vegetative powers, he fostered the materialism of the Enlightenment.

Thanks perhaps to the flux in religious regimes, England was fertile soil for energetic and atomistic ideas of matter as is evident in the work of Thomas Harriot and Walter Warner in the late 1500s and after, Bacon's atomistic speculations in the 1620s,

and the Cavendish circle at Welbeck in the 1630s (which included Hobbes) who knew Warner's work. Harvey's *De generatione*, which circulated for two decades before publication in 1651, was another source of energetic ideas. His physician disciples in Oxford and London were quick to take up the "Democritean and Epicurean system" associated with Hobbes and Gassendi. The latter had a considerable influence thanks to the flexibility of his system and close connection to chemistry.

The introduction of Gassendi's atomism in England has been credited to Walter Charleton's translations of Gassendi's early works on Epicurus in the 1650s. But the young physician and future economist William Petty provides an earlier source. Petty met and became friends with Gassendi during the mid-1640s in Paris when he served as Hobbes's scribe and tutor in anatomy. The two also took a course in chemistry. We can assume he introduced the ideas of Hobbes and Gassendi to his colleagues, including Robert Boyle, in Hartlib's circle in London, where he learned the chemistry of Cornelius Drebbel; in Oxford in 1649 where he led research meetings in his lodgings, and to his cousin, Nathaniel Highmore, who became a Gassendian.

Gassendi's synthesis of atomistic chemistry and Hippocratic physiology provided the basic framework that Harvey's physician disciples employed for four decades to develop the new circulation physiology. They conducted a remarkable series of experiments investigating combustion and the nutrition and respiration of plants and animals using nitre, sulfur, and other chemically active substances that brought them close to modern ideas about combustion and respiration. They also used ideas of fire as the active element of the nutritive, animate, and intellectual souls of living things, concepts Descartes had supposedly banned from biology.

The union of Harvey's circulation physiology and atomistic chemistry was quickly embraced by the scientists of the newly organized Academy of Sciences in Paris who initiated anatomical and chemical studies of plant and animals that were sustained, thanks to Royal patronage, over many decades. Early papers extended the circulation model to the preparation and movement of sap in plants. Long the neglected orphans of science, plants were now recognized as occupying the critical position in the physical and chemical circuits linking living organisms to the soil, air, and sunlight. They were the assemblers of the materials and energetic substances feeding and maintaining the other organisms. Cycles of production and consumption bound

all the parts of nature together in conditions of “mutual need.” Harvey’s circulation physiology extended to larger systems of circulating nutrients and materials provided a new vision of the unity and interdependence of nature and an obvious model for the human economy.

A second and closely related locus of energy ideas in 18th-century scientific and economic thought was the phlogiston chemistry formulated by the German chemist and physician, Georg Stahl, at the beginning of the century. Phlogiston, or “matter of fire,” was an imponderable (unweighable) substance that joined other substances in chemical reactions (synthesis) and was released in chemical decompositions (analysis). Combustible substances such as plants, alcohol, and coal were assumed to have high concentrations of phlogiston. One of Stahl’s key contributions was to emphasize the importance of the principle of conservation of matter in the methods of analysis and synthesis of compounds that became a basic tool of chemical research thereafter. Together with the “elective affinity” that different elements of matter revealed for undergoing reactions with other elements, phlogiston chemistry provided the main framework for chemical theorizing until it was supplanted by Lavoisier’s oxygen theory. As an early version of potential energy, it continued to provide an energetic explanation of nature’s operations into the next century.

Phlogiston chemistry was taught in the medical schools in Montpellier and Edinburgh and employed in conjunction with physiological ideas of sensitive matter. G.-F. Rouelle taught a version of phlogiston chemistry at the Royal Garden under Buffon. His students, some of whom wrote the chemistry articles in Diderot’s *Encyclopédie*, rejected mechanistic chemistry in favor of ideas of energetic and sensitive matter. This vision of sensitive molecules informed the theories of psychological sensibility (sentiment) advanced by the scientists of the radical Enlightenment. These ideas informed François Quesnay’s *Essay on the Animal Oeconomy* (1847) and were implicit in his doctrine of economic productivity. Adam Smith in turn praised the “philosophical chemistry” of his friends for the theoretical connections it made between the processes of the art of the furnace and the “vital flame” in animals. His associates justified their physiological and medical theories of sensitive substances by referencing the psychological theories of Hume and Smith. Phlogiston chemistry and its ideas of energetic substance were also central to the 18th-

century scientists who influenced Thomas Malthus and David Ricardo.

The third basic energy idea to shape economic thinking in the classical period was the new understanding and definition of work done in harnessing and using natural powers. The British engineer, William Smeaton, made the critical contribution in a study of the efficiency of different types of water wheels in converting a given head of waterpower into work. His study, published in 1759, established the superior efficiency of overshot wheels and enabled an increase in effective water power capacity in Britain by a third in a period when water power was critical to economic growth. He defined work as the capacity to lift a given weight a given distance in a given amount of time and made comparisons of the “work done” by horses, humans, and wind, water, and steam engines. Adam Smith adopted his definition of work but unfortunately confined its consideration to human labor, which he made the center of this theory.

The new mechanical theory of work was taken up by the French engineers Charles Coloumb and Lazare Carnot. Coulomb’s thrust was to assimilate machines to the work done by humans and to make machines “an economic object, a producer of work.” Carnot’s great study, published in 1783, was concerned with a theoretical analysis of machines as devices for transmitting motion from one body to another. Carnot was concerned with general principles irrespective of the source of power. He emphasized the notion of power over that of force and machines as agents for the transmission of power. His work received little notice until he was promoted to the top leadership in the new Republic and celebrated for his reorganization of the military as “the organizer of the victory.” His contributions to the theory of machines were taken up by J.-B. Say who recognized Smith’s neglect of the central role of machines employing natural powers in industrial productivity. Say’s construction of a general theory of production based on materials, forces of nature, and machines to multiply labor will receive important elaborations by the mathematician and engineer Charles Babbage, the economist Nassau Senior, and the physicist William Whewell. Their contributions do not, however, receive any sustained attention by subsequent economists. As happens to many good ideas, they were passed over in silence by the leading theorists of the time who followed J. S. Mill’s shift of focus from a physical theory of production to economizing, which brought the classical epoch to an end.

3. EARLY PHYSIOLOGICAL ECONOMICS: HOBBS, PETTY, BOISGUILBERT

Despite his considerable contributions to the psychological and political framework of early economics, Thomas Hobbes is considered to have made little or no contribution to the core ideas of economic theory. He is also widely seen as a mechanistic philosopher who reduced everything to Galileo's new mechanics. Most historians of philosophy and economics have not understood the central importance physiological ideas have in his system, his hierarchical conception of the sciences, and the use he makes of Harveyian physiology to extend the materials provisioning approach Aristotle initiated in his *Politics* but failed to complete. Hobbes in fact makes the first sketch the core ideas that will characterize preclassical and classical theory thereafter.

Like Gassendi, Hobbes constructs a conception of nature and society that proceeds from a basic energetic physics of matter in motion and avoids reductionism by treating each level of activity according to what he took to be the organizing principles and evidence for the topic at hand. In addition to the atomism of Democritus and Galileo, he knew and used Aristotle's biological and psychological ideas; Harvey's theories of generation, organic sensibility, and circulation; and the economic ideas of Aristotle's *Politics* and the pseudo-Aristotelian *Oeconomia*. This physics and physiology informs the production approach to economics he sets out under the heading of "The nutrition and procreation of commonwealths" in *Leviathan* in 1651. An economy, like living bodies, obtains the materials and "nutriments" it needs to produce a commodious existence from the land and sea by human industry. These material resources are passed through various stages of concoction (heating) and preparation and circulated via the conduits of trade to all parts of the body politic. His chapter also covers the institutions of property, contracts, and so on that were part of the evolutionary sociopolitical framework initiated by Democritus and his successors.

The energetic thrust of Hobbes's ideas is subtle but clear enough. He knew the atomism and active matter theories of Harriot, Warner, Bacon, and Gassendi but adopts the more dynamical approach of Galileo. He had met Galileo in 1635 and reports carrying away a vision of a world filled with motion. Every seemingly solid body was alive with the internal motions (agitations) of its smallest parts. He defines each specie of body by the imperceptible

motions of its smallest particles. These motions constituted the specific capacities of each body for the initiating motions (endeavors) triggered by interactions with other bodies (the explosion of gunpowder by a spark, etc.). Hobbes's idea of endeavor as an imperceptible but actual motion is recognized as an influence on the very different ideas of force developed by Leibniz and Newton.

The basic source of motion in nature was the undulating wave-like motions of sunlight. Hobbes did not attempt any explanation of the connection between sunlight, plants, and the food consumed by animals. In his *De homine* (of man) of 1658, he attributes the motions of the heart and muscles to motions of "the invisible atoms ... of a salt or niter" taken into blood via respiration. This idea, long popular among Harvey's and Hobbes's alchemically minded friends, was restated in 1654 by Ralph Bathurst, a supporter of Hobbes and associate of Petty at Oxford.

Hobbes's sketch was fleshed out by the brilliant William Petty (who may have contributed to their formulation in Paris). Petty had studied navigation, geography, physics, mathematics, and medicine in France and Holland, and he was familiar with the work of Hobbes and Gassendi and the scientific ideas of the Puritan reformers in the Hartlib's circle in London. He returned to take his medical degree at Oxford in 1650 where he organized research projects and quickly attained high positions in the university. The siren of political preferment and reform called him to Ireland. Briefly chief physician for the army, he reorganized and completed the Down Survey of Irish lands in breathtaking speed and became one of Ireland's largest landholders. Spending most of his time defending and managing his estates and seeking support for his political and economic projects, his scientific interests shifted towards the socioeconomic topics he called "political arithmetic." His Baconian-inspired works include a succession of acute but scattered theoretical insights central to classical economics, which were never drawn together in a single work.

His first contribution was the elaboration, in a *Treatise on Taxes* in 1662, of the idea of a surplus that the ancient atomists made the basis of social and economic evolution. Petty takes up the surplus in order to explain the source of taxable revenue. The surplus is the difference between the total food produced on the land and the portion that must be reserved to replace the inputs (food, fodder, seed, and equipment) used in production. It is effectively the energy available to support and feed the rest of

society, including the artisan and commercial sectors. His proposal was to tax the revenue going to the wealthiest land owners and direct it to projects increasing access to markets, land productivity, and the skills and habits of unemployed workers.

He sketches the implications of the surplus for a theory of competitive profits across sectors, which he terms “neat rents” (not the differential rents earned on more productive or favorably located soils). A “neat” or net rent is the ratio of net earnings from an activity to the cost of carrying it out. He applies this idea to agriculture and extends it to mining and casting coin. Higher returns in one sector, he suggests, lead to a flow of labor and capital goods to that sector, equalizing the rate of return between activities. In addition to this idea of an equal rate of profit (our term), he also developed a concept of differential rent for favorably located lands.

Energy ideas inform a production approach to prices in his *Political Arithmetic*. These are cost-based prices set by the food, materials, and tools needed plus profit. He reduces these in turn to labor, land, and food, given that humans and beasts must “eat, so as to live, labor and generate.” The ultimate determinate of prices is thus the “energy” required, directly and indirectly in production. He distinguishes between short-run prices (reflecting day to day movements of market forces) and the “intrinsic” prices determined by production costs. This distinction becomes a basic feature in classical economics via Richard Cantillon and Adam Smith.

The next significant development of a physiological economics was made by Pierre Boisguilbert, the Lieutenant-General of Normandy (active 1695–1714). Boisguilbert was obviously aware of Hobbes’s *Leviathan* and Petty’s theory of the surplus, which he made the basis of the emergence of the professions in modern society. A deep study of agriculture and its commerce inform his writings. He was also a close student of the new sciences of nature of the Academy of Sciences in Paris and the work of his younger cousin, Bernard Fontenelle, who was establishing himself as a leading interpreter of the new science. Neither he nor his more radical cousin were Cartesians. Boisguilbert’s commitments were to the Stoicism of his uncle (Pierre Corneille) and the Epicureanism of Gassendi and François Bernier. The latter was well known for his travels in the East and critique of Oriental Despotism—the aggrandizement of resources by the few and the absence of property rights and basic civil protections, which the Epicureans regarded as critical incentives for social intercourse, production, and prosperity. Bernier

implies a critique of France’s aristocratic economy, which Boisguilbert makes explicit.

The work of the Royal Academy of Science provided a new conception of nature linked together in a circuit of chemical exchanges. Plants were the prime producers of the energetic substances maintaining all organisms in “a state of reciprocal and mutual need.” Boisguilbert develops similar ideas in economics. Agriculture produces the “basic” goods (grain, wine, and plant and animal tissues) that, like the chemical elements, produce the rest. The fecundity of nature supplies the means that give birth to the 200 or so professions of the state.

The academy’s model of the circulation of sap in plants provided Boisguilbert (as Harvey’s theory did for Hobbes) a model for economic production and circulation. The plant model makes a clear separation of production sites. Primary materials assembled and processed at one site are perfected at others and move in distinct “circuits.” Boisguilbert likewise sees goods produced on the land and in artisan shops moving along distinct circuits between farms and towns. He extends a distinction between the sites of agricultural and manufacturing production to recognize and explain the very different price and quantity behavior between the two sectors.

He recognizes that the strict requirements that nutritional needs impose for maintaining the production and flow of food to consumers requires in turn the reciprocal needs of artisans for markets for their goods and services. Any break in this mutual circulation threatens a widening disequilibrium endangering the health of the whole society. Given the ferocious pursuit of self-interest by traders, equilibrium is maintained on the edge of a sword where a strong movement of supplies or fall in demand can become self-feeding. Such instability is magnified by the sectoral differences in conditions of production and market behavior. The price instability of agricultural markets has disastrous effects for maintaining productive investments and fertility of the soil. In urban industries, workers resist cuts in pay and employers respond to falling demand by laying off workers. The result exacerbates the crisis. His acute analysis captures an asymmetry in market behavior that continues to perplex modern equilibrium theorists.

Prices obviously play an important role in maintaining economic equilibrium but the analysis goes beyond the sphere of exchange to assume that nature’s conserving operations extend to the economy. Nature has the capacity to sense injury to the economy and the power to regulate and to heal

itself. His ideas reflect the Stoic vision of natural order and harmony that Gassendi and Bernier extend from animate nature to the universe and the Hippocratic medicine advocated by Gassendian physicians. Good policy can maintain health but when illnesses arrive they must follow a natural course. Nature not the physician heals, thus, “let nature alone.” These properties of self-regulation are the laws, God’s invisible hand, that maintain order in the world.

Boisguilbert’s psychological and social ideas also reflect Gassendi’s reinterpretation of Epicurean psychological and ethics. Living organisms are designed with natural capacities for sensing what is good for them and what to avoid. Pleasure and pain are part of God’s design to direct the activity of nature’s creatures to life, health, and reproduction. This naturalist psychology extends to the direction and regulation of human society where the desire for the goods of the world, property, and security induce individuals to establish collective agreements to respect and protect the property and rights of others. The ruler is bound in turn to protect property and individual rights and to advance the common good, which includes protecting the weak from the depredations of the strong and wicked. Boisguilbert’s use of this clandestine current of ideas links him to a lineage of economic thinkers going back to Hobbes and forward to Smith.

4. FRANÇOIS QUESNAY: THE ENERGETIC PREMISES OF PHYSIOCRACY

Before Quesnay became an economist, he was a surgeon and physician whose *Physical Essay on the Animal Economy* of 1747 was dedicated to establishing the physical and chemical foundation of the nutrition, generation, and sensitivity of living organisms. For Quesnay, this was the ancient physics revived by Gassendi, the energetic chemistry taught by Rouelle, and an Epicurean psychology extended by the physiological and psychological theories of sentiment of the Montpellier physicians—the chemical and psychological doctrines of Diderot’s *Encyclopedia*. Quesnay presents this radical energetic materialism in the form of a critique of ancient ideas, a skeptical strategy designed to avoid censorship (which has also misled historians of economics).

Thanks to his skills as a surgeon and systematic thinker, Quesnay rose from lowly origins to become

physician to Madam de Pompadour at the Court of Versailles. The grant of a rural estate by the king turned his attention to the problems of French agriculture. His articles in Diderot’s *Encyclopedia* on farming in the 1650s and his famous *Economic Table* at the end of the decade include his infamous economic doctrine of the exclusive productivity of agriculture and corresponding sterility of commerce and manufacturing. Developed within the context of an agrarian economy, his argument was that only agriculture is regenerative in the sense of replacing what is consumed. Artisan industry (assuming all inputs come from the land) only transforms and consumes things that have been produced by the farmers. These ideas can be seen to depend on the energetic theory of his physiology that plants are the only producers in nature and the rest of nature (and humans) only exist by consuming what plants have made.

According to the 1747 *Essay*, the cycle of nature’s production begins “in the womb of the earth” where elements from the atmosphere and soil are chemically combined in simple compounds and taken up by the plants where they are joined to the “very subtle and active matter” of fire provided by the sun. These organic compounds provide the active energetic substance of fire circulated throughout nature by the animals that eat the plants, the animals that eat animals, and so on. The materials are returned to the earth (recycled), but the subtle energetic matter is lost in the decaying compounds and has to be replaced by a new cycle of primary production.

Two basic criticisms have been leveled at Quesnay. The first is that he mistakenly assumed that only agriculture was capable of producing a surplus (and profit) and that he neglected the obvious importance of manufacturing and growing industrial productivity to human welfare. The criticism is valid. The second criticism is that he failed to apply the conservation of matter to agriculture in the same way he does in artisan manufacturing. Quesnay, however, would agree that production in both sectors obeys the conservation of matter. No output, let alone a surplus of grain, can be produced without the appropriate flow of material inputs from the soil and air. His point is rather that only plants can produce the vital energetic matter. His argument neglects the energy resources and inorganic materials used in industry that do not come from agricultural land. But within the assumption of a renewable resource economy limited to the use of plant biomass for materials and energy, his theory of the exclusive productivity of agriculture holds.

5. ADAM SMITH: THE WORK AND POWERS OF NATURE AND LABOR

Adam Smith is popularly regarded as the founder of classical economics, a judgment that owes more to the accessibility of his vision of a self-regulating market economy operating within an institutional framework offering something for everyone than to any fundamental contribution to theory. His *laissez-faire* vision excites some, his production orientation others. The heart of his production theory is the proposition that increases in the division of labor are the prime driver of productivity growth and the wealth of nations. He employs the new engineering treatment of work and natural powers to set out his theory. What is astounding, however, is the absence of any mention of the work done by natural powers other than labor (and land). This lacunae will have far reaching consequences for the treatment of energy and technology by later writers.

Smith wrote impressive histories of astronomy and ancient physics and reviews of Buffon, Linnaeus, Diderot, Rousseau, and others designed to introduce the ideas of the Enlightenment to his countrymen. He knew Gassendi's work and was good friends with Quesnay. His intimate circle included the philosopher David Hume; the chemist-physicians William Cullen, Joseph Black, and James Hutton (famous as a geologist); and the physicist John Playfair. Smith's early writings and *Theory of Moral Sentiments*, published in 1756, suggest theories of nature's design and operations, organic sensibility, and social evolution at the center of Enlightenment thought.

Smith's references to the system of the world as a great self-regulating machine have been taken to indicate an adherence to Newtonian orthodoxy. This inference is problematic. Smith, as did Hume, placed himself under Newton's mantle, but his ideas of nature's design and principles of operation are at variance with the theistic, antimaterialistic, and neoPlatonic metaphysics of immaterial principles connected with Newtonian "orthodoxy." In his "History of Astronomy," Smith interpreted gravity and inertia as balancing "material principles" that ensure a self-regulating universe. In his "Essay on Ancient Physics," he favorably compares the Stoic system of the world which used the idea of fire as the vital principle forming the world and living things to unify the design and the operation of the world, to the theism of Plato (and Newton), which artificially separated the principles of intelligence that formed and animated the world from those carrying out the ongoing operations of nature.

He also praises the "philosophical chemistry" of his associates for the "connecting principles" it provided for explaining the operation of fire in furnaces and the "invisible fire" or "vital principle" of plants and animals. He noted with regret how little this science was known outside the chemists who practice the "art of the furnace." Indeed, his idea of fire and what he calls the Stoical system was straight out of Bernier, which makes Gassendi a central source for the vision of nature's capacities of operation and self-regulation he both lauds and applies in his own work. Like Hutton and Playfair, who saw an eternal and self-regulating universe operating by natural laws, Smith appears closer to the deism of the Enlightenment than to Newton's theism.

The chemical ideas of fire and phlogiston continue to inform Smith's idea of the productivity of land in the *Wealth of Nations*. In this regard, he remains close to Quesnay. Agriculture is the source of the energetic matter feeding the human and animal populations. But he employs the new terminology of work and natural powers to displace Quesnay's view that "land is the unique source of wealth" and shift the emphasis of production theory to labor and the factors contributing to labor productivity as the source of wealth. The productivity differences characterizing nations and regions are not related to differences in nature but to the effects of the division of labor, which augments the efficiency of workers. Labor (work) is the primary source of growing wealth, which he contrasts to the "niggardliness of nature."

Smeaton's reports on the conversion of mechanical power by hydraulic and animate engines and definition of work as a universal measure of mechanical effect and value are the obvious source for Smith's ideas. He likely knew these ideas directly (the volumes are in his library) and discussed them with his friend James Watt in the 1760s when they were colleagues at the University of Glasgow. We know that Watt discussed his innovation of a separate condenser for the steam engine with Smith. Smeaton provides the engineer's concern with efficiency in the use of a given quantity of work capacity, the idea of work as a universal measure of effect and value in all times and places, and the notion of work as equivalent to effort or pain exerted over a defined period. All these ideas are prominent in Smith.

While this framework is pregnant with possibilities for a new foundation for production, Smith makes no mention of the work of machines in harnessing nature's resources and powers in his

economics. The new ideas of work are confined to labor and the related ideas of natural powers are confined to land and human labor (and working animals). The efficiency concerns of Smeaton's analysis picked up, but there is no discussion of the work done in the economy by natural powers of wind, falling water, or the immense sources of heat from coal, which had replaced wood in the industrial processes powering Britain's industrial growth.

Smith notes in passing the great investments made in "furnaces, forges, and slit mills." He would have known the great Carron ironworks near Glasgow—the first integrated iron works in Britain to employ coke in all its operations, whose blowing engines for the blast furnaces and water wheels and steam pumping engines supplying the mills with a continuous source of motive power were designed by Smeaton. He should also have known that nails and pins were made by power-driven machinery in rolling, slitting, and wire-drawing mills. Yet he illustrates the principle of the division of labor using the example of an artisan pin factory unassisted by machinery.

Modern economists tend to see Smith's division of labor and increases in the efficiency of labor in combination with increasing market size as an ever enlarging spiral of growth. It could not have been in Smith's theory since the original source of natural "powers" converted by labor in his theory continued to be, as it was in Quesnay, the energy fixed by plants. In Book II of the *Wealth of Nations*, he declares agriculture to be more productive than manufacturing because it employs two powers (nature and labor) in contrast to manufacturing which only employs one (labor). But the energy supplied by plants must be divided between workers in agriculture, manufacturing, transport, and a multitude of other uses and remains the fundamental limit on the economy. Hence his emphasis on the division of labor and productive employment as opposed to unproductive servants. His lack of clarity about the role of energy in agriculture and his neglect of industrial sources of power are major reasons for the neglect of energy in economics.

6. J.-B. SAY: GENERAL THEORY OF THE WORK OF NATURE, LABOR, AND MACHINES

The first general application of the new ideas of work in production theory was made by the French

economist Jean-Baptiste Say in 1803. Despite strong literary interests, his father sent him to a commercial school in England and then apprenticed him to a powerful French banker. Say left business to work for a newspaper run by Mirabeau (Quesnay's close associate), which led to his appointment as an editor for an important revolutionary journal that included a coverage of science. Aided by his brother, an engineer, he attempted to acquaint himself with "all of science." He knew Lavoisier's chemistry and the contributions of Lazare Carnot and Charles Coulomb to the new science of work. Reading Smith he was immediately critical of Smith's emphasis on labor and neglect of nature and machines. The greatest development of productive forces, he argued, came not from the division of labor but from machinery using the forces of nature. Smith did not understand the true theory of machines in producing wealth.

Say was even more critical of the physiocrats who failed to see that all production, whether it was the production of wheat in a field or the making of a watch in a workshop, did not create new matter but only rearranged or assembled existing matter. All production was subject to the conservation of matter and required the forces of nature. Quesnay, we have seen, would not have disagreed but was making an argument that only agriculture renewed the energetic properties of materials required for life. Say's contribution was to see that human production was exploiting other sources of power by inventing new machines and that production theory must be concerned with the fundamental contributions of materials and forces of nature in human production. His ideas were clearly shaped by Lavoisier's emphasis on the importance of the conservation of matter in chemical analysis and the energetic ideas employed by French engineers and scientists. He knew the energy mechanics of Lazare Carnot (a revolutionary hero) whose work was reviewed in Say's journal, Charles Coulomb's comparisons of the work of machines, animals, and labor, and most likely Lavoisier's ideas of energy conversions by animals and machines.

Say defined the work of humans and the work of nature as "the continuous action which is exerted for executing an operation of industry." He argues that all material production whether in the field or the factory requires the same basic categories of elements (i.e., materials, forces of nature, and actions of labor, and machines). He compares land to a machine but knows that labor and machines are powerless to produce on their own without materials and a source

of power. He understands the importance of production in economics and insists that economic theory must begin with production before taking up the topics of value, distribution, and consumption.

Say provided the first general physical theory of production based on the new theory of work and mechanical energy. Unfortunately, his achievement was obscured by a diffuse presentation (especially in the first edition of the *Treatise*) and his failure to develop the implications of his new approach for a production-based theory of prices. He incorrectly argues that production creates nothing new, only utilities. While production does not create new matter, a watch is more than the gross weight of its parts. It is a new physical entity, and its conditions of production as well as its utility are important for how it is priced or valued. Say's attempt to assimilate his theory to the French scarcity and utility approach to value ignored the interdependence between production and value and contributed to the misunderstanding of his important contribution to classical production theory by the English classicals who failed to take up his production approach and its specific concerns with materials, energy, and machines.

7. MALTHUS AND RICARDO: LIMITS ON POWERS OF LAND

Like Smith, Robert Malthus has been seen as a disciple of Newton in his scientific, economic, and theological ideas. Malthus took his degree in mathematics at Cambridge, which was basically the rational mechanics of Newton and he made good use of his training in formulating regulatory mechanisms for population and price theory. But the natural philosophy and theology he outlined in the first *Essay on Population* in 1798 is neither orthodox nor Newtonian. His ideas of an ongoing creation, the evolution of mind from "a speck of matter," and the forging of mental and moral qualities by the struggle to overcome evil have their sources in the associationist psychology and theology of David Hartley; the scientific and theological ideas of Joseph Priestley, Hartley's leading disciple; and the evolutionary physiology of the physician and biologist, Erasmus Darwin, grandfather of Charles Darwin.

Out of this heady brew, Malthus constructs a decidedly heterodox philosophy and theology which sees nature, mind, and human progress as products of a long struggle against adversity. He makes an important break with ideas of natural harmony by drawing out the implications of Priestley's theologi-

cal vision of God's imperfect but evolving creation and recognition of limited resources (even God faced limits in what he could do). Given the inherent drive to reproduce and the scarcity of resources, the capacities of the population of each species to outrun its food supply must eventually produce a struggle for existence in nature. Malthus recasts Priestley's progressive theology of struggle and growth into a theodicy of scarcity refracted through the mathematics of exponential growth and the focusing lens of regulatory mechanics. His ideas also reflect the absolute limits of an energetic perspective. Priestley was a phlogiston chemist and co-discoverer of photosynthesis. In a period when the nutrient limits of soils were not understood, the primary limit on agricultural production appeared to be the ability of plants to fix energetic matter as opposed to a shortage of particular (and augmentable) nutrients.

Malthus's preoccupation with the availability of food was reinforced by a theological vision of nature's design. Struggle against limits was necessary for progress. A preoccupation with food and limits, reflecting energy theory and theology, helps explain why Malthus's contributions to areas of production theory outside agriculture were so limited. Despite his acute observations on the central differences between returns and costs in industry (where techniques are replicated and constantly improved) compared to agriculture and an understanding of the energetic basis of production, he does not develop a general theory of the material and energetic transformations specific to industrial production. He certainly knows Say's production theory. It would appear that the idea of constraints and limits, which were rooted in theology and energy theory, carry implications for limits on human progress that explain his reluctance to engage Say's more general theoretical framework with its quite different vision of human progress.

The scientific underpinning of David Ricardo's powerful theorizing has received little attention from historians. What is known of his scientific interests and cast of his thinking suggests that natural philosophy played an important methodological and conceptual role in his economics. He had early interests in chemistry and geology. A convert to Unitarianism, we can expect he was familiar with the philosophical and chemical ideas of Priestley, the intellectual leader of the movement. The fact that Ricardo was a trustee of the London Geological Society, whose directors sought to advance James Hutton's uniformitarian geology, suggests he knew Hutton's powerful energetic arguments about the

geological and biological processes forming the earth and his method of dynamic (energy) equilibrium (and his tendency to ignore historical processes). Hutton was one of the leading energy thinkers of his generation and the likely source for Ricardo's similarly powerful focus on dynamic processes.

Energy ideas play a central role in Ricardo's early corn theory of profit, which is set out as a ratio of the surplus of corn to its costs of production (measured in corn). If land (and energy) increase relative to labor, profits rise. Corn continues to play a central role in the dynamics of his labor theory of production and value after Malthus convinced him that capital could not be limited to corn. The relative availability and cost of corn (the energy source running the organic machines of the agricultural and artisan economy) continued to provide the foundation of his theory of relative costs, profits, and rents, and thus the long-run dynamics of the economy.

Although Ricardo noted the merit of Say's division of economic theory into production, distribution, and value, he did not make the explicit treatment of the production principles that Say's plan suggested. Instead, he begins his *Principles* with a chapter on how the values of commodities are determined. His analysis is clearly informed by a vision of production but the physical perspective of contemporary scientific ideas that informs it must be reconstructed from the wording Ricardo employs, a footnote or two, and the tenor of his arguments. We can catch a glimpse of his acute understanding of the contribution of the natural powers employed in industry when he refutes Smith's notion that nature does nothing in manufactures. The passage reflects Say's language but is pure Hutton in its energetic focus:

Are the powers of wind and water which move machinery, and assist navigation, nothing? The pressure of the atmosphere and elasticity of steam which move the most stupendous engines—are they not the gifts of nature? to say nothing of the effects of the matter of heat in softening and melting metals, of the decomposition of the atmosphere in the process of dyeing and fermentation. There is not a manufacture ... in which nature does not give her assistance to man, and give it too, generously and gratuitously.

—Ricardo, *Principles of Political Economy*, [1817], chapter 2, n10

8. LATE CLASSICAL THEORY: BABBAGE, SENIOR, AND J. S. MILL

Until the mid-1830s, the energy ideas entertained by classical economists, apart from Say and Ricardo,

were effectively confined to plants and biological engines. Thanks to Babbage's pioneering study of the new industrial "Economy of Machinery and Manufactures" in 1832 and Andrew Ure's study of the factory system in 1835, the importance of coal and steam engines for industrial production and transportation and for Britain's industrial ascendancy were now in the public eye. In reviews of Babbage and Ure, the economist John McCulloch noted the link between Britain's industrial prosperity and the exploitation of coal technologies. He argued that the increase in the quantity of iron made with coal lowered the costs of machinery (and steam engines) leading to more output at lower unit costs and more demand for coal, iron, and machines. He did not, however, attempt to incorporate these ideas into a theory of production.

Babbage's great study of principles of a machine economy distinguished two types of machines, those which produced power and those designed to transmit force and execute work. The first class included water and wind mills. But it was the adoption and continued innovation in steam engines employing coal for fuel that captured Babbage's imagination (the heat-based industries need equal attention as well). Economies of scale and high fixed costs pushed mill owners to continuous operation and opened opportunities for new professions (accountants, etc.) and industries, such as gas lighting. New power sources and technologies created new products and materials and new uses for materials that formerly had little value. The principles of the machine (and coal and iron) economy extended beyond the factory to the larger political economy and society generating new knowledge and new models for administration.

The only economist to apply Babbage's theory of machines to a theory of production inputs and returns was Nassau Senior, whose contributions in this regard, however limited, were completely ignored by his contemporaries. Babbage's suggestion that "practical knowledge" would continue to grow without limit appeared in the third of Senior's "elementary propositions of economic science," which stated that "the powers of labor and the other instruments that produce wealth can be increased indefinitely." Senior's fourth proposition posited the existence of diminishing returns to additional labor employed on the land and increasing returns in manufacturing. While he understands the conservation of matter in manufacturing industry when he states that "no additional labour or machinery can work up a pound of raw cotton into more than a pound of manufac-

tured cotton,” his attribution of diminishing returns in agricultural production to the employment of additional labor “on the *same* material” does not make it clear that additional output here requires additional materials from the soil and atmosphere and additional solar energy to produce another bushel of corn or pound of cotton seed.

Senior used Babbage to advance Adam Smith’s distinction between fixed and circulating capital (structures and agents versus material flows). He distinguished tools and instruments from materials embodied in production and subdivided machines into those producing and transferring power. He separated natural powers from ordinary materials on the grounds the former are not embodied in output but did not know where to put food, coal, and other natural powers. Instead of distinguishing a subcategory of natural powers or energetic flows along side materials, he lumped food and coal with “fixed capital.” Clearly, more insight about the nature of power and the agents and machines applying and using power was needed. His laudable efforts to grapple with a physical classification of production inputs and their implications for production returns were not, unfortunately, taken up by subsequent economic writers. Indeed, J. S. Mill attempts to undermine both Senior’s classification efforts and treatment of production returns.

Another extremely important treatment of the ideas of the sources and application of power in human economy was that of the scientist and mathematician William Whewell in 1841, who extended Babbage’s classification of machinery to include moving power, trains of mechanism connecting power to work and work done. Whewell also recognized that the engine economy and its sources of power were dissipative and would eventually be exhausted (including the engine of the sun). He concluded, Norton Wise points out, that no balance existed in nature (or economy). In Whewell, the steam economy replaced the equilibrium of the balance with disequilibrium and decay, foreshadowing the second law of thermodynamics. Whewell embraced decay as part of a Christian eschatology but would have not been happy that evolutionary processes are the other side of a dissipative and temporal universe.

Despite his friendship with Babbage and adoption of J.-B. Say’s ordering of economic theory (Book I of his *Principles* is devoted to production), the socially progressive J. S. Mill made no attempt to apply physical or engineering insights to production theory. He follows Say (and Senior) in classifying production

inputs under the headings of labor, natural agents, and capital, but his concern throughout is to emphasize economic aspects of production. He shows no interest in the physical foundations of economic theory. Book I has three chapters on capital and at least two on labor but none on the characteristics of land or natural agents. He isolated “labor and natural objects” as the two “requisites of production.” But in contrast to his elaborate classification of indirect labor, and the separation he makes between materials and powers in nature, he refuses to make any distinction between materials, food, and fuels as circulating capital.

Ignoring Babbage, Senior, and Whewell, Mill argued that “political economists generally include all things used as *immediate* means of production either in the class of implements or materials to avoid a multiplication of classes [and] distinctions of no scientific importance.” He made the distinction between materials and machines turn on whether an “instrument of production can only be used once” and declared that a “fleece is destroyed as a fleece” in making a garment just as a fuel is consumed to make heat (which indicates he did not understand the distinction between energy and materials). He responded to the “able reviewer” of the first edition of his work (Senior) who objected to subsuming coals for engines and food for workers under the class of materials by admitting that while his terminology was not in accord with the physical or scientific meaning of material, the distinction between materials and fuels is “almost irrelevant to political economy.”

Mill correctly recognized that tools and machinery lack productive powers of their own. But after declaring that the only productive powers are labor and natural agents he denied this status to “food and other materials” (and thus fuels). The fact that steam engines burning fuels are productive and a worker lacking food is not exposes the weakness of his position. By undermining Babbage and Senior’s extension of Smith’s distinction between physical structures that process and store materials and flows of materials and energetic powers, Mill brought the development of engineering ideas about the machines and powers in production theory to a halt.

Mill recognized the existence of increasing returns in manufacturing and argued that the causes increasing “the productiveness of [manufacturing] industry preponderate greatly over the one cause [the increasing expense of materials] which tends to diminish that productivity.” Yet he treated increasing returns as a subsidiary proposition to “the general

law of agricultural industry,” that is, diminishing returns, which he calls “the most important proposition in political economy.” Increasing returns is not given an independent development. The statement of neither proposition includes a consideration of the essential inputs required to increase output according to the conservation laws which tie output to material and energy input. Mill was aware of the complementarity between capital and natural powers and materials and output but his lack of concern with accurate scientific terminology and any systematic consideration of the physical principles of production fail to preserve or advance the physical production framework advanced by Say, Ricardo, and Senior let alone Babbage or Whewell. Given the that Mill’s work was a common point of departure for neoclassical theorists, his failure was a critical factor in the neoclassical failure to seriously grapple with energy and scientific approach to production theory.

9. HOW ENERGY CAME UP MISSING IN NEOCLASSICAL ECONOMICS

The main early neoclassical economists, including William Stanley Jevons, Leon Walras, Alfred Marshall, and Wilfredo Pareto, took their inspiration from the rational and analytical mechanics of the 18th century and 19th century. Thanks to the construction of mechanics in terms of the equilibrium between forces, mechanics provided economics with a powerful mathematical language of analysis that included a model of self-regulating activity produced by the inexorable operation of opposing forces. Neoclassical economics abandoned the production approach of the classicals to reconstruct economics as a mathematical science of behavior and markets expressing the interplay of demand and supply based on theories of marginal utility, marginal productivity, and universal scarcity, the latter being the consequence of the substitution assumption or “law of variable proportions,” which assumes that any nonscarce good of interest to economics will become scarce by being substituted for other goods.

We have already noted how the standard formulation of diminishing marginal productivity violates the basic laws of thermodynamics. The adoption of the idea of marginalism necessarily pushed materials and energy and any scientific formulation of production from the scene. William

S. Jevons eliminated fixed capital from his theory and provided no treatment of the role of energy in his treatment of production despite having written and entire book on the critical importance of coal to British industrialization just a few years before. Leon Walras had initially specified raw materials as an intermediate input but then proceeded to vertically aggregate production so that output was a direct function of the “original factors of production,” land, labor, and capital. Capital was assumed to be like land so that adding more labor to a given amount of capital could be assumed to add more output albeit subject to diminishing returns.

But the theoretical problems of marginal productivity are not confined to a special assumption (individual inputs are productive everything else constant) that can be relaxed by adding more inputs. The assumption reflects a structural feature of the theory, which effectively prevents any realistic physical reformulation of the theory. Thanks to the adoption of the mathematics of field-theory as the structural framework informing utility and production theory, every input in the production function (which is based on the potential function of analytical mechanics) is effectively assumed to have productive capacities because of the existence of an energy gradient in the space over which the production function is defined. Each production input is effectively considered to be an axis in n -dimensional space. Thus, an increase in any input (an increase along one axis) automatically involves a movement to a higher energy gradient. Output, defined as the potential function of the field, increases. These movements, moreover, are always reversible since the field theoretical framework is nondissipative and conservative.

In the 18th and early 19th centuries, nature (economics included) had been assumed to be self-regulating as a metaphysical principle grounded in nature’s design. In the late 19th century, despite the development of a new energy physics based on the laws of thermodynamics and an understanding of the inexorable processes of entropic dissipation, economics had found a new and powerful mathematical and conceptual framework in field theory for expressing and developing its vision of order, stability, harmony, and rational allocation. For economics, the virtue of a theory of electromagnetic or gravitational fields is that there is no dissipation. Unlike other energy conversions, which are dissipative, the energy principles of field theory are conservative (no dissipation takes place). This provides a framework taken from physics that

isolates economics from the troubling questions about natural resources or sustainability. These issues have little meaning once the assumptions of the theory have been accepted.

SEE ALSO THE FOLLOWING ARTICLES

Conservation Measures for Energy, History of • Conservation of Energy Concept, History of • Early Industrial World, Energy Flow in • Economics of Energy Demand • Economics of Energy Supply • Energy in the History and Philosophy of Science • Geographic Thought, History of Energy in • Prices of Energy, History of • Sociopolitical Collapse, Energy and • Thermodynamic Sciences, History of

Further Reading

- Aspromourgos, T. (1986). Political economy and the social division of labour: The economics of Sir William Petty. *Scottish J. Pol. Econ.* 33, 28–45.
- Ayres, R. U. (1978). “Resources, Environment, and Economics: Applications of the Materials/Energy Balance Principle.” Wiley, New York.
- Christensen, P. (1989). Hobbes and the Physiological Origins of Economic Science. *History Pol. Econ.* 21, 689–709.
- Christensen, P. (1994). Fire, motion, and productivity: The proto-energetics of nature and economy in François Quesnay. In “Natural Images in Economic Thought, ‘Markets Read in Tooth and Claw’” (P. Mirowski, Ed.), pp. 249–288. Cambridge University Press, Cambridge, UK.
- Christensen, P. (2003). Epicurean and stoic sources for Boisguilbert’s physiological and Hippocratic Vision of nature and economics. In “Oeconomies in the Age of Newton” (M. Schabas and N. De Marchi, Eds.), pp. 102–129. Duke University Press, Durham, NC.
- Debeir, J.-C., Deleage, J.-P., and Hemery, D. (1991). “In the Servitude of Power: Energy and Civilization through the Ages” (John Barzman, trans.). Zed Books, Atlantic Highlands, NJ; London, UK.
- Dixon, B. L. (1988). “Diderot, Philosopher of Energy: The Development of his Concept of Physical Energy, 1745–1769.” Voltaire Foundation, Oxford, UK.
- Frank, R. G. (1980). “Harvey and the Oxford Physiologists.” University of California, Berkeley, CA.
- Garber, D., Henry, J., Joy, L., and Gabbey, A. (1998). New Doctrines of Body and its Powers. In “The Cambridge History of Seventeenth Century Philosophy, I,” pp. 553–623. Cambridge University Press, Cambridge, UK.
- Golinski, J. (2003). Chemistry. In “The Cambridge History of Science” vol. 4, “Science in the Eighteenth Century” (Roy Porter, Ed.), pp. 375–396. Cambridge University Press, Cambridge, UK.
- Vaggi, G., and Groenewegen, P. (2003). “A Concise History of Economic thought: From Mercantilism to Monetarism.” Palgrave, New York.
- Vatin, F. (1993). “Le Travail: Economie et Physique: 1780–1830.” Presses Universitaires de France, Paris.
- Wicken, J. S. (1987). “Evolution, Thermodynamics, and Information: Extending the Darwinian Program.” Oxford University Press, New York and Oxford, UK.
- Wise, N., and Smith, A. C. (1989). Work and waste: Political economy and natural philosophy in nineteenth century Britain. *History Sci.* 27, 263–301, 391–449.

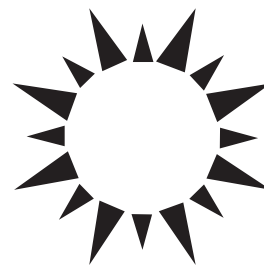
Ecosystem Health: Energy Indicators

DANIEL E. CAMPBELL and TINGTING CAI

United States Environmental Protection Agency
Narragansett, Rhode Island, United States

THOMAS W. OLSEN

Consultant
Kingston, Rhode Island, United States



-
1. Ecosystem Health and Ecological Integrity
 2. Historical Background on Ecosystem Health
 3. Energy Systems Analysis, Health and Emergy
 4. Energy and Ecosystems
 5. Direct Measures of Ecosystem Health
 6. Indirect Measures of Ecosystem Health

Glossary

ascendency A summary measure of the magnitude of the fluxes within a network and of the degree to which the paths of flow within the network are organized or preferentially directed.

available energy The potential to do work.

ecological integrity The state of structural completeness of components and maximum function of processes in an ecosystem operating under a given set of external forcing energies. In this state, the empower of the ecosystem network is near the maximum possible.

ecosystem An organized system of biotic and abiotic components including land, water, mineral cycles, living organisms, and their behavioral control mechanisms.

ecosystem health The functional integrity of an ecosystem as characterized by a state of vigorous self-renewal in the ecosystem as a whole and in its components (soil, water, animals, plants, etc.). One measure is the empower gained by the ecosystem from a given set of external energy sources.

emergy The available energy of one kind previously used up both directly and indirectly in making a product or service. Emergy has units of emjoules to connote energy used in the past.

empower Emergy per unit time or emergy flux.

energy flow network The interconnected energy flows in an ecosystem.

energy (emergy) signature The suite of external forcing energies from which an ecosystem can capture available energy. When the external energies are converted to emergy, the entire suite is called the emergy signature of the ecosystem.

enthalpy (heat content) The internal energy plus the product of the pressure times the volume. Relative changes in enthalpy are easily measured by calorimetric methods, which have been widely used to determine the heat content of biomass.

exergy The potential that energy has to do work relative to a reference state. As used by some ecologists, exergy includes the higher potential that energy associated with lower-probability structures has for the capture of additional available energy.

homeorhetic The tendency for a system to maintain constancy in its dynamic properties (e.g., energy flow).

maximum power (empower) principle The principle originally formulated by Alfred J. Lotka as “natural selection tends to make the [available] energy flux through the system a maximum, so far as compatible with the constraints to which the system is subject.” H. T. Odum realized that available energies of different kinds have different abilities to do work when used within an ecosystem network. Conversion to emergy flux (empower) expresses the energy fluxes of many kinds in units with the same quality (e.g., solar joules), allowing the maximum to be accurately determined.

pulsing paradigm The theory that alternating production/storage with consumption/release on all hierarchical scales maximizes power.

self-organization The use of net energy to build increasingly complex patterns of feedback and design to capture more available energy for the ecosystem; the properties of this process as they govern the creation of order in far-from-equilibrium systems.

solar transformity The solar energy required to make a unit of energy of a product or service. Its standard units are solar emjoules per joule.

specific energy The energy per unit mass of a flow, product, or service, its standard units being solar emjoules per gram.

Just as for human beings health is a concept that applies to the condition of the whole organism, the health of an ecosystem refers to the condition of the ecosystem as a whole. For this reason the study and characterization of ecosystems is fundamental to establishing accurate measures of ecosystem health. By definition, the idea of health is closely related to the optimal functioning of an organism or other entity. This idea can be applied to ecosystems, economies, and other organized systems in general and not simply to living organisms. Health is more than the simple absence of disease; rather it is a state that reflects the overall well-being of the system, whether it is a human organism, an ecosystem, or the economy of a nation.

1. ECOSYSTEM HEALTH AND ECOLOGICAL INTEGRITY

Integrity is a complex idea closely related to health that has two meanings relevant to ecosystems. The first important meaning of integrity is that it requires an entity to be unimpaired when compared to an original condition. This implies that a reference condition must be defined and that the presence or absence of integrity can be measured relative to this standard state. Health is also often measured with respect to a reference condition. A second relevant meaning of integrity is that it defines a state of organic unity, or wholeness within a system. Health and integrity are complementary ideas that together are capable of describing the well-being of any system. Because these ideas are so closely related (both share the meaning of soundness), they have not always been clearly distinguished in scientific and other writings.

2. HISTORICAL BACKGROUND ON ECOSYSTEM HEALTH

Early depictions of ecosystems as organized networks by A. J. Lotka, A. Leopold, G. E. Hutchinson,

and R. L. Lindemann and their insights into the role of energy flow, storage, and conversion efficiency in the healthy organization and functioning of ecosystems provided the basis for later assessment of ecosystem health in terms of the status and dynamics of various energy measures. However, the direct application of health, and its sister concept integrity, to ecosystems can be traced to the writings of Aldo Leopold during the late 1930s and 1940s. Leopold developed the concepts of a land ethic and land health that presaged the scientific interest in ecological integrity and ecosystem health, respectively. Leopold equates health of the land to its functional integrity, which he defines as a state of vigorous self-renewal in the land as a whole and in its components (soil, water, animals, plants, etc.). Leopold's concept of land is similar to our present idea of an ecosystem. Leopold's poetic description of land as a hierarchy of energy circuits implicitly called for the development of a quantitative tool like the energy circuit language, which H. T. Odum developed 30 years later.

The scientific interest in ecosystem health had its origin in the field of stress ecology that developed during the late 1970s and 1980s. Some research in stress ecology built on the earlier work of H. T. Odum, who showed that stress in ecosystems had the common property of acting as an energy drain, preventing available energy from flowing into useful structures and functions. Hans Selye discussed evidence for a nonspecific, generalized stress syndrome in humans. This work served as a basis for moving from the idea of stress in an organism to that of stress in an ecosystem, which led David Rapport and others logically to a consideration of ecosystem health. In the late 1980s, more research on ecosystem health, based on an analogy with medicine, began to appear. The First International Symposium on Aquatic Ecosystem Health was held in July 1990 and set the stage for a new publication, *Journal of Aquatic Ecosystem Health*. David Rapport characterized ecosystem health as a transdisciplinary science, which includes the study of ecosystem structure and function, along with socioeconomic factors and human health as they relate to the overall well-being of an environmental system.

3. ENERGY SYSTEMS ANALYSIS, HEALTH AND EMERGY

Robert Costanza proposed that the health of an ecosystem could be characterized by the vigor, network organization, and resiliency of the system.

These three general properties of ecosystems are essentially the total energy flow through the network, the arrangement of those energy flows into an organized network, and the excess energy flow or scope for growth that an ecosystem can use to recover from disturbance. These three characteristics relate to ecosystems and their functioning over time, and therefore systems analysis and simulation are fundamental tools needed to characterize ecosystem health. In this article the concepts of ecosystem health and ecological integrity are defined, measured, and interpreted from the perspective of energy systems theory and other systems approaches, such as information theory (ascendency), that are capable of analyzing networks of energy flow. Since the focus is on energy indicators of ecosystem health, the authors judge this perspective to be the most informative. Other perspectives on ecosystem health exist and are valid for certain purposes, but they are not considered here.

Energy systems theory provides tools and methods for describing and analyzing the hierarchical organization of ecosystem networks. Analyzing the energy hierarchy led to the insight that all energies are not equal in their ability to do work in a network. Emergy measures the relative ability to do work of the energy flowing through different components in an ecosystem network. Emergy values are always based on an underlying energy flow or storage times an appropriate emergy per unit coefficient. Emergy places energies of different kinds on the same scale so that they are directly comparable. The base unit for expressing emergy in ecosystems is the solar emjoule, sej. An emjoule is a joule of a particular kind of available energy that has been used in the past to make a product or service. Where the coefficient converts energy to solar emergy it is called the solar transformity and has units of sej/J. Also, mass in grams can be multiplied by the specific emergy of the item in sej/g to obtain emergy. Power must be converted to empower to gain a true picture of the relative importance of changes in different energy flows. The maximum power principle must be evaluated in terms of empower to obtain an accurate criterion for judging the relative health of ecosystems.

4. ENERGY AND ECOSYSTEMS

Available energy drives every known process at every level of organization in the universal hierarchy of natural phenomena. The existence of available energy potentials of sufficient magnitude is the fundamental

basis for all self-organizing structures that exist far from thermodynamic equilibrium (e.g., ecosystems, organisms, and societies). The use of available energy potentials enables ecosystems and their components (e.g., individuals, populations, and communities) to perform their natural functions. Ecosystems self-organize into hierarchical networks of components connected by energy flow. These flows interact to link ecosystem components into energy flow networks through pathways of power and control (Fig. 1). Some controlling feedbacks interact with external energy sources to capture more available energy for use in building the network. Hierarchical levels of organization are also linked by energy flow (Fig. 1). Energy is a common denominator among ecosystems that makes it possible to identify and characterize similar processes and functions in ecosystems of many kinds. Because of the fundamental role of power (energy flow) in ecosystem self-organization, it is logical to expect that measures of power will be useful in characterizing the functional integrity (health) of ecosystems and that changes in energy flow will measure possible disturbances in healthy ecosystem processes.

4.1 Energy Intake and Ecosystem Health

Ecosystems consist of biotic and abiotic elements organized as structures, gradients, and flows that are maintained far from thermodynamic equilibrium through the capture and transformation of available energy. The persistence of such organization is continuously threatened and disrupted by the dissipative tendencies of the established structures and gradients within the system and its surroundings. The second law of thermodynamics governs these tendencies toward disorder, and ecosystems as well as organisms must continuously draw in available energy to repair and counteract them. For ecosystems, many external forces such as solar radiation, heat, wind, nutrients, and invasive organisms can be either sources of available energy or disordering forces depending on their form, magnitude, or intensity. The similarity between the condition of ecosystems and organisms with regard to the maintenance of necessary structure has motivated the development of the various energy indicators of ecosystem health—for example, the ratio of the energy used in respiration to the structure that it supports is called the Schrödinger ratio, a measure of the energy cost to maintain a unit of structure.

Both the persistence and growth of ecosystems depend on the continuous influx of available energy

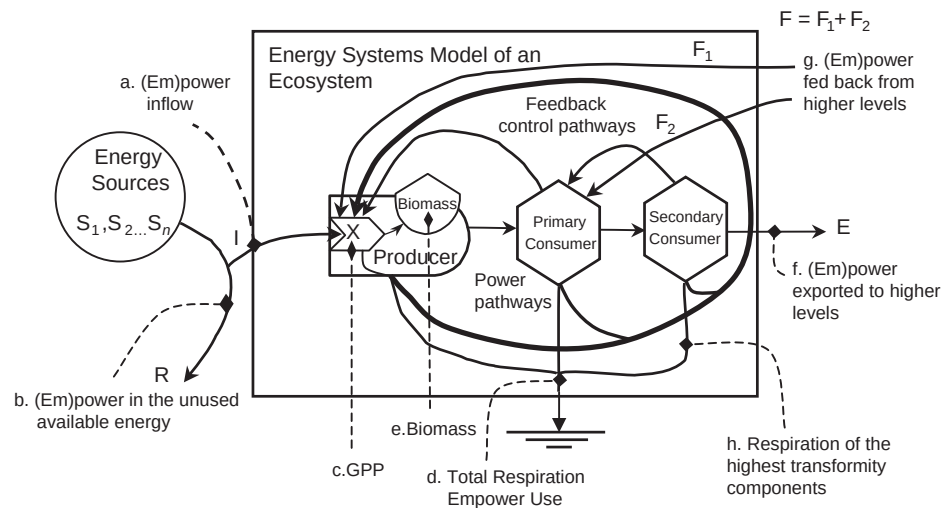


FIGURE 1 Energy systems model showing some of the general characteristics of ecosystems and the points where power and empower might be measured (labeled daimonds) to indicate ecosystem health. Symbols are briefly defined as follows: circles are energy sources, tanks are stored quantities, bullet symbols are producers, hexagons are consumers, and the block arrow is a work gate or interaction. Energy flows on the lines in the direction indicated by the arrowheads; light gray lines show flows of used energy to the heat sink (ground symbol), and the thicker, darker gray lines show the recycle of materials. S_i 's are sources, F_i 's are feedbacks from the larger systems, and E is the export.

and the transformation of this energy within the system to provide for its own self-renewal (i.e., structural maintenance and reinforcement of energy inflows) under the conditions imposed on it by its surroundings. When the structures, flows, and gradients within an ecosystem are organized to provide a capacity for maintenance and self-renewal that results in a sustained flourishing of the ecosystem, the system is healthy. For an organism or cell to remain healthy, both a sufficient metabolic rate *in toto* and an appropriate balance among the rates of specific metabolic processes are required to maintain the structural integrity and the functions on which its life depends. This same condition applies to ecosystems. Energy systems theorists have hypothesized that the maximum power principle governs the acquisition of energy by ecosystems. According to this principle, ecosystems prevail when the total rate of energy acquisition and conversion is sufficient to be competitive compared with the power acquisition rates of alternative ecosystem designs. If this principle is correct, ecosystem health depends on meeting this criterion for survival in evolutionary competition, and healthy systems will have total power flows that approach the maximum possible for a given set of inputs. To capture more of the available energy in external sources, ecosystems must develop an appropriate balance among the specific applications of this energy to maintain the internal structures and functions on which the viability of the ecosystem depends. Developing and maintaining the

capacity to respond to pulsing on several hierarchical scales (the pulsing paradigm) is a condition for maximizing empower when energy is stored. Using the energy systems perspective and substituting energy for energy and empower for power, the preceding concepts of ecological integrity and ecosystem health are defined and related in the following statement: ecological integrity is an emergent property of ecosystems that is greatest when all the structural components of a system that should be there based on the energy sources received by the ecosystem are there (i.e., structural integrity is complete) and all the processes operating within this system of components are functioning at the optimum efficiency for maximum empower (i.e., the ecosystem is healthy). Therefore, ecosystem health (maximum network empower) is one component of the overall ecological integrity of an ecosystem, and structural integrity (completeness) is the other.

4.2 Properties of Healthy Ecosystems

Healthy ecosystems can be described by a relatively few properties that accompany maximal energy but more specifically energy intake. Since ecosystems undergo a regular cycle of change, the expected values for these properties will be different depending on the stage of ecosystem development. One sign of stress for the ecosystem is a change in the expected course of succession and regression that characterize this cycle. The maximum power principle indicates

that healthy ecosystems will maintain optimally efficient energy transfers and nutrient cycling to maximize energy intake from available energy sources. This optimum ecological organization includes characteristic sets of species. These species sets vary with the stage of ecosystem development and with the particular ecosystem type, but in general a diverse species assemblage dominated by long-lived larger organisms characterizes the mature (climax) stage of ecosystem development. Healthy ecosystems also have the energy and information reserves to resist or adapt to change, and therefore they show resiliency in the face of perturbations or debilitating factors. The state of health involves optimization in terms of the allocation of resources because adaptation to any disordering factor entails tradeoffs with other factors including the maximum growth potential of the system.

4.3 Energy Indicators of Ecosystem Health

The characterization of the state of health of an ecosystem as defined earlier requires measures that express the condition of the ecosystem as a whole or some aspect of the system that can be quantitatively related to overall condition. Self-organization of ecosystems for maximum power has resulted in general design properties that are common to all ecosystems. The energy systems language model in Fig. 1 illustrates the following common design properties: (1) a set of external energy sources, (2) a web of interacting components, (3) convergence of successive energy transformations in a chain of increasing energy quality (emergy per unit energy), (4) storage of mass and information, (5) recycle of materials, (6) feedback control systems, (7) interactions of energy flows of different quality to maximize power by rewarding components that contribute, and (8) connection to other hierarchical levels of organization through imports and exports of energy. Important points where energy indicators of ecosystem health can be measured are marked on the diagram. The flows and storages of available energy that provide information about the health of ecosystems are (a) the energy inflow, (b) the albedo or unused energy, (c) gross primary production (GPP), (d) respiration of the whole system, populations, or groups of populations, (e) total biomass, the biomass of system components, and changes in biomass, (f) energy export from the system, (g) energy import into the system, and (h) the energy used by the highest transformity elements of the ecosystem. In the system diagram in Fig. 1, the

total inflow of empower in the available energy captured ($F + I$) will be close to a maximum in healthy systems operating under a given set of external energies. When the external energy sources are constant or remain within an expected range, ecosystem health might be measured by comparison to reference levels of ($F + I$) or reference values of any other flow that has a known relationship to ($F + I$). When the spectrum of external emergies (the emergy signature) is changing, health can only be judged by comparison of the network empower of alternative system designs.

5. DIRECT MEASURES OF ECOSYSTEM HEALTH

Figure 1 shows many points at which the energy flow in an ecosystem network can be measured. Other measurement points are also possible. Each point of measurement tells us something different about ecosystem function and all will have their usefulness. The indices that will be the most useful in characterizing ecosystem health remain to be proven by empirical testing.

5.1 Biomass: Energy Storage as an Indicator

The most direct and earliest measures of energy status and dynamics include the enthalpy stored by the biotic community and the rates of enthalpy capture by and transfer within this community. Measures of community and population biomass provide an estimate of energy storage and its organization when standard factors are used to convert biomass dry-wt. values to enthalpies. Based on representative carbohydrate, lipid, and protein fractions in the major biological taxa, typical conversion factors range from 19 kJ/g for terrestrial plants to 26 kJ/g for vertebrates on an ash-free dry-wt. basis. Because total community biomass increases during ecosystem succession and generally decreases in response to stress, this measure can serve as an indicator of ecosystem health, particularly if a reference for comparison can be established. Relatively little effort has been made to identify such reference values due to practical difficulties arising from the great variability among ecosystems in community biomass and in the factors affecting it. If these factors could be adequately accounted for, total community biomass might provide a reasonable

initial indication of an ecosystem's health in terms of its energy flow.

Biomass alone could be a very misleading indicator, if diversity and balance among ecosystem elements are important components of ecosystem health. The productivity that produces a high biomass could result from the predominance of a minimal number of energy flows and conversions that are highly efficient under a narrow range of conditions. When such efficiency is attained at the expense of the redundancy and diversity of flows and conversions that are critical to the long-term persistence of the ecosystem, high biomass does not accurately indicate a robustly healthy ecosystem. Diversity measures can also be derived from the stored biomass in an ecosystem's respective populations to provide an index of organization that has been proposed as a predictor of ecosystem stability and resilience. However, no consistent relation between these health-related properties and such diversity has been established.

5.2 Productivity: Energy Flow Rate as an Indicator

Gross primary productivity is the rate of conversion of the energy captured by the biotic community or its respective autotrophic populations to chemical energy in biomass. It is an indicator that is closely related to energy flow within the system and thus should be more sensitive than biomass is to changes in health. However, if gross primary productivity is measured only at the community level, the home-orhetic nature of biological production and respiration severely diminishes the sensitivity of this indicator. A decrease in the productivity in one population is often compensated for by an increase in the productivity of another. Productivity is also quite variable and measurements must be taken over a long enough period of time to obtain a representative average. The diversity (Shannon-Weiner index) of productivity (energy flows among an ecosystem's populations) has also been considered as a measure of ecosystem health based on a hypothesized dependence of stability (a characteristic of healthy systems) on diversity. The extent of this dependence has yet to be established. Some increase in home-orhetic potential with increasing diversity is expected at low diversities. In general, diversity appears to correlate with ecosystem but not population stability. Even though methods exist to estimate gross primary production from measurements of total system metabolism, the difficulty of measuring gross productivity directly (because respiration simultaneously

consumes the organic matter produced) has been an impediment to the successful development of productivity-related indicators of ecosystem health.

5.3 Respiration: Energy Dissipation as an Indicator

The community respiration rate has also been proposed as an indicator of ecosystem health because of its close association with processes of maintenance and repair. As one of the first effects of stress, the demand for these processes is expected to increase, and thus an increase in respiration could be an initial indication of the exposure of an ecosystem to an environmental change that adversely affects its health. In the longer term, community respiration typically declines as the productivity of an ecosystem is diminished by an adverse condition that reduces energy intake. Because processes of biotic self-organization and renewal, including growth and reproduction as well as maintenance and repair, are driven by the energy used in respiration, ecosystem health as described here is clearly related to respiration.

A healthy ecosystem might be characterized as one that maintains its respiration rate through periods of relative stability and in response to slow changes in its energy sources. Since stress acts as an energy drain, an increase in respiration rate over that expected for a particular ecosystem might indicate a decline in health. Optimal (or healthy) values for an ecosystem's respiration rate and for the time-dependent variability in this rate will differ with prevailing environmental conditions, and standards appropriate to any given set of conditions are not yet available. If such standards could be developed, deviations from the standard could provide a useful indication of a lack of healthy adaptation to extant stress. Where opposing effects of multiple stressors affect metabolic rates, simulation models would be needed to predict impairment. The difficulty and infrequency with which community respiration is measured and the complications involved in developing and validating diagnostic standards present obstacles to the use of community respiration rate as an indicator of ecosystem health.

5.4 Energy Absorbed: Energy Intake as an Indicator

The solar energy absorbed by terrestrial ecosystems and its complement the albedo are arguably the most direct measures of ecosystem health for these systems. This measure varies greatly among system

types, but it also varies within a system according to the overall condition of the system. Remotely sensed measures of ecosystem condition include the Normalized Difference Vegetation Index (NDVI), or greenness index, which shows clearly the effects of wet and dry years on the condition of forests and other terrestrial vegetation. Methods to determine the long-term changes in energy capture by an ecosystem show promise in documenting the regional effects of global climate change on ecosystem health. Because of the efficiency and power of remote sensing in characterizing large areas with high-resolution data, these methods may hold the most promise for developing future energy indicators of ecosystem functioning.

5.5 Energy Ratios as Indicators

Ratios among the preceding measures (i.e., P/R , P/B , and R/B , where P is gross community productivity, R is community respiration, and B is community

biomass) have also been proposed as indicators of ecosystem health or disease (i.e., lack of healthy adaptation to an extant stress). Production and respiration are the two essential processes within ecosystems, and the ratio of these two energy flows provides considerable information on the state of the ecosystem (Fig. 2). In ecosystems that are not subject to large net imports or exports of organic matter, the P/R ratio will eventually become balanced (at close to 1). Imbalances in the forcing energy signature as well as stressful factors that alter community production or respiration or that increase respiration in response to biotic damage can shift this ratio away from 1, indicating impaired ecosystem health for these systems. The P/R ratio is less dependent on a comparison to previous measures of community metabolism or established standards for ecosystem types than is an indicator based on P , R , or B considered independently. Its interpretation is dependent on knowledge of external factors such as imports and exports of organic matter and of any

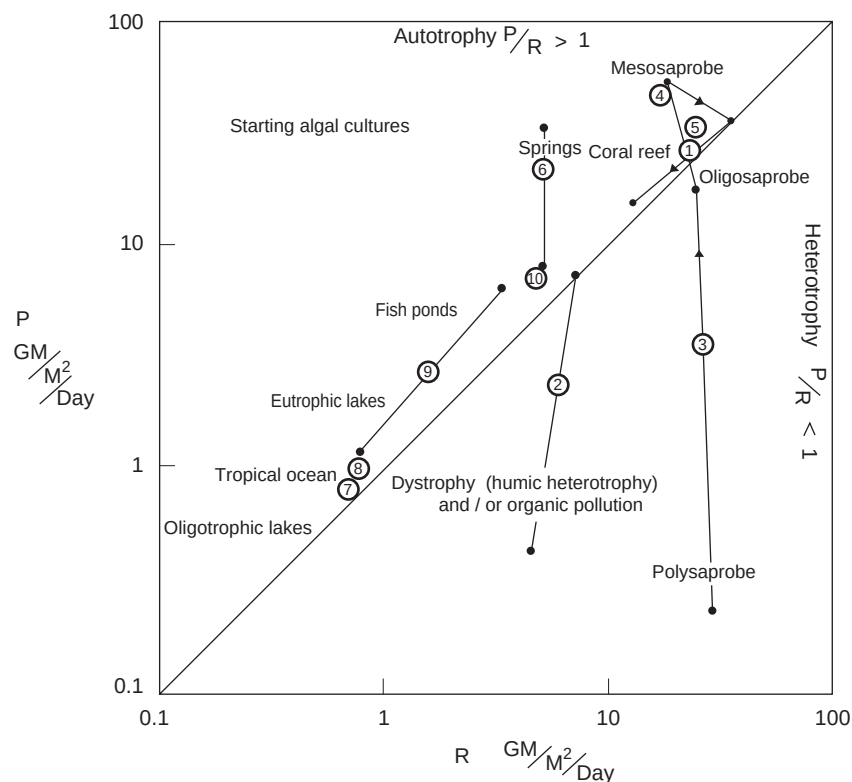


FIGURE 2 A diagram of the classification of ecosystems based on autotrophic (P) and heterotrophic (R) metabolism. Reprinted with permission from *Limnology and Oceanography* (1956), 1(2),115. The original caption read as follows: A diagram for the functional classification of communities according to the total and relative dominance of autotrophic and heterotrophic metabolism. Specific examples are indicated as follows: (1) Coral reefs; (2) Itchen River, England; (3) recovery sequence on the White River, Indiana; (4) Birs, Switzerland; (5) Turtle grass; (6) Silver Springs; (7) Tropical Ocean; (8) Linsley Pond, Conn. (Riley, 1940); (9) Little John Pond, Wisconsin (Juday, Blair, Wilda, 1943); (10) fish ponds (from Winberg, 1937, as quoted by Juday et al., 1943); Ryther (1954) described the behavior of the P/R ratio in algal cultures.

large disturbances to which the ecosystem may have been subjected.

The ratios P/B and R/B have been advanced in association with the hypothesis that the self-organization of healthy ecosystems results in a maximization of biomass subject to available energy constraints. Thus, decreases in these ratios under a stable energy regime would indicate a stress that had undermined in some manner the healthy self-organization of the ecosystem. From an energy systems perspective trends in biomass are viewed as secondary to the priority of maximizing empower in the network. Thus, ecosystems build structure and organization as the means of increasing the use of the available energies.

The R/B ratio (sometimes also called the metabolic quotient, especially when measured in units of carbon stored and respired, or the Schrödinger ratio) has been the more thoroughly investigated of the two, especially with respect to soil ecosystems. The Schrödinger ratio is considered by some to be a thermodynamic order function in an ecosystem. E. Schrödinger, a physicist, showed that biological systems with molecular order affected by thermal vibrations must continuously pump out disorder if they are to maintain their internal order. Thus, the larger the biomass, the greater the maintenance cost. Although the energy required for maintenance of biological structure relative to the energy stored in that structure might be expected to increase with the application of stress, the observed direction of change in this ratio in the presence of common stress factors has varied with the ecosystem and the stress.

6. INDIRECT MEASURES OF ECOSYSTEM HEALTH

There are many possible methods for analyzing the properties of networks to indirectly measure ecosystem functional integrity or health. The three methods discussed in this section have been used in the literature to characterize ecosystem health. In addition, the reader should examine the references given in the bibliography.

6.1 Emergy Measures and Models

Emergy addresses the differences in the effects of different forms of available energy when used in a network. When the direct energy measures presented earlier are converted to emergy before calculating

indices, a more accurate picture of the balance within the ecosystem should be gained. For example, the respiration of an ecosystem might decrease but the empower of respiration could increase if more of the energy in metabolism was due to higher transformity components. We expect that indices constructed using empower will be able to explain observations where power alone has failed. The fundamental theoretical connection between energy measures and the response of systems to selective pressures might make these measures particularly suitable for indication of the long-term health status of ecosystems. Three emergy indicators of ecosystem health are (1) the empower used to maintain a given quantity of stored energy in structure (similar to the respiration per unit biomass for an organism), (2) the dynamic transformity for the whole ecosystem or the ratio of the energy required to produce a unit of energy flowing along all the pathways of an ecosystem to that unit of energy, which is an intensity factor for emergy in the ecosystem just as the temperature of an organism is an intensity factor for heat in the body, and (3) the ratio of the emergy stored to the emergy consumed in an ecosystem or the emergy P/R ratio, which may have characteristic values for systems with different characteristic emergy signatures. Figure 3 shows an example of two of these indicators calculated for control ecosystems and ecosystems impacted by hot water at Crystal River, Florida. Although emergy indicators are theoretically promising, they have not been tested and evaluated in large-scale studies of ecosystem health.

Simulation modeling is a viable means to develop indicators of ecological integrity and to identify criteria for ecosystem health. Ecological integrity and ecosystem health can be characterized through an analysis of validated simulation models of energy flow, when it has been demonstrated that the network of interactions in the model captures those factors that are primarily responsible for the behavior of the real system. In model simulations, ecosystem health is determined by calculating the empower at various points in the network of energy flows being simulated. The empower at any point in the network is calculated by multiplying the energy flow along the pathway by a dynamic transformity for that pathway. The value of the transformity depends on the external sources delivering emergy to the system as well as on the pathways of energy transformation within the system. Relative ecosystem health has been hypothesized to increase with empower on the pathways. In some studies of simulated ecosystems, this value has been found to

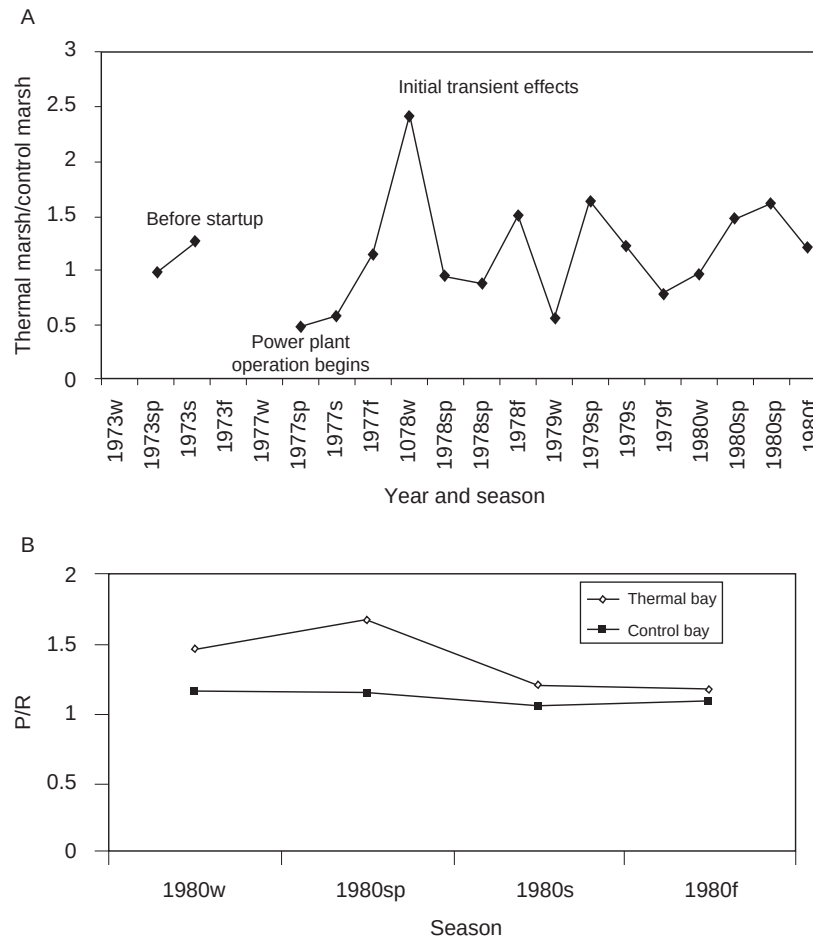


FIGURE 3 Emergy ratios illustrating patterns of health and disturbance. (A) Comparison of Schrödinger ratios for thermally impacted and control marshes. (B) The emergy P/R ratio of a thermally impacted versus a control bay. Reprinted with permission from Campbell, D. E. (2000), Using energy systems theory to define, measure, and interpret ecological integrity and ecosystem health. *Ecosystem Health*, 6(3), 195. The original caption read as follows: Emergy indicators of ecological integrity calculated using data from ecosystems affected by hot water effluent and control ecosystems at Crystal River, Florida. (A) The ratio of the fraction of energy in live biomass dissipated in daily maintenance for *Spartina alterniflora* from thermally affected and control marshes. (B) The emergy P/R ratio for the inner thermally impacted bay and its control bay.

approach a maximum for a given set of external energy sources that remain constant. There are special rules for determining the emergy of storages. The empower of the network includes information on the energy flow on each pathway as well as information on the position and connectivity of the pathways, which is supplied by the dynamic transformity. Empower can be determined at any point in the system where there is an energy flow.

The capacity to measure anthropogenic stress factors as emergy flows and to integrate these flows within a single dynamic systems model of the human-ecosystem interaction provides a basis for both ex ante and ex post assessments of the effects of human decisions and activities on ecosystem health. This

capacity also permits derivation of relative measures of ecosystem health that reflect dynamically changing human preferences for ecosystem services and functions. Within such an integrated dynamic systems analysis, not only can the subjective differences in human preferences be incorporated, but also the contributions of these preferences to the long-term likelihood of persistence and prevalence of both ecosystem and human economic subsystems can be estimated in terms of emergy acquisition and use. Furthermore, such dynamic analysis permits assessment of adaptation to fluctuating emergy inputs and the consequences of external or internal patterns of pulsing (i.e., emergy storage and release) that would not be possible through a static analysis.

6.2 Exergy

Exergy (in the sense in which this term has typically been used by systems ecologists) is a measure of the thermodynamic potential of an ecosystem with respect to a system at the same temperature and pressure and with the same elemental composition but in a state of thermodynamic equilibrium. This potential includes both the maximum work that could be acquired (through an entropy-free transfer of energy) from the ecosystem as it attained this equilibrium state and an estimate of the enhanced capacity for capturing or degrading a flow of available energy provided by low-probability structures maintained within the ecosystem. The first component of this potential (the maximum attainable work) has been estimated as the chemical potential of the organic matter in an ecosystem relative to that in the equilibrated system (i.e., using $-\Delta G_{org}$, the Gibbs free energy released by a unit of this organic matter—or of a representative combination of carbohydrates, fats, and proteins—upon complete oxidation).

There is no consensus on how (or in reference to what flow or process) the second component or capacity might be calculated, but the inclusion of this component reflects the aforementioned understanding that different forms of (available) energy can have very different capacities to affect ecosystem organization and function. One approach, which was proposed by S. E. Jorgensen, hypothesizes that the capacity of living organisms to increase the available energy capture by an ecosystem is positively correlated with the number of genes that they possess. The corresponding exergy contribution of an ecosystem compartment is then approximated based on the probability of producing the amino acid sequences coded by the DNA of the organisms in that compartment. The relevant compartments are selected by the ecosystem analyst based on their importance in maintaining or altering the thermodynamic gradients associated with the ecosystem. These probabilities are computed under two simplifying conditions: (1) the 20 amino acids used by organisms are all equally likely to appear at any point in the sequence and (2) each gene codes for a sequence of 700 amino acids. Given a typical number of genes possessed by the organisms within a compartment equal to n , the approximation for this probability is thus $20^{-700 \cdot n}$, and the corresponding exergy contribution is defined as $-R \cdot T \cdot c_i \ln 20^{-700 \cdot n}$, where R is the universal gas constant, T is the

absolute temperature, and c_i is the concentration (e.g., in g/L of an aquatic system), of carbon, for instance, in compartment i . The total exergy of an ecosystem with M organic compartments, as estimated using Jorgensen's formulation, can thus be expressed as

$$Ex = -\Delta G_{org} \cdot c_{org} + R \cdot T \cdot \sum_{i=1}^M c_i \cdot \ln 20^{700 \cdot n_i},$$

where c_{org} is the total concentration (of the chemical element or species of interest) in all organic compartments of the ecosystem (and ΔG_{org} is in J/unit of this element or species in the organic matter).

Due to the first term in this formula, exergy is highly sensitive to changes in available energy capture. Changes in structure related to shifts in species composition affect the second term, although a substantial shift in composition will not affect exergy if the number of genes and concentrations in the new taxa are similar to those in the former taxa. Furthermore, the sensitivity of an exergy index to important changes in ecosystem structure depends on the success of the analyst in selecting those compartments that are most closely related to these changes. The preceding formula is offered only as an approximation that provides one index for the assessment of the relative health of similarly analyzed ecosystems or of the relative effects of changes in an ecosystem or its environment on the health of that ecosystem.

Ecosystem resilience is not captured in any direct manner by the formula, but modeling studies have indicated a relation between exergy and the resistance of ecosystem state variables to perturbations. On average, resistance increases with increasing exergy, but exergy does not provide specific information on the resistance and resilience of critical ecosystem functions with respect to potential changes in particular forcing functions. The structural exergy, or that component of total exergy contributed by the internal structure of the ecosystem, rather than by its size, is more specifically sensitive to changes in diversity that are likely to affect ecosystem health. This index can be calculated as $Ex/(R \cdot T \cdot c_{tot})$, where c_{tot} is the total concentration of the element or species of interest in all compartments of the ecosystem. Increases in diversity will tend to increase the structural component of exergy if these diversity increases reflect a fuller utilization of resources within the ecosystem through the realization of additional potential niches.

6.3 Ascendency

Ascendency (an index proposed by R. E. Ulanowicz) is a summary measure of system activity derived from the network of flows associated with that activity, in which each flow is associated with a source compartment and a destination compartment within the system or its surroundings. As such it integrates a measure of the total flux through the system with an estimate of the extent to which the diversity of the fluxes among the respective compartments is organized by preferential paths of flow. (While the relevant compartments within the system must be determined by the ecosystem analyst, all flows from the system are assigned to one of two external compartments [i.e., either export or dissipation], and all flows into the system are assigned to a single import compartment.) When the flows are measured in units of available energy, the total flux through the system (T) is the sum of all available energy fluxes among all the compartments. As such T reflects the vigor of an ecosystem's processing of available energy and thus its capacity for growth and self-renewal under the prevailing conditions at the time of its measurement.

The (Shannon-Weiner) diversity of these fluxes, when multiplied by T , is suggested as an indicator of the total capacity (C) for organizational development. In other words, the number of compartments, evenness of flux among compartments (at any given scale of compartment aggregation), and T all correlate positively with the capacity for organization of flows within a system. In this context, organization refers to the statistical dependence of the destinations and sources of the flows in a network. Thus, in a system with no network organization, the likelihood of a flow of available energy being directed to any given compartment would be independent of which compartment supplied the flow. Because optimal ecosystem health requires that the flow of available energy be selectively directed to its most efficient and productive processes of maintenance, repair, and increase, optimal health will coincide with an optimal organizational development in the sense indicated here, particularly if the selection of compartments by the analyst reflects the intrinsic division of viability-sustaining functions among the ecosystem's critical processes.

The extent to which flow destination is independent of (or unconditioned by) flow source can be expressed as a function of the conditional influx probabilities of the compartments. The conditional

probability of flow to compartment j given that the source of this flow is compartment i is estimated as the flux from i to j divided by the total flux from i (f_{ij}/f_i). The conditional probability of flow from compartment i given that the destination of this flow is compartment j is similarly estimated as the flux from i to j divided by the total flux to j (f_{ij}/f_j). These terms are then used in the formula

$$O = - \sum_{i,j} f_{ij} \cdot \log_2 \left(\frac{f_{ij}}{f_i} \cdot \frac{f_{ij}}{f_j} \right) \quad \text{for } i, j = 1 \text{ to } N,$$

where O (the system overhead) is the product of T and a measure of the independence of destinations and sources that has a maximum value equal to the $\log_2$ Shannon diversity of fluxes within an unorganized network, and N is the total number of compartments. Thus, O attains its (theoretical) maximum value of C in a network in which destinations are completely independent of sources. It attains its minimum value of 0 in a network (with all flows supported by a single import compartment) in which flow destination is completely determined by flow source (i.e., in which knowing the source of a flow removes all uncertainty as to its destination). Accordingly, ascendency ($C - O$) is an indicator of the organization of flows within the network scaled by the total flux through the network. Increases in vigor, in evenness among the total fluxes to (or from) a network's compartments, and in the selective direction of available energy in the network all contribute to a higher ascendency. Increases in the number and evenness of the influxes to each compartment, on the other hand, decrease ascendency while increasing O .

Maximum ascendency requires complete determination of flow destination by flow source, which occurs when flow from each compartment is directed exclusively to a single destination compartment. All ecosystem compartments and processes (maintained as they are away from thermodynamic equilibrium) must contribute some flow to the dissipation compartment, however, in addition to whatever flow they contribute to any other compartments that they support. Thus, this theoretical maximum is not attained. Furthermore, although the direction of all flow from each compartment to a single process might be compatible with maximum long-term efficiency and productivity—and thus with peak ecosystem health—in the absence of any environmental variability, optimal health must be achieved under conditions of varying environmental provision, stress, subsidy, and

disturbance. Thus, the ascendancy attained at optimal health reflects the maintenance of the diversity of flows required to persist and thrive under this environmental variability.

Due to uncertainty with respect to the optimal balance between immediate gains in vigor through selective reinforcement of the most productive flows and the maintenance of other flows that might be critical to vigor or viability under possible or likely future conditions, no precise value for optimal ascendancy is suggested. It might be possible instead to establish ranges in ascendancy typically maintained by persistently thriving ecosystems within given environments and energy regimes. Alternatively, it might be necessary for ascendancy to be used in conjunction with O (or with O and T) in the assessment of ecosystem health. In a plot of ascendancy versus O for a given class of ecosystems, for instance, a region typical of persistently thriving ecosystems might be identifiable. Although ascendancy is sensitive to all of the components of ecosystem health discussed previously, further empirical investigation is required to determine whether sufficient information is retained by this summary measure (if used alone) to provide an accurate or useful indication of ecosystem health. Of the three related measures, ascendancy is most specifically associated with organization selectively reinforced for maximum efficiency and productivity, T is most specifically indicative of ecosystem vigor, and O of adaptive resiliency. A general use of any of these three measures to indicate health status will require comparison among ecosystems across similar scales and standards of aggregation. A more precise characterization of the optimal balances in the allocation of available energy between the most productive flows and potentially critical alternatives under a relevant range of energy and disturbance (or stress) regimes will also be needed.

Acknowledgments

The authors thank Environmental Protection Agency internal reviewers, H. Walker, S. Brandt-Williams, E. Dettmann, C. Strobel, and W. Munns for their helpful comments. This article

is Contribution No. AED-03-096 of the U.S. Environmental Protection Agency's Office of Research and Development, National Health and Environmental Effects Research Laboratory, Atlantic Ecology Division. Although the research for this article has been funded in part by the U.S. Environmental Protection Agency, it has not been subjected to agency review. Therefore, it does not necessarily reflect the views of the agency.

SEE ALSO THE FOLLOWING ARTICLES

Acid Deposition and Energy Use • Aquaculture and Energy Use • Easter Island: Resource Depletion and Collapse • Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development • Ecosystems and Energy: History and Overview • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Hazardous Waste from Fossil Fuels

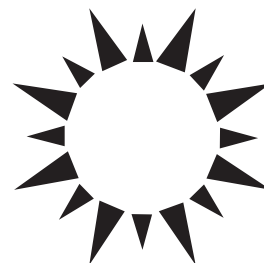
Further Reading

- Campbell, D. E. (2000). Using energy systems theory to define, measure, and interpret ecological integrity and ecosystem health. *Ecosystem Health* 6, 181–204.
- Costanza, R., Norton, B. G., and Haskell, B. D. (eds.). (1992). "Ecosystem Health". Island Press, Washington, DC.
- Jorgensen, S. E. (1997). "Integration of Ecosystem Theories: a pattern." Kluwer Academic, Dordrecht.
- Leopold, A. (1991). A biotic view of land [1939]. In "The River of the Mother of God and Other Essays by Aldo Leopold" (S. L. Flader and J. B. Callicott, eds.), pp. 212–217. University of Wisconsin Press, Madison, WI.
- Lotka, A. J. (1925). "Elements of Physical Biology." Williams and Wilkins, Baltimore, MD.
- Odum, E. P. (1985). Trends expected in stressed ecosystems. *Bioscience* 35, 419–422.
- Odum, H. T. (1994). "Ecological and General Systems." University Press of Colorado, Niwot, CO.
- Odum, H. T. (1996). "Environmental Accounting: EMERGY and Environmental Decision Making." John Wiley & Sons, New York.
- Odum, W. E., Odum, E. P., and Odum, H. T. (1995). Nature's pulsing paradigm. *Estuaries* 18, 54–555.
- Rapport, D., Costanza, R., Epstein, P. R., Gaudet, C., and Levins, R. (eds.). (1998). "Ecosystem Health." Blackwell Science, Oxford.
- Ulanowicz, R. E. (1997). "Ecology, the Ascendent Perspective." Columbia University Press, New York.

Ecosystems and Energy: History and Overview

CHARLES A. S. HALL

State University of New York, College of Environmental Science and Forestry
Syracuse, New York, United States



-
1. Introduction
 2. Energy and the Structure of Ecosystems
 3. Origin and Approach of Ecosystem Studies
 4. Biogeochemical Cycles
 5. Stability and Disturbance
 6. The Future of Ecosystems
 7. Conclusion

Glossary

biogeochemical Pertaining to the abundance and movements of chemical elements or compounds in and between various components of an ecosystem or between ecosystems.

density dependence The situation in which the number of animals or plants at a future time is dependant on the number at an earlier time.

ecosystem A unit of nature, generally from 1 hectare to some thousands of square kilometers, with boundaries defined by the investigator, that includes plants, animals, microbes, and the abiotic environment.

endogenous A factor from inside the system that influences the system.

exogenous A factor from outside the system that influences the system.

trophic Pertaining to food.

We may consider ecosystems as machines that have been selected to use solar and occasionally other energies to reorganize the raw materials of the earth into systems that support life. Ecosystems, and their component organisms, use this energy to rearrange the chemical molecules of the earth's surface and of the atmosphere into living tissue according to those patterns that have high survival potential. They are, in this context, antientropic—that is ecosystems (and life in general) work against the tendency for the

molecules of the world to degrade into relatively random arrangements. As such they are self-designed, in that their organization has emerged through natural selection for structures that capture, hold on to, and invest energy into further structure in a way that appears to be self-designed) the structure is generated to continue life and enhance over time its ability to capture and use energy). More generally, life is most explicitly nonrandom arrangements of molecules, and all organisms use energy to create and maintain these specific molecular arrangements. But individual organisms cannot survive in isolation because they need the supporting context of other species, the soil or other substrate, the atmosphere or water around them, proper hydrological cycling, and so on. Thus, to our knowledge, ecosystems are the minimum units of sustainable life outside of the laboratory. The proper functioning of the earth's ecosystems is also essential for human beings—for atmospheric and climatic stability; for the production of food, fiber, and clean water; and for physical and psychological well-being.

1. INTRODUCTION

1.1 Early Studies at the Ecosystem Level

Although the English term “ecosystem” was coined in 1935 by Tansley, ecosystems have been the subject of human inquiry for millennia. The first written records that survive include the work of Aristotle, an insightful and knowledgeable natural historian, and especially his student Theophrastus, who was interested in how different species of trees grew in different locations. Of course, practical people living on (or off) the land (fishermen, ranchers, hunters, trappers, many farmers) also knew and know a great deal of a practical nature about how ecosystems operate.

Well before Tansley, the formal scientific study of ecosystems was well developed by Russian and Ukraine scientists in the early part of this century. The serious student of ecosystems should return to the seminal, often brilliant earlier writings of Stanchinskii, Vernadaskii, Morozov, and their colleagues. Vernadsky in particular was interested in how ecosystems, and even the living world as a whole (the Biosphere), functioned. Unfortunately the work, and sometimes even the lives, of these early ecologists was terminated by Stalin and his henchman Lysenko because their view of nature's limits to human endeavors was in opposition to official communist social engineering.

The practical implications of understanding energy flow in ecosystems became obvious to a number of early investigators who often attempted to determine how much food could be harvested from a unit area of land or sea. Stanchinskii, mentioned earlier, attempted to examine the quantitative importance of the flows at each trophic (food) level in an Ukrainian grassland. Wells *et al.* examined qualitatively the flow of energy from the sun through the food chain to the herring fisheries of the North Sea as early as 1939. The first study in North America is usually attributed to Raymond Lindeman, who quantified explicitly the flow of energy from the sun through primary producers to higher trophic levels in a bog lake in Minnesota. Lindeman generated most of his numbers from preexisting information but was extremely creative in organizing the organisms of an ecosystem as trophic (or food) levels, rather than simply by species. Although Lindeman made some errors (for example, by subtracting predation after estimating production from turnover rate), his work remains extremely important for our conceptualization of how energy flows through nature.

An especially important group of studies on the energetics of ecosystems were done by Eugene and especially Howard Odum. The brothers worked together measuring photosynthesis in a coral reef at Eniwetok. The most important of these studies was done by Howard on the ecosystem at Silver Springs, Florida. The particular advantage of this site was that the water flowing out of the main spring boil was always at a constant oxygen concentration, so that the deviation of the oxygen content of water downstream (higher during the day, lower at night) could be used to estimate photosynthesis and respiration of the entire ecosystem. Odum also measured explicitly the energy fixed or used by each trophic level by, for example, putting fish into cages and measuring their growth. Odum found that by far the largest propor-

tion of the energy captured at a given trophic level was utilized by that trophic level for its own maintenance respiration, which was consequently unavailable to higher trophic levels. Lindeman had introduced the concept of trophic efficiency, defined as the ratio of production at one trophic level to production at the next. Trophic efficiency is commonly from 10 to 20% but occasionally may be very different. The concept is important and familiar in agriculture where beef or fish production per hectare is much less than the production of plants of that same area, due principally to the large maintenance respiration of both the plants and animals. Both Howard and Eugene Odum continued basic research on the relation of energy to ecosystem structure and function into the new millennia in astonishingly long and illustrative career in research and graduate training.

Research has emphasized that most trophic relations occur not as simple straight-line chains but as more complicated food webs, in which a given species and even different life stages of that species eat from different trophic levels. For example, a herring whose diet contained 50% algae and 50% herbivorous crustaceans would be assigned to trophic level 2.5. Many, perhaps most, organisms are omnivores rather than strictly herbivores or carnivores, and probably the majority eat dead rather than live food. The single most important attributes of food quality, other than its energy content, is its protein content, which is approximately proportional to the ratio of nitrogen to carbon in the food (Table I).

1.2 Ecosystems versus Species-Oriented Ecology

During this time something of a gulf developed between ecosystem-level ecologists and many other ecologists. This was due in part to the perception by

TABLE I
Properties of Food

Food type	Example	Kcal/gm	Percentage protein
Carbohydrate	Grains	4.25	2–14
	Beans	4.25	5–30
Fat	Butter	9.3	0
	Olive oil	9.3	0
Protein	Animal flesh	5.0	20–30

From Whittaker (1975) and USDA nutritional Web site.

some that ecosystem ecologists were paying insufficient attention to the basic laws and mechanisms of nature, including natural selection, that had been especially well developed by biologists at the individual and population level. Some ecosystem-oriented papers appeared to these ecologists to be almost mystical in their message that the system knows best. On the other hand, ecosystem-level ecologists could point to the failure of population dynamics to do a good job in prediction and management of fisheries, in large part because the population-level ecologists tended to ignore spatial and interannual variations in the ecosystem productivity that often determined year class strength. As part of this debate throughout the literature of ecology, there has been considerable discussion and even tension about the relative importance of exogenous (meaning external to the system, such as climate variability, another phrase meaning the same thing is forcing function) controls versus endogenous (meaning internal to the system, such as density dependent survival) controls for determining the structure and function of ecosystems. To some degree the argument is specious, for obviously both are important, although the endogenous aspects have been overly emphasized. Clearly, as described here, climate (an exogenous factor) is critical in determining the basic possibilities for life, and evolution (basically an endogenous factor) is important in determining the response of living creatures in a region to the climate.

Part of the problem leading to this dichotomy was the seemingly completely disparate perceptions of the different groups on what was the important currency of ecology. Energy was perceived as the currency of the ecosystem-level ecologists, and fitness (reproductive potential over time that includes the abilities to both survive and reproduce) was seen as the currency of the population-level ecologists, and each group tended to dismiss, argue against, or ignore the perspectives of the other. These were spurious arguments, and although many previous papers hinted at the importance of energy for fitness, this has been especially well done in papers such as that by Thomas *et al.* titled "Energetic and Fitness Costs of Mismatching Resource Supply and Demand in Seasonally Breeding Birds." This is one of the best papers in ecology, and there is more good science packed into the two-and-one-half pages of this paper than in many books. Simple observations and measurements, and double labeled isotopes, were used to examine energy costs and energy gains of various populations of European tits (chickadees).

Those birds that timed their migrations and reproductions appropriately to the availability of their preferred high-energy food (caterpillars, in turn dependent on the phenology of their oak-leaf food source) were much more fit in the evolutionary sense. That is, birds with optimal timing were much more successful at raising more and larger (and hence more likely to survive) offspring and were themselves more likely to survive and reproduce again the next year. Birds that timed their reproduction inappropriately relative to the caterpillar dynamics worked themselves into a frazzle trying to feed their less successful offspring. Finally Thomas *et al.* found that climate change was interfering with past successful timings because the birds were tending to arrive too late, since the phenology of the trees was responding more quickly to temperature increases than was the impetus for the birds to migrate. Presumably birds that had genetically determined behavioral patterns that were getting them on the breeding grounds too early in the past would be selected for if and as the climate continued to become warmer. This study also indicates the importance of the quality of energy. In this case, it was not simply total trophic energy available to the birds that was critical but also their quality as represented in their concentration as large food packages, showing again the importance of energy quality as well as quantity.

The paper by Thomas *et al.* shows elegantly and more explicitly than earlier studies what many suspected for a long time: probably the principal contributor to fitness is the net energy balance of an organism. The idea that energy costs and gains are related directly to fitness is not a new one, found, for example, as Cushing's match/mismatch hypothesis in fisheries, which in turn is a restatement of Hjort's 1914 paper about year class abundance in Scandinavian cod. Energetics, combined with climatic and other external forcing, can explain population dynamics much more powerfully than the density dependant population equations that still seem to dominate much of population ecology. Clearly density dependence exists. But how often is it the principal determinant of population year class strength? Perhaps the best way to consider the importance of density dependence is what Donald Strong calls density-vague relations—that is the empirical information shows that density dependence works sometimes but does not operate importantly every year. The way that we have set up many of our experiments in ecology virtually guarantees that we will find some density dependence even when it is not dominant.

1.3 General Principles of Energy for Ecosystems

Some general principles about energetics that have emerged from all of these studies are that (1) the world of nature is (of course) as beholden to the laws of thermodynamics as any other system; (2) ecosystems can be considered the products of self design—that is, given the environmental circumstances where they are found they “create” through natural selection (operating at various levels, from genes to organisms to species to coevolved groups of organisms) the systems that we see today; (3) virtually every ecological or even biological action or interaction has an important or even dominant energy component; (4) ecological efficiency, or transfer from one place to another, is low, from one or two to a few tens of a percent, principally because of the need for each level of a food chain to support its own maintenance metabolism; (5) each organism is an energy integrator that integrates through growth and reproductive output (i.e., fitness) all energy costs and gains; and (6) each specific organism is adapted to its own biophysical environment in a way that maximizes this difference between energy costs and gains. As a consequence, the collective energy captured and used by that ecosystem is thought to be maximized; (7) species are ways of packaging morphologies, physiologies, and behaviors that allow sufficient, or even maximum, reproductive output as a consequence of having more or less optimal patterns for that given environment. The ecosystems that have survived over time are thought by many to be maximizing the power (by operating at intermediate, not maximum, efficiency) obtainable under the existing physical circumstances.

More recent research on ecosystems and energetics (i.e., since about 1980) has focused on increasingly detailed examinations of the role and activity of particular species, a greater understanding of the factors that control ecosystems, the importance of disturbance, and recovery from disturbance and the role of microbial components in decomposition and regulation. An important component of this has been the 20 or so LTER (long-term ecological research) centers where basic studies are carried out for long time periods to see how, for example, ecosystem components respond to climate variability, long-term exclusion of herbivores, and so on. This approach has allowed the comparison of the basic patterns of ecosystem.

2. ENERGY AND THE STRUCTURE OF ECOSYSTEMS

Neither ecosystems nor their component species can exist without a constant supply of energy to maintain the biotic structures and their functions. The source of this energy is in almost all cases the sun. Clearly the sun runs the energy, carbon, and nitrogen fixation of green plants. Less obviously the sun does many other things for ecosystems: most important it evaporates and lifts water from the ocean and delivers it to continental ecosystems, replenishes carbon dioxide and other gases through winds, pumps water and associated minerals from the roots to the leaves via transpiration, provides nutrients through weathering, and so on.

Ecosystem studies tend to be different from much of the rest of ecology, and indeed science in general, in their emphasis on interdisciplinary and interactive studies. In most other subareas of science, including ecology, the emphasis is on understanding a specific phenomenon and on carefully controlled studies under exacting laboratory conditions. In contrast, when studying ecosystems it is generally necessary to consider aspects of physics, chemistry, geology, meteorology, hydrology, and biology, and in many cases some of the social sciences as well. Since it is impossible to replicate and control all of these factors carefully, as would be the case in the laboratory, ecosystem studies are more likely to be based on observation or correlation than experimental approaches. In these senses and in others, ecosystem science tends to be holistic—that is, it emphasizes the entirety of both a system and a problem rather than just some aspect, and it frequently uses the conceptual tools of systems analysis. For example, if we are to understand the relation between a particular ecosystem and the atmosphere, we obviously must consider the physics of the response of organisms to temperature, the chemistry of the relation of carbon exchange between plants and the atmosphere, the possible effects of changing climate on hydrology and hence the availability of the water needed for the plant to grow, the movement of gases through the atmosphere, and the social factors that are causing the carbon in the atmosphere to increase. In addition, we often need special integrative tools that are outside the traditional scientific disciplines that allow us to examine all or many of these factors in interaction. These tools include quantitative flow diagramming, computer simulation, and theories that are transcomponent.

Such a holistic, or, more accurately, systems perspective is increasingly important as human

impacts are changing many aspects of the earth rapidly. Such a perspective does not obviate other approaches to science; in fact, it tries to incorporate and integrate as possible all relevant levels and techniques of analysis. The problems that the increasing human population and its growing affluence are putting on the world require a whole different approach to science—one that is difficult if not impossible to do in the laboratory. Nevertheless it must adhere to the same basic standards as the rest of science in that the results of the science must make predictions that are consistent with the real world. This can lead to certain problems when scientists concerned with the destruction of ecosystems and their species are faced with attempting to apply their results to saving nature. The science of ecology is full of generalizations about the supposed virtues of undisturbed nature that have led to theories such as diversity and stability, diversity and productivity, and the intermediate disturbance hypothesis whose validity seems to wax and wane with each new study. Separating the robust ecological theory that will help with the genuine environmental crises we face from the good sounding but poorly extrapolateable theories has often been a difficult feat.

3. ORIGIN AND APPROACH OF ECOSYSTEM STUDIES

The word ecosystem is a combination of the words “ecology” and “system.” The word ecology is derived from the Greek word *Oikos*, which means pertaining to the household, especially to the management of the household. The academic discipline of ecology refers to the study, interpretation, and management of our larger household, including that of the planet that supports our species and that provides the basis for our economy. System refers to all of the (or all of the important) components of whatever is being considered and the pathways and rules of interaction among the components. Thus, ecological system, or ecosystem, refers to the structure and function of the components that make up household earth and the interactions among those components.

Ecosystems generally are considered from the perspective of their structure, that is their composition and changes in structure (in terms of geological land forms, number, diversity, and abundances of species, biomass, height of vegetation, abundance of critical nutrients, etc.), their function, that is the pathways and rates of energy and nutrient flows, and the regulatory or control processes that govern those

changes and flows. We might also want to ask many other kinds of questions about, for example, the diversity of the species that live there or how that diversity effects the function of the ecosystem. Ecosystem structure and function are related over evolutionary time. For example, tall trees functionally need to supply water and nutrients to growing leaf tips, which in the case of the rain forest might be 60 m above the ground. Consequently they have evolved elaborate structures, including a system of pipes, to pump the water to those growing tips. This in turn is possible only where the soil is generally wet and the air is dry, at least periodically.

An extremely useful basic assessment of the most important attributes of ecosystems is found in the classic paper “Relationships between Structure and Function in the Ecosystem” by E. P. Odum. Odum compared the basic attributes of an evergreen forest (such as might be found in the Southeastern United States) and a deep ocean ecosystem (such as might be found a few hundred kilometers to the west of that peninsula). Odum found that these ecosystems are very different in their physical structure—that is the physical characteristics of the ecosystem itself. The organisms themselves constitute the major physical structure in the rain forest. It is possible for the plants to be large here because the soil allows for an anchoring place for the tree, the climate is near optimal for plant growth, and there are no large grazing animals that can eat the tree. Therefore the trees can become very large and become themselves the principal determinant of the physical structure of the ecosystem.

In the ocean, by comparison, the lack of a firm substrate and the constantly changing position of a parcel of water makes it impossible for the organisms to change the nature of the physical medium (the water) significantly, and it is the water itself that forms the basic structure of the ecosystem. Since the organisms are constantly sinking or being swept away by turbulence, they must be small and be able to reproduce rapidly. The total biomass, or living weight, of the aquatic organisms is very small compared to that of the forest. Also, interestingly, the grazers (or animals that eat the plant) have nearly the same biomass as the algae in the water but are much less, as a proportion, in the forest.

One of the most interesting observations when comparing these two ecosystems is that although the biomass of the plants is thousands of times greater in the forest, the rates at which the plants capture the sun’s energy (their primary productivity) is much more similar, especially in coastal areas (Table II). In

TABLE II

Ecosystem Structure and Function of Some Typical Ecosystems as Measured by Approximate Biomass and Production Levels

Ecosystem type	Biomass (dry tons/Ha)	Net productivity (dry tons/Ha/yr)
Tropical rain forest	225 (60–800)	10–35
Tropical seasonal forest	175 (60–600)	10–25
Savanna	40 (2–150)	2–20
Temperate deciduous forest	200 (60–600)	6–25
Temperate evergreen forest	200 (60–2000)	6–25
Boreal forest	200 (60–400)	4–20
Temperate grassland	16 (2–50)	2–15
Desert	7 (1–40)	0.1–2.5
Ocean	0.03 (0–0.05)	0.02–4
Estuaries	10 (0.1–60)	2–35

Based on summary by R. L. Whittaker and G. Likens, in Woodwell and Pecan (1973), modified by Sandra Brown, personal communication.

fact, given the greatly varying quantity of plant biomass around the world, most ecosystems have broadly similar rates of production, at least if there is enough water. Thus, we can say that the structure appears much more variable than the function, at least with respect to these particular attributes.

From an energy perspective, we know that the structure of an ecosystem is costly to build and costly to maintain. “Good” structure—that is structure that has been selected for—will be well worth it energetically for it will draw maximum power (i.e., useful energy sequestered per time) given the environmental possibilities, and it will become more abundant than those components that draw less power. Hence, structure and function, and the controls that link them, are intimately and functionally linked through natural selection.

One of the fundamental ideas used to analyze ecosystems is the idea of food chains, which is based on the transfer of energy originally derived from the sun through the biological system. The word “trophic” comes from the Greek word meaning food, and the study of feeding relations among organisms is called trophic analysis. More specifically, the study of trophic dynamics emphasizes the rates and quantities of transfer over time. Trophic processes are the essence of ecosystem processes, for they are the ways that solar energy is passed from the abiotic world to plants and to animals and decomposers, and they allow nonphotosynthetic organisms, including ourselves, to exist. These pathways of food are also a

principal way that materials such as nutrients are passed from one location to another within an ecosystem. Trophic studies are an important component of ecology because trophic processes determine energy availability and hence what is and what is not possible for a given organism, and for all organisms collectively, within an ecosystem. In addition, many management concepts and objectives important to people, such as those relating to harvesting fish, timber, or crops, are oriented toward understanding and directing the trophic relations of nature. Energy is stored for a short time in the ecosystem in the biomass of the various organisms, soil, and so on, and some of this energy is transferred from one trophic level to the next (Fig. 1). For example, some of the free energy captured by the plants is eaten by heterotrophs, although a larger part is used by the plant itself. Thus, most of the energy once captured is lost from each trophic level as waste heat as organisms use that energy to fight entropy. Eventually, all the energy originally captured by autotrophs is degraded and returned to the surrounding environment as waste heat. Ecosystems do not recycle energy; instead, they capture a small percentage of the solar energy that falls on them, concentrate it, use it to do work, and then return it to the environment in the form of low-grade heat that is no longer available to do work. The principal work that has been done is the maintenance of life, and this has required energy to sequester and rearrange chemicals, overcome gravity, move either themselves (animals) or their chemical constituents (plants and animals), generate microclimates, maintain general homeostasis, and then replicate the whole structure and functions through reproduction.

There is constant selective pressure operating on organisms to maximize the useful energy that they capture and to use it in ways that contribute to their survival and to propel their genes into the future. Many think that this process leads to the collective maximization of power at the level of the ecosystem—that is, those organisms and arrangements of nature will be selected for that most fully exploit available energy and use it most effectively compared to alternative organisms or patterns. Obviously this is a difficult idea to test at the level of entire ecosystems and on a scale of human generations, but it does offer intriguing hints that natural selection operates in ways that generate and maintain (or self-organize) entire systems. For example, although it is unlikely that many of the same species were present in the wet montane tropics 300 million years ago as are there now, it is likely that a recognizable

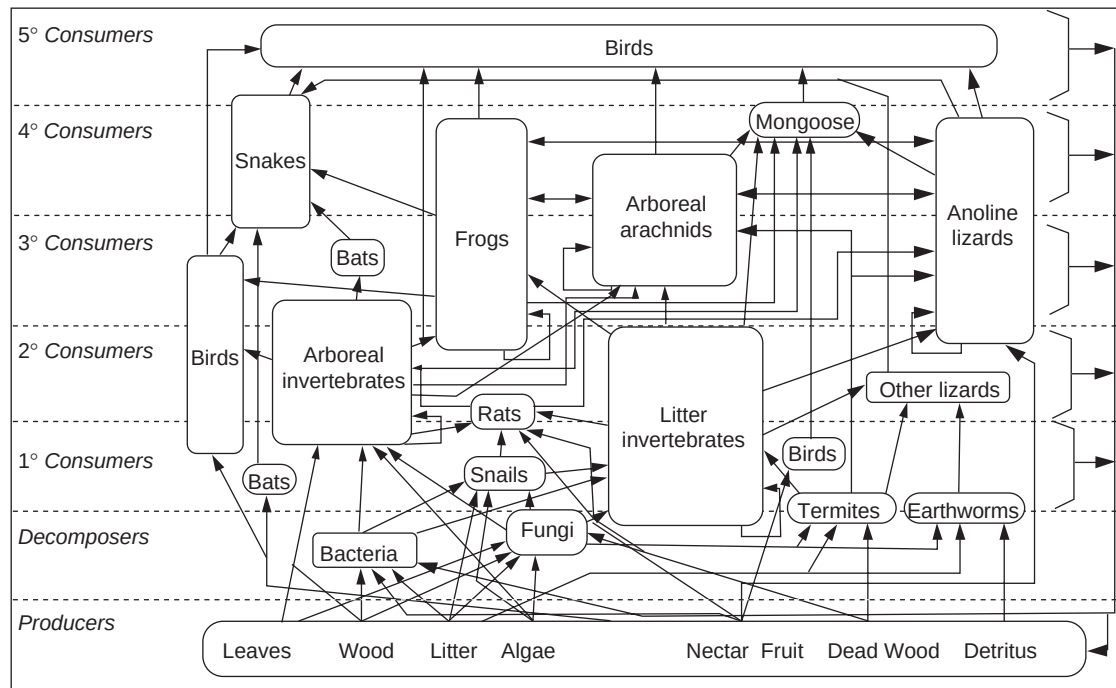


FIGURE 1 Energy flow diagram for the food web of the Luquillo Forest, Puerto Rica. Pictorial representation. Modified from Hall, C. A. S. (1999), *Ecosystems*, pp. 160–168, in D. E. Alexander and R. W. Fairbridge (Eds.), “Encyclopedia of Environmental Science.” Kluwer Academic, Norwell, MA.

tropical wet forest was there that had similar functions and rates of processes as modern tropical wet forests.

3.1 Trophic Processes in More Detail

The pathway of energy conversion and transfer (the eating of one organism by another) goes from the initial capture of solar energy by autotrophs, to the herbivores that eat plants, to the first-level carnivores (that eat herbivores), on to the top carnivores. The principal pathways within a food chain can be represented as energy transfer through a series of steps. The power flow of an organism per unit area, or of a trophic level, is called productivity and normally is expressed in units of kilocalorie per square meter per time.

3.1.1 Autotrophs

In the first step autotrophs, or green plants, use chlorophyll to capture energy from solar-derived photons and store this energy by restructuring the carbon atoms of carbon dioxide derived from the surrounding atmosphere or water into complex

organic compounds. Primary production is the fixation of solar energy by green plants. Gross productivity is total energy captured, whereas net production is that minus the energy required for respiration. Net energy is then allocated into tissue increments, leaf turnover, reproductive products, and, in time and through these processes, the replacement of the first plant with a second. In the second step herbivores or primary consumers obtain energy by eating autotrophs. Secondary production is the accumulation of animal or decomposer living tissue. Heterotrophs obtain metabolically useful energy from the consumption of the organic molecules in the food they obtain from other organisms.

3.1.2 Grazers

Since a substantial amount (perhaps 80 to 90%) of the energy captured by an ecosystem is used by the first trophic level for its own necessary maintenance metabolism, food chains are inherently and necessarily “inefficient.” Thus, there is and must be much more production of biomass of deer than wolves or cougar in a forest, for the deer must use a large

portion of the energy in the leaves it eats for its own metabolism rather than for growth; much less is left for their own growth and the potential exploitation by the predators.

3.1.3 Decomposers

The world is covered with living creatures, and these creatures are continuously dying. Yet the world is not littered with carcasses, so obviously they disappear in some way. We call this process decomposition, and it is mediated principally by single-celled organisms, especially bacteria and fungi. In addition, the decomposition process gets a start initially by larger animals, from buzzards to maggots to protozoans, and in earlier days, even our own ancestors. Although humans are often disgusted by decomposition (and, in fact, have probably been selected for this quality), without decomposition the biological world would come to a stop, as all of the earth's available materials would be tied up in various carcasses!

Some food chains are complicated indeed. Think of a salmon feeding on the plankton of the mid-Pacific ocean and then swimming far up an Alaskan river, fueled by energy reserves built on this plankton, to spawn and die—or to be eaten by a bear. The offspring of the salmon that did spawn spend their first year of life in the freshwater environment, eating zooplankton that ate phytoplankton whose growth was fertilized by phosphorus leaching from the bones of the dead parents. Another way that food chains are complicated is that much of the energy flows not through live but through dead organisms. For example, in the rain forest most of the plant material is not eaten directly by live leaf munchers but instead by various insects and other small organisms that consume the dead leaf material called detritus.

3.2 Trophic Dynamics and Biomass Pyramids

Thus, of all the energy captured by green plants, progressively less and less flows to the next consumer, or, as it is often called, trophic level. When the rate of flow of energy is graphed, it nearly always will look like a pyramid, with the large base representing the energy captured by the autotrophs and each successive layer representing higher trophic levels, each further removed from the base (Fig. 2). This is called the pyramid of energy. Biomass plotted by trophic level may look like this but often looks

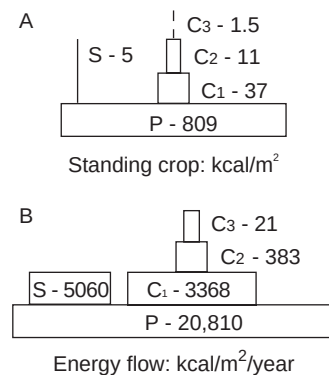


FIGURE 2 The trophic categories given in Fig. 1 represented as (A) an energy pyramid and (B) a biomass pyramid. Data from Silver Springs, FL. Modified from Hall, C. A. S. (1999), *Ecosystems*, pp. 160–168, in D. E. Alexander and R. W. Fairbridge (Eds.), “Encyclopedia of Environmental Science.” Kluwer Academic, Norwell, MA.

more like a square than a pyramid. How can this be? The reason is that larger organisms often found in higher trophic levels generally respire (use up) energy more slowly; in a sense they get less energy, but they hold on to what they do get for a longer time. So in the ocean there is roughly the same biomass of algae, zooplankton, small fish, and large fish. The reason is that the small algae “turn over” much faster, meaning that they have a much higher rate of metabolism and loss to other organisms that eat them.

4. BIOGEOCHEMICAL CYCLES

A second critical area for the study and understanding of ecosystems is the structure and function of nutrient cycles. Nutrients can mean all of the chemical elements that an organism or an ecosystem needs to grow and function, but most commonly the term is restricted to a consideration of nitrogen, phosphorus, potassium, and less commonly calcium, sulfur, iron, and micronutrients such as copper, cobalt, and molybdenum. Sometimes, but not generally, CO_2 is considered a nutrient, for it too can limit growth of plants. Finally, it also (rarely) can mean complex organic materials necessary for growth, such as vitamins. We refer to the movement of a particular nutrient through an ecosystem, often changing from one form to another, as a nutrient cycle. Nutrient cycling, like all other processes, occurs because of energy forces including evaporation, precipitation, erosion, photosynthesis, herbivory decomposition, and so on, all run by the sun.

4.1 Limiting Nutrients

The study of nutrients in ecosystems was first undertaken systematically by the German chemist Liebig, who found that the growth of plants tended to be limited by only one element at a time. For example, if a plant had insufficient phosphorus relative to its needs, its growth would stop until phosphorus was added. At some point adding more phosphorus had no further effects, but at that time some other element, perhaps potassium, might be limiting the plant's growth, as could be determined by noting the growth response, or lack thereof, of adding potassium. This led to the formulation of Liebig's law of nutrient limitation, which states that the growth of plants tends to be limited by a single nutrient at any one time, that which is least abundant or available relative to the needs of the plant. Nutrient cycles at the global level were initiated by Vernadsky and greatly encouraged by the growth of atomic energy in the 1950s, which provided radioactive tracers that helped follow nutrient pathways, funding to do that research, and also a reason as to why we might be concerned about them—that is, if there were releases of radioactive elements into the environment.

Nitrogen is especially important because plants need it in large quantities and it is often restricted in its availability and energetically expensive to get into a useful form. Plants and animals use nitrogen extensively because it is, after carbon, the most important component of proteins. Plants extract it from the soil, and animals get it by eating plants or other animals. All organisms must concentrate the normally relatively rare nitrogen to get enough to build their proteins. Waste nitrogen is released from animals as ammonia or other compounds in their urine. Since N_2 , ammonia, and nitrous oxide are volatile, nitrogen is different from most other nutrients in having a gaseous phase.

The cycle of nitrogen in ecosystems is also complicated because nitrogen is a useful and versatile atom with a valence state that allows for many complex interactions with other atoms. All organisms use nitrogen to make proteins. Some use ammonia as a fuel to gain energy. Others use nitrate for an electron receptor in the same way that humans use oxygen. Still others invest energy to take it out of the atmosphere for their own use. Even though nitrogen is presently about 80% of the atmosphere, plants cannot use it in this form, only after it is fixed (i.e., incorporated as nitrate or ammonia). In nature, this can be done only by the bacteria and blue-green

algae. Thus, all members of all ecosystems are dependant on these two types of life for their nitrogen, or at least they were until about 1915, when humans learned how to fix nitrogen industrially through the Haber-Bosch process.

Evolution has found phosphorus to be an especially useful element, and hence it has a particular importance in ecosystems. Phosphorus is used by all organisms for their genetic material (DNA and RNA) and for energy-storage compounds (ATP), and in vertebrates for teeth and bones. It seems to be in a rather peculiar short supply in the world's biogeochemical cycles relative to the importance given it by evolution. Phosphorus tends to be especially limiting for the growth of plants in freshwater, and even where abundant it may be held very tightly by soils so as to be virtually unobtainable by plants. Phosphorus cycles rapidly between plants and the environment and is passed along food chains as one organism eats another (Fig. 3). When organisms die, they decompose, and we say that the nutrients such as phosphorus are mineralized (returned to their nonliving state). Phosphorus is easily dissolved in water, and in many respects the cycle of phosphorus at the ecosystem level is the cycle of water.

4.2 Redfield Ratios

It is not only the abundance of each nutrient that is important, but also their concentration relative to each other. The importance of this idea was first worked out by the oceanographer Alfred Redfield, who sailed around the world's seas on the first oceanographic ship, the *Challenger*, meanwhile taking samples of the oceans and its various life forms. Redfield's chemical analysis was a rather difficult undertaking, since the surface of most of the earth's seas are low in nutrients because the constant sinking of organisms and their feces tends to deplete nutrients in the surface of the sea. Nevertheless, Redfield was rather astonished to find that both the sea and the small plants of the sea had nearly constant ratios of nitrogen to phosphorus, about 15 atoms of nitrogen for each one of phosphorus, in both the living and the nonliving environments. Since we know that aquatic plants need roughly 15 atoms of N for each one of P, then we can determine whether the environment is N or P limited by whether it has more or less than this ratio. If the water is 20 to 1 it is likely to be P limited, and if it is 10 to 1 it is likely to be N limited. Likewise agricultural extension agents can analyze different soils and determine which nutrient is limiting, saving

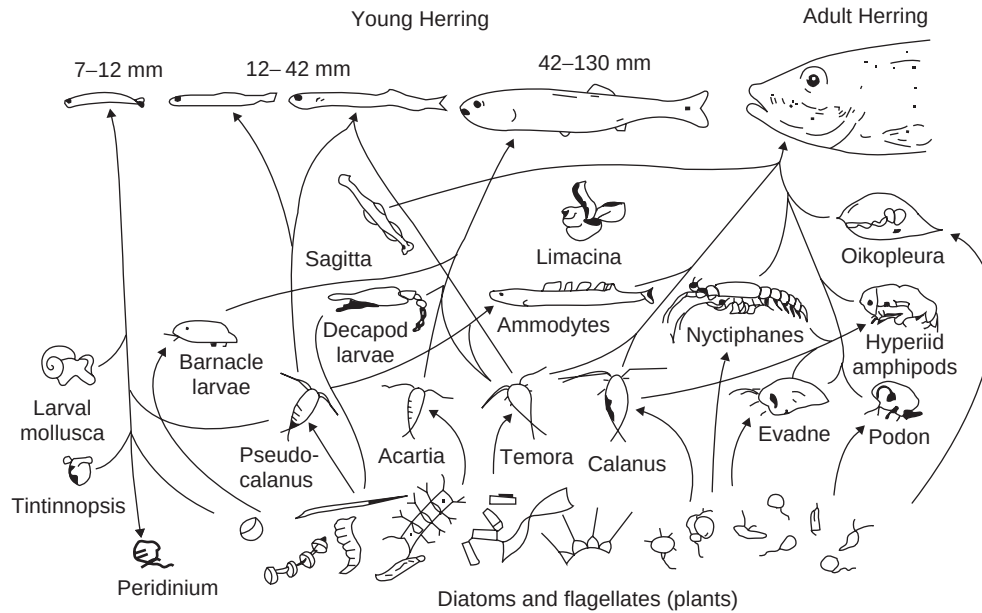


FIGURE 3 An early example of a food chain (from Wells *et al.*, 1939). In this case, the energy is represented as flowing from right to left, the opposite of later convention.

the expense and environmental problems of pouring on unneeded nutrients.

5. STABILITY AND DISTURBANCE

A question that has intrigued many ecologists is the degree to which ecosystems are stable—that is relatively homeostatic or self-correcting. It seems to many early investigators that natural ecosystems tended to have self-correcting properties, so if they were disturbed they would bounce back toward their original conditions.

5.1 Succession

Hence earlier ecological studies tended to emphasized the equilibrium nature of ecosystems. The idea was that there was a strong tendency for ecosystems to “evolve” toward a “climax” state determined by the climatic conditions of the site. In this view, succession was an orderly process that would bring the system back to climax, a condition of stability with no more species change. Climax was considered the normal state of ecosystems, and considerable virtue was associated with this process, a sort of “mother nature knows best.” The view is evolving that ecosystems are continuously subjected to disturbance, and are continuously responding in various

ways to the various disturbances they are subject to. For example, tropical ecosystems were once thought particularly unchanging. Research has shown that many tropical forested ecosystems are continuously in disequilibrium—that is, recovering in some way from some externally imposed stress or pulse, such as a hurricane, a landslide, a drought or, locally, a tree fall.

5.2 The Role of Species Diversity

An unresolved issue in ecology is to what degree all of the different species found in an ecosystem are required for its proper functioning. Twenty-five years ago, the idea that diversity (of species) and stability (of populations and ecosystems) were positively connected was very popular. At this time, the best that we can say is that certainly an ecosystem needs many species to function properly and that through coevolution many species are tightly linked to each other. But we do not know in a general sense whether ecosystems *need* all, or even most, of the plethora of species that are found there in order to continue their basic structure and function. At the same time, we are finding an increasing number of examples where animals once thought rather superfluous with respect to ecosystem function are in fact critical for some important function such as distributing or germinating the seeds of a dominant

plant. Species that have a disproportionate effect on the entire ecosystem are called Keystone species. An example is the sea otter of California-to-Alaska coastal waters, which maintains kelp beds by eating sea urchins that would otherwise overgraze the kelp. Additionally we are understanding better that ecosystems themselves can be much more resilient than many of their component species, since if one species is knocked out, another (although perhaps one less desirable from a human perspective) can sometimes take over that function. This is an important question as all over the world humans are reducing the diversity of ecosystems, and they are especially impacting the large animals.

5.3 Disturbance

An important aspect of ecosystems that we have learned is that although we once thought of ecosystems as, generally, stable and focused on the tendency toward equilibrium that occurred through the process of succession, we now view ecosystems as much more transient phenomena, generally subject to, and responding to, external forces, which we normally term disturbances. For example, graduate students are often taught that tropical forests were “stable,” due principally to a lack of winter and glacial-cycle impacts. In reality, many tropical forests, for example the Luquillo experimental forest in Puerto Rico, are routinely subject to three types of natural disturbances: tree falls, landslides, and hurricanes, so that the forest is almost always responding to, or recovering from, disturbances.

Some important nutrient cycling studies were done at the U.S. Forest Service experimental watersheds at Hubbard Brook, New Hampshire. Here entire watersheds were cut, and vegetation regrowth suppressed, to examine the role of the vegetation in determining nutrient retention by ecosystems (it was substantial). Even so, the scientists there also found that long-term trends in rainfall intensity were equally important in determining the nature of the ecosystem. In other words, even in a relatively undisturbed forest there was a large amount of what we might call disturbance introduced simply by the normal decade-scale cycles of drought and flood, and this greatly interfered with the evolution of the ecosystem toward what we might have once called mature. Thus, at the level of the ecosystem it is clear that both endogenous and exogenous factors are important in determining the structure of the ecosystems and the nature of nutrient movements.

5.4 The Role of Human Disturbance

We previously discussed the importance of natural disturbance in determining the structure and function of existing ecosystems. Unfortunately, often it has been frustratingly difficult to unravel the impact of natural factors on ecosystem structure and function because almost all ecosystems of the world have been heavily impacted by prior human activity. For example, it was widely believed that New England forests were in some kind of pristine condition prior to the American Revolution, or even prior to European colonization. Instead we now know that the New England landscape had been modified severely by the activity of both Native Americans and very early colonists.

Likewise Perlin has documented the intensive deforestation that has taken place in virtually all portions of the world over the past thousands of years due to human activities. The basic cycle, repeated over and over again, has been that initially human settlers had relatively little impact on the forested lands except to clear small areas for agriculture. Often this land was abandoned after fertility declined and a new patch cut while earlier patches reverted to forests. This was not too different from the patterns of natural disturbance. But then, characteristically, some metal was discovered. Huge areas of forests were cut to provide the charcoal to smelt the metal, and the net effect was extensive deforestation. Subsequently there tended to be an increase of agricultural production on these newly cleared and fertile lands, followed by large human population increases, and then substantial soil erosion, river and harbor sedimentation, and eventual agricultural and population collapse. For example, ancient Greece was deforested 4000 years ago and then recovered, only to be again deforested, recover, and be deforested again in relatively modern times. The use of chemical fertilizers has allowed agriculture to be maintained over a longer time period, so that the land is not abandoned to reforestation. Curiously, some few parts of the world, notably the Eastern United States and Western Europe, are now undergoing reforestation because the exploitation of fossil fuels has allowed the forests a respite, albeit perhaps only temporarily.

There are other ways in which people have and continue to impact natural ecosystems. Humans have changed the relative abundance of species through the introduction of new species and through the exploitation of natural populations. For example, the introduction of goats into the Canary Islands 800 years ago

resulted in wholesale destruction of the natural vegetation, and the unique flora and fauna of islands such as the Galapagos and Hawaii are being changed dramatically by human-caused introductions, many of them centuries old. Curiously the flora and fauna of the island of Puerto Rico seem to be much less impacted by invasive species, and when the vegetation is allowed to recover from human disturbance, native species predominate.

One consequence of all of this is that it is difficult for ecologists to measure what is “natural” in nature, for almost all nature that we have had an opportunity to observe is already quite different from what might have existed some tens of thousands of years ago. Meanwhile, climate change, at both a local and global level, are impacting all organisms in every ecosystem, subjecting them to different temperature and rainfall regimes and to different carbon and nitrogen levels than those within which the organisms evolved. We are learning that ecosystems can adapt surprisingly well to at least some of these changes, but we do not know how long this will continue into the future.

6. THE FUTURE OF ECOSYSTEMS

The natural ecosystems of the earth are tremendously and increasingly impacted by human activities. For example, Vitousek and his colleagues have estimated that fully 40% of the earth’s primary production is directly or indirectly exploited by our own species, and others believe this proportion is higher. This is remarkable considering that we are but one of millions of species on this planet, and that about two-thirds of the planet is water and as such is difficult for humans to access. People are probably not doing anything conceptually different from other organisms in nature, which also tend to eat or otherwise exploit whatever resources are available as much as possible. Natural selection almost certainly has hardwired us for lust and something resembling greed. The difference is that agricultural, medical, and especially fossil fuel technology have allowed people to be enormously more effective at exploiting resources compared to other species, or indeed compared to our earlier selves. At the same time, the earth’s ability to check our own population levels through predation and disease has been greatly reduced. The net result is that the human impact has grown enormously.

The consequences of this tremendous rate of exploitation are only beginning to be understood.

They are not necessarily bad from the perspective of ecosystems, at least if we measure good and bad in terms of their biomass and productivity. For example, Nemani and his colleagues assessed the impact of changing climates and nutrient cycles on global vegetation and found that there had been a probable increase in productivity of ecosystems in the temperate regions and a decrease in the tropics. But in a conference on the sources and sinks of carbon, Sampson and his colleagues concluded that human activity was turning what had been natural sinks of carbon for most of the world’s ecosystems into carbon sources, probably exacerbating the greenhouse effect. Although various scientists argue whether or not human activity has impacted the earth’s climate to date, that is not really the issue. We have changed the concentration of carbon dioxide in the atmosphere by only about 40% so far. If civilization continues along its present path, the change could be 500 to 1000%. Thus, natural ecosystems, and the services that they provide, are likely to continue to decrease in area and naturalness, since the human population continues to grow—in fact, to increase more rapidly than ever before—and most of the earth’s human citizens have increasing desires for material goods, which can come only from the earth itself. Young people reading this article will almost undoubtedly see a disappearance of most of the world’s natural ecosystems during their lifetimes unless there is an extraordinary effort to reduce human impact or unless that occurs involuntarily.

7. CONCLUSION

Ecosystem science got a relatively slow start in America because of an early species-oriented, stamp collecting attitude toward the large, fascinating, and largely unknown diversity of nature in the new world. In addition, there was competition for funding, attention, and prestige from other levels of ecological and biological inquiry, a largely false sense of the power of formal mathematical analysis that came into vogue in ecology, and the fact that ecosystem studies often were expensive. But in the 1950s, ecosystem studies began to prosper, propelled in part by new techniques for studying the energetics of whole aquatic systems and also by curiosity and funding related to how radioactive materials might move through ecosystems if released by accident or war. Important studies of grasslands, forests, and lakes in the 1960s quantified many basic aspects of

ecosystems and began the process of constructing ecosystem models. In the 1970s, an ecosystems approach to watersheds in New Hampshire (Hubbard Brook) and North Carolina (Coweeta) showed how an experimental approach could be used to address many basic aspects of understanding ecosystems, and Howard Odum began his analyses of humans as components of ecosystems. The importance of ecosystems-level studies is becoming more and more obvious, and ecosystem studies have changed from a relatively obscure part of biology to a healthy and vibrant science all of its own. For example, the U.S. Forest Service in the early 1990s changed the entire focus of its management for our national forests from one based on timber yield to one focused on managing the forests as ecosystems. Another important trend is that ecosystem approaches are moving from the temperate latitudes increasingly to other regions, such as the tropics and tundra. One next phase, barely started, is to use ecosystem approaches to study cities and the relation of cities to natural or seminatural systems. Perhaps our understanding of how natural systems have evolved to prosper in the circumstances of relatively limited and stable energy supplies can give us some useful information as humans, presumably, will have to adapt to a more energy-restricted future.

Acknowledgments

This article is based in part on Hall, C. A. S. (1999), *Ecosystems*, pp. 160–168, in D. E. Alexander and R. W. Fairbridge (Eds.), “Encyclopedia of Environmental Science.” Kluwer Academic, Norwell, Massachusetts.

SEE ALSO THE FOLLOWING ARTICLES

Easter Island: Resource Depletion and Collapse • Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development •

Ecosystem Health: Energy Indicators • Environmental Gradients and Energy • Heterotrophic Energy Flows

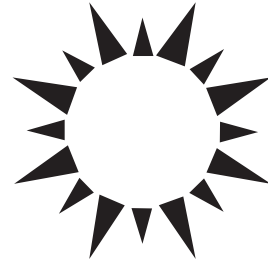
Further Reading

- Brown, M. T., and Hall, C. A. S. (eds.). (in press). An H. T. Odum primer. “Ecological Modeling.”
- Craig, B. J. (2001). Eugene Odum. “Ecosystem Ecologist and Environmentalist.” University of Georgia Press.
- Golley, F. (1993). “A History of the Ecosystem Concept in Ecology: More Than the Sum of the Parts.” Yale University Press, New Haven, CT.
- Hagen, J. P. (1992). “An Entangled Bank: The Origins of Ecosystem ecology.” Rutgers University Press, New Brunswick, NJ.
- Hall, C. A. S. (ed.). (1995). “Maximum Power: The Ideas and Applications of H. T. Odum.” University Press of Colorado, Niwot, CO.
- Hall, C. A. S. (1988). An assessment of several of the historically most influential models used in ecology and the data used in their support. *Ecol. Modelling* **43**, 5–31.
- Likens, G., and Borman, H. (1995). “Biogeochemistry of a Forested Ecosystem.” 2nd ed. Springer Verlag, New York.
- Lindeman, R. (1942). The trophic-dynamic aspect of ecology. *Ecology* **23**, 399–418.
- Odum, E. P. (1962). Relationships between structure and function in the ecosystem. *Japanese J. Ecology* **12**(3), 108–118. Reprinted in Bobs Merrill College Division reprint series in the Life Sciences.
- Odum, H. T. (1957). Trophic structure and productivity of Silver Springs, FL. *Ecological Monographs* **27**, 55–112.
- Odum, H. T. (1988). Self-organization, transformity and information. *Science* **242**, 1132–1139.
- Odum, H. T. (199X). *Systems Ecology*. Wiley Interscience. Reprinted as “General and Ecological Systems.” University Press of Colorado, Niwot, CO.
- Thomas, D. W., Blondel, J., Perret, P., Lambrechts, M. M., and Speakman, J. R. (2001). Energetic and fitness costs of mismatching resource supply and demand in seasonally breeding birds. *Science* **291**, 2598–2600.
- Wiener, D. (1988). “Models of Nature: Ecology, Conservation and Cultural Revolution in Soviet Russia.” Indiana University Press, Bloomington, IN.
- Woodwell, G. M., and Pecan, E. V. (eds.). (1973). Carbon and the biosphere. Brookhaven Symposium in Biology, 24. National Technical Information Service. Springfield, VA.

Electrical Energy and Power

JOSEPH PRIEST

Miami University
Oxford, Ohio, United States



1. Atomic Structure
2. Electricity
3. Electrical Resistance
4. Electric Power and Electric Energy
5. Magnetic Force and Magnetism
6. Electric Generators
7. Transformers

Glossary

alternating current (AC) An electric current that alternates direction at regular intervals.

ampere (A) Unit of electric current. One ampere is a rate of flow of charge equal to 1 C/s ($I = Q/t$).

atom An atom consists of a dense, positively charged nucleus surrounded by a system of electrons equal in number to the number of protons in the nucleus. The atom is bound together by electric forces between the electrons and the nucleus.

atomic number The number of electrons (or protons) in a neutral atom.

coulomb (C) The metric unit of electric charge. In terms of the charge of the electron, the sum of the charges of approximately 6 1/4 billion-billion electrons equals 1 C.

direct current (DC) An electric current for which charges flow in one direction.

electric current Electric charges in motion. The ampere is the measuring unit.

electric field Qualitatively, a region of space in which an electric charge feels a force.

electric force A force between two objects that each have the physical property of charge.

electrical resistance The resistance offered by the structure of a conductor to the flow of electric charges.

hertz (Hz) Unit of frequency equal to one cycle per second. Named for Heinrich Hertz (1857–1894), a German physicist.

kilowatt-hour (kWh) A unit of energy equal to a power of 1 kilowatt (1000 W) acting for 1 h.

magnetic field Qualitatively, a region of space in which the pole of a magnet or a moving charge feels a force.

magnetic field line An imaginary line in a magnetic field along which a tiny compass needle aligns.

nucleus (of an atom) The positively charged central region of an atom. It is composed of neutrons and protons and contains nearly all the mass of the atom.

ohm Unit of electrical resistance. If a potential difference of 1 V across some electrical element causes a current of 1 A, the electrical resistance is 1 ohm ($R = V/I$).

Ohm's law If the electrical resistance of a device (e.g., a toaster element) does not depend on the current in the device, then it is said to obey Ohm's law.

potential difference The work required to move an amount of charge between two positions divided by the strength of the charge ($V = W/Q$). The volt is the measuring unit.

power Rate of doing work. The watt is the metric measuring unit.

transformer An electromagnetic device used to change the strength of an AC voltage.

volt (V) Unit of potential difference. If 1 J of work is required to move 1 C of charge between two positions, the potential difference between the positions is 1 V ($V = W/Q$).

voltage A common expression for potential difference.

watt (W) The metric unit of power. A rate of doing work of 1 J/s is a watt ($P = W/t$).

Electric charge is a property of matter in the same sense that mass is a property of matter. These two properties produce two distinct types of force. Two objects experience a mutual gravitational force because of the property of mass. We see the result of a gravitational force when a pencil is dropped and pulled toward Earth. An electric force is felt by each of two objects possessing the property of charge. Unlike the gravitational force, which always tends to pull masses toward each other, electric forces can tend to pull objects together or push objects apart. If the electric force is attractive, we say the charges acquired by the two objects are unlike. The charges are like when the two objects repel. Two unlike charges are distinguished by calling one positive and the other negative. A toy balloon acquires an electric charge when rubbed on one's hair. The charge on the

balloon is negative. Any charge repelled by the balloon is negative; any charge attracted is positive.

It is easy to relate to the property of mass because mass is a measure of the quantity of matter. Charge is a much more subtle property because we cannot see charge; we see only the results of forces produced by charge.

In 1784, Augustin de Coulomb discovered that the force between two charges labeled q and Q and separated a distance r obeys a law of the form

$$F = k \frac{qQ}{r^2}. \quad (1)$$

In the metric system of units charges q and Q are measured in Coulombs (C), r is measured in meters, and the constant k has the value

$$k = 8.99 \times 10^9 \frac{\text{newton meter}^2}{\text{coulomb}^2}.$$

Electric charge, electric force, and energy associated with the electric force play roles ranging from the structure of atoms to the liberation of energy in either a nuclear-fueled or coal-fired electric power plant.

1. ATOMIC STRUCTURE

The charge acquired by rubbing a balloon as well as electric current in a wire has its roots in atoms, which are the basic structural components of matter. An elementary model of an atom incorporates features of an Earth-satellite system. Electrons are satellites and a nucleus acts like Earth. An electron has a mass of $9.11 \times 10^{-31} \text{ kg}$ and a negative charge of $-1.60 \times 10^{-19} \text{ C}$. A nucleus is positively charged and is much more massive than an electron. The mass of the nucleus is confined to a volume approximating a sphere of radius of approximately $1 \times 10^{-15} \text{ m}$. The closest electron is approximately $1 \times 10^{-10} \text{ m}$. In perspective, if the radius of the nucleus were approximately the thickness of a dime, the nearest electron would be approximately the length of a football field away. Having both mass and electric charge, electrons and a nucleus experience gravitational and electrical forces that tend to bind the electrons to the nucleus. However, the overwhelming binding is due to the electric force.

In a normal atom, the total negative charge of the electrons balances the positive charge of the nucleus and the atom is said to be electrically neutral. The number of electrons in a neutral atom is called the atomic number, which distinguishes different atomic species. For example, a carbon atom has 6 electrons

and an oxygen atom has 8 electrons. Most electrical wires in homes are made of copper. Each atom of copper has 29 electrons. As the atomic number increases, the mass of the atom also increases. This results from the additional electrons and by an increase in the mass of the nucleus.

2. ELECTRICITY

A current is defined as “a steady and smooth onward movement, as of water.” A steady movement of cars on a highway would be a car current. The rate at which cars pass a point on the highway is a measure of the car current. To determine the car current, one would count the cars, measure the time for the cars to pass, and divide the number by the time to obtain the rate in units such as cars per minute. If each car had two passengers, then multiplying the car rate by two yields a passenger current measured in passengers per minute.

An electric current is a flow of electric charges. In wires and appliances the flow is due to electrons because they are much freer than the positive charges in the nucleus. Electrons moving through a wire are analogous to cars moving on a highway. One cannot see the electrons but in principle one can stand at the edge of the wire and count the electrons passing by in a measured time interval and express the rate in units of electrons per second. Multiplying the electron rate by the charge per electron yields the rate at which charge flows in units of coulombs per second. As an equation,

Electric current

$$= \frac{\text{coulombs of charge flowing through a wire}}{\text{time required for the flow}}$$

$$I = \frac{Q}{t}. \quad (2)$$

Recording charge (Q) in coulombs and the time interval (t) in seconds yields coulombs per second as the units of electric current (I). A coulomb per second is called an ampere (A). If 100 C of charge flow by some position in a wire in 10 s, the current is 10 A. The electric current in a lit flashlight bulb is approximately 1 A. If the bulb were on for 10 min, the total charge passing through the bulb in 10 min is

$$1 \text{ C/s} \times 10 \text{ min} \times 60 \text{ s/min} = 600 \text{ C}.$$

A force must be exerted on the electrons to cause them to flow through a wire just as a force must be applied to water to keep it moving through a pipe. The electric force has its origin in the force between

electric charges. Electrons pulled toward a positive charge create an electric current. Electric forces do work on the electrons just as work is done on an automobile pushed some distance by a person. It is the role of a battery in a flashlight or an electric generator in an electric power plant to supply electrical “pressure” for moving electrons. This electrical pressure is called a potential difference (voltage is a less formal word describing the same idea). A potential difference is analogous to a water pressure difference for moving water through a pipe. Potential difference is defined in terms of the amount of work done by electric forces on a charge as it moves between two positions:

Potential difference

$$= \frac{\text{electrical work on the amount of charge moved}}{\text{amount of charge}}$$

$$V = \frac{W}{Q}. \quad (3)$$

Work (W) is measured in joules (J) and charge (Q) is measured in coulombs, making potential difference (V) joules per coulomb. A joule per coulomb is called a volt (V). If 100 J of work is done by electric forces on 1 C of charge when it moves between two positions, the potential difference between the two positions is 100 V. Batteries for most automobiles provide a 12-V potential difference between two terminals. If 2 C of charge flow between the terminals, electrical forces do 24 J of work on the 2 C of charge.

The notion of a difference in potential is very important. If a building is said to be 500 ft high, this probably means the difference in “height” between the top and bottom is 500 ft. Clearly, however, if the building were on top of a mountain, its height would be more than 500 ft relative to the base of the mountain. Potential differences, like heights, are also relative. The potential difference between the terminals in a household electrical outlet is 115 V. This means that one of the contacts is 115 V relative to the other contact. This other contact in a household outlet is at the same potential as any other part of the room, including a person in the room. You can touch this contact and not get an electrical shock because there is no potential difference between you and the contact. This contact is called ground because a wire connects it to a metal stake driven into the ground. Generally, this ground wire can be seen alongside an electric utility pole holding wires that bring electricity to a house. The other contact in an outlet is referred to as “hot.” Electrical shocks occur when you touch both ground and the hot contact.

All electrical appliances have two electrical connections. To operate an appliance, an appropriate potential difference must be connected to these two connections. If two wires were connected to a flashlight bulb and the other ends of the wires touched to the terminals of a flashlight battery, the bulb would light. The lit condition signals a complete electrical circuit. It is complete in the sense that a complete (or closed) path is provided for the flow of electrons. Cutting one wire with a pair of scissors breaks the circuit and the light no longer glows. A switch is a device that allows opening and closing electrical circuits. A switch is analogous to a faucet that can be opened or closed to control the flow of water. An example of a water circuit is shown in Fig. 1. A pump pulls water from the container, forces it through a valve, past a water wheel, and back into the container. In a household electrical circuit, an

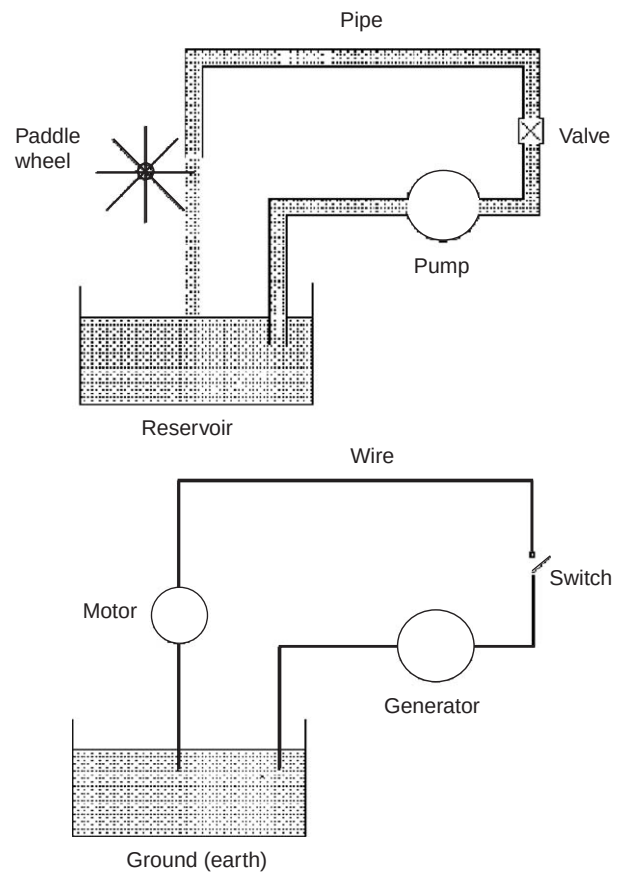


FIGURE 1 Schematic comparison of a fluid circuit and an electrical circuit. The water pump forces water to circulate through the water circuit. The electrical pump, a battery or generator, forces electrons to circulate through the electrical circuit. The sum total of electric charges never changes in the process, just as in the absence of leaks the sum total of water never changes in the water circuit.

electric generator provided by an electric utility pulls electrons from the ground, forces them through a switch, through a motor, for example, and back into the ground. The wire plays the role of a water pipe.

3. ELECTRICAL RESISTANCE

Electrons moving through a wire migrate through a network of atoms. This network impedes the flow, and we say the wire has electrical resistance. The resistance depends on the composition of the wire. For example, an aluminum wire having the same length and diameter as a copper wire offers more resistance than the copper wire. Just as water flows easier through short pipes than through long pipes, electrons flow easier through short wires than through long wires. Also, just as water flows easier through large-diameter pipes than through small-diameter pipes, electrons flow easier through large-diameter wires than through small-diameter wires. Accordingly, the resistance of a wire increases as the length increases but decreases as the diameter increases.

A current is produced in a wire when the ends are connected to a potential difference. The current increases if the potential difference (the electrical pressure) increases. If the resistance increases, the current decreases. This reasoning is modeled as an equation by writing

$$\text{Current} = \frac{\text{potential difference}}{\text{resistance}}$$

$$I = \frac{V}{R}. \quad (4)$$

Resistance has units of volts divided by amperes. This unit is called an ohm. If 1.5 A is produced in a bulb connected to a 12-V battery, the resistance of the bulb is

$$R = V/I = 12 \text{ V}/1.5 \text{ A} = 8 \text{ ohms}.$$

The resistance need not be constant. Doubling the potential difference across the bulb may produce a current of 2 A and a resistance of 6 ohms. If the resistance does not depend on the potential difference, the device obeys Ohm's law. Wires used for electrical conductors and heating elements in toasters and irons obey Ohm's law. Transistors do not obey Ohm's law.

In a household in which the potential difference is fixed at 115 V, the current in an appliance depends on its resistance. As the resistance decreases, the current increases. The energy used to overcome this electrical resistance appears as heat. For example, the heat liberated in a toaster is produced by electrons

overcoming the electrical resistance provided by the metallic elements of the toaster.

4. ELECTRIC POWER AND ELECTRIC ENERGY

Like water pulled over a dam by a gravitational force, electrons moved by an electric force acquire kinetic energy as a result of work done on them. Electrons convert this energy in a variety of ways. In a lamp, the energy is converted to heat and light. In an electric motor, the energy is converted to mechanical rotational energy. The rate at which moving charges convert energy is called electric power:

$$\text{Power} = \frac{\text{energy converted in joules}}{\text{time required in seconds}}$$

$$P = \frac{E}{t}. \quad (5)$$

Electric power is expressed in joules per second, or watts. Electric clothes dryers use electric power in the range of a few thousands of watts; electric power plants produce power in the range of hundreds of millions of watts. Thousands of watts and millions of watts are usually expressed as kilowatts (kW) and megawatts (MW), respectively.

A useful relation for electric power follows from combining the equation for the definition of power, $P = W/t$, and the definition of potential difference, $V = W/Q$. Substituting for W in the equation $P = W/t$, we have $P = V(Q/t)$. Recognizing that Q/t is current I , we find $P = VI$. Power equals voltage times current is a general relationship for any electrical device. For a resistance, we can use $V = IR$ to arrive at $P = I^2R$ and $P = V^2/R$.

Companies that provide the voltage for forcing electrons through our appliances are often referred to as power companies. It is reasonable to expect that the companies are paid for power. However, they are actually paid for energy, not power.

Power is the rate of converting energy. A 100-W light bulb requires energy at a rate of 100 J/s regardless of how long it is on. However, the cost of operation depends on how long the bulb is lit. Using Eq. (5), it follows that energy, power, and time are related by

$$\text{Energy} = \text{power} \times \text{time}$$

$$E = Pt.$$

It follows that any unit of power multiplied by a unit of time yields a unit of energy. Watts (or kilowatts)

TABLE I
Power Requirements of Common Household Appliances

Appliance	Power (W)
Cooking range (full operation)	12,000
Heat pump	12,000
Clothes dryer	5,000
Oven	3,200
Water heater	2,500
Air conditioner (window)	1,600
Microwave oven	1,500
Broiler	1,400
Hot plate	1,250
Frying pan	1,200
Toaster	1,100
Hand iron	1,000
Electric space heater	1,000
Hair dryer	1,000
Clothes washer	500
Television (color)	330
Food mixer	130
Hi-fi stereo	100
Radio	70
Razor	14
Toothbrush	7
Clock	2

and hours are units of power and time. The watt-hour (or kilowatt-hour) is a unit of energy. It is peculiar because the units of time (hours) do not cancel the time units of seconds in power.

Table I lists the power requirements for several household appliances. Note that those involving the generation or removal of heat require the most power. These devices tend to have low resistance.

5. MAGNETIC FORCE AND MAGNETISM

We have discussed electric current and the idea that an electrical pressure is required to establish electric current. Nothing has been said about how the power company generates the pressure or how it is transmitted to homes and factories. To understand this important aspect of producing electricity, we must discuss some fundamentals of magnetic force and magnetism.

Images on a TV are produced by electrons impacting with the screen. The picture distorts if a magnet is

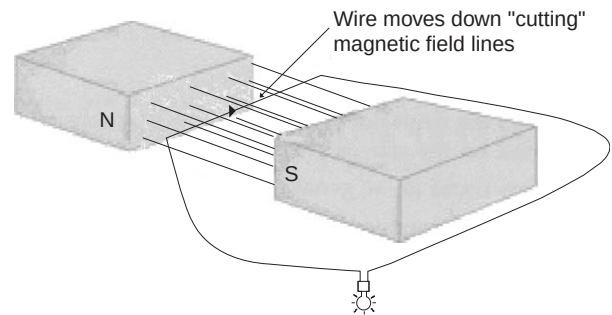


FIGURE 2 A current is established in a wire when it moves in a magnetic field so as to “cut” the magnetic field lines. If the wire is moved directly from the N to the S pole, no magnetic field lines are cut and no current results. The arrow indicates the direction in which electrons move in the wire.

placed near the face of a black-and-white TV. The distortion results from an interaction of the moving electrons with a magnetic field produced by the magnet. As a general principle, any moving charge “cutting” magnetic field lines experiences a magnetic force. In Fig. 2, magnetic field lines are directed from the N pole to the S pole. A wire moved vertically in the magnetic field “cuts” the magnetic field lines and all charges in the wire experience a force by virtue of moving in a magnetic field. However, only the free electrons in the wire acquire motion. If the ends of the wire are connected to a light bulb, for example, an electric current is produced in the circuit. Had the wire been moved horizontally parallel to the magnetic field lines, there would have been no current in the circuit. Work done by the agent moving the wire is responsible for the energy acquired by the electrons. This is the principle of an electric generator such as one might find on a bicycle or in an electric power plant. On a bicycle generator, the agent moving the wire is a wheel rotating wires mounted on the shaft of the generator. A large steam turbine is the agent in an electric power plant.

The generator principle was illustrated by holding a magnet fixed and moving the wire in the magnetic field. A current would also be produced if the wire were held fixed and the magnet moved. As long as the wire cuts the magnetic field lines, a current results. If the wire were moved in a direction opposite to that shown in Fig. 2, the electric current changes direction. Cutting magnetic field lines with an oscillatory motion produces a current whose direction alternates. Such a current is called an alternating current (AC). The current produced by a battery is called direct current because the charge flows in one direction only.

6. ELECTRIC GENERATORS

A practical electric generator contains a coil of wire that is rotated in a magnetic field. To understand the operation, we examine a single loop as shown in Fig. 3. Metallic rings, called slip rings, are mounted on the shaft of the rotating loop and the ends of the loop are bonded to the slip rings. Metal contacts, called brushes, rub against the slip rings. The loop, slip rings, brushes, and connected electrical device form an electrical circuit for a current. The magnet produces a magnetic field whose field lines are directed from the N pole to the S pole. Only the sides of the loop labeled A and B are able to cut magnetic field lines. The two segments of the coil labeled A and B always move in opposite directions. Accordingly, the current is always in opposite directions in these two segments. The maximum current results when the wires move perpendicular to the magnetic field lines. There is no current in the circuit when the wire segments move parallel to the magnetic field lines. When the segment designated A in Fig. 3 is in the upmost position of its rotational path, it is moving horizontally to the right. Because the magnetic field lines are directed to the right, the wire is instantaneously moving parallel to the magnetic field lines. At this instant, there is no current in the circuit. As the loop rotates, the segments begin cutting the magnetic field lines and a current develops in the circuit. When the portion of the loop designated A has moved one-fourth of a turn down, instantaneously it moves perpendicular to the magnetic field lines. In this position, the electron current is a maximum. The electron current decreases to zero when the coil makes another one-fourth turn. As the coil rotates further, the electron current increases and changes direction. It reaches a

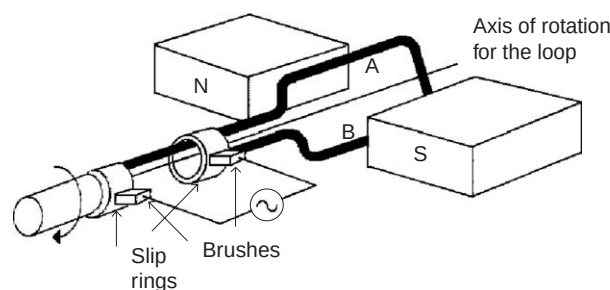


FIGURE 3 Principle of operation of an AC generator. A current develops in the rotating coil when it moves through the magnetic field created by the magnet. The slip rings rotate with the coil and make electrical contact with the brushes that are connected to an electrical device, such as a light bulb or a motor.

maximum when the coil makes a third one-fourth turn. Finally, the electron current decreases to zero when the coil returns to the initial starting position. A plot of electron current versus position of the coil, or equivalently time since the position depends on time, is shown in Fig. 4. The cycle of current and time shown in Fig. 4 repeats with each revolution of the coil. The number of cycles produced each second is called the frequency. A cycle per second is called a hertz (Hz). In the United States, the frequency of the AC voltage produced by commercial electric power plants is 60 Hz.

When a toaster is plugged into an AC household outlet, the current produced in the wires of the toaster continually changes in magnitude and direction. Nevertheless, the current causes the toaster wires to warm. Otherwise, one would not be able to toast bread. Household voltage is usually labeled 120 V, even though the voltage is continually changing. The 120-V rating means that the voltage is as effective at producing heat as a 120-V battery connected to the toaster. The 120-V AC label on a light bulb is the effective value of an AC voltage needed to power the bulb.

A generator that provides electricity for a headlamp on a bicycle functions very much like the one illustrated in Fig. 3. Mechanical energy is provided by the bicycle wheel rubbing against and rotating a shaft connected to the coil of the generator. The magnetic field is provided by a magnet. This type of generator produces approximately 5 W of electric power. Although the generator in a commercial electric power plant is much larger in size and often produces

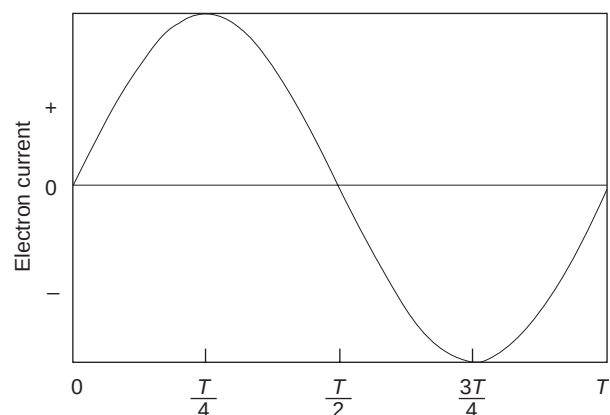


FIGURE 4 Plot of induced current in the rotating coil of an AC generator. Positive and negative currents are used to distinguish the two directions of charge flow. $T/4$, $T/2$, $3T/4$, and T denote times for one-fourth of a rotation, one-half of a rotation, three-fourths of a rotation, and a full rotation of the coil, respectively.

1 billion watts of electric power, the basic physical principle is the same. The commercial generator uses a steam turbine for the input mechanical energy and a massive stationary coil with a circulating electric current to provide the magnetic field.

7. TRANSFORMERS

An AC generator and a battery both deliver energy to such things as light bulbs connected to them. If a light bulb requires 12 V and one has a single battery that can provide 2 V, it is difficult to use the battery to power a device that can change the voltage from 2 to 12 V. If AC voltages are involved, it is reasonably easy to tailor voltages using an apparatus called a transformer. Transformers play very important roles in the transmission of electric power and in adapting voltages for industrial and household uses. The transformer principle is straightforward.

An electric current produces a magnetic field proportional to the current. If the current is constant, as would be the case if a battery were connected to the ends of a wire, the magnetic field is constant. The direction of the magnetic field lines depends on the direction of the current. If the battery is replaced by an AC source, the magnitude of the magnetic field changes as the current changes and the direction of the magnetic field lines changes when the current changes direction. There exists in the vicinity of the wire a changing magnetic field. If a coil of wire is placed in this changing magnetic field, an alternating current develops in it because the magnetic field lines as they change cut the wires of the coil. It is much like changing the magnetic field lines by moving a magnet. This principle is used in transformers to change the size of an AC voltage. The small black box often connected to an electrical outlet and to a portable computer to recharge the computer's battery is a transformer that typically reduces the outlet voltage from 115 to 20 V.

A transformer has two distinct coils of wire wound onto an iron core (Fig. 5). One of the coils, the primary, is connected to an AC voltage such as at the outlet in a house. The other coil, the secondary, is connected to an electrical device such as a light bulb. The AC current in the primary coil produces magnetic field lines that are guided into the secondary coil by the iron core. The interception of the magnetic field lines by the secondary coil produces a current in the secondary coil. The ratio of the primary and secondary voltages is equal to the ratio of the number of turns of wire in the primary

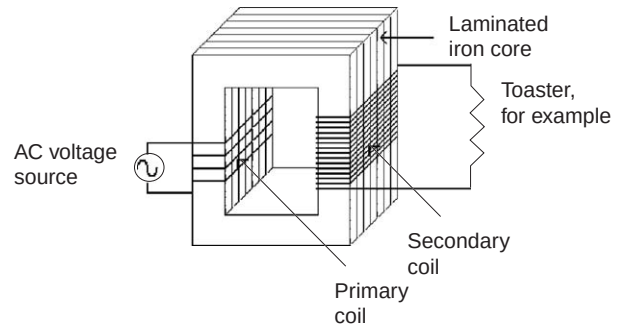


FIGURE 5 Essentially, a transformer consists of two coils of wire (primary and secondary) and an iron core to guide the magnetic field lines from the primary coil to the secondary coil.

and secondary coils of the transformer:

$$V_s/V_p = N_s/N_p. \quad (6)$$

Door chimes in houses often require 6 V for operation. Suppose that the voltage available at an outlet is 120 V. If a transformer has 20 times more turns on the primary than on the secondary so that $N_s/N_p = 1/20$, then the secondary voltage will be 6 V when the primary voltage is 120 V. The primary is connected to the available 120 V and the secondary is connected to the chimes.

The secondary of a transformer delivers electric power to whatever is attached to it. This electric power has its origin in the electrical source connected to the primary. The principle of conservation of energy prohibits obtaining more power in the secondary than was supplied in the primary. Transformers tend to warm, indicating that some of the power supplied in the primary appears as heat so that the secondary power is always less than that in the primary. In a well-designed transformer, the efficiency for energy transfer is approximately 99%. The ability of a transformer to tailor voltages and its very high efficiency makes it an extremely useful device.

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Electricity Use, History of • Electric Motors • Electric Power Measurements and Variables • Electric Power Systems Engineering • Electromagnetism • Mechanical Energy

Further Reading

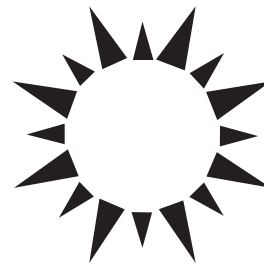
Hobson, A. (2002). "Physics: Concepts and Connections." 3rd ed. Prentice Hall, Englewood Cliffs, NJ.
Priest, J. (2000). "Energy: Principles, Problems, Alternatives." 5th ed. Kendall/Hunt, Dubuque, IA.

Serway, R. A., and Beichner, R. J. (2000). "Physics for Scientists and Engineers." Brooks/Cole, Pacific Grove, CA.
Serway, R. A., and Faughn, J. S. (1999). "College Physics." Brooks/Cole, Pacific Grove, CA.

Electricity, Environmental Impacts of

ALEXANDER E. FARRELL

University of California, Berkeley
Berkeley, California, United States



-
1. Activities
 2. Impacts

Glossary

coal combustion products (CCP) Solid waste left after burning the carbon and hydrocarbons in coal; made up of flyash, bottom ash, slag, and sulfur dioxide scrubber sludge. CCP are often used as construction materials.

direct cooling The transfer of waste heat from power plants to water bodies using water-to-water heat exchangers. This approach does not consume significant amounts of water (most is returned) but can cause significant impacts on aquatic species due to contact with screens or pumps.

electromagnetic field The combination of electric and magnetic fields that surround moving electrical charges (e.g., electrons), such as those in electric currents. Electromagnetic fields apply a force on other charges and can induce current flows in nearby conductors.

indirect cooling The use of a closed loop water system to reject waste heat to the atmosphere. A significant amount of water is evaporated in this process, which must be replaced, but total flows are less than with direct cooling, and water bodies are not heated.

radionuclide An isotope of an element that is radioactive, that is, a radionuclide undergoes spontaneous fission (splitting), releasing energy when it does so. The energy released from radionuclides can cause various health effects, including increased risk of cancer.

submergence Permanent flooding of land behind (upstream from) a dam.

Electric power is simultaneously one of the cleanest energy technologies and one of the dirtiest; it has virtually no environmental impacts at the point of use, but generation and transmission create enormous impacts. This article provides an overview of the environmental impacts of electricity, the most

important of which are air pollution, water use, solid waste, and land use.

1. ACTIVITIES

Most of the environmental impacts of electricity are associated with generation and transmission, including equipment manufacture and construction, fuel supply, fuel combustion, and waste disposal. Currently, electric power systems rely on fossil (coal, oil, and gas) powered generation for to ensure adequate electricity is always available and to follow load; but if fossil generation is limited in the future, other technologies will be needed. If so, the environmental impact of the entire system, including the manufacture and use of new technologies such as energy storage, must be considered. Here, only direct effects are discussed.

1.1 Generation

Enormous quantities of fossil fuels are used to generate electricity, so even relatively minor contaminants found in them (especially in coal) lead to very significant emissions of pollutants such as sulfur dioxide (SO₂), nitrogen oxides (NO_x), carbon dioxide (CO₂), toxics such as mercury (Hg) and hydrogen chloride (HCl), and radionuclides such as uranium (U) and thorium (Th). If electricity generation occurs near the point of use, emissions of carbon monoxide (CO) may be an issue.

Another major environmental impact of fossil and nuclear fuel electricity generation is water use, which includes consumption, impacts on aquatic species, and discharge of heat into nearby water bodies. Electricity generation by coal and nuclear plants also creates significant solid wastes that must be dealt

with. The majority of the environmental impacts of nuclear technologies occurs before and after generation, especially in fuel production and spent fuel disposal. Nuclear power plants are also vulnerable to accidental release of radionuclides.

Aesthetic concerns are also important; many people find electricity generation and transmission equipment unpleasant to be near or even to look at. While aesthetic concerns may seem relatively unimportant compared to health and environmental quality issues, they can be more difficult to resolve than pollution (which generally has a technical solution).

Renewable electricity generation technologies usually have quite different impacts; they typically create little or no emissions or waste during generation, but their manufacture may have important environmental impacts and they often require the use of large land areas to collect relatively diffuse renewable energy resources. This is especially true of hydroelectric power, which can require flooding vast areas. Fuel cells may also fall into this category because while they emit only water, producing the fuel cells and manufacturing and delivering hydrogen fuel may have significant environmental impacts. In this entry, however, only the direct impacts of operation are considered.

1.2 Transmission, Distribution, and Storage

The primary environmental impacts of transmission and distribution systems are associated with land use and aesthetics, but their scales are quite different. Distribution systems are ubiquitous in developed areas where they affect land use little (if at all), and are largely ignored by the public. In highly urbanized areas, electricity distribution infrastructure may be largely underground. The public often strongly opposes proposals to build transmission lines, especially when they will be close to population centers or highly visible. One reason for this opposition is concern about possible health effects of electromagnetic fields (EMFs), although it is very uncertain what (if any) risk EMFs create. On the other hand, transmission systems reduce the amount of generation capacity that's needed for a reliable system and allows for power plants and their emissions to be located far from populated areas, which tends to decrease their health and aesthetic effects.

Renewable energy sources are often located far from demand centers (cities and factories), extensive transmission systems may be needed to deliver

renewable power from where it's available to where it is needed. In addition, wind and solar energy are intermittent (they are not always available—such as on a calm night), so if they are to supply electricity whenever consumers need it, these technologies need some sort of energy storage. Nuclear power plants may rely on storage in the future as well; it can be hard to adjust reactor power up and down rapidly (over the course of a few hours) to follow demand.

Energy storage is extremely challenging, only two practical methods now exist—pumped storage and compressed air. Pumped storage facilities are like hydroelectric dams; energy is stored by pumping water from a lower reservoir to a higher one, and withdrawn by letting it flow back down through turbine generators. At compressed air energy storage facilities, electrically driven compressors fill caves or other underground formations with pressurized air to store energy, which is released later through a turbine. Neither of these technologies is widely used, and the environmental impacts of significant deployments of them are not well known.

1.3 End Use

There are essentially no environmental impacts of electricity in its end-use phase. Thus, as electric technologies replaced end-use fuel consumption for heating, cooking, lighting, and mechanical work, many environmental problems disappeared at the cost of introducing environmental impacts from electricity generation and transmission. This has helped homes and workplaces in many countries around the world become much cleaner and safer since the start of the 20th century, vastly improving public health and environmental quality. The only exception is the potential risk of EMF effects from household and workplace exposures, especially appliances.

2. IMPACTS

2.1 Air Pollution

2.1.1 Gases: SO₂, NO_x, and CO₂

Electricity generation causes a significant fraction of anthropogenic emissions of three major air pollutants, CO₂, NO_x, and SO₂, although the fraction varies significantly around the world, as shown in Table I. Anthropogenic emissions of SO₂ are relatively large compared to natural emissions of sulfur into the atmosphere, and are highly concentrated in the northern hemisphere. Electricity sector

TABLE I
Annual Emissions of Major Gaseous Pollutants from Global Electricity Generation

Pollutant Region	CO ₂		NO _x		SO ₂	
	Mass (Tg-CO ₂)	Percentage of anthropogenic	Mass (Tg-NO ₂)	Percentage of anthropogenic	Mass (Tg-SO ₂)	Percentage of anthropogenic
Canada	103	20	0.3	8	1.5	53
United States	2045	37	6.2	29	11.0	61
OECD Europe	966	26	2.1	15	5.0	30
Oceania	138	37	0.5	21	0.7	38
Japan	377	29	0.3	9	0.2	8
Eastern Europe	448	45	1.0	37	5.8	55
Former USSR	1278	41	4.4	45	6.9	43
Latin America	185	9	0.6	6	1.8	18
Africa	246	20	1.0	9	1.7	25
Mid-East	258	22	0.7	14	1.8	31
South Asia	410	34	1.7	25	2.5	33
East. Asia	1139	25	4.6	29	12.0	30
Southeast Asia	150	12	0.5	10	1.3	34
Total	7744	29%	23.8	21%	52.1	37%
Percentage of natural flows to the atmosphere	2		6		80	

Source. Oliver and Berdowski (2001).

SO₂ emissions are expected to grow in the future and become a larger fraction of anthropogenic emissions due to more use of coal for power production, compared to flat or declining uses in other sectors. In contrast, electricity sector NO_x emissions are relatively small compared to anthropogenic and natural NO_x emissions. Power sector NO_x emissions can have important effects because they are often concentrated near populated areas and because the nitrogen cycle can be significantly altered by relatively small perturbations. In the future, NO_x emissions from electricity generation are expected to grow, but not as fast as other sources, especially transportation.

The environmental effects of these emissions from power plants are diverse, including acidification, reductions in visibility (also called haze), and eutrophication. Acidification is a change in an ecosystem's natural chemical balance caused by an increase in the concentration of acidic elements and can lead to severe damage to trees and the elimination of fish species from lakes and streams. Eutrophication is excessive plant growth due to the addition of compounds like NO_x, which leads to

noxious algae and loss of fish and other aquatic wildlife. Visible haze is increasingly viewed as an important issue in parks and other natural areas, and it is driving some regulatory processes to control power plant emissions.

The most important human health impacts of SO₂ and NO_x emissions are their contributions to urban air pollution. In photochemical smogs, increased concentrations of ozone, polycyclic aromatic hydrocarbons, and fine particles (less than 2.5 micron diameter) can lead to increased morbidity and mortality. These compounds have many sources, but anthropogenic NO_x is among the most important.

Emissions of CO₂ from the electricity sector are not due to fuel contamination, CO₂ is the intended result of the combustion of fossil fuels. Unfortunately, it is also the primary cause of climate change.

2.1.2 Smoke and Particulate Matter

One of the main environmental impacts that electricity has helped lessen over the past century is smoke in urbanized and industrial areas. Smoke is a combination of unburned carbonaceous material (both solid and gaseous), nonburnable ash, sulfur

dioxide, nitrogen oxides, and trace amounts of metals. In the early 20th century, cities like Pittsburgh and London could experience dustfall of 10 to 30 g/m² each month. This material was certainly a nuisance and possibly a health hazard as well. Electricity replaced industrial boilers for mechanical work and in some cases as a heat source, reducing some of the smoke burden.

At first, however, electric power plants simply changed the location and character of smoke, or as it has come to be known particulate matter (PM) emissions. Early power plants had little or no PM controls and some power plants in developing countries lack them. In the 1960s, power plants in North America and Western Europe began to be fitted with PM control equipment and now virtually all coal-fired power plants in industrialized countries capture 99% or more of fly ash. In the United States, electricity generation accounts for about 1% of anthropogenic PM emissions of 10 microns diameter or greater (PM-10). In contrast, electricity generation contributes between 15 and 45% of PM emissions in some cities in India.

The greatest health risks of PM are due to the smallest particles, those 2.5 microns and smaller (PM-2.5). These fine particles also scatter visible light, causing haze. PM-2.5 can be emitted directly but are also formed in the atmosphere from gaseous emissions, especially SO₂ and NO_x. Electricity generation contributes similar fractions of direct PM-2.5 as it does PM-10, but it is a much larger contributor to global SO₂ and NO_x emissions and therefore to secondary PM-2.5 in the atmosphere. However, fine particles may have different health risks depending on their chemical composition, so the importance of electricity sector emissions to the health impacts of PM-2.5 pollution is not completely clear.

2.1.3 Toxics

Because many coals contain toxic contaminants and almost two-thirds of global coal production goes to

power electricity generation, the sector has substantial toxic emissions. Petroleum products and to a much lesser extent natural gas, contain some toxic contaminants as well. These pollutants are persistent (i.e., they do not degrade) and some of them, notably mercury (Hg), bio-accumulate. The most important compounds that coal-fired power plants may emit include Hg, hydrogen chloride (HCl), sulfuric acid (H₂SO₄), and hydrogen fluoride (HF). These materials can cause health and environmental effects at lower concentrations than more conventional pollutants like NO_x, and can be a significant local problem. Table II contains approximate annual global emissions of metals from worldwide electricity production and shows how these emissions compare to global anthropogenic emissions and of global natural emissions. Natural emissions of these metals are very small, so electricity generation creates very large perturbations to natural flows.

Geothermal resources (hot water and steam) can contain significant amounts of some of these toxics, as well as the toxic gas hydrogen sulfide (H₂S). In installations with very small emissions these may be vented to the air, in larger facilities they are reinjected into the ground along with most of the water or steam that has been withdrawn, and so they are. However, the contribution of geothermal electricity to all toxics emissions is trivial compared to that of fossil fuels.

Except for Hg and Se, most metals in coal tend to be captured by the fly ash, and their ultimate fate corresponds to that of the ash from the power plant. Many of these metals can also be captured in SO₂ scrubbers, and if so, their environmental impact is then largely determined by the final disposition of scrubber sludge. However, up to two-thirds of the Hg that enters a power plant in the coal remains a gas and escapes to the atmosphere. Controlling Hg emissions may include the conversion of gaseous Hg in power plant exhaust into compounds such as HgCl that can be more easily captured. Another

TABLE II
Approximate Annual Emissions of Metals from Global Electricity Generation

	Cr	Cu	Hg	Ni	Pb	Sb	Se
Emissions (Tg/year)	0.013	0.008	0.0011	0.042	0.013	0.0013	0.0038
Percentage of anthropogenic	40	20	50	80	15 ^a	40	60
Percentage of natural flows to the atmosphere	30	30	40	150	110	50	40

Sources. Nriagu (1994) and Pacyna and Pacyna (2002).

^aNot including leaded gasoline.

strategy may be to inject a sorbent (such as activated carbon) into the exhaust gas to capture the Hg and allow the particle control device to remove it.

Probably the most important public health impact of toxic emissions from electricity generation is neurotoxicity risk to unborn children of Hg. The primary exposure route is through consumption of fish by pregnant women, since aquatic bacteria create the most harmful form (methylmercury), which then accumulates in edible fish such as bass and perch. By 2002, Hg contamination had caused almost all U.S. states to warn women of child-bearing age to limit consumption of some fish from some water bodies. Wildlife may also be at risk due to Hg exposure, especially fish-eating birds and mammals.

As a liquid at typical temperatures, Hg can be transported very long distances through deposition and revolatization. Thus, Hg is a global pollutant, and for many countries more than one-third of Hg deposition originates outside of their borders. The hemispheric temperature gradient drives deposition toward the poles, although most Hg emissions occur in northern midlatitudes, health and ecological effects in the Arctic may be most severe.

2.1.4 Radionuclides

Radionuclides present a human health risk by increasing the risk of cancer, genetic changes, and birth defects, and present a similar risk to wildlife. Globally, the background radiation dose is typically about 2.5 milli-Sievert (mSv) per year, of which about one-third is from cosmic radiation and two-thirds from naturally occurring radionuclides, mostly uranium (U), thorium (Th), and especially potassium (K). Medical equipment can add measurably to this amount (for instance, 0.5 mSv/yr. in the United States, on average), as can lifestyle choices (e.g., flying in airplanes frequently). Nuclear power plants are often designed to give a dose of no more than about 0.05 mSv/year (about 2% of background) to the most exposed nonworker group.

Radionuclide contaminants in coal used to generate electricity total about 1.5×10^{14} disintegrations per second (or Becquerels, Bq) each year. The isotopes most important isotopes (of U, Th, and K) are solids by the time they leave the combustion chamber, so radionuclide emissions and exposure from coal generation are dependent on PM controls and solid waste disposal. Significant direct contact with ash or poor disposal techniques may create a slight hazard, but radionuclide concentrations in most coal ash are lower than in many soils and building materials (e.g., limestone). Interestingly, the

total radiation doses due to coal-fired electricity and due to nuclear electricity are virtually identical under normal operating conditions (unless the fuel is reprocessed, in which case the nuclear electricity dose more than doubles).

Nuclear power presents the unique problem of potential accidental release of radionuclides. Operating reactors contain significant amounts of radioactive Iodine (I-131), which tends to be absorbed readily by the body and concentrated in the thyroid, increasing the effective dose to that gland. This radionuclide has a short half-life (about 8 days) and so is essentially eliminated as a risk 2 months after a release. Nonetheless, I-131 can cause sufficient genetic damage during this short period to induce thyroid cancers later. Iodine tablets are an inexpensive way to minimize exposure in case of accidental release.

Almost a dozen nuclear reactors have suffered meltdowns since the mid-1940s, in addition to fires and other accidents. Most of these accidents were associated with experimental or weapons reactors, but two large electric power reactors have had catastrophic meltdowns, Three Mile Island Unit 2 (TMI-2) in 1979 in the United States and Chernobyl RBMK Unit 4 in 1986 in the Ukraine (which was part of the Soviet Union at the time of the accident).

The TMI-2 accident resulted in a release of 6×10^{11} Bq I-131 to the environment and resulted in a dose of less than 1 mSv to the closest neighboring civilians. The containment system prevented further releases despite most of the core melting to the bottom of the steel reactor vessel. The TMI-2 accident has had no observable effect on the health of either workers at the plant or the public. However, the total does for the United States was 20 to 40 person-sieverts, which will cause one or two excess cancer deaths in the United States. Cleanup of TMI-2, which will be completed some time after 2010, will eventually cost about \$1 billion, more than the plant's construction cost.

The Chernobyl accident was far worse; inadequate design, incompetent operation, and negligent disaster management combined to cause a large steam (thermal) explosion that exposed the core directly to the environment and caused great human suffering. Contemporary reports talk of a scarlet glow in the sky that could be seen 10 km away due to the light of the red-hot core being reflected in the clouds. The accident released large quantities of radioactive gases and particles, including 3% to 4% of the fuel. The greatest risk was from the release of isotopes that are readily taken up by the body,

including 1.7×10^{18} Bq of Iodine (I-131) and 8.5×10^{16} Bq Cesium (Cs-137).

About 200 tons of fuel escaped as liquid metal from the concrete cavity in which the core sat and mixed with other building material to form several piles that are now crumbling into dust. This dust presents a significant potential radiation risk, as does the 20,000 liters of contaminated water under the remains of the reactor. The reactor is contained within a massive but unstable “sarcophagus” that was constructed quickly and under extremely difficult conditions. To date, cleanup of Chernobyl has cost more than \$3 billion; more will likely be needed to replace the sarcophagus and for continued monitoring and cleanup.

More than 30 employees and emergency response personnel working onsite were killed due to radiation exposure during the first several days of emergency response to the accident, and more than 100 other workers were hospitalized with acute radiation sickness. Evidence has been found that considerable numbers of neighboring civilians probably also had some acute radiation sickness, but this was not acknowledged by Soviet officials at the time of the accident.

In the subsequent decade, approximately 800,000 people were involved in cleanup of Chernobyl, most of whom received <25 mSv, but one quarter of whom received >200 mSv. Among these workers, increased (approximately double) rates of various cancers have been observed, as well as genetic damage (although it is not clear if these have hereditary implications). Evidence of genetic mutations in the DNA of children of cleanup workers is mixed. One of the more common problems that Chernobyl cleanup workers have faced is psychological distress, which has caused various neurological symptoms.

Average exposure to the public was lower but still significant. Most of these doses were from I-131 and so were delivered primarily to the thyroid. Most people received doses below 20 mSv, but some went as high as 1000 mSv, with small children and infants being most affected. People in this age cohort who lived within a thousand kilometers of the plant at the time of the accident have experienced about a five-fold increase in thyroid cancers (typically nonfatal), probably totaling a few thousand through their lifetimes. Some researchers have found evidence of additional leukemia among this group as well, while some have found no relationship. There is some evidence of increases in birth defects in places as far away as Berlin and Greece. Estimates of the overall effects are difficult to calculate, but it appears that the

Chernobyl accident will cause between 10,000 and 30,000 fatal cancers, most in the Ukraine and Belarus.

Approximately 325,000 people were evacuated or relocated from the vicinity and the area within 30 km of the site is now a permanent exclusion zone. Two of the other reactors at the Chernobyl site operated until 2000 but are now shut down. The area around the site has become economically and socially depressed due to migration and other reasons. Local ecosystems were severely affected immediately after the accident, but are recovering without observable damage (although the extent of any genetic damage has not been determined).

About a dozen reactors with the same design as the reactor in Chernobyl (RBMK) exist elsewhere in the former Soviet Union, but all have received operational and engineering changes that make a repeat extremely unlikely. All other nuclear plants in the world are more similar to the TMI-2 design with massive steel containment vessels and other features (especially negative coefficients of reactivity for temperature and voids) that make an accident exactly like the one at Chernobyl impossible. However, the root causes of both the TMI-2 and Chernobyl accidents were a combination of design flaws, operating errors, and equipment failures, which can afflict any complex system. For instance, in 2002 a massive hole was found in pressure vessel of the Davis-Bessie reactor. Several inches of steel had corroded away without being noticed, leaving only a thin layer of stainless steel in place. No accident occurred, but this new type of situation demonstrates that the complexity of nuclear power plants continues to be a safety challenge.

2.2 Water Use

Water is consumed during virtually all forms of electricity generation, although fossil and nuclear plants that use steam turbines generally use far more than other technologies. Water is used to cool the plant, to make up feedwater, in emission control equipment, for ash handling, and for cleaning. By far the greatest volumes are used in cooling, although most of these withdrawals are returned to the water body (but at a higher temperature). In industrialized countries, water use by electricity generation is significant. In the United States, for example, 708 electricity generating plants (half of the installed capacity in the country) use cooling water, making up about half of all large industrial uses of cooling water in the country. These power plants have a combined average flow of $10,000 \text{ m}^3/\text{s}$. About half are located

TABLE III
Water Use by Power Plants

Cooling technology	Plant type and size							
	350 MW CCGT				350 MW coal			
	Direct	Indirect	Hybrid	Air	Direct	Indirect	Hybrid	Air
Withdrawal (m ³ /s)	5.26	0.122	0.102	0	11.7	0.388	0.324	0
Return (m ³ /s)	5.26	0.061	0.051	0	11.7	0.194	0.162	0
Net withdrawal (m ³ /s)	0	0.061	0.051	0	0	0.194	0.162	0
Withdrawal per MW (m ³ /s-MW)	0.0150	0.0004	0.0003		0.0334	0.0008	0.0006	

Source. Martin and Todd (1999).

on lakes and reservoirs, a quarter on streams and rivers, and a quarter on oceans and estuaries. Discharge temperatures can be as high as 40°C.

Cooling can be either direct or indirect. In the direct approach, cooling water enters a once-through system with heat exchangers that condense turbine exhaust steam to make return feedwater. All the waste heat goes into the water body, but little water is actually consumed. In addition to heating water bodies, electric power plants with direct cooling kill many marine organisms that come in contact with cooling water systems. Organisms may enter the cooling water system (entrainment) or be trapped on screens (impingement). For example, estimates for individual nuclear power plants in the 1600 to 2200 MW size indicate annual losses of 3 billion to 4 billion larvae, equivalent to 23 million adult fish and shellfish, or 17 to 45 tons of adult fish (creating observable changes in populations).

In the indirect approach, a closed loop of water absorbs heat from the steam heat exchanger to a cooling tower and gives up heat to the atmosphere, partially evaporating. This water must be made up with input from a nearby source. Tall (100 m) natural-draft cooling towers may be used (wet cooling); however, they are unsightly and may cause a visible plume or fog. Dry cooling replaces the cooling tower with finned radiators over which air is blown with fans. Hybrid systems combine wet and dry cooling. The water use and efficiency implications are shown in Tables III and IV. Hybrid cooling also tends to cost less than indirect wet cooling or air cooling and have less visual impact than wet cooling towers.

Overall, direct cooling involves larger flows of water, resulting in larger numbers of marine organisms killed and significant heating, while indirect cooling consumes larger amounts of water.

TABLE IV
Effect of Cooling Technology on Plant Efficiency

Cooling technology	CCGT efficiency (%)
Direct	56.4
Indirect—Dry	55.5
Indirect—Hybrid	56.1

Source. Martin and Todd (1999).

Technologies that do not use a steam cycle (and thus require little cooling), such as gas turbines, need little water. However, if water or steam injection is used for NO_x control, gas turbines can consume considerable amounts of water. Flue gas desulfurization to control SO₂ emissions also requires a considerable amount of water; a 1000 MW unit burning 1.5% sulfur coal would need approximately 0.1 m³/s for this purpose alone. All solar electricity technologies need some water for cleaning, and those that use sunlight to drive a steam cycle may need cooling water as well.

Large dams can also impact water quality by vastly changing natural flow regimes in terms of flow timing, volume, turbidity, and temperature. Such changes can have very large impacts on riverine species as well as on terrestrial species that depend on aquatic sources of food. Of course, large dams pose insurmountable barriers for species that spawn in rivers and streams but live most of their lives in the ocean, such as salmon.

2.3 Solid Waste

Besides nuclear waste, the major solid wastes created by electricity production are coal combustion pro-

ducts (CCP), which consist of flyash (ash that leaves with the flue gas and can leave the stack), bottom ash (ash that falls to the bottom of the combustion chamber), slag (ash that has melted and resolidified), and sludge from flue gas desulfurization scrubbers. The ash and slag consist of impurities of silica, iron, alumina, and other incombustible matter that are contained in the coal, while the sludge is largely gypsum (CaSO_4). Some ash is also produced by petroleum coke or fuel oils. Many toxics are solids by the time they leave the combustion chamber (exceptions include Hg, Cl, and F) and end up wherever the ash goes.

Assuming 10% ash content on average, about 500 million metric tons of CCP are created each year. Most of this ash is dumped in large landfills; sometimes a small valley will be dammed and ash slurry pumped behind the dam until filled. Increasing pressure to protect the environment in addition to increasing interest in controlling costs have led the electricity and coal industries to find alternative uses for CCPs. The largest use is in making cement and concrete; followed by wallboard (gypsum), road construction materials, mine fill, blasting grit, soil additive, and structural fill. Importantly, the use of CCP avoids the need to mine or otherwise produce raw materials for these products. In Europe, almost 90% of CCP finds a use, while only about one-third is in the United States.

Environmental regulations have made the reuse of CCP more difficult. First, low NO_x burners increase the unburned carbon in the ash, which concrete manufacturers limit to 6%. Second, NO_x reduction systems that inject urea or ammonia into the flue gases can contaminate the ash, creating handling difficulties. Looking ahead, requirements to produce renewable fuels may lead to cofiring of biomass and coal, which will add simple organic salts to the mostly alumino-silicate CCPs. The importance of this change has yet to be determined.

Power plant operators have learned how to optimize their scrubbers to produce high-quality production by forcing the oxidation of calcium sulfite and adding chemicals that help stabilize the dry mass. Gypsum can be used to make wallboard or cement. It can also be used to mitigate acidification and acid mine drainage due to its alkalinity, or as a fertilizer and soil enhancer in some locations.

Petroleum fuels do create small amounts of ash, which will have more vanadium and nickel. Ash and slag from the combustion of municipal solid waste (MSW) contains much higher quantities of toxics, as well as potentially valuable metals (like copper). The

solid waste from fluidized bed combustors will contain unreacted limestone and products of SO_2 and Cl capture, mostly as calcium compounds.

2.4 Land Use

All forms of electricity production as well as electricity transmission require large equipment and considerable amounts of land. High-voltage transmission lines can have a number of important land use implications and the siting of transmission lines in industrialized countries has become extremely difficult. However, transmission lines are necessary to use some resources effectively, and they can reduce the cost of system operation by allowing distant systems to share resources. Land use is the major environmental impact of renewable technologies. Good locations for hydropower and wind turbines are rarely located at demand centers (e.g., major cities), necessitating long-distance transmission and the associated environmental effects.

The amount of land different electricity technologies require depends on how the calculation is done. Key questions include these: Are the impacts of fuel production and transportation included? (Natural gas production can be very disruptive to wilderness areas and biomass may require vast areas for planting.) Are the impacts of waste disposal included? (Nuclear waste storage removes some areas from other uses.) Are the impacts of pollution measured in area? (Coal fired power plants can cause acidification across large regions.) Are the impacts of associated transmission included? (Wind power will often have to be transmitted long distances.)

2.4.1 Aesthetic, Cultural, and Equity Concerns

Many people find electric power plants and transmission towers ugly and undesirable to live or work near. For this reason, property near power plants or power lines have a lower value than comparable properties not near power lines. Electricity infrastructure is typically prohibited in parks, wilderness areas, scenic areas, and so forth as well as historic or culturally important locations. An important and related aspect is equity: Why should one community accept a power line (or power plant) when that facility will serve another community? This problem can create tensions between urban dwellers and rural/suburban dwellers.

Concerns about aesthetics and equity are among the largest challenges facing developers of wind power farms and transmission systems. This is exacerbated by the fact that many of the best

potential wind energy sites are highly visible and prized mountain ridges. This problem has forced many new installations offshore.

As income levels rise, the desire for political freedoms and preferences for environmental amenities tend to rise as well, suggesting that aesthetics and equity will play larger roles in electricity projects in the future. This trend may boost technologies that can reduce the visual impact of transmission systems, such as more compact high-voltage direct current transmission systems and underground cables. It may also boost small, distributed generation technologies, which typically are connected to an underground gas distribution system and may require fewer transmission lines.

2.4.2 Land Submergence and Relocation

The environmental impact of hydroelectric facilities differs greatly from other types of electricity generation in that it often requires the submergence of large land areas and relocation of residents and other ecosystem inhabitants. This may be true of both facilities that use impoundments (dams) and run-of-river facilities, since the latter are often built downstream of a dam and rely on moderated river flows of the dam. Unlike other renewable energy sources that utilize a lot of land for energy collection, submerged land is lost completely, along with rivers and creeks. However, a new, lake-like reservoir is created, with its own amenities. Dams can also change the aquatic ecosystems dramatically. Importantly, dams can reduce the hazard of flooding, which can be considerable for settlements in floodplains.

Globally, about 45,000 large dams (>15 m high or 5 to 15 m high with a reservoir of 3 million m³) exist, along with 800,000 smaller ones. About 100,000 dams worldwide generate electricity, mostly the larger ones. (The remainder are used for irrigation or water supply only.) The total area covered by reservoirs is 0.5 Million km², less than 1% of the land used for human habitation and agriculture. Although dam construction in North America and Europe has largely halted (or even reversed), several nations in Africa, Asia, and Latin America have considerable potential and are involved in large-scale dam development.

Reservoirs can submerge homes, farmland, wilderness, polluted sites (possibly liberating the pollutants), cultural features, and appealing landscapes, especially dramatic river valleys. Larger hydroelectric projects tend to have larger reservoirs, but submerge less land per unit of capacity. Dams under 100 MW generally require over 200 hectares per MW of

capacity (ha/MW), those from 100 MW to 500 MW less than 100 ha/MW, and those above 500 MW less than 50 ha/MW.

Relocation of large numbers of people can be very difficult and expensive, and finding new homes (and possibly farms) for them may cause environmental stress. There are often tensions between different groups, such as between urban dwellers (including elites) who will typically get most of the power from a hydro project and rural dwellers who will be displaced. Tensions between different ethnic and political groups have arisen as well. In addition, dam construction often reduces flood risk, inducing settlement on floodplains for the first time. Dams also reduce sediment loads downstream, which can weaken dikes and levies. These conditions can worsen outcomes when flood control measures of dams fail, which does occasionally happen.

2.4.3 Wilderness and Wildlife

Electric power systems can have major impacts on wilderness and wildlife, mostly because fuels or hydroelectric facilities can be located in wilderness areas or because electric power systems in different areas may find it beneficial to interconnect themselves to share resources. In particular, dams and wind turbines can affect wildlife. Submergence of habitat is the most obvious impact. In addition, the roads and power lines needed to construct, maintain, and connect hydroelectric facilities sometimes run through wilderness areas, fragmenting them. The vegetation below power lines must be trimmed, burned, or controlled with herbicides, and there must be access roads for maintenance, all of which have environmental impacts. These effects may reduce the habitat for threatened species, or cause new species (including threatened species) to move into an area. Birds are known to strike transmission towers and lines, but the numbers doing so are usually not very large, and proper routing and design can reduce bird deaths.

Birds have also been known to strike wind turbines, sometimes in large numbers. This problem can be greatly reduced by modern designs that spin slower and by not placing wind turbines in areas that birds use heavily as flight corridors.

Underground and underwater electric cables are becoming more common, most often due to space restrictions, such as in major cities, and partly because of aesthetic concerns. These cables are filled with dielectric oil and are armored. Breaks that leak the dielectric oil are rare and are usually identified quickly. Underwater cables are typically laid 1 or 2 m

below the seafloor using a mechanical plow or hydraulic jet to create a trench, which is then filled in. This alters the seabed, and there are concerns that it might have permanent effects on fisheries and marine wildlife. Trawling and other fishing techniques, repair of the cable, and retrieval at end of life must all be considered.

2.5 Electromagnetic Fields

When electricity is generated, transmitted, or used, electromagnetic fields (EMFs) are created as a direct consequence. These fields can induce small electrical currents in the body, which have been suspected of causing cancer and other health effects. This idea remains controversial, however, since the currents induced by EMFs seem far too small to cause such effects, leading to a suspicion that contact currents and not EMFs may account for the observed effects.

The main sources of exposure to EMFs are wiring and appliances in the home and workplace, and electric power lines (both transmission and distribution). The use of electricity increased for 100 years without any observed environmental effects from EMFs until a 1979 study reported that children living near power lines might have an increased risk for developing cancer. Since then, studies of human populations, animals, and isolated cells aimed at understanding the potential health effects of EMFs have been conducted, but experts are divided about their interpretation.

Three types of scientific evidence have been examined in the effort to understand the possible health effects of EMFs: human epidemiological studies, mechanistic (test-tube based) research, and animal toxicology. The epidemiological studies have shown a small risk of increased childhood leukemia, and possibly adult brain cancer, amyotrophic lateral sclerosis, and miscarriage. No evidence for other health risks has been found. However, epidemiological research can only show associations, it cannot determine cause and effect. The mechanistic and toxicological research has never found any consistent effect. The failure of these two approaches means scientists cannot explain how EMFs might cause health effects.

The degree to which experts believe that EMFs present a health risk depends on how they weigh these three lines of inquiry. Generally, however, scientists who study EMFs feel they cannot be recognized as entirely safe, sometimes listing EMFs as possible carcinogens. If there is an effect, the risk

probably falls just within the range that is considered sufficient to warrant minor regulatory action in industrialized countries. Various bodies have set voluntary guidelines that are designed to limit the induced current density in the body to 10 mA/m^2 , sometimes reduced to 2 mA/m^2 for exposure to the general public. Exposure in the home is typically less than 0.01 mA/m^2 , while contact with electrical appliances (e.g., razors, electric blankets) can create currents up to 30 mA/m^2 .

2.6 Other Impacts

Electric power systems can affect nearby people and the environment in a number of less obvious ways. Some equipment creates noise, including wind turbines, transformers, and power lines. There is some concern that this noise could limit the suitability of land under wind turbines for wildlife or livestock. Wind farms can occupy from 4 to 32 ha/MW of capacity, although they typically only use a small percentage of that land for towers and other equipment, leaving the rest free for other uses. Wind turbines can interfere with television and radio signals. Corona effects on transmission system components can create electrical interference, but this is a highly localized effect.

Transmission lines can interfere with other infrastructures and with aircraft, possibly a significant problem in areas where aerial application of herbicides is used. Wind turbines and tall exhaust stacks can also affect aircraft. Dams will change a region's hydrology and may induce excess seismicity. Geothermal electricity production will generally cause subsidence, although this is usually controlled by reinjection of the highly saline water and steam that is brought to the surface for electricity production.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Clean Air Markets • Clean Coal Technology • Electric Power Generation: Valuation of Environmental Costs • Electromagnetic Fields, Health Impacts of • Entrainment and Impingement of Organisms in Power Plant Cooling • Hazardous Waste from Fossil Fuels • Nuclear Waste

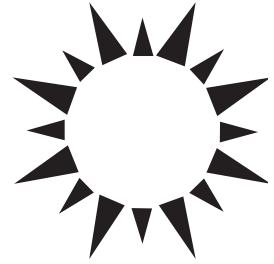
Further Reading

- Boorman, G. A., Bernheim, N. J., Galvin, M. J., Newton, S. A., Parham, F. M., Portier, C. J., and Wolfe, M. S. (1999). "NIEHS Report on Health Effects from Exposure To Power-Line Frequency Electric and Magnetic Fields." National Institute of Environmental Health Sciences, Research Triangle Park. www.niehs.nih.gov/emfrapid/home.htm.
- Garwin, R. L., and Charpak, G. (2001). "Megawatts and Megatons," p. 412. University of Chicago Press, Chicago.
- Hall, M. L., and Livingston, W. R. (2002). Fly ash quality, past, present and future, and the effect of ash on the development of novel products. *J. Chem. Technol. Biotechnol.* 77(3), 234–239.
- Hammons, T. J. (2001). Environmental implications and potential of international high-voltage connections. *Electric Power Compon. Sys.* 29(11), 1035–1059.
- Martin, P. C., and Todd, I. J. (1999). The changing picture of water consumption by power stations. *J. Chartered Inst. Water Environ. Management* 13(5), 369–374.
- Neutra, R. R., DelPizzo, V., and Lee, G. M. (2002). "An Evaluation of the Possible Risks From Electric and Magnetic Fields (EMFs) from Power Lines, Internal Wiring, Electrical Occupations and Appliances." Oakland, CA, California EMF Program. www.dhs.cahwnet.gov/ehib/emf/RiskEvaluation/riskeval.html.
- Nriagu, J. (1994). Industrial activity and metal emissions. In "Industrial Ecology and Global Change" (R. Socolow, C. Andrews, F. Berkhout, and V. Thomas, Eds.), pp. 277–285. Cambridge University Press, New York.
- Olivier, J. G., and Berdowski, J. J. (2001). Global emissions sources and sinks. In "The Climate System" (J. J. Berdowski, J. R. Guicheir, and B. J. Heij, Eds.), pp. 33–78. A. A Balkema, Lisse, The Netherlands.
- Pacyna, E. G., and Pacyna, J. M. (2002). Global emission of mercury from anthropogenic sources in 1995. *Water Air Soil Pollution* 137(1–4), 149–165.
- Sinha, S. (2001). Environmental guidelines for power plants in India and other nations. *Environ. Qual. Management* 11(1), 57–69.

Electricity Use, History of

DAVID E. NYE

University of Southern Denmark
Odense, Denmark



-
1. Inventing and Promoting Electrification
 2. The Telegraph, Alarm Systems, and the Telephone
 3. Public Lighting
 4. Industrial Applications
 5. Electrified Transportation and the City
 6. Personal Uses of Electricity
 7. Domestic Electrification
 8. Rural Electrification
 9. Air-Conditioning
 10. Electricity Use and the Environment

Glossary

alternating current This current is produced when flows of electrons are made to shift directions in a constant pattern. It is easier to transmit long distances than direct current, and it is the form of electricity in most homes and businesses.

arc light An early, brilliant form of illumination produced by passing an electric arc between two carbon rods.

direct current This current is created when electricity moves in one direction along a wire. It was most common in the late 19th century; for example, it was used in Thomas Edison's first light and for early streetcars. It still has many applications, for example, in batteries.

incandescent light Light produced from an enclosed bulb by heating a filament until it glows.

In the 19th century, electricity was at first exotic, and then a luxury. Since then it has become a necessity, but paradoxically both the level of demand and the level of use vary considerably from one nation to another. The telegraph (1838), telephone (1876), and broadcast radio (1919) at first spread slowly; as with many later applications, consumers needed time to discover their uses. An interconnected electrical system grew rapidly from c. 1880, with lighting followed in rough sequence by commercial establish-

ments (such as department stores), public buildings, monuments, streetcars, railways, subways, factories, homes, airports, farms, and stadiums. The precise sequence of adoption and degree of penetration varied from one nation to another, particularly in rural areas. Likewise, the intensity of use varied. In c. 1960, Norway, Canada, and the United States consumed more than twice as much per capita as England and France. Continually improved generating efficiencies did much to meet increasing demand while lowering the cost per kWh, until reaching engineering limits in the 1970s. Since then, new efficiencies have been in the appliances and light bulbs consumers use. Nevertheless, rising demand for electric power creates environmental problems that alternative energy sources such as wind mills can only partially solve.

1. INVENTING AND PROMOTING ELECTRIFICATION

Until the many practical applications of electricity that became available in the late 19th century, no single energy source could provide light, heat, and power. Electricity's versatility made it an enabling technology that allowed people in all walks of life—including architects, designers, homemakers, industrialists, doctors, and farmers—to do new things. These new and often unanticipated uses found for electricity mean that its history must include far more than such matters as dam construction, the extraction of fossil fuels, or improvements in dynamo efficiency. One must also consider energy history in terms of demand. Consumers have been an active force behind the expansion of electricity use. If one considers the rapid increase in electricity demand from the middle of the 19th century to the present, one is struck by how this increase parallels the growth of the middle class, the expansion of cities,

the development of mass consumption, and increased demands on the environment. The uses of electricity are more than mere applications; they are inseparable from rising demand, which at times has outstripped the capacity of utilities and led to major programs in demand side management. Indeed, the same pattern repeated itself with the electric light, the electric iron, or the electric vacuum cleaner and a host of appliances and later applications. A new device first became a luxury for the rich, then a desirable item of conspicuous consumption for the middle classes, and, within little more than a generation, a necessity for all.

Practical applications were only possible in the 19th century and after. The understanding of electricity was so rudimentary in the late 18th century that Benjamin Franklin, a clever self-taught amateur in the American colonies, made an important contribution to knowledge by discovering that lightning and electricity were the same thing. The phenomenon of electricity had remained a curiosity since classical times, when static electricity was known but scarcely investigated. Few scientists focused attention on it until after the invention of the Leyden jar (1746), which facilitated a century of experiments that prepared the way for practical applications. Charles Coulomb, Michael Cavendish, Benjamin Franklin, Luigi Galvani, Alessandro Volta, Humphrey Davy, and Andre Ampere developed a rich store of observation and theory that made possible Faraday's studies of electromagnetic induction, which in turn made possible the dynamo (1831). When Faraday was asked whether his experiments would ever result in tax revenue, he boldly declared that they would.

Indeed, during the next half century a series of industries and businesses emerged based on the special properties of electricity. Its movement was so rapid as to be instantaneous, a property exploited by the telegraph, various alarm systems, and the telephone. It could produce electrolysis, making possible metal plating. Electricity could also produce light when passed through certain conductors, eventually making possible arc and incandescent lighting. Electricity provided a way to transmit power over considerable distances, allowing falling water or burning coal to be translated into motive force by distant electric motors. Electricity could also be used to produce precisely controlled heat without fire or the consumption of oxygen, leading to applications in baking, enameling, and industrial heating and burning. Each of these innovations was eventually incorporated into devices used in the home and on the farm.

Given the many potential uses of electrification, it might seem that its adoption was virtually automatic. Such was not the case. As late as 1912, less than 3% of electrical power was residential, whereas almost 80% was used in manufacturing, with the remainder used commercially. Although it might be tempting to treat the history of electrification as the story of inventors producing a stream of practical devices, it was a far more complex developmental process that included government, manufacturing, marketing, and several different groups of consumers. Utilities quickly discovered that large, urban customers were far more profitable than smaller customers in suburbs and the countryside. The spread of electrical service therefore moved from larger to smaller customers, and from urban to rural, in a series of distinct stages. This process began in the 1870s in Europe and the United States and has since spread throughout the world. However, even in the early 21st century electricity has by no means become the dominant source of lighting or power in every location, and in most cases other sources of heat are more economical. If the world's capital cities are electrified, some of the countryside remains in darkness, as revealed by satellite photographs. Indeed, even in the nations in which the process began—England, France, Germany, and the United States—in 1900 almost all factory power still came from falling water or steam, and less than 5% of all homes had electricity. Although in retrospect electrification might appear to have been a steady and inevitable process, closer inspection reveals uneven growth, at times thwarted by competition from gas, coal, and other energy forms. In general, electricity replaced gas for lighting because the incandescent bulb was safer and cleaner than an open flame and because it was more like daylight. In contrast, gas remains a strong competitor to electricity for heating and can also be used for air-conditioning. Yet such generalizations do not apply universally because energy use was inflected by local conditions and by social class. For decades, English workers clung tenaciously to gas lighting, leaving electricity to the middle and upper classes. In contrast, wealthy Swedes were reluctant to give up traditional forms of lighting that they considered genteel, whereas Swedish workers more rapidly embraced the electric light. In 1931, less than 30% of British homes were wired, and few were all electric, at a time when 90% of Swedish homes had been wired. In the United States, most urban but few rural dwellings were electrified by that time, and wiring was less a marker of social class than of geographical location.

From a technical standpoint, electrification was not a unified process that spread like ripples in a pond from certain nodal points. Rather, there were several interrelated processes, which began at many different locations, driven not only by utilities but also by the demands and interests of consumers. Particularly in the early decades, separate systems developed to serve different needs, including small stand-alone generating plants for some hotels, department stores, theaters, and wealthy homes and large, direct current power houses owned by street railways. Until the 1930s, many factories also had their own power stations.

The degree of such fragmentation varied, being considerably greater in Britain than in the United States, for example. In the early 20th century, London had 65 electrical utilities and 49 different types of supply systems. Anyone moving within London risked entering a new grid because there were 10 different frequencies and 32 different voltage levels for transmission and 24 for distribution. Once consumers had converted all the lamp and appliance plugs and made other appropriate changes, they discovered that each utility also had its own pricing system. In the United States, where utilities worked harder to coordinate their systems and interconnect their lines in order to achieve economies of scale, consumers still faced considerable inconsistencies in plug and socket sizes. These were only ironed out in the 1920s with the help of the federal government. Even today, as every tourist knows, the electrical current and the plugs needed to access the system vary among the United States, Britain, and continental Europe, and these three different standards have been reproduced throughout much of the rest of the world.

The large electrical manufacturers, which sold generating equipment and provided technical expertise to new markets, exported these standards. With so many potential customers, the manufacturers did not attempt to enter all markets. Rather, each focused on certain regions, to a considerable degree based on political spheres of influence and colonial empires. For example, General Electric (United States) sold lighting and traction systems to Brazil and much of Latin America, whereas in most cases General Electric (United Kingdom) transferred the same technologies to the British Empire. German and Austrian electrical expertise spread throughout central Europe and the Nordic countries. To some extent, France developed its own systems but relied heavily on investment and partial ownership from the United States.

On the demand side, in all the industrial countries the middle classes developed a demand for electrifi-

cation in many guises. On the supply side, the electrical industry was an expansive investment opportunity that was more important and longer lasting than the Internet bubble of the 1990s. Between 1875 and 1900, electrical equipment makers and utilities formed the most dynamic sector of the American economy, growing from a few small adjuncts to the telegraph companies into a \$200 million industry. After 1900, it grew into a colossus that was inseparable from the health of the national economy. The electrical industry's potential was early recognized by J. P. Morgan and other leading investors who financed Edison's work. Once commercial development began, a flurry of mergers reduced the field from 15 competitors in 1885 to only General Electric and Westinghouse in 1892. Before this time, railroads had been America's largest corporations, demanding the greatest capital investment, but they were a mature industry compared to the emerging electric traction companies, local utilities, and equipment manufacturers.

2. THE TELEGRAPH, ALARM SYSTEMS, AND THE TELEPHONE

The invention of the telegraph in 1838 heralded the first widespread practical application of electricity. However, potential consumers were not certain what to use the telegraph for. Conceived and patented by Samuel F. B. Morse, the telegraph astounded incredulous crowds, and the first telegraph offices often provided seating for the public, who could scarcely believe that it was possible to sever language from human presence. Twenty years after its invention, when thousands of miles of lines linked the United States, the *New York Times* declared that "the telegraph undoubtedly ranks foremost among that series of mighty discoveries that have gone to subjugate matter under the domain of mind" (August 9, 1858).

However, even the inventor of the telegraph did not understand its enormous commercial potential, and he tried to sell it to the federal government. Many European countries chose to make the telegraph part of the postal service, but the U.S. Congress refused, and only after 5 years granted Morse \$30,000 to build a line from Washington to Baltimore. People seemed to have little use for it, however, and even when the first line connected New York, Philadelphia, and Washington business still developed slowly. Six months after it started operation, the line averaged less than 2000 words a day, and it was generating

only approximately \$600 a week. By 1848, the telegraph linked Chicago to New York, however, and grain traders with immediate access to information soon made fortunes trading in midwestern wheat. In less than a generation, Western Union became one of the largest corporations in the United States. During the same years, newspapers realized the commercial value of Mexican war news and formed an association that inaugurated the wire services. As these examples suggest, the potential of the various forms of electrification was only gradually understood.

During the second half of the 19th century a series of technologies based on the telegraph, such as burglar alarms, stable calls, and door bells, familiarized people with the idea that a current carried over wires could be instantaneously transmitted to virtually any point. However, other practical applications were few because most electricity still came from inefficient batteries. Further use of electricity was limited until large and efficient generators became available in the 1870s. Until then, “impractical” electric lighting was confined to special effects in theaters.

Like the telegraph, the invention of the telephone (1876) at first seemed a curiosity, and Alexander Graham Bell was unable to sell his invention even to Western Union. Setting up on his own, Bell concentrated on businessmen and wealthy households and did not expand service widely until his patent ran out and competition emerged in the 1890s. It took more than a generation before phone companies realized that people wanted phones not just for business and emergencies but also for gossipy conversations. Even so, the telephone took a long time to become a necessity to consumers compared to the electric light. Only after World War II did most American homes have telephones. Even then, they were all black, heavy apparatuses that the subscriber did not own but leased. They were not stylish elements of décor that varied from one home to another but utilitarian objects that were engineered to last decades. In contrast, a greater variety of models were marketed in Europe, despite the fact that universal service only came to some countries during the 1960s, and Spain and Portugal only achieved it at the end of the 1980s. On the other hand, during the 1990s Europeans adopted mobile phones far more quickly than did Americans.

Again, as late as 1919 the U.S. government failed to see the broadcasting potential of radio, understanding it only as a superior form of the telegraph, for wireless point-to-point communication. In the early 1920s, a few amateurs initiated the first public radio broadcasts. Yet even when the potential of a

device is grasped, not all societies have made the same choices about how to embed electrical technologies within the social structure. Should suppliers be public or private or a mix of the two? Should pricing encourage extensive or intensive use of a system? What is the best balance between regulation and open competition? Between 1880 and 1900, such questions had to be answered first with regard to public lighting and then street traction systems.

3. PUBLIC LIGHTING

In both Europe and the United States, regular public electric lighting appeared first in the late 1870s as a spectacular form of display in city centers and at expositions and fairs. The first installations were arc lights, which shed a powerful glare over streets and the insides of some stores. Department stores and clothiers quickly adopted such lights because they were cleaner, brighter, and safer than gas and because they proved popular with customers. By 1886, 30 miles of New York’s streets had arc lights, installed approximately 250 ft apart. In 1879, Thomas Edison demonstrated his enclosed incandescent bulb, which had a softer glow that was well suited to most indoor and domestic settings. Both arc and incandescent lighting competed with already existing gas systems until the first decades of the 20th century. Edison recognized that his light, its distribution system, and its cost had to be designed to compete with gas. For as much as a generation, some new buildings had both systems. The Edison system was rapidly adopted wherever convenience and fissionability were highly valued, and it spread rapidly into the wealthiest homes, the best theaters, and the most exclusive clubs in New York, London, Vienna, and Paris. However, 20 years after its introduction, the electric light was almost entirely restricted to public places. Less than 1 home in 10 had been wired, and the average person encountered electricity at work, in the street, and in commercial establishments.

From its inception, the public understood lighting as a powerful symbolic medium. The history of lighting was far more than the triumph of the useful. Before gas and electric streetlights people had to find their way with lanterns, and the city at night seemed fraught with danger. Public lighting made the city safer, more recognizable, and easier to negotiate. However, mere functionalist explanations cannot begin to explain why electric lighting had its origins in the theater or why spectacular lighting emerged as a central cultural practice in both Europe and the

United States between 1885 and 1915. As early as 1849, the Paris opera used an electric arc lamp to create a sunrise, and showmen of every description seized on the new medium to attract crowds. During the 1870s, thousands of people turned out to see the first arc lights in Cleveland, Boston, London, and Paris. The new light fascinated them not only because it was so much brighter than existing gas but also because it seemed to violate the natural order. For the first time in history, light was separated from fire. It needed no oxygen. It was unaffected by the wind. It could be turned on in many places simultaneously at the turn of a switch. In 1883, Niblo's Garden featured "novel lighting effects by the Edison Electric Light Company," which included an illuminated model of the Brooklyn Bridge and chorus girls with lights flashing from their foreheads as they flourished electrically lighted wands. Promoters and public alike demanded ever-greater public displays as use far exceeded necessity.

Intensive lighting became a form of symbolic expression at not only theaters but also world's fairs and other public events. Every exposition attempted to outdo all previous efforts by escalating both the scale and the completeness of displays. Intensive arc lighting at Chicago in 1894 gave way to thousands of incandescent lights at Omaha (1898), arrays of colored bulbs in Buffalo (1901), and artificial sunrises and electrical fireworks in San Francisco (1915). Many fairs made illuminated towers their central symbols, notably Buffalo's Electric Tower, San Francisco's Tower of Jewels, and New York's Trylon and Perisphere in 1939. Electricity became more than a theme; it provided a visible correlative for the ideology of progress, and it was placed at the apex of an evolutionary framework, as opposed to "ethnological" villages—compounds representing the life of Mexicans, Cubans, blacks in the Old South, Filipinos, or Africans in the jungle—where the visitor could see staged versions of earlier forms of social evolution. Electrification thus became embedded in a social Darwinist ideology of racial superiority, as darkness represented the "primitive," whereas artificial light exemplified Christianity, science, and progress. At the Buffalo Pan-American Exposition, "the strongest, crudest colors" were nearest the entrances, and there was "a progression from warm buff and ochre walls ... through more refined and brilliant hues to ... ivory white, delicate blue, greens, and gold" on the central tower.

During the same years, electrical advertising signs spread from a few isolated installations in approximately 1890 to nearly ubiquitous evening displays.

Particularly in the United States, main streets became "great white ways" that were more alluring at night than during the day. In each nation, particular locations became famous tourist attractions because of their advertising signs, notably New York's Times Square and London's Piccadilly Circus. In these public squares, the electric sign, the spotlight, and even the streetlight became economic weapons. The competition for attention did not create the majestic serenity of an exposition, dominated by beaux-arts architecture bathed in white light. Commercial lighting produced a potpourri of styles and colors flashing against the night sky. Thousands of individual decisions produced this electric landscape, but it nevertheless had an unanticipated collective impact. Literally millions of lights seen together, particularly from a distance or from the top of a skyscraper, had a strong popular appeal. Electric lighting was also used to dramatize important monuments. One of the most effective early displays was the lighting of the Statue of Liberty in 1886, which made the statue stand out against the darkness of the New York harbor. In the following decades, search lights would be permanently trained on many of America's most important symbols, including many skyscrapers, the White House, Niagara Falls, Old Faithful, Virginia's Natural Bridge, and Mt. Rushmore.

Between 1880 and approximately 1920, dramatic lighting transformed the night city into a scintillating landscape that was a startling contrast to the world seen by daylight. Street lighting, illuminated buildings, and electrical advertising simplified the city into a glamorous pattern that signaled more than the triumph of a new technology. By 1920, spectacular lighting had become a sophisticated medium of expression that governments and corporations used to sell products, highlight public monuments, commemorate history, encourage civic pride, and even simulate natural effects.

4. INDUSTRIAL APPLICATIONS

Industry found many uses for electricity, and from 1880 to 2000 it used the preponderance of power in most nations during any given year. Factory electrification began with lighting. Munitions plants and flour and cotton mills quickly adopted electric light because it was far safer than any other form of illumination. Printers, artists, and others whose work required accuracy and true color relations found gas inferior to incandescent light. Factory owners soon realized that light made possible the construction of

large interior spaces without regard for windows. However, if lighting came quickly to the factory, adoption of electric drive required a generation. Existing steam and water power functioned well and cost a great deal to replace. These power sources could not easily be transmitted over more than a few hundred feet, but there were other alternatives. In 1891, to transmit water power 2 km or less, wire rope cost half the price of electricity. At that time, electricity was only cost competitive for networked power systems more than 5 km long. For example, high-pressure hydraulic transmission was a well-established alternative. Introduced by Joseph Bramah in Dublin in 1795, it was applied to cargo handling in many British ports and to centralize urban hydraulic power supply. Hydraulic power was also applied to heavy industry, such as steel production, and water motors using a city water supply could drive small appliances requiring less than 2 horsepower as well as church organs and many elevators in commercial buildings. Milwaukee alone had 600 hydraulic elevators connected to the city water system. Power from illuminating gas also briefly flourished in the 1890s. Clean, safe, compressed air, however, was the most successful alternative to electric power and was well developed in Paris, where in 1899 more than 100 miles of pipe supplied 24,000 horsepower to a wide variety of customers.

Nevertheless, electricity proved so versatile and generation costs declined so quickly that after 1900 it was in the ascendant. An investigation by the Franklin Institute showed why it was preferred. Interestingly, savings in energy costs were considered to be of no great importance since power usually comprised only 1–3% of a workshop's budget. Furthermore, electrical equipment was somewhat more expensive than the familiar mechanical drive, but savings in maintenance and depreciation offset this expense. Electricity's advantages were primarily in creating freedom in the location of machinery, improvements in lighting and ventilation, cleanliness (there were no overhead oil drippings from shafting), reductions in shutdowns, greater precision of electrical equipment, and increases of 20–30% in output as a result of all these factors.

Adoption of electricity in industry varied considerably by nation, however, with the United States taking the lead, closely followed by Germany. Although accurate statistics are difficult to obtain, in 1928 one British expert estimated that 73% of manufacturing power was electrical in the United States compared to 67% in Germany, 48% in Britain, and less for other European countries.

Britain particularly lagged behind in iron, steel, and chemicals, but this gap was partly closed during the 1930s, when its economy performed better than that of the United States.

Electricity's versatility extended into virtually all areas of manufacturing. For example, electrolysis enabled silverware manufacturers to coat materials with silver. In coal mines and rock quarries, electric drills were far less cumbersome to move around than drills driven by compressed air or steam. By 1920, many steel rolling mills had installed electric drive and controls that made it possible for one man to guide steel from the ingot furnace to the finished rail. An observer noted, "Every motor replaced a man, but the work is done better and more quickly." In food processing plants, aluminum factories, brass foundries, and potteries, electric furnaces and ovens produced no smoke, consumed no oxygen, and maintained high temperatures precisely. Such factories reduced spoilage and increased throughput.

Electric motors proved especially useful in moving materials, particularly in mines and other closed environments. Electric motors were also adapted to overhead hoists and cranes, and these began to move their loads more freely once overhead drive shafts were removed in favor of electric drive. Indeed, by 1920 electricity had become the basis for a revolution in factory design, brightening and opening up the work space and making possible new work flows, most notably on the assembly line, where electrically powered machines were arranged in combinations that were impossible using older power-delivery systems. This is not to say that electrification necessarily led to a particular form of mass production. Rather, it facilitated experiments and new designs. At the same time, electrification altered the siting requirements of factories, which no longer had to be immediately accessible to rushing streams or coal supplies. Indeed, after the successful demonstrations of alternating current in the 1890s, the short practical transmission range of direct current no longer placed a limit on where power could be delivered. Whereas early generating plants such as Edison's Pearl Street Station in New York had to be near the city or town to reach customers, during the 20th century power stations increasingly were located on the edge of the city or in the countryside. Because transmission lines could reach virtually anywhere, factories and businesses could also move away from railroads and canals to the edge of town or along main highways. Thus, electrification was used to transform both factory layouts and the location of industry.

From a worker's point of view, incandescent lighting reduced the danger of fire, reduced pollution, and made it easier to see, but it also made possible round-the-clock shifts. Furthermore, as electric motors and cranes provided more horsepower to production, they made possible radical changes in the layout of production, most strikingly in the development of Henry Ford's assembly line in 1912, an innovation partly anticipated by Edison's experiments with automating iron mining in the 1890s. The assembly line was literally impossible in any complex industry before machines were freed from the rigid organization required by overhead drive shafts and given the flexibility of electrical devices.

The connection between electrification and mass production at times has been overlooked by historians, who have emphasized scientific management, particularly the work of Frederick Winslow Taylor. It is true that Taylor systematized work and made tasks such as shoveling more efficient. In contrast, Ford's assembly line used electrified machines to eliminate shoveling altogether. Taylor retained piecework, a concept that makes no sense on the assembly line, where all share the same pace. Taylor maximized efficiency in existing production technologies; Ford transformed the means of production. Taylor saved time; Ford sped up time. In 1914, after Taylor imposed his system on the Packard Motor Company, its 4525 workers produced 2984 cars a year, less than 1 per man. This was essentially artisanal production of a luxury vehicle. Ten years earlier, well before the assembly line, Ford produced 12 cars for every worker. By 1914, 13,000 Ford employees, using conveyor belts, electric cranes, and 15,000 specialized machine tools, manufactured 260,000 cars or 20 per worker. Ford's system outperformed Taylor's by more than 2000%.

5. ELECTRIFIED TRANSPORTATION AND THE CITY

As industrial plants grew larger and became concentrated in cities, the urban scale expanded and electricity supplied many forms of transportation. Indeed, it made possible faster and easier movement not only at the street level but also into the sky and down into the earth. The Columbian Exposition of 1894 introduced visitors to electric boats and electrified trains and even installed some electrified moving sidewalks. Although similar sidewalks are found in few places outside airports today, a closely

related technology, the escalator, has become almost ubiquitous in department stores and large public buildings.

The clean and quiet electric car initially seemed to be a promising form of urban transportation. A French coach builder, Charles Jeantaud, introduced an electric car in 1894, but it never sold well. In contrast, many American inventors enjoyed initial success, including an "Electrobat" introduced in Philadelphia in the same year. The early leaders were two Hartford-based firms, Pope Manufacturing, which made 500 vehicles in 1897, and the Electric Vehicle Company, which made 2000 electric taxis and several electric trucks the following year. For approximately a decade such vehicles were almost as numerous as fledgling gasoline cars. Even Henry Ford bought an electric car for his wife Clara in 1908. Such vehicles were easier to control than gasoline cars, and with their slower speeds (25–30 mph) and shorter range (60–80 miles) they seemed ideal for women making social calls or shopping but not for men, who demanded faster, long-distance transport. The electric car could easily strand drivers in the countryside, far from any recharging station, and even if one were found recharging took hours. If electric trucks and cars largely disappeared after the early 1920s, however, smaller electric gaffle trucks, golf carts, and indoor delivery vehicles held a niche market. Eight decades later, electric vehicles are still being designed and tested, and hybrid cars are now sold that switch automatically from gasoline motors to electric drive, depending on driving conditions.

Other forms of electric transportation were more widely adopted. Vertically, the elevator made the skyscraper practical, as it carried executives to their preferred offices high above the streets. These offices had electric lights and telephones, their plumbing was supplied with water by electric pumps, and their air could be heated or cooled electrically. The electrified office literally detached executives from the noises, smells, and temperatures of the street.

Horizontally, electric trolleys made a much larger city feasible. In the 1890s, streetcars were faster and cleaner than horsecars, and they could be stopped more precisely and climb steeper grades than cars pulled by steam. Electric cars could easily be lighted at night and heated in winter, and overall their operation cost less than any alternatives. Real estate promoters often invested in or owned traction companies, whose spreading lines defined much suburban and regional development, especially in new cities such as Minneapolis and Los Angeles but also in older metropolises, such as Boston, Copenhagen, or Amsterdam. By

1902, large cities averaged more than 200 trips per citizen per year, and Americans alone took 4.8 billion journeys by streetcar.

In the larger cities, notably Paris, New York, Budapest, and London, the technology of the electric tram was adapted to underground use. Construction began in the 1890s, and millions of commuters and shoppers were literally hidden from view. The subway showed that electric light, power, and ventilation made possible an entire new urban setting beneath the pavement. New department stores added several brightly lighted basement stories, and underground arcades connected some buildings. In Toronto and Montreal, where winter temperatures prevail, such urban designs became particularly popular.

Electrified transport transformed the city, bringing an unprecedented number of rural and urban consumers to the center, which stimulated new forms of entertainment. Vaudeville theaters and movie houses were able to fill thousands of seats several times a day, and new sports arenas (electrically illuminated) drew even larger crowds to see professional baseball, football, and basketball. Traction companies drew some of the crowd to their own amusement parks, which consumed some of their excess electrical output on evenings and weekends. The technology of electrical traction was the basis of the roller coaster, the subterranean tour, and many other rides. Trolley parks spread rapidly after 1888, and by 1901 more than half of all street railway companies operated at least one (often several). These typically generated as much as 30% of the line's overall traffic as well as provided a market for electricity. The trolley made the city the center of an easily accessible network, from the downtown shopping district to streetcar suburbs, the amusement park, and the fringe of rural hamlets on interurban lines.

Urban electrification was not merely a matter of improving existing structures. The electrified city had a fundamentally different built environment than the city before it. At the center were large office buildings, department stores, theaters, and brilliantly lighted public buildings accessible on an expanded scale reaching miles further into surrounding countryside. Practical electrical lighting, elevators, and air-conditioning likewise made it possible to create self-enclosed environments, including the later developments of indoor sports stadiums and shopping malls. The modern city is unimaginable without electricity, which enabled enormous changes in its design.

Yet these changes were by no means the same in all cultures. In the United States, where the automobile became a central form of transportation in

the 1920s, the electrified city was quickly modified, sprawling out into the suburbs. In contrast, European cities retained mass transit and bicycles for a generation longer, and as a result the electrified cityscape developed and became naturalized. The urban landscape made possible primarily by electrification (i.e., without intensive use of the automobile) is therefore far more visible in a city such as Oslo than in Los Angeles.

6. PERSONAL USES OF ELECTRICITY

The public became excited by electrical medicine well before Edison had perfected his incandescent light. Electricity seemed to be a mysterious fluid that permeated living organisms, leading Walt Whitman to write of "the body electric." Electricity seemed capable of recharging the body, helping it to overcome a host of diseases and conditions. Neurasthenia and neuralgia became fashionable complaints apparently caused by lack of nerve force, and millions bought patent medicine cures for electrical imbalances. Respected doctors hooked their patients up to generating machines and passed currents through diseased limbs or sought to (re)animate sexual organs. Patent medicines such as Dr. Cram's Fluid Lightning claimed to contain electricity in liquid form so that one could ingest it and revitalize the whole body. By the mid-1890s, magazines such as *Popular Science Monthly* asked, "Is the Body a Human Storage Battery?" So many believed the answer was yes that electrical devices were mass marketed. The *Sears and Roebuck Catalog* for 1901 sold electric belts designed to restore men's sexual powers and to help either sex restore lost energies. Some electrotherapeutic electric belts retailed for as much as \$40 and promised a "triple silver plated battery of great power, silver plated nonirritating electrodes, and a very powerful electric sack suspensory." Women bought electric corsets fitted with "magnetic steel" with the expectation that they could mould their figure to any desired form. Even the U.S. Congress had a cellar room in the Capitol building fitted with an electrical medical apparatus. The *Electric Review* noted, "The members say it is splendid after they have exhausted their brain power by speechmaking or listening. A great many members take electricity, and some go to the basement of the Capitol for it every day during the season."

Although these "uses" of electricity are clearly bogus, other applications just as clearly were not. By

the 1890s, it was firmly established that ultraviolet light, either from the sun or from arc lights with special filters, could kill bacteria, and in 1896 Niels Finsen discovered that lupus could be cured by repeated exposure to ultraviolet rays. He won the Nobel Prize for this work in 1903, and such successes seemed to be only the beginning. The Belgian chemist Ernest Solvay proposed that electrical stimulation might accelerate plant growth and increase the yield of crops. Similar ideas inspired experiments by the Nobel Prize-winning chemist Svante Arrhenius. He “tested the effect of electrical stimulation on the growth of children,” who, he reported, grew 20 mm more than those in the control group.

When prestigious scientists made such claims, the gullibility of laymen is not surprising, and the enthusiasm for electrical cures lasted well into the 20th century. Despite the efforts of the American Medical Association to discredit them, people bought electrical cures in the 1920s and 1930s. For example, thousands of people paid \$60 for an I-ON-A-CO electric collar in hopes of rejuvenation. As this popular faith in electrical medicine suggested, people regarded electricity to be a magical fluid, a nerve-tingling “juice,” or an invisible tonic that could restore normal health. From the 1880s until 1920, when few people had wired their homes, electrical medicine offered direct contact with the mysterious force that was transforming work and public life. For approximately half a century, the body literally became the contact zone where men and women explored electricity’s possible meanings.

By the time the electrical medicine craze was over, electricity had begun to be used to enable the body in many ways as battery technology improved. Small, lightweight batteries could deliver small but constant amounts of power for a year or more, and starting in the 1950s they first made possible electric watches and hearing aids, and later pacemakers and a host of ingenious medical devices. Slightly larger batteries combined with the new transistors made possible radios that could be carried in a shirt pocket. Consumers have not given up on using electrical devices to reshape their bodies. Some still buy electric belts, advertised on TV in the wee hours of the morning, and attach electrodes to their stomachs with the expectation that they can effortlessly tone up their muscles and look thin. Millions of people pay to use private gyms with Stairmasters and other electrical devices that tell them how long the workout has lasted, monitor heartbeat, and count calories lost.

By the 1990s, improved batteries, weighing less and lasting longer, had become an assumed part of

most people’s everyday life. They drove electric toothbrushes, cordless drills, and other tools. They powered tiny portable radios, disc players, mobile telephones, and palm desktop computers. They had become essential to thousands of children’s toys, even if they were not usually included with those toys when purchased. In outer space, satellites and exploring machines relied on solar-powered batteries to keep them operational for years. In little more than a century, electrical devices had penetrated into the most intimate parts of the body and reached millions of miles into space and, in the process, extended human sight and hearing.

7. DOMESTIC ELECTRIFICATION

In the 1880s, the few domestic consumers of electricity were almost exclusively wealthy, and they chose the new form of illumination as much for its social prestige as for convenience. Some even bought personal generating plants, especially in the countryside. Utilities did not focus on domestic customers until after a generation had passed, during which time a host of appliances appeared that made higher levels of consumption possible. Electric lights alone did not create much demand, but with the marketing of electric irons, fans, toasters, coffee makers, and other devices after 1900, the domestic consumer became more interesting. In the United States, gas continued to supply the vast majority of light and heat to private residences until after 1910, with the major shift to electricity occurring between 1918 and 1929. In Britain, the change occurred later, and as late as 1919 less than 6% of all homes were wired.

In the decade before World War I, the new profession of home economists championed the electrified house, but the domestic market only became particularly attractive to utilities once their overall load curve peaked in the day, creating a need for compensating night demand from homeowners. However, the conversion to electricity was neither automatic nor easy, and in many areas gas remained the stronger competitor, for example, in refrigerators until at least the end of the 1920s and in stoves and home heating until the present. Furthermore, it was not always easy to determine what the consumer would buy. Electric Christmas tree ornaments were a success in the United States but less so in some European countries. Even today, most Scandinavians prefer to put lighted candles on their trees. Electric cigarette lighters caught on in automobiles but not in the home. The air-conditioned bed never sold well,

although there was limited success with electric milk warmers for babies.

Given the uncertainties of public demand, could the electrical utilities create this demand in practice? In Chicago, Samuel Insull showed other utilities that it was possible to build up the load through innovative pricing policies designed to attract a wide variety of customers, whose variable energy use evened out the demand curve and permitted production efficiencies. His variable pricing persuaded large customers, such as factories and traction companies, to abandon their own power systems, creating daytime demand that decreased at night. This created the generating capacity for a large domestic market, where use peaked after working hours.

In addition to Insull, Henry L. Doherty in Denver was another innovator in promotion. Their salesmen marketed electrical devices systematically, selling door to door. They also employed extensive newspaper advertising, built model electric homes that were extremely popular in approximately 1920, and paid for follow-up demonstrations by home economists in schools and at public events. Many people were afraid to use an electric iron, for example, and salesmen were trained to educate consumers and to dispel their fears of the new technologies. They emphasized cleanliness, comfort, convenience, and economy of operation—themes particularly appealing to those responsible for public buildings, especially schools. By approximately 1915, educators were convinced that the air temperature should be 68–70° and that sickness could be controlled through good ventilation. Schools also prepared a new generation of consumers. Students experienced the superior light, heat, and ventilation that electricity made possible, and in class they learned about the scientific background of the new domestic technologies. There were also gender-specific classes: newly invented home economics for girls and shop for boys. Additional domestic appliances made the home an even more attractive market as vacuum cleaners and washing machines became common during the 1920s, followed by radios, stoves, electric phonographs, and refrigerators in the 1930s.

The widespread adoption of labor-saving machines soon spread in Europe. In Britain, by the late 1920s domestic power consumption had outpaced commercial use. France lagged slightly behind and experienced the appearance of electric kitchen appliances as Americanization, although its level of consumption nevertheless was only 40% that of the United States as late as 1937. In that year, Norwegians and Canadians were by far the most avid domestic consumers. They faced long, cold

winters but possessed ample hydroelectric resources and as a result used more than twice as much electricity per inhabitant as the United States and four times as much as Germany. Although electrification spread more slowly in some areas than others, once it entered the home it was seldom removed, and by the 1930s it had begun to be seen as a necessity.

U.S. consumption continued to double every decade. *Fortune* noted in a 1969 article on the dangers of brown-outs and utility overloads that “the American likes his home brilliantly lit, of course, and he has a passion for gadgets that freeze, defrost, mix, blend, toast, roast, iron, sew, wash, dry, open his garage door, trim his hedge, entertain him with sounds and pictures, heat his house in winter, and—above all—cool it in summer.”

The European pattern of development was much different than the American pattern. In part because of the denser population and in part because electricity was more often considered to be a state service rather than a private business, lighting was widely dispersed because it seemed fair to do so, although the cost was generally higher than in the United States. As with gasoline, the expense of the service discouraged Europeans from adopting a wide range of appliances during the first half of the 20th century. Since approximately 1960, however, the patterns of consumption have converged, and by 2000 a French or German home had much the same repertoire of electrical appliances as one would find in Canada or the United States. Nevertheless, differences remain. What are considered normal and acceptable levels of illumination in many European homes may seem dim to American visitors. Many a temporary resident complains that the owners must have removed most of the floor lamps. In fact, Scandinavians prefer a living room that is dim by American standards. They generally find it cozy to have only small pools of light and a few candles.

8. RURAL ELECTRIFICATION

One of the most striking changes in the past two centuries is the shift away from agricultural life, which rural electrification abetted, although it had begun long before. In 1800, approximately 90% of all Europeans and Americans were farmers; today, farmers comprise less than 4%. Half the United States was still rural until the 1920 census, so much of the decline had little to do with electricity. However, the disappearance of millions of additional farmers was due in large part to the increased

efficiency brought by electrification. Although the scale and organization of individual farms changed continually for 200 years, rural electrification exemplifies the substitution of capital equipment for labor, bringing both reductions in waste (e.g., through refrigeration) and improvements in productivity using such devices as milking machines, incubators, and electric pumps.

Not all applauded these developments. British thinkers, such as G. M. Trevelyan, Vaughan Cornish, and Patrick Abercrombie, developed environmental ideas about the relationship between man and nature. Such thinkers believed that periodic experiences of wilderness could assuage the wounds of urban civilization and therefore attacked the march of electrical pylons across the countryside. In contrast, most agrarians in the United States embraced electrification as a way to reverse flight to the city by bringing new amenities to farmers and small towns. John Crowe Ransom, for example, realized that electrification could make farming more attractive, and Herbert Agar argued that it could bring some urban amenities to the countryside and help preserve the Jeffersonian ideal of equality and independence.

Although the long view makes the process of electrification seem rational and even inevitable, in fact rural electrification was long ignored by utilities, unless prodded by government. Some of Edison's early utilities were in smaller cities, but he quickly realized that the distribution costs for scattered customers were too high. Indeed, the costs were much higher for direct current than for alternating current, which only became commercially available in the 1890s. After that time, in places where rural services were nationalized or subsidized, such as in New Zealand, the countryside was rapidly brought into the grid. In the United States, however, where power was largely privately controlled, more than 90% of all farmers still did not have electricity as late as 1935. Although equipment manufacturers and hardware stores advertised electrical wares to farmers in magazines and displayed appliances at county fairs, few people could buy them because before the 1930s private utilities used most of their resources to electrify urban homes and sponsored few experimental rural projects. The long distances between farms, with only three or four customers per mile, made transmission costs high, and farmers had little money due to the long agricultural depression that persisted throughout the 1920s.

For utilities, this rural poverty was not an incitement to action but a caution against expansion.

As a result, by the 1930s electricity had become naturalized as a normal part of daily life in cities, creating a vivid contrast with a countryside still plunged in darkness. New Deal politicians, encouraged by Franklin D. Roosevelt, made rural power a national issue and created a Rural Electrification Administration (REA) that established rural cooperatives throughout the United States. Indeed, one of the selling points for this program was that the nation had fallen behind the successful efforts in Europe to provide electricity to the entire population.

Organizing farmers into REA cooperatives and getting electric lines to their homes were only first steps. A self-sustaining cooperative required higher electrical consumption, and REA coops soon were being visited by "utilization experts," Washington's bureaucratic term for salesmen. Their mission was to convince farmers to buy motor-driven equipment and larger appliances. Customers who never installed more than a few electric lights did not create enough business to enable coops to repay government loans. Washington agencies at first misjudged the farmers, however, and the utilization experts initially pushed heavy farm equipment. In practice, however, only dairies and chicken ranches discovered immediate advantages of electrical equipment. Raising wheat, tobacco, cotton, and many other crops could be done profitably much as before. However, the farmhouse was quite a different story. Farmers adopted washing machines and refrigerators more quickly than urban consumers, and they did so for sound economic and practical reasons. Cities had laundries, but in the country all washing had to be done by hand after water was pumped from a well and heated on a stove. Little wonder half of the REA's customers in the first 3 years bought washing machines, which had a much higher priority than indoor toilets. A 1939 study of the first cooperative in Indiana showed that after 2 years of service 68% had washing machines, whereas only 10% had utility motors; 45% had vacuum machines but less than 1% had bought any machines for agricultural production. Only during World War II did farmers purchase large amounts of such equipment due to manpower shortages and high crop prices. Productivity and profits rose.

Rural electrification was not implemented to drive people off the land but to keep them there by bringing to the country the amenities enjoyed by the city. Paradoxically, however, rural electrification meant fewer farmers were needed. In the United States, people as diverse as Henry Ford, Frank Lloyd Wright, Lewis Mumford, southern Agrarians, and New Dealers all agreed on the desirability of rural electrification

as a means to regenerating life in the countryside. However, in practice relocation programs and rural electrification instead helped make farmers so productive that many became redundant. In the United States, 40% of the population was on the farm in 1930, 15% remained in 1950, and there was only 8.7% by 1960. The percentage kept declining for the rest of the century and fell below 2% in 2000. This transformation cannot be attributed to electrification alone, of course, because mechanization of farm work and transport also played a role. As farmers moved out, however, others began to move beyond the suburbs into the countryside. Although some were hobby farmers, most wanted to escape the city, but they demanded the urban amenities that only electricity made possible, such as indoor running water, washing machines, vacuum cleaners, radio, and television. After 1975, migration into rural America was greater than out-migration, and the trend shows no sign of abating. Thus, rural electrification underlay not only the decline of the farm but also the deconcentration of cities and suburbs.

Most European countries did not repeat this development to nearly the same degree. During most of the 20th century, cities remained more attractive to most citizens, and until at least 1965 those who commuted were more likely to use bicycles, trains, and trolleys than automobiles. At the same time, European Union subsidies made it possible for a higher percentage of farmers to remain on the land, and national land legislation, notably in The Netherlands, often discouraged sprawl into the countryside. In Europe, where rural electrification was first realized, the electrified city remains the focus of life far more than in the United States, suggesting how much cultural values shape the consumption of electricity. Indeed, overall in the 1990s, a North American family was using twice as much energy as its Western European counterpart.

9. AIR-CONDITIONING

The skyscrapers of Dubai or Mexico City would not be habitable without air-conditioning, which has transformed possibilities of life in large areas of the world that are hot during much of the year. Stuart Cramer coined the term air-conditioning in 1904. He developed humidity controls for Southern textile factories, where excessive dryness made threads break easily, creating havoc in production. Cramer focused on ventilation and humidification more than cooling. Willis Carrier, a Cornell engineering grad-

uate, did the most to promote the acceptance of a precisely controlled indoor environment. His Carrier Air Conditioning Company (1908) sold devices that could regulate both a room's humidity and its temperature, and these spread rapidly from textile mills to food processors and breweries.

Engineers used air-conditioning to increase their authority and control in factories. They first insisted on control of the building and later over the activities within. At times, the new technology displaced skilled workers whose tacit knowledge of materials and their response to different atmospheric conditions previously had been crucial to production of such varied things as textiles, macaroni, candy, beer, and cigarettes.

In the 1920s, air-conditioning spread into stores, skyscrapers, theaters, restaurants, and trains, familiarizing the consumer with its benefits and making people more accustomed to an enclosed, artificial environment. By the 1930s, many firms saw enormous potential for domestic sales and began a fierce competition for a market that only developed after World War II. Many early domestic machines did not regulate humidity but focused on cooling only. As a result, the public perception of air-conditioning remained confused, and in 1938 less than 1 home in 400 used it in even one room. After World War II, however, the industry overcame consumer resistance by working with builders and architects and by achieving economies of scale in producing the equipment.

Air-conditioning was not merely something added to existing structures; its use transformed housing design. Builders partially offset the cost of air-conditioning by eliminating attic fans, window screens, sashes, and high ceilings. New homes were often designed to minimize the purchase price, not the operating costs. Sealed windows combined with the heat produced by appliances, television, and the modern kitchen made air-conditioning a necessity, and the resulting electricity demand eventually helped create the national energy shortages of the 1970s.

In 1955, the federal government adopted a temperature policy that in effect required air-conditioning in most of its new buildings. The Federal Housing Association also permitted new mortgages to include it. Two forms of air-conditioning competed in the marketplace. Selling to both home and workplace, the Carrier Corporation led old engineering firms committed to central systems, whereas General Electric and other appliance manufacturers sold smaller units designed for individual rooms. Central systems began to predominate in hot

cities such as Las Vegas, Dallas, and Bakersfield by the 1960s. Air-conditioning reached 24 million homes by 1970. Thereafter, it rapidly changed from being a luxury to a necessity, creating enormous summer demand for electricity. At office buildings such as the United Nations headquarters, floor-to-ceiling windows increased air-conditioning costs by 50%. By 1982, just after the end of the energy crisis, most office towers and more than 60% of all homes had some form of air-conditioning, half of which were central systems for cooling the entire structure. This was high market penetration considering that this technology was unnecessary in areas such as the Pacific Northwest, the northern tier of states, higher elevations generally, and many coastal communities.

10. ELECTRICITY USE AND THE ENVIRONMENT

Compared to burning wood or coal, electricity seems to be a clean and odorless form of energy. Indeed, for generations electricity provided a better alternative to older, more polluting technologies. Unlike gas and kerosene lamps, electric lights left no soot. Clean electric trains eliminated the cascades of soot from steam locomotives. In factories, electric motors emitted far less oil and raised far less dust than overhead drive systems. Hydroelectric plants create energy with no smoke, but unfortunately falling water provides less than 5% of U.S. energy needs and even less in some other nations. Generating electricity in coal-fired power plants causes pollution, even if it is often invisible to consumers. Such plants are usually located far from the sites of consumption, but their smoke, and the acid rain that eventually results, falls on distant consumers.

Fortunately, during the 20th century utilities developed economies of scale, transforming a larger percentage of coal into energy. Gas-fired plants are even more efficient. During the 20th century, the production of electricity became increasingly efficient, which lowered prices and reduced pollution. In New York City, for example, the average residential customer in 1930 paid \$29.23 for 404 kilowatt-hours (kWh) of electricity per year, slightly more than 7¢ per kilowatt-hour. By 1956, the average New Yorker's consumption had almost quadrupled to 1536 kWh, but the cost had only doubled to \$61.09. For the nation as a whole by that date, the average consumer used much more, 2969 kWh, but paid only \$77.19 for it. Electricity became cheaper because engineers found ways to extract more from a ton of coal or barrel of

oil. By the 1970s, however, it was difficult to achieve further production efficiencies. Thereafter, the most cost-effective savings lay not in production but rather in more efficient consumption.

Industrial research therefore focused on developing highly efficient products, notably light bulbs, motors, washing machines, stoves, and refrigerators. As a result, during the two decades after 1973 total U.S. energy use increased only 10%, an achievement based more on technological ingenuity than on changes in lifestyle. People did not use fewer appliances, but their new devices were more efficient. Home refrigerator manufacturers tripled the energy efficiency of refrigerators during these two decades, and by the 1990s washing machine electricity use had been halved. However, efforts to become more efficient resembled the mad queen in *Alice in Wonderland*, who had to run as fast as she could just to stay where she already was. Even greater efficiencies or lower per capita consumption will be necessary in the 21st century as electricity reaches all the world's population. Recognizing this, in 1994 the World Energy Council proposed the idea of national limits in electricity use. It noted that annual electricity use per capita higher than 6 MWh did not seem to improve life expectancy. It therefore suggested that the 30 highest consuming nations should reduce electrical consumption to this level, whereas the rest of the world should not allow consumption to rise higher than this level. Even if these targets are achieved, however, the increased use of batteries in toys and other consumer products poses another environmental problem because the discarded batteries contain highly poisonous materials. The public tends to discard them in ordinary dumps, allowing lead, cadmium, and other toxic wastes to filter into the groundwater.

Overall, demand for electricity poses environmental problems that lay at the heart of the quest for sustainable development. However, whereas coal and oil-fired power plants contribute to global warming and acid rain, new forms of wind mills, solar power, and passive solar heating promise not only to reduce pollution but also to decentralize power production, moving it closer to the consumer. In the future, as in the past, production of electricity will prove inseparable from its use.

SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Coal Industry, History of • Electrical Energy and Power • Electromagnetism • Hydropower, History and Technology of • Manufac-

tured Gas, History of • Natural Gas, History of • Nuclear Power, History of • Oil Industry, History of • Refrigeration and Air-Conditioning • Solar Energy, History of • Wind Energy, History of • Wood Energy, History of

Further Reading

- Beltran, A., and Carré, P. A. (1991). "La fée et la servante: La société française face à l'électricité xix-xx siècle." Éditions Belin, Paris.
- Hannah, L. (1979). "Electricity before Nationalisation: A Study of the Development of the Electricity Supply Industry in Britain to 1948." Macmillan, London.
- Hughes, T. P. (1983). "Networks of Power Electrification in Western Society, 1880–1930." Johns Hopkins Univ. Press, Baltimore.
- Marvin, C. (1988). "When Old Technologies Were New." Oxford Univ. Press, New York.
- Nye, D. E. (1990). "Electrifying America: Social Meanings of a New Technology." MIT Press, Cambridge, MA.
- Nye, D. E. (1998). "Consuming Power: A Social History of American Energies." MIT Press, Cambridge, MA.
- Platt, H. L. (1991). "The Electric City." Chicago Univ. Press, Chicago.
- Thomas de la Peña, C. (2003). "The Body Electric: How Strange Machines Built the Modern American." New York Univ. Press, New York.

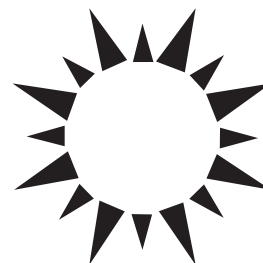
Electric Motors

ANIBAL DE ALMEIDA

University of Coimbra
Coimbra, Portugal

STEVE GREENBERG

Lawrence Berkeley National Laboratory
Berkeley, California, United States



-
1. How Electric Motors Work
 2. Development of Electric Motors and Their Role in Industrialization
 3. Current Status of End-Use Applications of Electric Motors
 4. Energy Efficiency Trends

Glossary

centrifugal The motion away from the center of rotation. Centrifugal fans, pumps, and compressors use centrifugal action to create pressure differences that result in the flow of fluids.

harmonics Electrical voltages or currents with frequencies that are integral multiples of the fundamental frequency, e.g., 150 cycles/second is the third harmonic in a 50-cycle system.

power factor The ratio between real power (measured in watts or kilowatts) and apparent power (the product of the voltage times the current measured in volt-amperes or kilovolt-amperes), expressed either as a decimal fraction or a percentage (0–100%). In the case of pure sinusoidal waveforms (i.e., not distorted by harmonics), the power factor is equal to the cosine of the phase angle between the voltage and current waves in an AC circuit. This value is known as the “displacement” power factor (because it deals with the displacement in time between the voltage and current). If the power factor is less than 1, more current is required to deliver a unit of real power at a certain voltage than if the power factor were 1. In the case of waveforms that include harmonics, the harmonic current adds to the total current without contributing to the real power, so the power factor is reduced. This higher current is undesirable because the energy lost to heat in the wires supplying power is proportional to the square of the current. In motors and other inductive loads operating in AC circuits, the current wave lags behind the voltage wave.

three-phase alternating current (AC) The flow of electrons back and forth along the conductors making up a circuit; the frequency of the alternation is 50 cycles/second in most of the world and 60 cycles/second in North America. Three-phase AC circuits have three conductors, each having an alternation at the nominal frequency, but the voltage waveforms in the three phases are offset from each other by 120° (equally out of phase around the 360° cycle).

Electric motors, devices that convert electrical energy to mechanical energy, power a wide range of applications in all sectors of the economy. Electric motors take over half of the world's electrical energy as their input. In this article, how the most common types of electric motors work, their development and application, and their trends in energy efficiency are described.

1. HOW ELECTRIC MOTORS WORK

All electric motors convert electrical energy into mechanical energy, typically in the form of a rotating shaft; they operate through the interaction of magnetic fields. The operation of the most important motor types is described in this section.

1.1 Induction Motors

1.1.1 Squirrel-Cage Induction Motors

The vast majority of the motors used in industry are squirrel-cage induction motors (Figs. 1 and 2) due to their low cost, high reliability, and fairly high efficiency. These motors also use more of the total electrical energy than any other motor

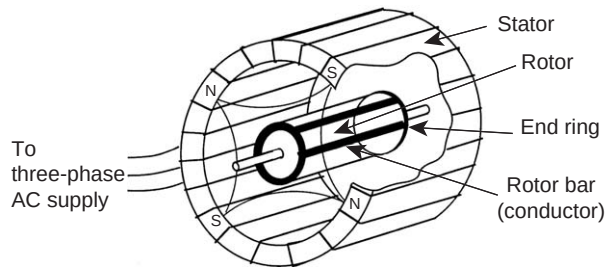


FIGURE 1 Diagram of squirrel-cage induction motor.

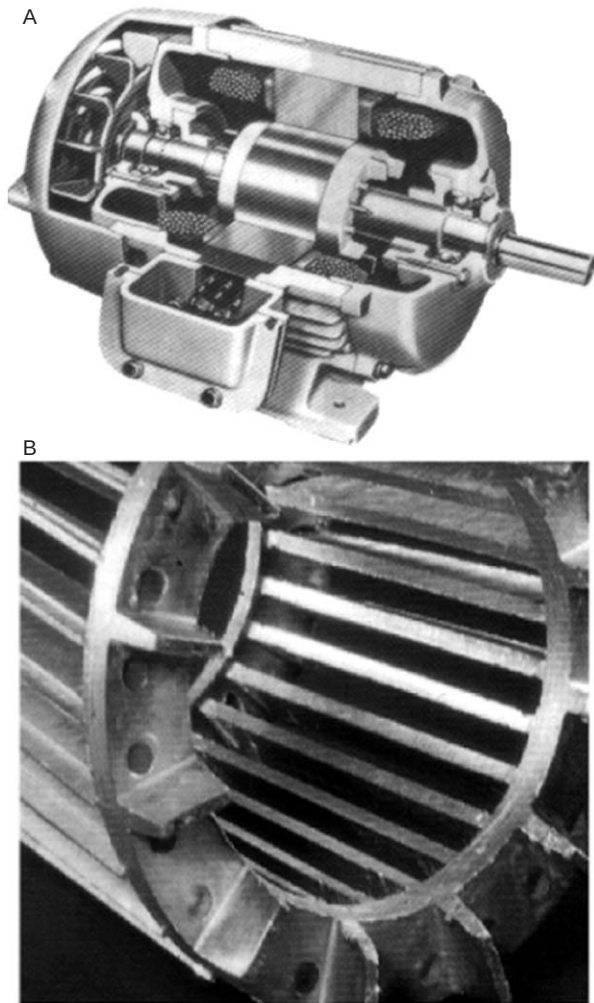


FIGURE 2 Squirrel-cage induction motor. (A) General structure. (B) Rotor squirrel cage (bars and end rings).

type in homes and buildings. The three-phase induction motor, invented in the 1890s, is composed of three sets of stator (the stationary part of the motor) windings arranged around the stator core (a stack of steel laminations with slots for the windings).

There are no electrical connections to the rotor (the rotating part of the motor), which means that there are fewer wearing parts to maintain and replace. When the three stator windings are fed by a three-phase alternating current (AC) supply, a rotating magnetic field is generated, the rotating speed of which is determined by the frequency of the power supply and by the number of magnetic poles of the motor. If the motor is fed by 60 Hz and has two pairs of poles as in Fig. 1, the speed of the rotating field will be 1800 revolutions per minute (rpm). When the rotating field crosses the squirrel-cage rotor (so named because the cylindrical shape of the conducting rotor bars resembles a squirrel cage), the variable magnetic flux induces currents in the rotor bars (hence the name induction motor), which leads to a pulling torque in the direction of the rotating field. The rotor accelerates to a speed close to the speed of the rotating field speed (synchronous speed), but never is able to reach it. If the two speeds were equal, there would not be a change in the magnetic flux (of the stator relative to the rotor), and thus no induced currents, and therefore no torque. The induction motor speed decreases by a few percent when the motor goes from no load to full-load operation. The full-load speed is a function of the motor design and power and is usually in the range of 94–99% of the synchronous speed; the difference is the “slip.” Figure 2 shows the typical construction of a squirrel-cage induction motor. The rotor squirrel cage, detached from the steel laminations can also be seen.

1.1.2 Wound-Rotor Induction Motor

Large induction motors can also have a wound rotor. As the name suggests, these motors feature insulated copper windings in the rotor, similar to those in the stator. This type of induction motor was once used in industrial applications in which the starting current, torque, and speed needed to be precisely controlled. In such applications today, squirrel-cage motors with variable-speed drives are much more commonly used.

1.2 Synchronous Motors

Synchronous motors, as the name implies, rotate at a constant (synchronous) speed. The rotor of this type of motor is a wound rotor, which receives the excitation (magnetizing) current from its excitation system (a separate direct current source with controller). Synchronous motors, although they are more costly and require more maintenance compared

to induction motors, are used in applications requiring constant speed (such as in the textile fiber and paper industries), a high degree of operating efficiency, and a controllable power factor. The last two factors are particularly important above 1000 kW. Synchronous motors are often used in large plants to drive the central air compressor. A large synchronous motor can be used to control the power factor of the whole plant, compensating the lagging power factor of a large number of medium and small induction motors.

1.3 Motor Efficiency

It must be emphasized when considering the energy efficiency of electric motors that the motors are part of a system. The loads (cooling, movement of fluids or materials, etc.) imposed on the system are worthy of attention, as well as the motor-driven process or equipment designed to meet those loads: both influence to a large extent the overall efficiency (units produced or service provided/kilowatt-hour) of the system. Automatic controls in industry and other sectors can reduce labor intensity and thus increase productivity, as well as increase energy efficiency.

1.3.1 Motor Losses

Part of the electrical energy supplied to the motor is not converted into useful mechanical energy, ending up instead in losses released to the outside of the motor in the form of thermal energy (Fig. 3). There are four different kinds of losses occurring in a motor: electrical, magnetic, mechanical, and stray. These losses can be reduced by using quality materials as well as by optimizing the design. The electrical losses are of the type I^2R (the product of the square of the electrical current I and the resistance R of the stator and rotor), and they increase rapidly with the motor load (because current and load are roughly proportional). These electrical losses can be decreased by increasing the cross section of the stator and rotor

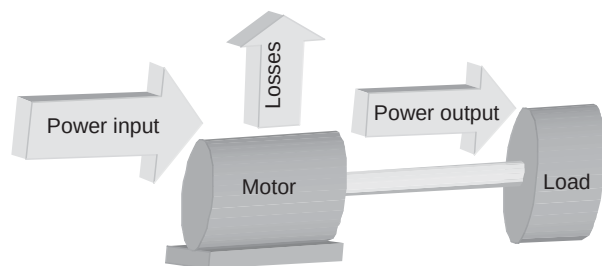


FIGURE 3 Power flow in a motor system (efficiency = output power/input power).

conductors. Magnetic losses occur in the steel laminations (“iron”) of the stator and rotor. They can be reduced by increasing the cross section of the iron in the stator and rotor and by using thinner laminations and improved magnetic materials. Mechanical losses are due to friction in the bearings, ventilation, and windage (air-drag) losses. They can be decreased using low-friction bearings and improved fan design. Stray losses are due to leakage flux, nonuniform current distribution, mechanical imperfections in the air gap, and irregularities in the air gap flux density. They can be reduced by optimal design and careful manufacturing.

1.3.2 Energy-Efficient Motors

After World War II and until the early 1970s, there was a trend to design motors that minimized the use of raw materials (namely copper, aluminium, and silicon steel). These motors, although inefficient, had lower initial costs and were more compact than previous generations of motors, but due to their low efficiency, their running costs were higher. By the mid-1970s, electricity prices started escalating rapidly. Most of the large motor manufacturers started offering a line of energy-efficient motors in addition to standard-efficiency motors. Efficiency improvements have now been achieved in the standard motor range. There is competition in the market based on motor efficiency. Energy-efficient motors (EEMs) feature optimized design, more generous electrical and magnetic circuits, and higher quality materials. Incremental efficiency improvements are still possible with the use of superior materials (e.g., amorphous silicon steel) and optimized computer-aided design techniques. EEMs typically have 30–50% lower losses (this translates into 2–10% better efficiency, depending on motor horsepower) compared to the equivalent standard motors. EEMs normally carry a price premium of around 20–25% in relation to standard motors. This translates into a price premium of \$5–22/kW (U.S. dollars). In a new application, and for a motor with a large number of operating hours, the payback time is normally under 2 years.

Because the reduced losses result in lower temperature rise in the motor, the lifetime of the motor winding insulation, of the bearings, and of the bearing grease increases. Reliability therefore increases, leading to less down time and reduced maintenance costs. Lower losses also provide the following advantages:

- Better tolerance to thermal stresses that typically result from stalls or frequent starting.

- Increased ability to handle overload conditions.
- Better resistance to abnormal operating conditions, such as under- and overvoltage or phase unbalance.
- Higher tolerance to poorer voltage and current wave shapes (e.g., harmonics).
- The ability to use smaller cooling fans, allowing quieter motor operation.

1.3.3 Proper Motor Sizing and Selection

Motors rarely operate at their full rated load. Field tests in the United States and Europe indicate that, on average, motors operate at around 60% of their rated load. Induction motor efficiency typically peaks near 75% of full load and is relatively flat down to the 50% load point. Motors in the larger size ranges can operate with reasonably high efficiency at loads down to 30% of rated load. In Fig. 4, the efficiency and power factor values at partial load points are shown. The adverse effect of the reduction of power factor due to oversizing is often neglected. Motor oversizing is a major reason for the low power factor of some industrial operations. Unless the reactive power is compensated for each motor, the additional line losses from higher currents due to oversizing are an additional incentive for proper motor selection. The original equipment manufacturers (OEMs), who are responsible for choosing the motor in the first place, as well as the users, who influence the buying phase or the replacement of failed motors, should

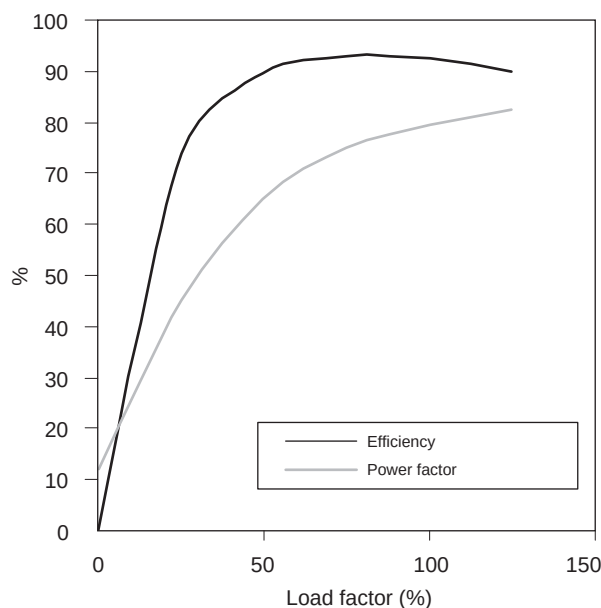


FIGURE 4 Typical efficiency and power factor versus motor load.

consider that the design criteria leading to oversizing can result in higher first and operating costs, especially when a motor has a working point far below 100% of rated power. Oversizing can also lead to an increase in the mechanical power used by the load. The operating speed is dependent on the applied load, and the loading imposed on a motor is in turn dependent on its size. For example, a 25% loaded 100-kW motor could be replaced by a 50-kW motor with a 50% load factor or a 25-kW motor with a 100% load factor. The oversized motor operates at higher speed, due to having a smaller slip. For centrifugal fans and pumps, even a minor change in the motor's operating speed translates into a significant change in imposed load and annual energy consumption. For most fans, the belt drive between the motor and the fan can be changed to obtain the correct fan speed; most pumps are direct drive, so the motor size, slip, and efficiency must all be carefully considered to make the best selection.

2. DEVELOPMENT OF ELECTRIC MOTORS AND THEIR ROLE IN INDUSTRIALIZATION

2.1 Early Motors

The first practical electric motors date back to the late 1800s. They were driven by direct current (DC) supplied by nearby generating plants. The early DC systems were superseded by alternating current, due to the ease of transmitting AC power generated centrally and to the simplicity of the AC induction motor, which was invented in the 1890s. Soon, DC motors were relegated to niche roles, especially when speed and motion control were needed (e.g., elevators and electric transportation).

2.2 Motor Evolution

2.2.1 Effect of Materials and Technology

The ability of the original motor winding insulation systems to withstand high temperatures was limited, so motors needed to be large, efficient, or both in order to keep the insulation from being damaged. As synthetic insulation materials were developed in the mid-20th century, higher operating temperatures were acceptable, allowing motors to be designed with lower operating efficiency, as noted in Section 1. In the 1970s and later, with renewed interest in efficiency, special steels were used in the magnetic

core to reduce energy losses there; computer-aided design and other techniques were used to design and build more efficient motors. The advent of inexpensive power electronics in the late 20th century made controlling the speed of AC motors easy and made DC motors with permanent magnets practical and efficient in applications ranging from residential furnace fans to drive motors for electric cars (see Section 4). Ongoing work includes motors that use high-temperature superconductors in the windings to further increase efficiency and motors that are extremely small (so-called nanomotors, because their size is on the scale of a nanometer).

2.2.2 Energy and Labor Considerations

For most of the 20th century, decreasing electricity costs and the evolution in insulation materials together put downward pressure on motor first cost and efficiency. Decreasing motor first cost and increasing labor rates meant that repairing motors in the United States and Europe gave way to simple replacement, especially for small motors. The electric motors that drive factory and construction equipment have been fundamental contributors to increases in worker productivity.

2.3 Role in Industrialization

2.3.1 Fluid Motion

The movement of fluids (collectively, liquids and gases) is the largest application for electric motors throughout the world.

2.3.1.1 Fans Fans are used for moving air at relatively low pressures; they are sized from a few watts for cooling small computers to megawatts for forcing combustion air through boilers at electric power plants. Fans are commonly used in systems that heat, cool, and ventilate buildings in which people live, learn, and work.

2.3.1.2 Pumps Pumps are used for moving liquids, most commonly water; they are sized from a few tens of watts for circulating water for heating homes to tens of megawatts for supplying drinking water to large cities. Industry uses large numbers of pumps for moving cooling water and a wide variety of liquid chemicals.

2.3.1.3 Compressors The most common use of compressors in industry is for making compressed air for operating air-powered (pneumatic) tools and

processes. A very common type of compressor through all sectors of the economy is the refrigeration compressor, used for ice making, air-conditioning, and food refrigeration. The scale of these compressors ranges from below 100 W for small household refrigerators to over 1 MW for large building air-conditioning systems.

2.3.2 Materials Transportation

Electric motors power conveyor belts, slurry pipelines, and pneumatic and air-blown systems used in the transport of all types of materials.

2.3.3 Materials Processing

In industry, materials are processed by a wide variety of electric motor-driven equipment such as grinders, saws, drills, crushers, stampers, shears, and lathes.

2.3.4 Transportation

Electric motors are widely used throughout the transportation sector. Electric railroad locomotives and diesel-electric locomotives both use electric motors in the low-megawatt size range. Likewise, streetcars and subways typically use DC motors, though there is a trend toward the use of AC induction motors with inverter drives. Electric and hybrid electric automobiles use AC induction motor/inverter drives or permanent magnet motors in the range of tens of kilowatts. Elevators, escalators, and people movers provide vertical and horizontal transportation in buildings, all powered by electric motors. Many ships are propelled by electric motors, supplied by onboard generating plants.

3. CURRENT STATUS OF END-USE APPLICATIONS OF ELECTRIC MOTORS

3.1 Common End Uses

3.1.1 Industrial Sector

In the industrial sector, motor-driven loads use about three-quarters of the total sector electricity use (74% of 2007 TWh/year for the United States and European Union combined). Common end uses are dominated by pumps and material processing, with air compressors and fans making significant contributions (see Fig. 5).

3.1.2 Commercial Sector

In the commercial sector (known as the tertiary sector in Europe), 37% of the electricity is used in

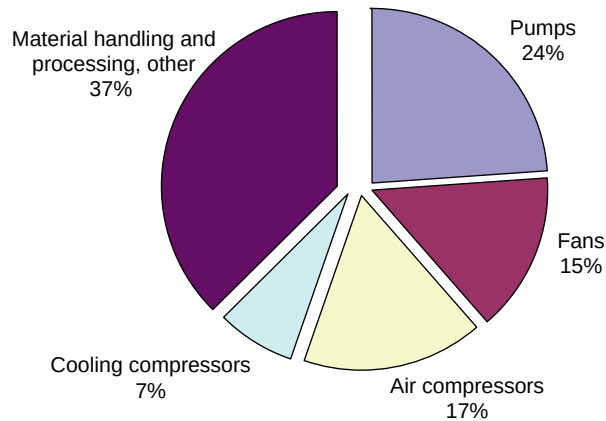


FIGURE 5 Industrial sector motor electricity use in the United States and European Union. Data from Nadel *et al.* (2002) and the University of Coimbra.

motor systems (out of a sector total of 1540 TWh/year for the United States and European Union combined); the dominant end uses are fans, air-conditioning, and refrigeration (see Fig. 6).

3.1.3 Residential Sector

In the residential sector, the dominant end uses in the United States are space conditioning (heating, fans, and air-conditioning) and refrigeration; motor-driven end uses require 38% of the residential sector's total of 1192 TWh/year (see Fig. 7). In Europe, a slightly higher fraction of the residential electricity is estimated to flow into motors.

3.2 Evolving End Uses

3.2.1 High-Performance Motion Control

In many industrial applications (e.g., robotics), there is increasing use of high-performance permanent-magnet or switched-reluctance drives (see Section 4 for more on these drives). With the advent of vector drives (inverters that separately control the speed and torque of AC induction motors), many applications (e.g., elevators) that were the stronghold of DC motors are giving way to simpler and more efficient AC motors.

3.2.2 Increasing Use of Speed Control

The cost of inverters for speed control (see Section 4) is continuing to decrease, and the reliability is increasing as the circuitry becomes more integrated (more functions on fewer components). As a result, the controls are becoming standard in many applications in homes, commercial buildings, and industry

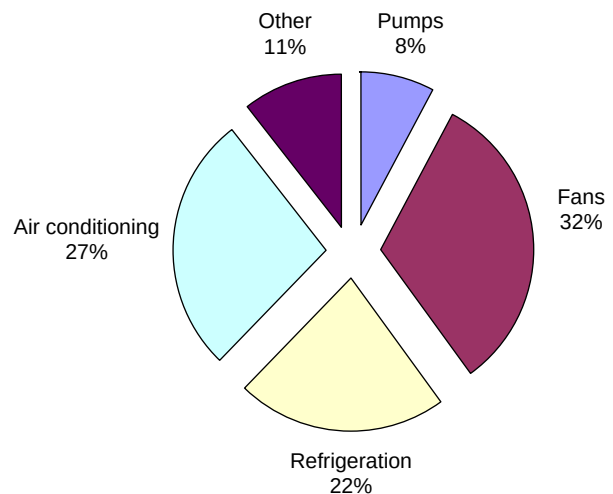


FIGURE 6 Commercial (tertiary) sector motor electricity use in the United States and European Union. Data from Nadel *et al.* (2002) and the University of Coimbra.

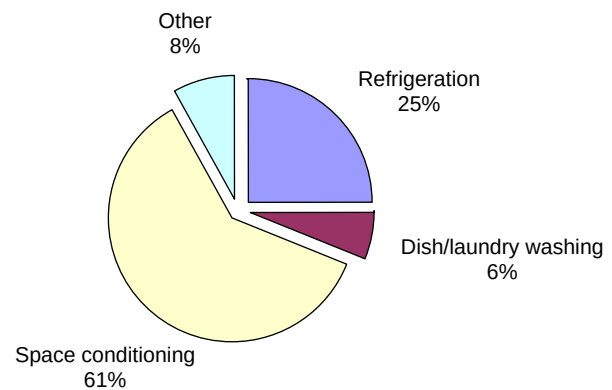


FIGURE 7 Residential sector motor electricity use in the United States. Data from Nadel *et al.* (2002).

that formerly used on/off cycling or throttling to match output to load. Examples include residential furnace fans, ventilation fans, and air-conditioning compressors in office buildings, and many industrial applications. A continuing challenge is the control hardware and especially software to take full advantage of the improved process control and energy-savings opportunities offered by speed control.

3.2.3 Electric Transportation

3.2.3.1 Rail Due to the availability of powerful and compact inverters, there is an increasing tendency to use AC induction motors with inverter drives rather than DC drives. In the future, high-speed rail will include magnetic levitation trains (using magnetic repulsion to float the cars above the

guideway) that use a linear motor to propel the cars along the guideway.

3.2.3.2 Road Electric and hybrid electric vehicles use sophisticated water-cooled drives that include regenerative braking to increase energy efficiency. Some designs use AC induction motors and some use permanent-magnet motors. These will continue to evolve as hybrid vehicles move from internal combustion engines to fuel cells as the primary power source.

3.2.3.3 Air The National Aeronautics and Space Administration in the United States is flying unmanned airplanes powered by numerous electric motors that gather their electricity from onboard solar–electric panels and fuel cells. These can fly at very high altitudes for long periods and could eventually perform many of the functions of satellites at much lower cost.

4. ENERGY EFFICIENCY TRENDS

4.1 Technology

4.1.1 Permanent-Magnet Motors

In the last decades of the 20th century, new technologies led to the introduction of new motor types that have some performance and efficiency advantages over induction motors. Permanent-magnet (PM) motors have a stator winding configuration similar to the three-phase induction motors, but they use permanent magnets in the rotor instead of a squirrel cage or wound rotor. The permanent-magnet rotor tracks with synchronism the stator rotating field, and therefore the rotor speed is equal to the rotating magnetic field speed. A wide variety of configurations is possible, and the two-pole version can be seen in Fig. 8. Motors of this sort have output ranging from about 100 W up to 100 kW. The key advantage of the permanent-magnet motor is that no excitation power is required for the rotor and therefore its efficiency is higher, compared to the induction motor. Early permanent-magnet motors, besides being very expensive, suffered from the tendency of the magnets to be demagnetized by the high stator currents during starting, and from a low maximum allowable temperature. Much improved versions using rare-earth magnets have been developed since the 1970s to overcome these problems. Rare-earth alloys (neodymium, iron, and boron),

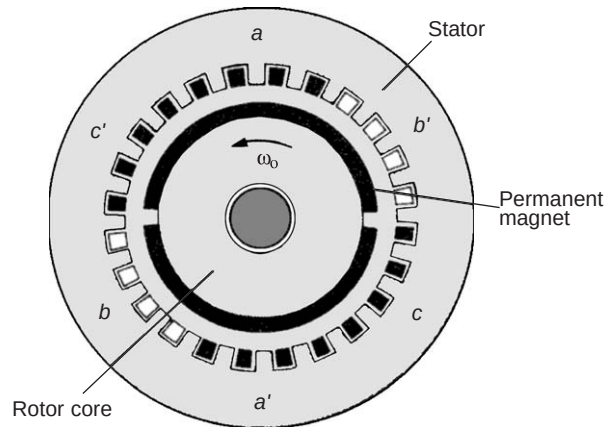


FIGURE 8 Permanent-magnet motor: two-pole version. See text for discussion.

developed in the mid-1980s, have allowed an increase in performance with a decrease in costs. For starting from a fixed-frequency supply, a rotor cage is required; such motors are referred to as “line-start” motors. Because there are no electric and magnetic losses in the rotor, there is less heating than in a conventional motor, so higher specific outputs can be achieved. The rotor inertia can also be less than that of an induction motor rotor, which means that the torque/inertia ratio is better, giving a higher acceleration. The torque-to-weight ratio, the steady-state efficiency, and the power factor at full load of PM motors are in most cases better than those of the equivalent induction motor, and they can pull in to synchronism with inertia loads of many times rotor inertia.

Although better efficiency and low-speed performance can be obtained with permanent-magnet synchronous motors, there would need to be a significant reduction in the cost of rare-earth permanent-magnet material for these motors to replace induction motors in most applications. Permanent-magnet motors are in most cases used with electronic speed controls, normally called brushless DC (BLDC) motors. In the BLDC motor, the stator winding currents are electronically commutated (switched) by digital signals from simple rotor position sensors; these motors are also known as electronically commutated motors (ECMs). The stator winding also creates a rotating field that creates a rotating torque by pulling the permanent-magnet rotor. This combination permits the motor to develop a smooth torque, regardless of speed. Very large numbers of brushless DC motors are now used, particularly in sizes below 10 kW. The small versions (less than 0.75 kW) are

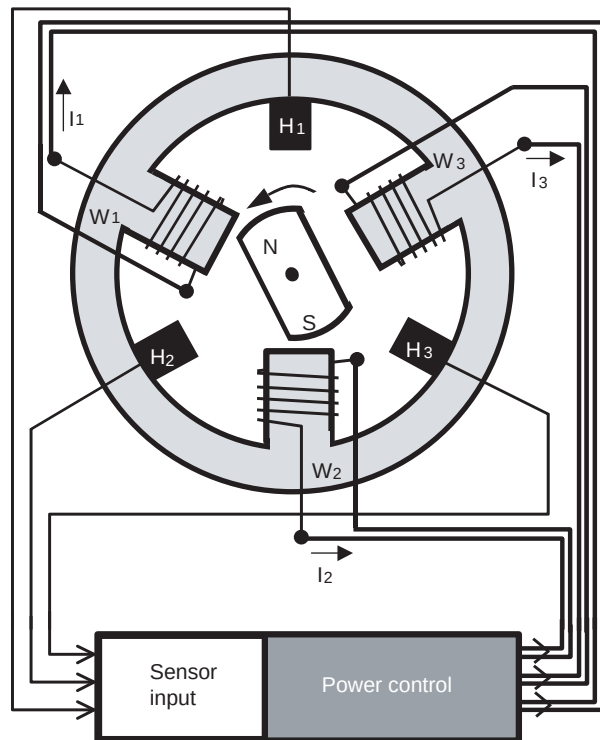


FIGURE 9 Control system scheme of a brushless DC motor. H_1 , H_2 , and H_3 are magnetic position sensors.

increasingly made with all the control and power electronic circuits integrated at one end of the motor. The typical stator winding arrangement and the control scheme are shown in Fig. 9.

Many brushless motors are used in demanding servo-type applications (e.g., robotics and high-performance systems), in which they need to be integrated with digitally controlled systems. For this sort of application, complete digital control systems that provide torque, speed, and position control are available. New applications such as energy-efficient elevators with direct-drive (gearless) permanent-magnet motors are also entering the market.

4.1.2 Switched-Reluctance Motor

The switched-reluctance (SR) drive is a recent arrival (ca. 1990s) on the drives scene, and can offer advantages in terms of efficiency, power density, robustness, and operational flexibility. The drawbacks are that it is relatively unproved, can be noisy, and is inherently not well suited to smooth torque production. Despite being recent, SR technology has been successfully applied to a wide range of applications, including general-purpose industrial drives, traction, domestic appliances, and office and business equipment.

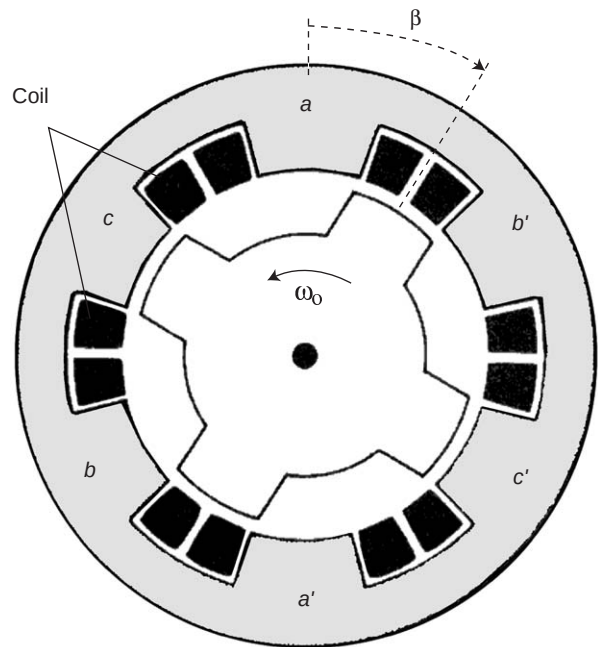


FIGURE 10 Switched-reluctance motor configuration.

In the switched-reluctance motor, both the rotor and the stator have salient (protruding) poles. This doubly salient arrangement proves to be very effective as far as electromagnetic energy conversion is concerned. The stator has coils on each pole, the coils on opposite poles being connected in series. The rotor, which is made from machined steel, has no windings or magnets and is therefore inexpensive to manufacture and extremely robust. The example shown in Fig. 10 has six stator poles and four rotor poles, and represents a widely used arrangement, but other pole combinations are used to suit different applications. The motor rotates by energizing the phases sequentially in the sequence a - a' , b - b' , c - c' for clockwise rotation or a - a' , c - c' , b - b' for counter-clockwise rotation, the “nearest” pair of rotor poles being pulled into alignment with the appropriate stator poles by reluctance (the alignment of a metal object with a magnetic field) torque action. In this way, similar to PM motors, the rotor synchronously tracks the stator rotating magnetic field.

As with the PM motor, the SR motor has no electric or magnetic losses in the rotor. Therefore, the overall efficiency is generally higher compared to induction motor efficiency. The SR motor is designed for synchronous operation, and the phases are switched by signals derived from a shaft-mounted rotor position detector. This causes the behavior of the SR motor to resemble that of a BLDC motor.

Because the direction of torque in a reluctance motor is independent of the direction of the current, it means that the power converter can have fewer switching devices than the six required for three-phase bipolar inverters used in BLDC motor. Some of the early SR motors were deemed to be very noisy, but improved mechanical design has significantly reduced the motor noise.

4.1.3 Speed Controls

Most induction AC motors operate at fixed speed. However, most motor applications will benefit if the motor speed is adjusted to match the process requirements. Motor speed controls are the devices that, when properly applied, can tap most of the potential energy savings in motor systems. Motor speed controls are particularly attractive in applications in which there is a variable fluid flow. In many centrifugal pump, fan, and compressor applications, the mechanical power grows roughly with the cube of the fluid flow; e.g., to move 80% of the nominal flow, only half of the power is required. Centrifugal loads are therefore excellent candidates for motor speed control. Other loads that may benefit from the use of motor speed controls include conveyers, traction drives, winders, machine tools, and robotics.

Conventional methods of flow control used inefficient throttling devices such as valves, dampers, and vanes. These devices, although they have a low initial cost, introduce high running costs due to their inefficiency. Several speed-control technologies can be used to improve motor system operation. Electronic variable-speed drives (VSDs) are discussed in detail here because they are the dominant motor speed-control technology.

As was mentioned in Section 1.1, the speed of the rotating field created by the induction motor stator windings is directly linked with the frequency of the voltage applied to the windings. Electronic variable-speed drives can produce variable-frequency, variable-voltage waveforms. If these waveforms are applied to the stator windings, there will be a shift of the torque-speed curve, and the same slope of the linear operation region of the curve. In this way, the motor speed is proportional to the applied frequency generated by the VSD (Fig. 11). Figure 12 shows the general configuration of most VSDs. The three-phase alternating current supply is initially converted to direct current, then filtered, and, finally, the DC/AC inverter converts the DC voltage to the variable voltage and variable-frequency output applied to the motor. The adjustment of the motor speed through the use of VSDs can lead to better process control,

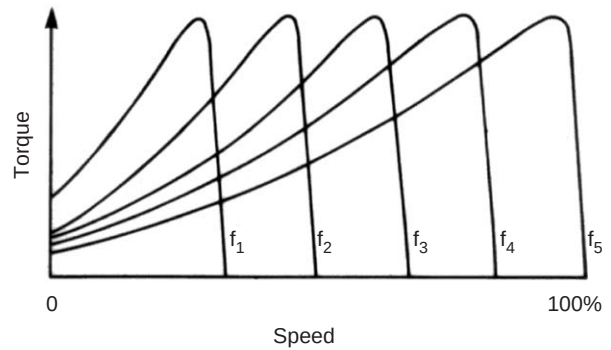


FIGURE 11 Speed-torque curves for an induction motor ($f_1 < f_2 < f_3 < f_4 < f_5$ and f_5 = normal line frequency).

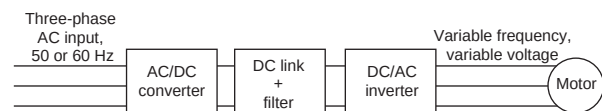


FIGURE 12 General configuration of inverter-based variable-speed drives.

less wear in the mechanical equipment, less acoustical noise, and significant energy savings. However, VSDs can have some disadvantages, such as electromagnetic interference (EMI) generation, current harmonics introduction into the supply, and the possible reduction of efficiency and lifetime of old motors.

There are five basic types of power electronic VSDs: two voltage-source (pulse-width modulation and six-step) types, two current-source (force-commutated and load-commutated) types (typically for large motors), and cycloconverters (typically for large-motor, low-speed applications). Of these, the pulse-width-modulated VSD has come to dominate the market.

4.1.3.1 Pulse-Width-Modulated VSDs The pulse-width-modulated (PWM) voltage source inverter is widely accepted as giving the best overall performance below power levels of 1 MW. The PWM inverter maintains a nearly constant DC link voltage, combining both voltage control and frequency control within the inverter. The objective of the sinusoidal PWM is to synthesize the motor currents as near to a sinusoid as economically possible. The lower voltage harmonics can be greatly attenuated, and therefore the motor tends to rotate more smoothly at low speed. Higher order harmonic motor currents are limited by the motor inductance (the tendency to resist changes in current). Torque

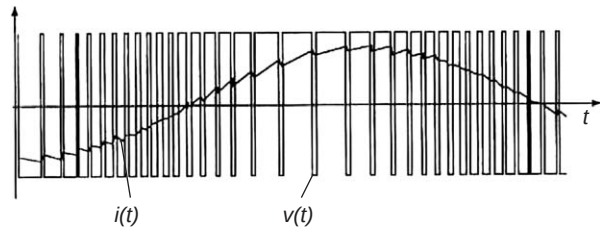


FIGURE 13 Illustration of a sinusoidal pulse-width modulation technique.

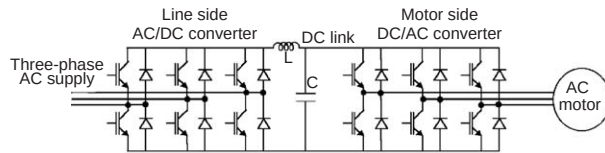


FIGURE 14 Power circuitry of a voltage source inverter pulse-width modulator with regenerative capacity and power factor control. Whereas conventional VSDs use diode rectifiers in the input stage, regenerative units use insulated gate bipolar transistors (IGBTs) at both the input and the output stages to allow bidirectional power flow.

pulsations are virtually eliminated and the extra motor losses caused by the inverter are substantially reduced (see Fig. 13).

The advantages of the PWM VSD are the good displacement power factor throughout the speed range, low distortion of the motor current, wide speed range, and, with proper circuitry, regeneration capability, which allows the motor to absorb energy from the load, generate electrical power, and feed it back into the electrical supply. When the power rating is large and/or when the motor may have to operate for long periods of time in the regenerating region, the line-side converter can be modified to allow regeneration, feeding power back to the line by reversing the direct current in the DC link. New (ca. 2000) PWM circuitry (Fig. 14) allows for the braking energy to be injected back to the source. This feature can be a way of saving a significant amount of energy in applications with high speed/high inertia (e.g., centrifugal separators) and/or long time operations in the braking mode (e.g., elevators, downhill conveyor belts).

4.2 Policies to Promote Energy-Efficient Motor Technologies

Experience from many energy-saving initiatives around the world shows that the most successful programs to achieve the potential savings are based on a combination of technical information, financial

incentives, voluntary agreements, and, in some cases, the use of regulation. Strategies involving education, technical support, financial aid, and legislative standards can be pursued.

4.2.1 Education and Training

More education on energy efficiency is needed on many fronts. The information gap should be filled through several means:

- Media, including advertising, trade press releases, direct mail, electronic mail, Web site, and exhibitions; the mix of these depends mainly on the available budgets, with care taken to examine the possible impacts.

- Publications, including a selection of materials from short and simple introductory leaflets and brochures to guides, handbooks, journals, and software; detailed technical information may be needed to satisfy more experienced personnel. Literature should be issued and distributed by manufacturers, trade associations, distributors, energy agencies, etc.

- Training and workshops.

- Dissemination of educational materials, not only through existing energy agency databases and via the media (paper-based products, software programs, and Internet information), but also third parties such as trade associations and equipment suppliers, groups with tremendous potential leverage.

4.2.2 Provision of Technical Assistance

Decision support tools, such as user-friendly software tools, are required for more standardized comparison of efficiency and economics of system configurations. These include support tools for product selection, such as the U.S. Department of Energy's Motor Master and (for Europe) the European Database of Energy-Efficient Motors (EURODEEM) for motors, guides that help users to select particular equipment for an application.

On-site assessment and assistance can provide motor users with information describing their savings potential. Energy-saving publications frequently use comprehensive measurements to ensure the reliability of the figures presented, but in practice most site personnel will have the time and resources only for much simpler measurements. Demonstration projects and pilot actions can be very helpful. One key factor that limits the adoption of energy-efficient technologies is the scepticism about their viability. People doubt equipment durability, performance, and real savings. There already are many demonstration projects and case studies on the

application of EEMs and VSDs; additional demonstration projects aimed at new technologies as well as easier access to the results of existing actions are needed.

4.2.3 Financial Tools

Tax incentives, including tax reduction and accelerated depreciation on efficient equipment, e.g., with VSDs, would be very effective measures to promote efficiency. Another market mechanism for improving energy efficiency involves the development of the energy service company (ESCO) industry. Energy service companies are often privately owned, non-regulated organizations that contract with a utility or an end user to assess, finance, and install energy conservation measures. The financial mechanism used by ESCOs for encouraging the use of their services is performance contracting. The ESCO takes as payment a share of the energy cost reduction, based on either engineering estimates or measurements. Yet another mechanism, rebates, is now seen primarily as a short-term measure to help stimulate the market for energy-saving products; this served in North America as a very useful precursor to legislation on minimum motor efficiency standards.

4.2.4 Efficiency Standards

The establishment of minimum efficiency standards for motors, for end-use devices, and for motor systems is an important market transformation tool. These standards are particularly effective for OEMs, obliging them to incorporate efficient devices in their equipment. A case in point in the United States is the mandatory motor minimum efficiency standards (Energy Policy Act), enforced in October 1997, as well as minimum efficiency standards for appliances dominated by motor use (e.g., refrigerators and air conditioners).

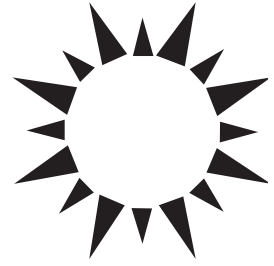
SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Batteries, Transportation Applications • Electrical Energy and Power • Electricity Use, History of • Electromagnetism

Further Reading

- De Almeida, A., Fonseca, P., Falkner, H., and Bertoldi, P. (2003). Market transformation of energy-efficient motor technologies in the EU. *Energy Policy* 31(6), 563–575.
- De Almeida, A., Fonseca, P., and Bertoldi, P. (2003). Energy efficient motor systems in the industrial and in the services sectors in the European Union: Characterization, potential barriers and policies. *Energy Int. J.* 28(7), 673–690.
- European Commission (2000). “Improving the Penetration of Energy-Efficient Motors and Drives.” Project report, Institute of Systems and Robotics—University of Coimbra. European Commission, Brussels.
- Institute of Systems and Robotics (ISR)—University of Coimbra. (2001). “VSDs for Electric Motor Systems.” Project report, ISR—University of Coimbra, May 2001. Coimbra, Portugal.
- Joint Research Center (JRC), European Commission (2004). European Database of Energy-Efficient Motors (EURODEEM) and Motor Challenge Programme. Information available on the JRC Web site at <http://www.energyefficiency.jrc.cec.eu.int/eurodeem/>.
- Nadel, S., Elliot, N., Shepard, M., Greenberg, S., Katz, G., and de Almeida, A. (2002). “Energy-Efficient Motor Systems.” 2nd ed. American Council for an Energy-Efficient Economy, Washington, D.C.
- U.S. Department of Energy, Energy Information Administration (EIA). (2004). “Annual Electricity Sales by Sector.” Available on the DOE Web site at <http://www.eia.doe.gov/>.
- U.S. Department of Energy, Office of Industrial Technologies (OIT). (2004). Motor programs available on the OIT Web site at <http://www.oit.doe.gov>.
- U.S. Environmental Protection Agency (EPA) and U.S. Department of Energy (DOE). (2004). Energy Star award rules and instructions; continuously updated information available on the EPA/DOE Web site at <http://www.energystar.gov>.

Electric Power: Critical Infrastructure Protection



ALEXANDER E. FARRELL

University of California, Berkeley
Berkeley, California, United States

HISHAM ZERRIFFI

Carnegie Mellon University
Pittsburgh, Pennsylvania, United States

-
1. Defining the Problem
 2. Turning Out the Lights
 3. Other Attack Modes
 4. Mitigation

Glossary

cascading failure A disruption at one point in a network that causes a disruption other points, leading to a catastrophic system failure.

critical infrastructure protection (CIP) Activities and procedures designed to avoid, prevent, detect, resist, and recover from deliberate attacks on electric power systems.

distributed generation (DG) Electricity production near point of use, typically with small (< 10 MW) generators using natural gas. One form of embedded generation.

energy management system (EMS) System of sensors, actuators, communications, and computation hardware and software used to operate an electric power system. Key functions include generation scheduling and control, transmission system analysis and control, and operator training.

outage A temporary suspension of electric service; a loss of electric power.

security The ability of electric power systems to withstand accidental, sudden disturbances such as electric short circuits or unanticipated loss of system elements.

stress Electric power system conditions more severe than those considered in reliability planning analysis, that is, outside of typical operating conditions in industrialized countries. These may include deliberate attacks, multiple failures, persistent failures (or attacks), and conditions that make system restoration difficult (i.e., a lack of spare parts).

survivability The ability of a system to fulfill its mission in a timely manner, despite attacks, failures, or accidents by detecting, resisting, and recovering from attacks. Survivable systems are likely less vulnerable to stress.

The rise of terrorism around the world and the persistence of inadequate investment or maintenance in some areas have given rise to concerns about how to safeguard key infrastructures (energy, communications, banking, roads, etc.). Electric power systems are among the most difficult to protect because of their complexity. The concept of critical infrastructure protection is still relatively new and is likely to evolve over time, possibly away from a “guards, gates, and guns” defensive approach and toward a design approach that yields systems that are inherently harder to successfully attack. Such survivable systems may feature distributed intelligence, control, and operations.

1. DEFINING THE PROBLEM

Deliberate attacks on electric power systems occur around the world, although their character and magnitude vary greatly. In the industrialized world, such events are small and infrequent; they pale in comparison to the largest weather-related events. Modern war-fighting includes tactics and weapons designed to damage electric power systems, but preparation for war is outside of the power system planning processes. Thus, attention paid to deliberate attacks on electric power systems is relatively scarce. However, the growth of terrorist acts within

developed countries since the late 20th century has raised concerns about critical infrastructure protection (CIP) for power, fuel, telecommunications, transportation, and other systems.

Even less attention is paid to the reconstruction of power systems damaged by conflict and in places where political tensions may result in frequent deliberate attacks in the future. The most obvious terms for these issues, reliability and security, have different, well-established meanings in electricity system planning. Thus, we will use CIP to refer to activities and procedures designed to avoid, prevent, detect, resist, and recover from deliberate attacks on electric power systems. The understanding of the CIP implications for electric power systems is rather limited, and the ability to manage the threat of deliberate attacks even more so. Significant research and development efforts are expected over the coming decade, with subsequent deployment of new approaches and technologies.

Electricity cannot be stored economically, so generation and consumption must be matched at every instant. In addition, electric power systems must be managed so that several other parameters (including voltage, reactive power, and frequency) remain within specifications at all times. Not only can these parameters change extremely quickly, electric power systems are typically large interconnected grids that are vulnerable as a system due to the possibility of cascading failures in which a small number of failures in a network cause other failures, leading to widespread disruption and outages. Another key feature of electric power systems is that they are inherently vulnerable to attack since they extend over great distances, often through relatively isolated areas. Further, transmission systems are extremely complex with many custom-designed and hard-to-replace components, including transformers, switches, and relays. These components are monitored and operated by technicians both in the field and at central control centers using electronic systems while other groups form electricity markets that use different computer networks. These applications open up the possibility of cyber attack on electric power systems.

To better understand CIP, it is necessary to introduce three key concepts. The first is reliability and how electric power system planners have dealt with normal failures in the industrialized countries. The second concept is stress, and why planning for reliability must change when the system is under stress. The third concept is cascading failures of the electric power system and how they relate to reliability planning and stress.

1.1 Reliability

These constraints make CIP a greater challenge in electric power systems than in any other part of the energy sector. However, the necessity of dealing with these constraints during normal operations has resulted in a long history of reliability planning, which includes consideration of both normal operation under reasonable expectations of intended operating conditions (adequacy) and of limited disturbances due to weather and occasional equipment failure that are unplanned but not unexpected (security). These conditions fairly well describe most conditions in high-income countries, but they do not describe a variety of situations now considered important in both high- and low-income countries (see Table I). Frequent deviations from intended operating conditions in industrialized countries is defined as a condition of stress, which includes deliberate attacks, multiple failures, persistent failures or attacks, and conditions that make system restoration difficult (e.g., a lack of spare parts). Cascading failures are still largely excluded since stress is a planning concept and concerned with improving reliability under stress, not an operating concept, unless the stress is directly linked to a change in the probability and magnitude of cascading failures.

Generally, reliability concepts are thought of first when CIP is considered by the electric power industry. However, traditional concepts related to electricity reliability may not be appropriate for achieving protection of critical infrastructures from deliberate attack. There are two main reasons for this; first, the assumptions made in reliability analysis about single point failures may not be realistic; second, some attack modes do not necessarily result in service outages, which is the focus of reliability planning.

In the past, reliability concerns have been handled from the supply side under the fairly restrictive assumption of a single mode failure. The goal that planners adopted was to protect generators and transmission and distribution networks from the ill effects of an equipment failure or natural disturbance (e.g., a tornado). Thus, practice is to have available at all times more capacity than actually needed in order to deliver power to the end users under both normal and a limited set of abnormal conditions. The reserve capacity typically required under the single failure mode assumption is sufficient for the largest single generator or transmission system component to fail without causing either damage to other equipment or a loss of power to any customers. This rather brute force, reliability-through-redundancy approach has

TABLE I
Modes of Disturbance in Electric Power Systems

Mode of disturbance	Possible causes	Likely characteristics	Likely impacts
Weather related damage ^a	Hurricanes, tornadoes, floods, ice storms	Random, not repeated, not targeted, regional	Impacts T & D primarily. No long term impacts on failure probabilities, magnitudes, or durations; recovery only hampered by environmental conditions
Equipment failure ^a	Trees, animals, or vehicles contacting power lines, unanticipated wear	Single events occur randomly, but at a known, acceptable rate	Reduces the equipment available to the system by one component; managed to obtain desired balance of cost and reliability
System-wide direct conflict damage	Civil war (e.g., Bosnia), guerilla movement	Persistent, system-wide, impacts all levels of system	Both failure probabilities and magnitude of damage high, recovery difficult and expensive due to continuing conflict
Regional direct conflict damage	Regional insurgency	Persistent but localized, impacts all levels of system	Failure probabilities and magnitudes increase in affected region, recovery difficult
Localized direct conflict damage	Terrorism or sabotage	Targeted, repeated (lower frequency), less damage per attack on average, less damage to large generators	Failure probabilities increase, magnitudes do not increase greatly except for the most extreme acts, recovery relatively unhampered
System-wide indirect conflict damage	Civil war (e.g., Bosnia), guerilla movement	Mobility hampered, increased nontechnical losses creating financial problems	Failure probabilities increase, magnitudes of failures do not increase, recovery more difficult
Regional indirect conflict damage	Regional insurgency	Regional mobility hampered, increased nontechnical losses, financial problems	Failure probabilities increase, magnitude of failures do not increase, recovery more difficult
Lack of investment in new capacity	Capital access, investment uncertainty	Units need to be run more often and for longer as reserve margins decline, transmission system strained	Possible increase in failure rates over time
Poor maintenance	Capital and spare parts access		Failure rates increase over time, repair times increase

^aThese disturbances are part of reliability planning. All others are stress conditions.

proven effective in the past, but its efficacy is challenged both by the advent of restructuring and the potential for terrorist or other military attacks.

The rationale for a supply-side approach to reliability was that by protecting the electricity supply system; the concerns of most end users would automatically be accommodated. These concerns were quantified with reliability indices. However, these indices do not differentiate among the very different needs of the end users, nor are supply-side measures alone likely to meet the needs of all

customers. For instance, a sag in voltage that lasts no more than a fraction of a second would go unnoticed by most end users but could be a serious event for some. Eliminating such sags from the supply side is difficult, if not impossible. Similarly, procedures for measuring indices ignore outages less than a minute long, even though a few milliseconds of lost power may be sufficient to interrupt or damage electronic equipment.

No matter what the cause, power outages impose significant costs on electricity users, while there are

significant costs to electricity companies for avoiding outages. Lowering outage rates raises electricity costs, especially for high levels of reliability. Choosing an acceptable level of outages based on supply costs as well as the costs of outages is a key decision for power system planners and regulators. For instance, typical average outage rates in the United States are 1.4 outages per year, and a total time without power of about 3.5 hours per year. About half of the time without power is due to problems with the local distribution system and about half are associated with the transmission system.

In the past, a balance between outage costs and reliability costs has been struck through a political/legal process of some sort, whether by a Ministry deciding on power system design and prices or by a public service commission setting rates for private power companies who operate within a set of (usually voluntary) reliability guidelines. While this approach has been effective, it has also served the interests of utility managers in both public and private systems. In public systems, reliability through redundancy meant higher budgets and more employees, generally seen as markers of status in government bureaucracies. In private systems, reliability through redundancy served shareholders by tending to increase the size of the ratebase, and thus profits.

However, it is not clear how this balance will be struck if and when privatization and competition are used to guide most investment and operating decisions in the power sector. Voluntary reliability requirements could be harder to implement, although competition may lead to flexible, lower cost alternatives. CIP adds a new concern to this balancing act; actions to prevent or mitigate deliberate attacks may raise the cost of electricity service directly through new expenses or indirectly by limiting the flexibility of firms to innovate or economize. Minimizing cost increases due to CIP is an important goal—protecting a country's homeland security should not slow its economic growth unduly. A further complication will be deciding who pays for what fractions of the cost of CIP. This will be an area of research, development, politics, and practice for some time.

1.2 Stress

Stress differs from the modes of disturbance analyzed in typical reliability planning in several important ways, as seen in Table I. Since electrification occurred first in industrialized countries and has been studied most extensively there, analysis of the electricity

sector reflects conditions in industrialized countries. Typical conditions in low-income countries have not been well studied. In addition, deliberate attacks on infrastructure have not been a significant concern in industrialized countries until relatively recently, so these conditions have not been well studied either. This article focuses on deliberate attacks, which will include one or more of the following general characteristics (to different levels of intensity).

1.2.1 Coordination of Attack

Unlike equipment failures or even extreme weather events, deliberate attacks are not short-term, random events. Both militaries and terrorists alike have the capabilities to coordinate attacks and intentionally maximize damage to the system. For example, the Farabundo Marti National Liberation Front (FMLN) was able to interrupt service to up to 90% of El Salvador at times and even produced manuals on disrupting electricity systems. Attacks can also be repeated as system components are repaired.

1.2.2 Scope of Impacts

The failure of distribution equipment within a half mile of the customer is the cause about half of the outages in the United States. However, in conflict situations, remote transmission lines, generating stations, and transformer substations can also become primary targets or be impacted by indirect damage stresses. The nonconflict stresses discussed earlier (e.g., poor access to spare parts) would also have a wider scope of impacts.

1.2.3 Persistence of Outage

Deliberate attacks may not be single events, and they may occur under conditions (or create conditions) that make restoration of service and repair more difficult. Alternatively, conditions could be created that make even necessary routine maintenance more difficult. When outages do occur, a number of factors can cause them to persist for days or even longer. Such factors would include risks to personnel, impeded transportation of personnel and parts, lack of funds for replacement parts, absence of technical expertise, and the possibility of subsequent sabotage. Long outages are not considered normal by electricity planners in industrialized countries (a sustained interruption is generally classified as one that lasts more than 1 hour, there is no classification level for longer outages).

1.3 Cascading Failures

Large-scale blackouts result from cascading failures, which occur when equipment failure or other disruption cause further failures or disruptions in large portions of the power grid. These events are relatively rare but potentially very costly. Reliability planning is meant to avoid creating situations in which a cascading failure is even possible while accepting that some localized (normal) outages will occur. However, such cascading failures do occur in industrialized countries and are not unique to stressed systems or to concerns about CIP. Reliability improvements may reduce the probability of cascading failures but will not eliminate them, and it may be difficult to determine the precise impact of reliability improvements on the frequency and size distribution of cascading failures.

Changes in electricity system planning to account for stress should reduce the higher incidence of normal outages that would otherwise be expected under these conditions. Avoiding cascading failures is still a goal, but planning for stressed systems is considered with wider impacts of stress and over much longer timescale than any individual cascading failure. In some places, rolling blackouts may already be the norm rather than the exception. These are not the result of cascading failures but are imposed in a controlled fashion on a regular basis due to inadequate generation. In fact, cascading failures may not even be a major concern for many stressed systems. In a tightly coupled, interconnected, and generally well-functioning system like in the United States, cascading failures are a real possibility and concern. On the other hand, the infrastructure in many places may be only loosely interconnected, which may make widespread cascading failures less likely. As a result, electric power systems under stress may or may not suffer cascading failures at a higher rate depending on the context, but they will suffer more and possibly longer localized outages.

2. TURNING OUT THE LIGHTS

One of the primary concerns for CIP in electric power systems is preventing outages, particularly large blackouts. There is much to be learned from the study of past outages. Contingency planning, spare equipment preplacement, emergency preparedness training, and temporary personnel transfers tend to be the keys to recovery from large outages. Power companies often loan trained workers in emergencies under fairly informal arrangements, knowing that

they are likely to need to make such requests themselves someday. Several aspects of successful management of electric system outages have been identified: planning, training, analysis, and communications. However, communication during and after an event might be difficult, especially if, as discussed later, the attack is not on the physical part of the electric power system but on the information technology (IT) components.

The experience of several high-profile outages reinforces the need for planning and leadership. One of the largest such event in U.S. history was the Great Northeastern Blackout of 1965, which shut down New York, as well as much of Ontario and New England after an equipment failure. New York City was blacked out for an entire night, and about 30 million people were directly affected. The National Opinion Research Center studied this event and discovered the following:

An outstanding aspect of public response to the blackout was the absence of widespread fear, panic, or disorder. There is probably little question that this absence is largely due to the ability of individuals to interpret their initial encounter with power loss as an ordinary event.... Of equal importance in maintaining order was the rapid dissemination of information about the blackout.

A smaller outage in New York City in July 1977 was far more troubling. Although outright panic did not emerge, there was considerable looting in some parts of the city. At the time, crime had been increasing and there was a general sense of insecurity in New York City. Some of the looting occurred due to poor response by the police, leaving some areas of the city more vulnerable. The 1977 blackout was frightening in ways that ordinary electricity outages are not, suggesting that there *can* be troubling social responses to turning out the lights. In contrast, the August 2003 blackout in the northeastern United States, the largest outage to date, did not generate terror. Commentators who had lived through both the 1977 and 2003 events noted that a more positive attitude in New York pervaded the second time.

There are three primary concerns in protecting the electric power system from an intended failure: physical security, cyber security, and fuel security. Each will be discussed in turn.

2.1 Physical Security

Many CIP concerns are about possible disruption of electric service, or turning out the lights. In places where deliberate attacks on the power system and

interruption of service are frequent, both everyday life and economic development can be damaged. However, it is not clear if an attempt to turn out the lights would permit terrorists to create widespread panic or the shock associated with violent attacks on symbolic objects that cause loss of human life. Electricity customers in most industrialized countries lose power for an average of couple of hours per year. For the most part, individuals and society cope with these outages well. Power companies respond rapidly to restore service, and facilities that have special needs for reliability (e.g., hospitals and airports) typically have backup generators.

There is some experience with large physical damage from natural forces. For example, in January 1998 an ice storm struck southeastern Canada, New York, and New England, felling 1000 transmission towers and 30,000 distribution poles while sending thousands of tree branches into power lines. This event left 1.6 million people without power, some for more than a month. Almost a quarter-million people had to leave their homes to take shelter and at least a dozen people died (mostly due to loss of heat or suffocation from faulty temporary heating setups). Insurance claims reached approximately \$1 billion Canadian. This event, although massively destructive, did not cause panic or shock.

It is difficult to imagine a terrorist organization amassing the capability to accomplish damage on this scale, but a terrorist attack could nonetheless be larger and more long lasting than typical equipment- or weather-related problem, and multiple attack modes could be coordinated for maximum effect. Critical points exist in the electricity infrastructure (especially among transmission and substation assets) where attacks could cause more damage. A U.S. National Research Council panel found that “a coordinated attack on a selected set of key points in the system could result in a long-term, multistate blackout.” Furthermore, the ability of the electric power sector to deal with coordinated attacks on the system is lacking. As noted by the U.S. National Electricity Reliability Council, “most (electricity sector) members do not have in place business recovery plans that include procedures for dealing with widespread, well-planned attacks on physical facilities. Likewise, they do not include procedures for dealing with deliberate attempts to hamper repair and restoration activities.”

High-voltage transformers have been long identified as a key vulnerability in electric power systems, but due to the cost of stockpiling, few spares are available. It might be possible to develop modular,

flexible transformers, but this goal is frustrated by the tendency of different power companies to insist on unique standards and often on customized equipment for every situation. The trend toward restructured markets may exacerbate this problem, as noted by a U.S. National Research Council panel, “Deregulation has encouraged efficiency, invested-cost utilization and return on investment rather than redundancy, reliability and security. For the same reasons, power companies keep fewer spares on hand.” Other problems have also been identified. Spares are often kept at the substation site, increasing the possibility that an attack will also result in damage to the spares. If spares have to be ordered from the manufacturer, the customized nature of the transformers may mean the units could take a year or more to manufacture. Delivery could also be lengthy given that transformers are increasingly being manufactured in a small number of countries (e.g., China).

Even so, it is not clear if a highly successful physical attack on an electric power system would have the same social effect as terrorist attacks that claim lives or focus on symbolic targets. Responses would likely include outrage and a sense of vulnerability, but evidence from large-scale blackouts indicates that panic and shock do not generally ensue.

2.2 Cyber Security

Traditionally, information technology (IT) in the electric power sector has been considered less vulnerable to attack (at least by outsiders) than the transmission grid. This is partly because power sector IT systems often consisted of dedicated systems without interconnection to other computer networks and with built-in redundancies. However, systems may be more vulnerable to attack due to their use of dialup modems and the Internet. For instance, in January 2003 an Internet-transmitted computer virus infected and disabled several computers at the Davis-Bessie nuclear power plant, although this problem apparently presented no safety concerns.

In the electricity industry, energy management systems (EMS) monitor and control the production, transmission, and distribution of electrical energy. This includes scheduling and controlling generators and monitoring and analyzing the transmission network in order to provide power while maintaining reliability of the system. Competition is leading to more and more use of electricity markets to make decisions about what power plants will be operated when. Since markets are typically open, it may be

impractical (or simply expensive) to verify the identity and reliability of all participants, exacerbating the problem.

Concern about EMS arises because of their use of open protocols, increased use of intelligent devices, and lack of encryption. Remote systems receiving control signals from the control center often do not verify such signals for authenticity. Instead of relying on traditional means to protect these systems, information security typically has relied on uncommon, often proprietary protocols and the need for considerable hard-to-obtain knowledge about operating systems and so forth. However, such measures are insufficient, particularly when one considers the insider threat problem. An employee with the proper knowledge could create serious problems and face minimal resistance from the system due to lack of encryption and authorization requirements.

2.3 Fuel Security

Electric power systems vary in how vulnerable they are to fuel supply disruption. Four major fuels are used: coal, natural gas, uranium, and renewables (wind and water). Each has a different vulnerability to deliberate attack. Several important distinctions between fuel distribution and electricity transmission systems are relevant. First, it is possible to store fuel but not electricity. Second, system stability concerns in electric power systems have no real analog in fuel supply systems. Third, fuels usually have a number of alternate delivery paths, while electricity transmission systems often have few, and some electricity generators can switch fuels. Fourth, almost all of the components of electricity transmission and distribution systems are aboveground, while much of the natural gas delivery system is buried.

2.3.1 Coal

Coal is the fuel least vulnerable to deliberate attacks because it is produced in many locations, can be shipped by multiple means, and can be easily stored. In addition, coal supply systems are made up of large and relatively robust capital equipment. Coal is transported by a variety of means. Rail dominates, and other major modes include river barge and truck. Coal-fired power plants may often have multiple supply routes available to them, although one mode will tend to dominate and generally only that mode will have coal-handling capacity sufficient for long-term supply. Coal can be readily stored on site and coal-fired power plants typically have three to eight weeks of fuel supply on hand. For power plants built

at coal mines, fuel security is obviously less of an issue, although these plants may require more transmission assets.

2.3.2 Petroleum

Overall, petroleum supplies are considered more vulnerable to deliberate attack than coal supplies, but less than natural gas. The most important vulnerabilities are in refineries and pipeline pumping stations, only a limited number of which typically supply any particular region. Although petroleum production (wells) and shipping (tankers) can and have been attacked (for instance, the destruction and ignition of many of Kuwait's oilfields by retreating Iraqi forces during the 1992 Gulf War), they are numerous and operate through a global market, so substitutes may be relatively readily found. An additional problem is that several chokepoints vulnerable to mines or other types of attack exist on ocean petroleum shipment routes, including the Straits of Malacca, Bosphorus Straits, and Strait of Hormuz.

Table II shows the percentage of electricity generation that is fueled by imported oil for the 10 nations with the greatest fraction. (This table and the next two only consider countries with large imports.) The United States accounts for over 24% of world oil imports; however, those oil imports contribute less than 2% to the total electricity generation and so is not included. In contrast, for the next largest oil importer, Japan, electricity generation accounts for

TABLE II

Percentage of Electricity Generation from Imported Oil, Select Countries 1998–1999

Importer	Percentage of world oil imports	Percentage of generation from imported oil
Singapore	3	65
Italy and San Marino	5	32
Philippines	0.9	28
Portugal	0.7	25
Greece	1	16
Japan	11	16
Thailand	2	16
Belarus	0.5	8
Korea, Republic of	6	7
Spain	3	7
Netherlands	3	7

Source. United Nations.

16% of oil imports. Seven countries rely on oil imports for over 10% of their electricity generation.

2.3.3 Natural Gas

The gas supply system is more vulnerable to deliberate attack than coal or petroleum supplies, but it is more robust than the electric system. It is less vulnerable than the electricity system because it is mostly buried underground, it can be used to store gas (line pack), and it can use relatively secure depleted oil and gas wells for storage. Like the electricity industry, companies in the gas business have long recognized and effectively planned for contingencies that mitigate terrorism. Spare parts are generally kept on hand to effect quick repairs. However, the potential vulnerabilities in gas system supply are illustrated by the August 2000 rupture and explosion caused by corrosion in a 30-inch gas pipe near Carlsbad, New Mexico, killing 12 people. This explosion led to significant increases in gas prices in California, exacerbating the electricity crisis there. The U.S. National Transportation Safety Board subsequently determined that decades of inadequate testing and maintenance by the pipeline company caused the mishap.

Table III shows the percentage of electricity generation that is fueled by imported gas for 11 countries with the highest percentages. Due to the regional nature of gas supply, generation is shown for each major gas supply region. Eleven countries rely significantly on imported natural gas for electricity generation. Russia is a key supplier to many of the European countries that rely on natural gas imports for electricity generation. Algeria is a significant supplier for Italy, Belgium, Spain, and Turkey; while the Mideast is the key supply region for Japan and South Korea.

Considering Tables II and III together, two countries rely on imported oil and gas for over half of their electricity generation (Belarus and Italy). For comparison, the United States uses imported oil and gas for about 4.3% of its electricity generation, over half of which is from Canada.

2.3.4 Nuclear

Although nuclear power plants are operating in about 30 countries, uranium fuel is not available in all of them. Although refueling is infrequent (every 18 months or so for most reactors) it does involve risks of deliberate attack associated with production and shipping. The Table IV shows the countries most reliant on nuclear power for their electricity generation and the countries with the greatest known

TABLE III

Percentage of Electricity Generation Resulting from Imports of Natural Gas, Select Countries, 1998–1999

Importers	Percentage of world gas imports	Percentage of electricity generation from imported natural gas (by supply region)				
		Russia	Canada	Algeria	Europe	Asia (incl. Middle East)
Belarus	3	87				
Turkey	2	19		8		2
Italy and San Marino	8	9		12	2	
Belgium	3			7	15	
Japan	13					16
Kazakhstan	0.6	14				2
Finland	0.7	14				
Ukraine	10	13				
Hungary	2	13				
Korea, Republic of	3					11
Slovakia	1	11				

Source. United Nations.

reserves of uranium. The top seven countries, ranked in order of nuclear contribution to electricity generation, have no uranium resources whatsoever (e.g., France, Ukraine, Korea, Germany, Japan, Spain, United Kingdom). However, uranium deposits exist in many countries and are available in significant quantities in several stable democracies.

Considering oil, gas and uranium, six countries import more than half of the energy they use for electricity generation. If coal imports were considered, this number could be larger. Further, given likely policies to limit greenhouse gas emission, gas imports are likely to rise substantially in many nations.

2.3.5 Renewables

Renewable energy sources have little, if any, fuel supply vulnerabilities. However, water impounds behind dams may be a source of concern due to the potential for damage downstream if a large dam is breached. The typically large distances between major renewable energy sources (e.g., rural or offshore locations where wind turbines are both economic and politically acceptable) and major electricity demand centers (i.e., cities) produces an

TABLE IV

Nations with Large Nuclear Energy Sectors and with Recoverable Uranium Resources

Percentage of generation provided by nuclear power		Percentage of global known recoverable uranium resources	
France	77	Australia	28
Ukraine	45	Kazakhstan	15
Korea	37	Canada	14
Germany	30	South Africa	10
Japan	30	Namibia	8
Spain	28	Brazil	6
United Kingdom	23	Russia	4
United States	20	United States	3
Russia	15	Uzbekistan	3
Canada	12		

Columns 1 and 2 are from IEA key world indicators report; columns 3 and 4 are from the World Nuclear Association.

inherent risk due to the likelihood of producing energy in remote, uninhabited areas and the necessity of long-distance transmission.

3. OTHER ATTACK MODES

Deliberate attacks on some parts of the electric power system could cause damage or terror without necessarily creating a power outage. Nuclear power plants, hydroelectric dams, and cooling towers are the major concerns.

The most obvious threat is a direct attack on nuclear reactors themselves. One possible attack mode is by airplane. Most nuclear reactor containment buildings are massive structures designed to withstand significant impacts (exceptions include plants like the Chernobyl facility in the Ukraine). However, no conclusive analysis shows that all containment buildings and the nuclear systems they house can safely withstand a direct hit by a jumbo jet or by a smaller plane loaded with high explosives. In addition, the spent reactor fuel stored in water-filled pools or aboveground containers at all operating reactor sites may be far more vulnerable, especially if the water that cools and shields hot core elements is lost. These temporary facilities have grown around the world due to the failure of any country to develop a permanent reprocessing or disposal plan for nuclear waste. Attacks on nuclear power plants, even if they have little direct impact, could raise public alarm and lead to the closure of facilities out of fear.

Due to the dread the public has of nuclear radiation (whether scientifically supported or not), it might be possible to create significant public distress (or even terror) even with an unsuccessful attack on a nuclear power plant. The core meltdown of the Three Mile Island plant in Pennsylvania in 1976 had a devastating effect on the nuclear power industry even though no one was hurt and only insignificant amounts of radioactivity were released. Even an unsuccessful attack could lead to nuclear plant closures.

Concerns about direct attack by truck bombs or commando squads have led to physical barriers and armed guards that protect all nuclear power plants. However, it is not clear how effective these steps are; a 1999 review by the U.S. Nuclear Regulatory Commission found "significant weaknesses" in 27 of the 57 plants evaluated.

Nuclear power plants also present a unique threat labeled "proliferation," which is the potential for nuclear power plant fuel (or waste) to be used in creating nuclear weapons. These could be either fission bombs, which would be relatively hard to produce, probably requiring the resources of a nation, or much cruder "dirty bombs," which even small groups could construct from conventional explosives and almost any highly radioactive material. One concern is that reprocessing fuels so that fissile material (plutonium) is separated from highly radioactive fission products could increase the proliferation potential of nuclear waste. New reactor designs that are more resistant to physical assault or require no refueling might be developed, but the cost of doing so is unclear.

Typically, the costs of mitigating the CIP risks of nuclear power plants, and other security risks such as proliferation, are not fully included in the cost of electric power these plants produce. For instance, after the terrorist attacks of September 11, 2001, many state governors put National Guard troops at nuclear plants, so the taxpayer in effect subsidized electricity consumers. As the owners of nuclear power electricity facilities have begun to face competition and the bottom line becomes more important, the costs of security are likely to receive greater scrutiny. This subsidy is similar to the support for private nuclear power plant owners in the United States from the Price-Anderson Act, which limits their liability due to accidents.

The second major threat is to hydroelectric dams, which can store significant amounts of energy in the form of water retained behind them. If they were suddenly breached, both property and lives could be lost. It is not very easy to successfully attack most large dams, however, and physical access to them can

be controlled relatively easily. One complicating factor is that smaller upstream facilities such as locks or small dams may be more vulnerable to attack. Breaching one or more of these could send a pulse of water downstream, possibly causing a larger dam downriver to fail catastrophically.

Third, cooling towers, used in many large power plants, could be used in a deliberate attack as a means of dispersing biological or chemical agents. The magnitude of this threat is not well understood since these facilities have some physical security and the effectiveness of cooling towers to disperse various agents is not well studied.

Finally, large electromagnetic pulses (EMP) can induce instantaneous voltages of hundreds of thousands of volts in conductors, creating large disruptions in electric power systems and destroying electrical equipment components such as motors, backup generators, and microprocessors. Concerns about EMP have traditionally been focused on the impact of high-altitude nuclear detonations, but relatively simple, inexpensive devices that can deliver an EMP have been designed and tested by the United States and other nations. Concerns have emerged that even crude devices could be deployed with significant localized impacts. Such weapons could damage the electric power systems and also use the electric power system as a conduit to attack other infrastructures (e.g., telecommunications) without physically penetrating facility perimeters. It is not clear how vulnerable the system is to such assaults, nor what could be done to defend against them. There is very little information from credible sources on the EMP threat from nonstate actors, so this concern is somewhat speculative.

4. MITIGATION

Mitigation efforts are designed to reduce the probability or magnitude of disruptions to electric power systems or the use of the power system for other attack modes. Mitigation measures can take a variety of forms ranging from improved security procedures to changes in the physical and institutional structure of the system. Before describing specific mitigation strategies, it is necessary to introduce the concept of survivability, which should inform any mitigation strategy.

4.1 Survivability

The traditional reliability planning approach does not lend itself well to responding to unexpected,

deliberate, and potentially very high impact attacks. For one thing, it is difficult to imagine all possible attack modes or strategies. Even if this problem could be overcome, practices in reliability analysis allow for essentially only one strategy—provide more reserve capacity—a costly solution that competitive firms are not likely to implement, and one that may divert resources from other important uses in CIP or economic growth. Thus, the advent of restructuring and the possibility of terrorism require the reformulation of the reliability problem. Supply-side solutions are not likely to be sufficient any longer; more attention will need to be paid to the economic and social implications of end-use disruptions. One concept that may serve to link reliability and CIP and also be compatible with a competitive electricity industry is survivability, also known as the self-healing grid or the resilient network.

Survivability is the ability of a system to fulfill its mission in a timely manner, despite attacks, failures, or accidents; survivability can be contrasted with the fortress model of security that tries to prevent or counter all attacks but has disastrous outcomes (e.g., cascading failures) when it does inevitably fail. A fundamental assumption of survivability analysis and design is that no individual component of a system is immune from attacks, accidents, or design errors. Thus, a survivable system must be created out of inherently vulnerable subunits, making survivability an emergent property of the system rather than a design feature for individual components.

Due to the complexity and speed of electricity system operations, it may be difficult to recognize attacks until there is extensive damage. Thus, ways must be found to recognize attack early and to protect the system without taking the time to discover the cause of the attack. Survivable systems must be able to maintain or restore essential services during an attack and to recover full service after the attack. Thus, the system must fail gracefully, shedding low priority tasks and then adding tasks in a priority ordering during recovery.

For example, in most cities, traffic signals are powered by the same circuits that provide service to much less critical loads like billboards. During blackouts, injury and property loss may occur due to blank traffic signals. Worsening the problem, blackouts cause gridlock that prevents police and fire vehicles, as well as emergency response crews from reaching their destinations. This demonstrates the fortress aspect of traditional reliability planning – it creates a system in which frequent or large-scale blackouts are not supposed to occur, but when they

do, the consequences are severe. In contrast, a system designed around survivability concepts might use low-power light-emitting diode traffic lights with battery backup to ensure that a blackout did not interrupt traffic flow.

4.2 Guards, Gates, and Guns

One approach to dealing with these issues is to assess, and if necessary and justified, improve the physical security of installations in the electric power system. This approach also probably requires relying on surveillance technologies to detect attack, ranging from automated systems such as sensors and alarms to detect intrusion to manned surveillance. The primary purpose of such technologies is to detect an attack and mobilize the resources (i.e., private security force, local police, military) to repel the attack and prevent damage. A secondary purpose would be to collect evidence to be used in the investigation of any attack.

While there exist a wide variety of monitoring and surveillance technologies, they are not considered adequate for protecting the electricity infrastructure. In the case of remote facilities, monitoring technologies may not be of use in preventing an attack if the response time of an intervention force is too long.

Unfortunately, any technology used for monitoring or surveillance has the risk of failure. Another problem has to do with nuisance alarms. Surveillance and monitoring systems are not always able to distinguish between an intruder and an animal or between a line being deliberately cut and one that has been struck by a falling tree. Guarding against failures and false positives may require backup systems or complementary technologies that can confirm or deny events, increasing costs.

4.3 Increased Backup

The traditional reliability approach has also led to a localized version of reliability through redundancy at facilities that are particularly concerned about electric power reliability: the use of backup generators. Especially since the recognition of a potential Y2K problem (the probability that older computer programs, especially those embedded in operating systems, would have failed on midnight of January 1, 2000), hospitals, airports, water supply systems, and sometimes office buildings have invested in onsite diesel generators. In some cases, building codes or operating licenses require backup generators. These are typically designed to provide a minimal level of

electricity supply for up to a few days before needing refueling. Such approaches may not be satisfactory under attack situations, depending on the level of functionality supported by the backup generator, its own reliability, and the feasibility of long-term refueling. Furthermore, such approaches are typically too expensive for residences and small businesses.

4.4 System Architecture and Control

4.4.1 System Architecture

There may be fundamental limitations to the improvements that can be made in the vulnerability of the electric power system solely through improved physical or cyber security of current systems. One possible method of addressing this limit would be a change in the overall physical architecture of the system. Such changes in the system architecture would be designed to provide the electric power system with more of the properties of survivable systems.

One such approach would be to use significant decentralized (or distributed) generation, often called DG, which places energy supply at the location of its use rather than relying on large, central station plants and long-distance transmission. DG facilities are smaller (though they can range widely in size from the few tens of kW to tens of MW) and more dispersed. Until recently, DG technologies have generally found use in either niche applications (e.g., in certain industrial contexts) or as backup power. However, there are areas (e.g., the Netherlands) in which significant portions of generation is supplied by grid-connected distributed units that both supply local power and feed power into the grid.

However, there has been little analysis presented to indicate either the gain in robustness expected by using DG, nor the impact on system cost. Work has been conducted to quantify some of the costs and benefits of a more distributed system by modeling hypothetical centralized and distributed systems under stress. Generally these efforts find that distributed systems could prove to be significantly more robust under stress conditions and results in lower reserve requirements, both of which can reduce costs. In addition, in many DG applications it is possible to use waste heat, resulting in higher overall efficiencies and lower costs. The potential effect of DG on cascading failures is unknown.

Advanced renewable energy sources (solar, wind, hydro, and biomass) have the potential for significantly increasing the diversity and flexibility of

electricity generation, but they face limitations due to their cost and their dependency on transmission. (Many renewable resources are located far from demand centers.)

Another approach to improving survivability might be demand response technologies that can reduce and make more flexible demand for electricity. Demand reduction could conceivably come either through energy conservation or through demand response technologies that could adjust demand if necessary.

At its most extreme, demand response could lead to shedding of load, otherwise known as rolling blackouts. Intelligent load shedding under emergency situations could help in preventing cascading grid failures while maintaining service to the most critical loads. Such an approach would provide for inherent deterrence from attack by offering fewer large, obvious targets. It will also emphasize differences in importance among end uses during emergencies (subways and traffic lights might be considered more important than billboards and air conditioners, for instance). Such a system might be able to reconfigure itself to isolate faults (adaptive islanding).

4.4.2 Control

The traditional approach to reliability has led to the automatic control of power system components in response to very small, short-term (seconds to minutes) deviations in demand by adjusting power generated to maintain locally measured frequency and voltage within the acceptable limits. Over longer terms (minutes to hours), human operators start and stop generators to meet forecast demand changes. When major equipment outages occur, decisions are made by human operators intimately familiar with the response of each specific system to the given (previously studied) outage. In spite of these efforts, cascading failures have occurred as a result of events that were not easy to detect by the system operator. An important issue is how system control strategies would perform under attack and how they could be improved.

The primary controllers on power plants are tuned to respond to small deviations in system conditions. It is, however, possible, to make these much “smarter” by equipping them with additional sensing and communications effectively resulting in a more observable system whose dynamics are controlled. In addition, it is possible to design these controllers with adaptive, nonlinear logic. Equipping the existing power plants with such power system stabilizers

has been known to greatly enhance transient stability of a power grid under stress.

Problems in protective relaying are frequently at the heart of major blackouts. A typical problem is that a protective relay is not set correctly or it fails and incorrectly senses a fault on line it is protecting. If that happens, the breaker disconnects the line, which, in turn, causes line overflows in other transmission lines creating a sequential blackout situation. The fundamental problem with protective relaying is that it only uses local information, and the settings are tuned for a certain range of operating conditions. If these are exceeded, relay functions may fail.

4.4.3 Institutional Design

The robustness of electric power systems, and the ability to restore power quickly after accidental outages, is no accident. Government regulation and private institutions have been established to promote reliability in the electric power sector. In the United States, voluntary coordination councils were set up by the electric utility industry following the 1965 blackout, partly to avoid government regulation of reliability. State public utility commissions play a role as well. All countries have some form of electricity reliability institution(s). However, industry restructuring is drastically changing institutional arrangements, leading to serious questions about whether such private approaches will be adequate to ensure reliability, let alone provide CIP.

Both reliability and CIP share an important characteristic—they are what economists call public goods, services that if supplied to one person is available to others at no cost (i.e., nonrival consumption and nonexcludable access). Public goods are not provided in socially desired quantities by market forces alone, government action of one sort or another is always needed. Thus, government will probably have to act in some way to ensure sufficient reliability and CIP. Following the 2003 blackout in the United States, pressures to move toward regulation and mandatory reliability rules mounted, although the industry still largely preferred self-regulation.

Institutional issues also play out in apparently purely technical domains. For instance, several studies have shown that barriers are often raised to the interconnection of a distributed source to the grid. These barriers could come from market structure issues, the relative roles of DG owners and utilities, and from standards that are developed

for the interconnection of distributed resources. Similar concerns exist about the incentives for cooperation and investment in new control strategies and technologies.

One of the key challenges for the future is ensuring that institutional solutions to emerging reliability and CIP problems are created for whatever new structure the electricity industry takes. This includes looking at threats that go beyond those considered in traditional contingency planning. However, the potential for creating terror through attacks on the electric power system seem limited (with the exception of an attack on a nuclear facility). Loss of power, even if large-scale and frequent, is costly and annoying, and if caused by malicious attacks can be disconcerting, but it pales in comparison to the effect of the large-scale loss of life that is often the purpose of terrorist acts.

SEE ALSO THE FOLLOWING ARTICLES

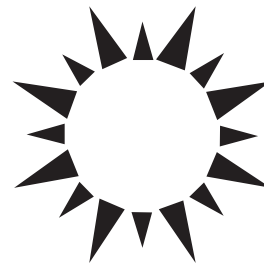
Electrical Energy and Power • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate

Making • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

- Amin, M. (2001). "Towards Self-Healing Infrastructures." IEEE Computer Applications in Power, p. 20–28.
- Billington, R., and Li, W. (1994). "Reliability Assessment of Electric Power Systems Using Monte Carlo Methods." Plenum Press, New York.
- Council on Foreign Relations. (2002). "America Still Unprepared—America Still in Danger: Report of an Independent Task Force Sponsored by the Council on Foreign Relations." Council on Foreign Relations, New York.
- Cowart, R. (2002). "Electrical Energy Security: Policies for a Resilient Network." Regulatory Assistance Project, Montpelier, VT, www.raponline.org.
- Farrell, A. E., Lave, L. B., and Morgan, M. G. (2002). Bolstering the security of the electric power system. *Issues Sci. Technol.* XVIII(3), 49–56.
- Lipson, H. F., and Fisher, D. A. (1999). Survivability—A new technical and business perspective on security. In "New Security Paradigms Workshop." ACM, Caledon Hills, OR.
- National Research Council (2002). "Making the Nation Safer: The Role of Science and Technology in Countering Terrorism." National Academies Press, Washington, DC.
- Seger, K. (2003). "Utility Security: A New Paradigm." PennWell, Tulsa, OK.
- Thorp, J. S., and Phadke, A. G. (1999). Protecting power systems in the post-restructuring era. *IEEE Comp. Appl. Power* 12(1), 33–37.
- Zerri, H., Dowlatabadi, H., and Strachan, N. D. (2002). Electricity and conflict: Advantages of a distributed system. *Electricity J.* 15(1), 55–65.

Electric Power Generation: Fossil Fuel



JÁNOS M. BEÉR

Massachusetts Institute of Technology
Cambridge, Massachusetts, United States

-
1. Background
 2. Fossil Fuel Utilization Systems
 3. Coal-Fired Power Generation
 4. Pollutant Emission Control in Coal-Fired Plants
 5. Natural Gas-Fired Plants
 6. Power Cycles of Increased Thermodynamic Efficiency
 7. Conclusions

Glossary

coal pyrolysis The process of heating coal in an oxygen-deficient atmosphere.

combined cycle A combination of Brayton (gas turbine) and Rankine(condensing steam turbine) cycles; exhaust heat from gas turbines generates steam for steam turbines, and electric power is produced in both cycles. The efficiency of the combined cycle is high because of the complementary temperature ranges of the Brayton and Rankine cycles.

cycle efficiency The energy output of an energy conversion cycle expressed as the percentage of the energy input.

diffusion flame The flame produced when fuel and oxidant are separately injected through a burner, are only partially mixed prior to ignition, and their mixing continues during combustion. The overall rate of combustion of the fuel is determined by the rate of the fuel/oxidant mixing in the flame.

excess air The air required in addition to the theoretical combustion air to burn the fuel completely (usually expressed as a percentage of the theoretical combustion air).

flue gas desulfurization The postcombustion reduction of SO_x emissions. Most systems involve wet scrubbing of the flue gas with lime to produce calcium sulfate as the waste. Capable of 95% or higher sulfur removal.

fluidized bed combustion The combustion of millimeter-sized coal particles in a dense bed of limestone (SO_2 sorbent) particles. The particles are suspended in a column of air supplied through a distributor plate at the

bottom of the bed. Steam-generating tubes immersed in the fluidized bed extract heat and maintain the bed at a low ($<1100\text{ K}$) combustion temperature. This favors the reduction of SO_2 and NO_x emissions.

premixed flame A combustion system in which fuel and oxidant are mixed prior to ignition. The rate of combustion is determined by the chemical rates of combustion reactions, and hence by the reactant concentrations and the temperature.

Power generation can be fueled by coal or natural gas. Pulverized coal and fluidized bed coal combustion are used for steam generation in the Rankine cycle; natural gas is used in gas turbines for power generation in the Brayton cycle. These techniques, including the problems of combustion-generated pollution and the progress in pollution control technologies, are discussed in this article. The current status of advanced, high-efficiency power-generating cycles, capable of lowering significantly emissions of sulfur oxides, nitrogen oxides, mercury, and CO_2 , is highlighted and their potential for deployment and the critical needs of research and development are examined.

1. BACKGROUND

The provision of electric power is one of the prerequisites of prosperity; there is strong correlation between electric power-generating capacity and per capita gross domestic product (GDP) (Fig. 1). Across the world, economic indicators signal that there will be continued growth and increased electricity demand; during the next 15 years, 1200 GW of new capacity is projected to be added to the world's present electric generation capacity of about 3500 GW.

The U.S. Energy Information Administration (EIA) data for 2001 show that 54% of the electricity supply in the United States was generated by coal,

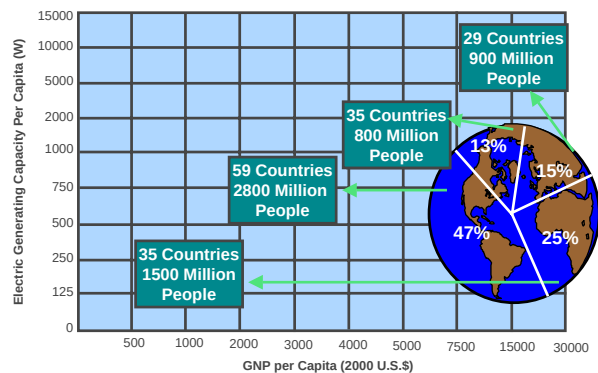


FIGURE 1 Prosperity and electricity in 2003. Generating capacity (watts) per capita as a function of gross national product (GNP; U.S. dollars) per capita. From EPRI (2003).

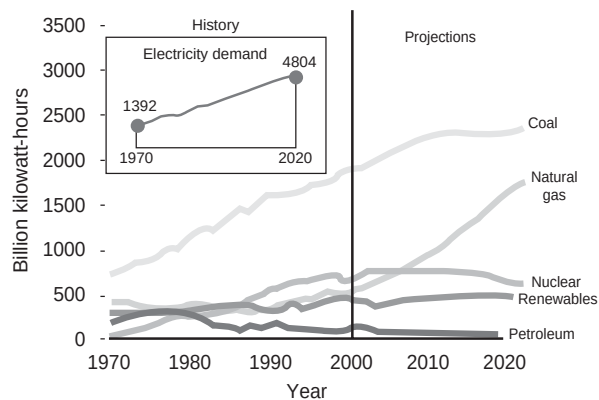


FIGURE 2 United States electricity supply by fuel type from 1970 through 2020. From Annual Energy Outlook (2000).

12% by natural gas (increasing to an estimated 32% by 2020), and about 2% by oil, with the rest produced by nuclear power (21%), hydro power (9%), and by renewable solar or wind (2%). Electricity generation (billion kilowatt-hours) by fuel type from 1970 to 2020 in the United States is illustrated in Fig. 2. The increase in the use of coal and, especially, natural gas for power generation projected for the period from 2000 to 2020 is to meet growing demand for electricity and to offset the projected retirement of nuclear power plants.

Because of the large coal reserves in major developing countries such as China, India, and Indonesia, where most of the new power-generating plants are being installed, it can be expected that coal will remain the dominant source of power generation worldwide during the first half of this century. There is also a strongly growing demand for natural gas, a clean fuel capable of being used in power generation with high efficiency.

Coal as an energy source has the attraction of broad availability (there are large reserves in several countries around the world), low cost, and economic utilization by mature technologies. The disadvantages are due mainly to environmental and health impacts at the mining phase and in the course of coal utilization. Both the mining and the utilization of coal are subject to increasingly tightening standards by governmental regulations. Because there is no presently acceptable practical alternative to coal and natural gas, industrial research and development programs are concentrated on methods of reducing pollutant emissions, especially SO_2 , NO_x , and fine particulates, from coal-fired plants, and NO_x from natural gas-fired combustion turbines. New environmental technologies developed and demonstrated under the Clean Coal Technology Program of the U.S. Department of Energy (DOE) during the past 15 years have become commercial and are being applied worldwide. Major advances have also been made in stationary gas turbine technology by increasing the inlet gas temperature to the turbine and by significantly reducing nitrogen oxide emissions.

The reduction of CO_2 emission remains a challenge. Research aimed at CO_2 capture and sequestration is receiving governmental and industrial support, but it may take a decade before an economically feasible solution is found. In the mean time, the most practical approach to this problem is the improvement of electric power-generating efficiency. Improved efficiency will reduce the amount of fuel used for the generation of a given electric energy output and hence will reduce the emissions of all of the pollutants.

2. FOSSIL FUEL UTILIZATION SYSTEMS

The steam (Rankine) cycle and the gas turbine (Brayton) cycle are the two main thermodynamic cycles in utility power generation. In the Rankine cycle, steam is created in a boiler, and part of the enthalpy of the high-pressure superheated steam is converted to shaft work in a steam turbine that drives an electric generator. In the Brayton cycle, high-pressure and high-temperature gaseous combustion products enter directly a gas turbine, where part of their enthalpy is converted to shaft work. The gas turbine drives a compressor and the electric power generator. Because of the direct entry of the combustion products into the turbine, the gas turbine

cycle is highly sensitive to the quality of fuel and is restricted to the use of natural gas or distillate fuel oil. Boilers in the steam cycle are, by contrast, fairly tolerant to fuel quality and are capable of being fueled even by low-quality coal or waste fuels. Combustion is the prevailing fuel utilization technology in both of these cycles. In future central power plants, there are good prospects for application of gasification, which enables coal use in gas turbines, because of the environmental advantages and the promise of higher cycle efficiency.

During the past decade, there has been a growing trend toward distributed generation. Compared to large, centralized utility plants, power plants of smaller capacity, can be sited close to consumers, thereby reducing electricity transmission losses and improving operational flexibility. Being situated near population centers, these smaller plants have to be highly efficient and use a clean fuel, usually natural gas. Natural gas is free of sulfur and ash, and because of a carbon content lower than that in coal or oil, the CO₂ emission from natural gas combustion is also lower.

High-efficiency utilization is achieved by combined gas turbine–steam cycles. Also, being close to industrial and population centers opens the opportunity for combined heat and power (CHP) generation, which further enhances the plant's energy efficiency. An example of distributed generation and CHP is the Massachusetts Institute of Technology's natural gas-fired gas turbine/steam Cogen plant, which provides the campus with 22 MW of electric power and 80 t/hr of steam for heating or cooling at an energy efficiency above 80% and with very low emissions. This plant replaced a low-efficiency, more polluting, and more costly energy system in which steam was produced on campus in oil-fired boilers and electric power was purchased from a utility.

3. COAL-FIRED POWER GENERATION

3.1 Pulverized Coal Combustion

Pulverized coal (PC) combustion is presently the system of choice for coal-fired power-generating plants. In PC combustion, the coal is dried and is ground to a specified fineness, with a maximum particle size of 250–300 μm, depending on the reactivity of the coal. Coal preparation, which involves feeding, drying, and grinding of the coal, and the pneumatic transport of the pulverized coal to

the burners, is fully integrated with the boiler. For lower reactivity coals, the fineness of grind is increased to create a larger specific surface area so as to improve conditions for ignition and combustion. The powdered coal is pneumatically transported to burners and injected via particle-laden jets into the combustion chamber. The transport air that carries the coal from the mill to the burners is a small fraction of the total combustion air. It is kept at low temperature, limited to about 373 K, for reasons of safety against ignition and explosion in the mill and in the powdered coal transport pipeline between the mill and the burners. The rest of the combustion air, which can be preheated to higher temperatures, is injected separately and admixed with the already ignited coal particle-laden jet in the combustion chamber. A schematic illustration of a PC combustion boiler is shown in Fig. 3. The combustion chamber is typically of parallelepiped shape; the dimensions of a 300-MW coal-fired boiler would be approximately be $15 \times 15 \text{ m}^2$ of cross-sectional area and 45–50 m in height. The combustion chamber walls are completely

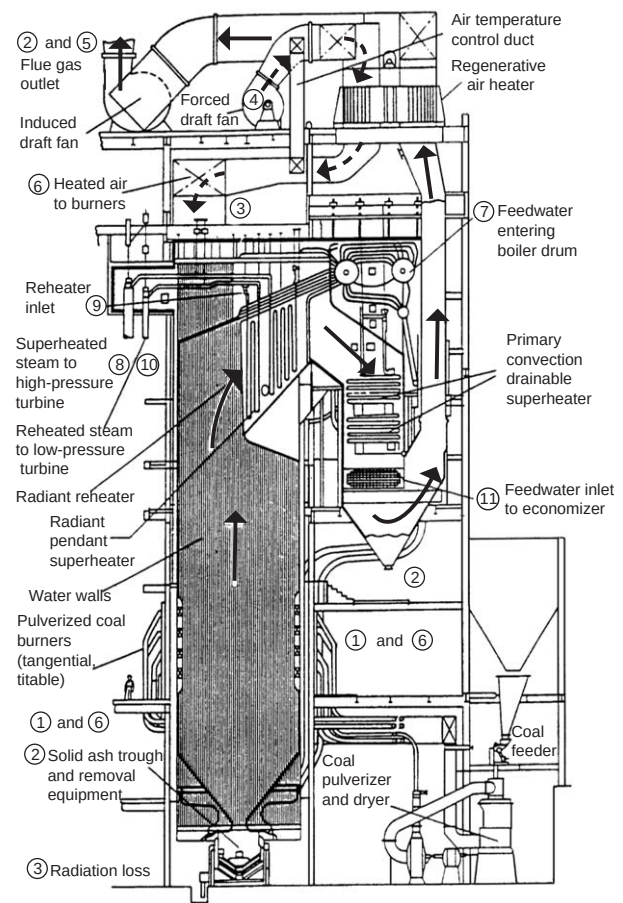


FIGURE 3 Utility boiler fired with pulverized coal.

cooled by steam-generating tubes. As the pulverized coal particles burn, the flame transfers heat, mainly by thermal radiation, to the steam-cooled tube walls of the boiler. The design of the combustion chamber has to provide for sufficient residence time of the burning particle to complete combustion, and for the cooling of the residual fly ash particles to below their “softening temperature” before entry to narrow convective heat exchanger passages, to prevent the build-up of sticky ash deposits. At regular intervals, friable deposits are removed by soot blowing using steam.

Although there are a great variety of burners in PC combustion boilers, the most widespread types are the circular burners and the vertical nozzle arrays. Circular burners are usually positioned perpendicularly to the combustion chamber walls, whereas vertical nozzle arrays are in the corners, firing tangentially onto the circumference of an imaginary cylinder in the middle of the combustion chamber. Designs utilizing circular burners are more concerned with the tailoring of the individual burner flame, whereas vertical nozzle arrays in tangentially fired furnaces are designed to rely more on the furnace volume for the mixing of the fuel and air streams. Complete combustion with a minimum (typically less than 15%) of excess air is a challenging task for a coal combustion system. The excess air, usually expressed as the percentage of the theoretically required combustion air, is the additional air needed for complete burnout of the fuel. The significance of the excess air is that the heat carried by the stack gases into the atmosphere, the largest heat loss (about 10%) of boiler operation, is directly proportional to the volume of the flue gas and, hence, to the excess air.

The fly ash formed in pulverized coal combustion is removed from the flue gas in the form of dry particulate matter; a small proportion (~10%) of the coal ash that falls off the tube walls as semimolten agglomerated ash is collected from the bottom hopper of the combustion chamber (“bottom ash”). Ash particles are molten in the high-temperature regions of the flame, but cool and solidify as they reach heat exchange surfaces of the boiler, forming friable deposits that are easily removed by soot blowing.

3.2 Fluidized Bed Combustion

In fluidized bed combustion (FBC) (Fig. 4), crushed coal, 5–10 mm in particle size, is burned in a 1000-mm-deep bed of 0.5- to 3.0-mm particles consisting mainly of additive sorbents (limestone or dolomite) and coal ash. The bed is fluidized by a blast of air through a distributor plate at the bottom of the bed.

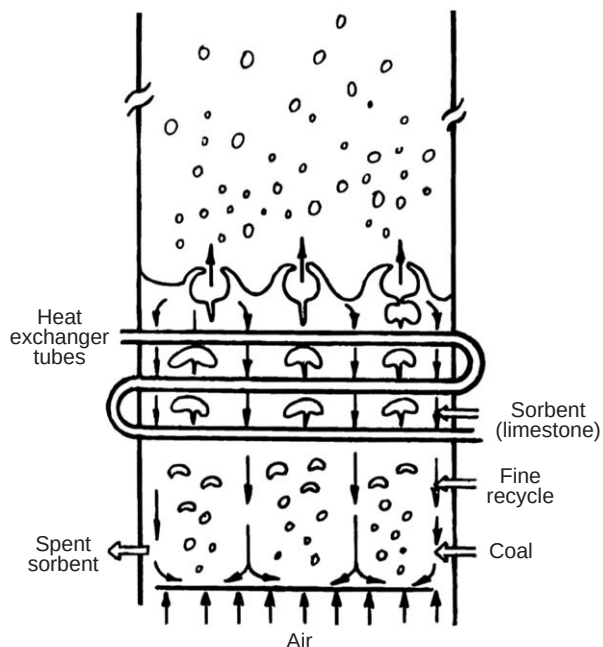


FIGURE 4 Fluidized bed coal combustion with sorbent injection for sulfur capture, and heat extraction by means of steam-generating tubes immersed in the bed. From Sarofim and Beér (1978).

As the airflow through the bed increases, the drag lifts and separates the particles from one another, and the pressure drop across the bed becomes sufficient to support the weight of the bed; the bed is incipiently fluidized. Further increase of the velocity causes the air to form bubbles that rise in the bed like gas in liquid. The bubbles stir the bed by carrying pockets of particles on their tops and in their wake. The particles are released and are returned to the bed as the bubbles burst on the top of the fluidized bed. Less than 2% of the bed material is coal; the rest is coal ash and additive limestone, capable of capturing sulfur after the calcination of the additive limestone and the formation of lime (CaO). Due to the excellent gas–solid contact, the fluidized bed is a favorable reactor for the combustion of millimeter-size coal particles and for sulfur capture in the combustion process. The bed is cooled to a temperature of 1050–1170 K by immersed steam-generating tubes. This relatively low combustion temperature prevents the softening of the coal ash and the chemical decomposition of CaSO_4 , the product of sulfur capture.

3.2.1 Circulating Fluidized Bed Combustion

In circulating fluidized bed combustion (CFBC) (Fig. 5), the gas velocity is higher than in FBC, and a large part of the solid particles is carried over from

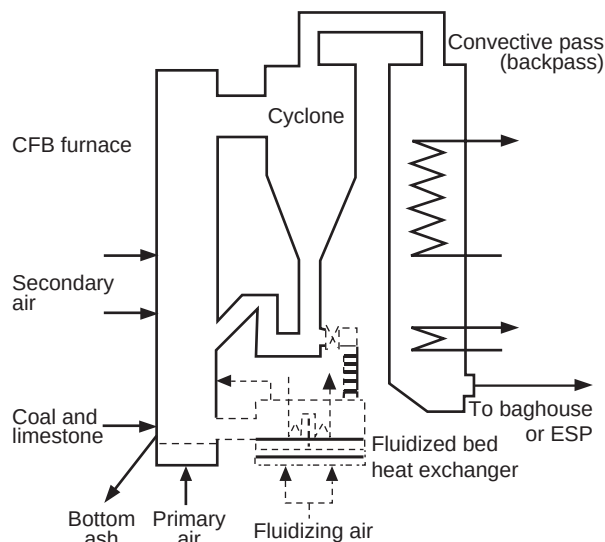


FIGURE 5 Circulating fluidized bed (CFB) furnace. ESP, electrostatic precipitator. From Scott and Nilsson (1999).

the bed into a tall riser, where the particles aggregate into clusters that reflux at the wall. Particles that are carried over from the riser are precipitated by a high-temperature process cyclone and are recirculated into the riser. With the higher gas velocity, the bed cross-sectional area becomes smaller than in FBC, for a given performance. This, in turn, leads to a smaller number of coal feed points required for maintaining uniform coal concentration over the bed cross-sectional area, a significant operational convenience. Also, because of the higher gas velocity and the somewhat smaller sorbent particle size, less additive sorbent is required in CFBC for a given percentage sulfur capture.

FBC and CFBC are commercial, mature technologies; primarily applied to the use of low-grade fuels in smaller unit sizes. The main advantages of CFBC are the capability of capturing sulfur in the combustion process by additive sorbents and the lack of sensitivity to fuel type or fuel quality variation. Due to the refractory lined process cyclone circulating fluidized bed combustors do not lend themselves easily to scale-up to the larger utility boiler sizes (≥ 660 MW).

4. POLLUTANT EMISSION CONTROL IN COAL-FIRED PLANTS

The combustion-generated pollutants of concern are mercury, fine inorganic particulates, and oxides of sulfur and nitrogen. Figure 6 is an illustration of the pollutants and their effects.

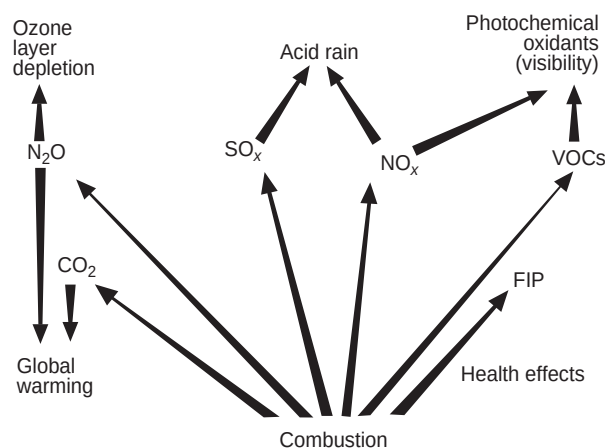


FIGURE 6 Combustion-generated pollutants; environmental and health effects. VOCs, volatile organic compounds; FIP, fine inorganic particles.

4.1 Sulfur

Sulfur in coal will oxidize to SO_2 , a precursor of acid rain. Methods of sulfur capture in the combustion process involve the reactions of sorbents such as calcined limestone (CaO) with SO_2 to produce CaSO_4 , a stable, disposable solid waste. In the high-temperature, fuel-lean environment of pulverized coal flames ($T_{\text{peak}} \sim 2000$ K), however, CaSO_4 is unstable; it decomposes, which leaves flue gas desulfurization (FGD) as the viable option for sulfur capture from PC combustion. The development of fluidized bed combustion provided the opportunity to effectively retain sulfur in the combustion process, because CaSO_4 is stable at the lower FBC operating temperature of 1050–1170 K. Sulfur capture in fluidized combustion ranges between 70 and 90%, with even higher capture efficiencies obtained for subbituminous coals of high calcium content. FGD technologies are mature and commercially proven; 99% SO_2 reduction has been achieved at some plants, with 90–95% reduction as routine. Sulfur dioxide emissions from utility plants in the United States have been reduced 40% since 1980, even though electricity production increased by 35% during the same period. This was achieved by switching to coals of lower sulfur content and by retrofitting some of the existing coal-fired power plants with scrubbers. This process will have to continue to satisfy the constraints of the Clean Air Act Amendment of 1990, which requires further reductions in SO_2 emissions in the Phase 2 implementation that began in the year 2000.

4.2 Nitrogen Oxides

Nitrogen oxides as pollutants deserve special attention, because of their wide-ranging effect on the environment, including contribution to acid rain, reduction of atmospheric visibility, production of tropospheric (ground level) ozone, and, in the case of N_2O , depletion of stratospheric ozone. It is also noteworthy that NO_x emissions are amenable to reduction by combustion process modifications. The main routes of NO formation in coal flames are the attack of atomic oxygen on molecular nitrogen in high-temperature, fuel-lean flames (“thermal NO”), and the oxidation of the organically bound nitrogen in the coal (“fuel NO”). It is noteworthy that fuel-bound nitrogen and NO can be converted to molecular nitrogen and, hence, rendered innocuous for NO_x emission by coal pyrolysis, a process in which coal is maintained at high temperature in an oxygen-deficient atmosphere. To complete combustion, the rest of the combustion air, called “overfire air” (OFA), ought then to be admixed to the products of the pyrolysis reactions. In this burnout stage of the flame, the temperature needs to be controlled to levels sufficiently high to complete combustion but insufficient for significant thermal NO formation.

NO_x emission control by combustion process modification is accomplished by OFA, low- NO_x burner, and NO reburn technologies. In low- NO_x burners (LNBs) (Fig. 7), air staging is achieved by aerodynamically tailored mixing of the fuel jet with airstreams supplied through each individual burner, rather than through the splitting of the combustion chamber into fuel-rich and fuel-lean combustion zones by the use of OFA. LNBs and OFA represent the most cost-effective methods of achieving reduced NO_x emissions from new plants and also from existing boilers by retrofit.

NO reburning involves a secondary fuel, usually natural gas; the gas, amounting to 10–20% of the total heat input to the boiler, is injected into the

combustion chamber and mixed with the NO-bearing coal combustion products. The NO is reduced by reactions with hydrocarbon fragments, CH and CH_2 , that are formed in the course of the pyrolysis of the natural gas. Control of NO_x emissions by combustion process modifications and their combinations can yield up to 70% reductions. These technologies are successfully applied in the utility industry; more than 188 GW of electric power-generating capacity currently in operation internationally has been fitted by these combustion measures.

Further reductions in NO_x emissions can be obtained by the injection of additive ammonia into the combustion products. At high temperature (1060–1350 K), no catalyst is needed in the selective noncatalytic reduction (SNCR) process for the removal of up to 40% of the NO formed during coal combustion. In selective catalytic reduction (SCR), the ammonia vapor is injected over a catalyst bed situated at the boiler economizer outlet, where the flue gas temperature ranges from 600 to 650 K. In the presence of the catalyst, the ammonia chemically reacts with NO_x to form water vapor and N_2 . The cost of SRC installation is higher than that of SNCR, but much less ammonia has to be injected, and NO_x reductions in excess of 90% can be obtained, depending on the type of coal burned.

4.3 Fine Particulates

Fine particulates are removed by electrostatic precipitators (ESPs) or by baghouse (BH) filters. In the ESP, particles with an electric charge are attracted to collecting electrodes in a high-voltage field. The precipitation efficiency is high, in the 10- to 300- μm size range, but it falls off for smaller particles, especially those in the submicrometer range. These ultrafine particulates have important health effects and their emissions will likely be regulated. Baghouses are fabric collectors in which the dust-laden gas flows

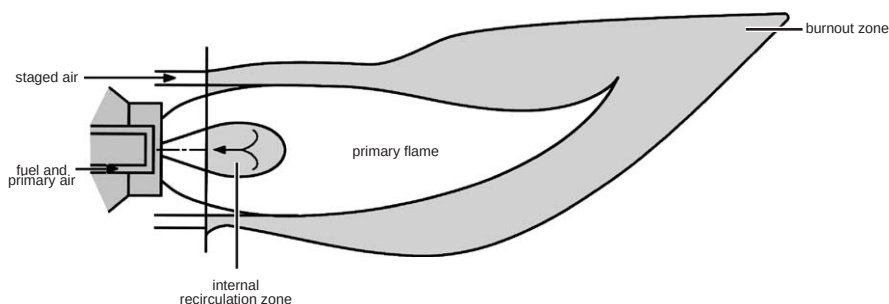


FIGURE 7 Schematic of a low- NO_x burner. From Soud and Fukusawa (1996).

through a cloth tube, where particles are captured by sieving action. As a “filter cake” forms by deposition, particles much smaller than the interstices of the cloth are captured. The particles are then removed by reversed air flow or shaking, and are collected in a bin below the cloth filter tubes. Baghouses are capable of precipitating submicrometer particles and fumes with high efficiency. Their flow resistance is higher than that of the ESP and their operation is, therefore, somewhat more costly.

4.4 Mercury

Mercury has been identified by the U.S. Environmental Protection Agency (EPA) as a toxic substance of great concern, because of its adverse health effects. Regulations for mercury removal from coal-fired utility plants are expected to be issued by the end of 2004. Significant research and development efforts by the U.S. DOE and industry are in progress to find cost-effective ways for the removal of mercury from stack gases. Potential solutions include SO_2 scrubbers that will also capture a high percentage of soluble (oxidized) mercury, SCR catalysts for NO_x removal to oxidize elemental mercury for subsequent capture by electrostatic precipitators or fabric filters, and the absorption of vapor-phase mercury by beds of activated carbon or particles injected into the flue gas and subsequently removed by gas–solid separation devices. The cost of mercury removal is significantly lower in integrated coal gasification combined cycle (IGCC) plants than it is in PC plants because the treated volume of syngas in IGCC plants is much smaller, compared to the flue gas in PC plants. However, this affects mainly the choice of new power plants, because of the small number of IGCC plants presently in operation.

5. NATURAL GAS-FIRED PLANTS

Natural gas is a premium fuel; because of its relatively high price, in the past it was used primarily for “peak-shaving” in gas turbine plants. Reassessment of the long-term availability of natural gas, the development of gas turbine–steam combined cycle (GTCC) plants with high cycle efficiency, environmental pressures for “decarbonization” of the fuel supply, and the rapid growth of distributed power generation have made natural gas the choice fuel for base load power generation. Natural gas has no sulfur, and the carbon:hydrogen ratio of its composition is much lower than that in petroleum fuels or in coals. Because

of the lower C:H ratio, CO_2 emissions from the combustion of natural gas are significantly lower compared to the heat equivalent of other fossil fuels.

In the gas turbine cycle, the gaseous combustion product is the direct working medium of power generation. The gas turbine (GT) consists of a compressor, combustor, turbine, and heat exchanger (Fig. 8). Combustion takes place at elevated pressure in the combustor. Large amounts of excess air are used as diluent, to ensure that the combustion product temperature at entry to the gas turbine does not exceed the limit set to protect the structural integrity of the turbine. However, because the cycle efficiency is proportional to GT entry temperature, concerted research and development efforts are in progress to permit higher inlet temperatures by means of improved construction materials and more effective turbine blade cooling technologies. As a result, gas turbine entry temperatures have risen almost steadily over the past two decades to a state-of-the-art value of 1570 K. Further increases can be expected by the introduction of advanced ceramic materials to GT technology.

5.1 Nitrogen Oxides Emission Control in Natural Gas-Fired Gas Turbine Plants

There is no nitrogen organically bound to fuel molecules in natural gas; NO forms primarily by the “thermal NO ” process. In the conventional method of combustion, fuel and air are separately injected into the combustor and mix in the course of combustion (diffusion flame). This process is prone to high rates of thermal NO formation because near-stoichiometric (theoretical air:fuel ratio) mixtures prevail on the boundaries of fuel-rich and fuel-lean turbulent eddies in such flames. In contemporary

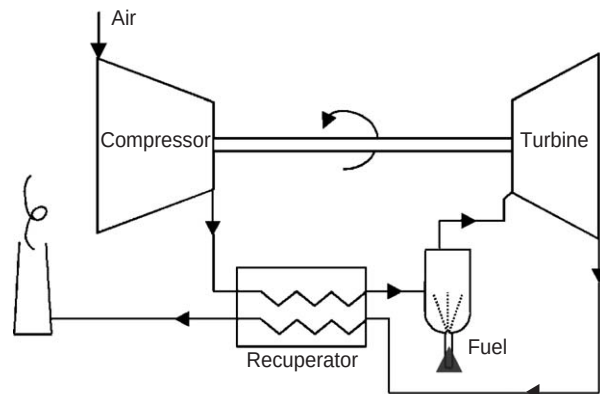


FIGURE 8 Gas turbine cycle with regenerator.

combustion technology's response to this problem, the fuel gas and air are premixed prior to their entry into the combustor, creating a strongly fuel-lean mixture that corresponds to the combustor exit gas temperature. This, so-called ultralean premixed combustion results in very low NO_x emissions, but because the combustor is operated at mixture ratios close to the lean blow-off limit of the fuel gas/air mixture, the flame stability may be impaired. For restoring flame stability, a small amount of non-premixed fuel, say 10%, is injected into the combustor to form a pilot flame and act as a stable source of ignition. Further reduction of NO can be obtained by steam injection into the combustor. Although steam injection will depress the cycle efficiency on account of loss of the latent heat of steam, this effect is somewhat compensated for by the fact that the steam represents additional mass flow through the turbine without the need for expending additional compression work.

6. POWER CYCLES OF INCREASED THERMODYNAMIC EFFICIENCY

Pollutant emissions from electric power-generating plants are inversely proportional to the thermodynamic cycle efficiency of power generation. Some power plants with advanced thermodynamic cycles are in commercial use, others have the potential to be applied in the near and medium term, and there are also long-term research and development programs for very high-efficiency and "zero-emission" cycles in power generation. It has to be borne in mind that in addition to technical feasibility, a novel electric power-generating system needs to be successfully demonstrated at a commercial scale for years, and must be economically viable, before it is accepted for application by the utility industry.

6.1 Pulverized Coal Combustion in Supercritical Steam Cycle

Pulverized coal-fired supercritical steam boilers (e.g. 250 atm, $2 \times 853 \text{ K}$), have been in use since the 1930s, but improvements in materials and increasing demand for higher efficiency are making this system the choice of new coal-fired utility plant worldwide. Their increased efficiency is due to the higher mean temperature of heat addition in the supercritical steam cycle. Because of a strong increase in the moisture content of high-pressure steam as it

expands through the last stages of the steam turbine, the steam has to be reheated by taking it back from the turbine to the boiler. Reheating, once or multiple times, will raise also the thermodynamic efficiency of the power cycle because it further increases the mean temperature of heat addition. Advanced supercritical steam power plants with steam parameters of 300 atm pressure, temperatures of 866 K, and cycle efficiencies up to 43% are expected to play a major role in new power generation in the near future. As the steam temperature approaches 973 K, the efficiency may reach 47%, but new advanced designs and materials for the boiler, the steam turbine, and the associated piping will be required; development will likely yield commercial applications after 2010.

6.2 Gas Turbine–Steam Combined Cycles

In gas turbine–steam combined cycle plant, the gas turbine exhaust, typically 750 K and containing 13–15% O_2 , raises steam in a heat recovery steam generator for producing power in a steam turbine and/or heat in cogeneration. Because of the complementary temperature ranges of the Brayton (1600–900 K) and Rankine (850–288 K) cycles, their combination can produce significantly improved thermodynamic cycle efficiency, approaching 58% in modern commercial plants. When coal or residual fuel oil is added as a supplementary fuel in the heat recovery steam generator of a GTCC plant (Fig. 9), there is considerable gain in operational flexibility, albeit at the expense of a small reduction in the cycle efficiency. Due to a low first cost (\$400–500/kW, less than one-half of the approximate first cost of a PC plant), high efficiency, and low emissions, GTCC plants are the favored choice for new power plants, despite the gas/coal price differential and the uncertainty of future gas prices.

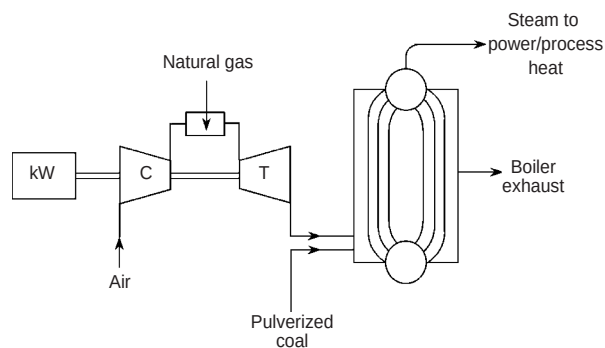


FIGURE 9 Gas turbine–steam combined cycle with fired heat recovery steam generator. C, compressor; T, turbine.

Gas turbines can be used also for retrofitting existing steam plants by injecting the gas turbine (GT) exhaust through the boilers' coal burners ("hot wind box repowering"), or applying it to partial boiler-feed water heating, thereby allowing more of the steam flow in the steam turbine to expand to condenser pressure instead of being bled for regenerative-feed water heating. This will then result in both higher efficiency and increased electric power-generating capacity.

6.3 Pressurized Fluidized Bed Combustion

In pressurized fluidized bed combustion (PFBC), the advantages of sulfur capture by additive sorbents in the bed and an operation insensitive to fuel quality are combined with increased efficiency of the combined steam and gas turbine cycle. Compared to FBC, the heat release rate per unit bed area in PFBC is several times higher, making the "footprint" of the pressurized power plant smaller. The bed height in the PFBC is 3–4 m instead of the typical bed height of 1 m in FBC. Under atmospheric pressure, the acceptable pressure drop of about 100 mbar limits height across the bed. In the PFBC–GT cycle, a 300 mbar pressure drop represents less than 3% of the total pressure ratio. A consequence of the increased bed height is a larger carbon inventory in the bed and lower NO emission due to the reduction of the NO formed by solid carbon. The high carbon load, however, does not reduce the emission of nitrous oxide, N_2O , which is still stable at the relatively low temperature (~ 1050 K) in the PFBC. A further disadvantage of this temperature is that it is also low for efficient gas turbine applications. The efficiency of the PFBC combined cycle units in operation is about 41%. Further increases in efficiency can be obtained by raising the turbine inlet temperature using a topping combustor.

6.3.1 The Second Generation of PFBC

The second generation of PFBC is a response to the need for increasing the gas temperature at the inlet to the gas turbine. In this cycle (Fig. 10), coal, usually in the form of a coal–water slurry, is injected into a pressurized fluidized bed carbonizer, where it undergoes mild gasification to produce char and a low-calorific-value gas. The char is burned in a pressurized circulating fluidized bed combustor (PCFBC) with high levels of excess air. Sulfur is captured in the PCFBC by additive dolomite and the flue gas is cleaned of particulates and alkali at high temperature before entry to the GT. Sulfur is also captured in the

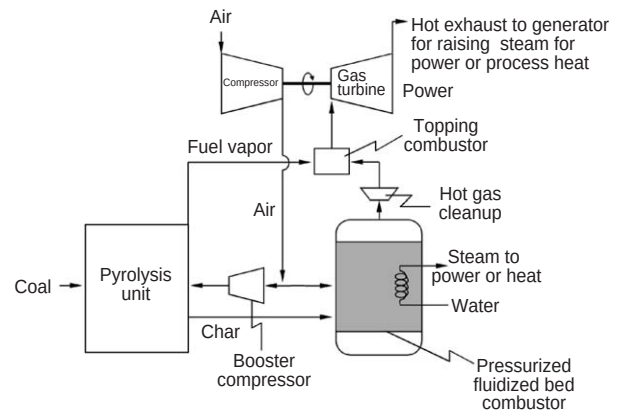


FIGURE 10 Pressurized fluidized bed combined cycle with topping combustion. From Robertson *et al.* (1989).

fluidized bed carbonizer, and the fuel gas, after being cleaned of particulates and alkali, is burned in the topping combustor to increase the temperature at the inlet to the gas turbine to 1470 K. This temperature rise increases the cycle efficiency up to about 45%. Further improvements in efficiency can be obtained by the application of advanced gas turbine technology and, on the steam side, by advanced steam parameters. An additional advantage of this cycle is that the N_2O emission is eliminated because the N_2O formed in the pressurized fluidized combustor decomposes at the elevated temperature in the topping combustor. The successful application of this cycle requires further demonstrations of the hot gas clean up and the topping combustor. The PCFBC with topping combustor is expected to be ready for commercial application after 2010.

6.4 Integrated Gasification Combined Cycle

IGCC involves the total gasification of coal, mostly with oxygen and steam, to produce a medium-calorific-value fuel gas for use in a gas turbine cycle. The fuel gas, after being cleaned of particulates, alkali, and sulfur compounds, is burned in a low- NO_x combustor. The gasifier also produces steam for a steam power cycle. A schematic arrangement of IGCC equipment (gasifier, gas-cleaning tanks, steam-raising tanks, and power-generating apparatus) is illustrated in Fig. 11.

IGCC is the cleanest presently available advanced coal technology. It is working with no major operational problems. The future of IGCC depends on whether it will be possible to reduce its first cost and to increase the cycle efficiency. The cost is

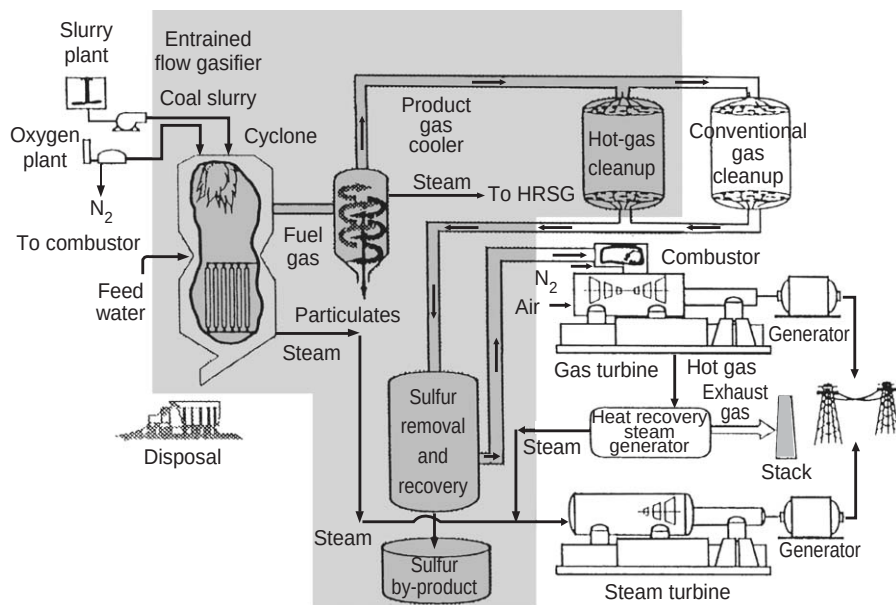


FIGURE 11 Integrated gasification combined cycle. HRSG, heat recovery steam generator. From Ruth (1998).

presently high, mainly because of the oxygen plant needed for the oxygen-blown gasifier and also because of the less than complete integration of the various subsystems, such as the gasifier air separation system, fuel gas cooler and cleanup, and gas turbine and steam plants. Existing IGCC demonstration plants in the United States have efficiencies under 40%, but two newly commissioned European IGCC demonstration plants have higher design efficiencies of 43 and 45%. These higher cycle efficiencies are mainly due to improved gas turbine and steam plant efficiencies and also to better subsystems integration. Further improvements yielding cycle efficiencies up to 50% are possible.

Even with improved cycle efficiencies, however, IGCC plants are presently not competitive with other advanced coal-burning systems, such as PC-fired supercritical steam plants, because of the higher installation costs of IGCC plants. Nevertheless, the following considerations may eventually tilt the balance in favor of IGCC applications:

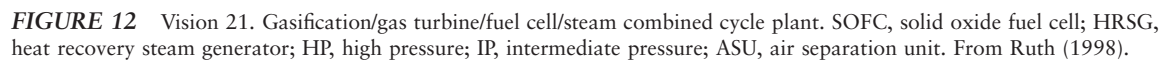
- The cost of mercury removal in IGCC plants is lower because the treated volume of syngas is much smaller than that of the flue gas in PC-fired plants.
- IGCC lends itself to more efficient removal of CO₂ from high-pressure fuel gas.
- By broadening the fuel supply to the ever-increasing volume of refinery wastes (heavy residual oils and petroleum coke), IGCC could become attractive for clean and efficient central

power generation by using fuels of very low or even negative cost (the waste fuel cost is negative if it stands against the cost of disposal).

6.5 Hybrid Gasification/Fuel Cell/Gas Turbine/Steam Combined Cycles

6.5.1 The DOE's Vision 21 Cycle

One of the promising coal-fired advanced cycles expected to be ready for demonstration in the period between 2010 and 2015 is the Vision 21 Cycle of the U.S. DOE. (Fig. 12). The fuel gas produced in an oxygen-blown gasifier at elevated pressure is cleaned at a temperature of 1273 K. The gas, composed mainly of H₂ and CO, enters a solid oxide fuel cell (SOFC) on the anode side, while air from a compressor exhaust, preheated in a recuperator, enters on the side of the cathode. The hydrogen is used to generate electricity in the SOFC, and the CO burns in a combustion turbine that drives the compressor. Electric power is produced in another SOFC and gas turbine, at a lower pressure downstream of the high-pressure turbine, with more power added by a bottoming steam cycle. The efficiency could reach 60% in this "zero-emission" scheme. Concerted research and development efforts by a DOE/industry/university program aim at the development of individual modules of this cycle and the complex control system required for flexible load following.



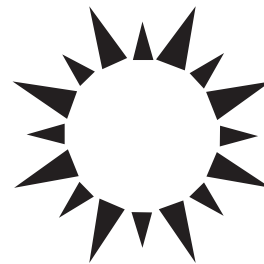
Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional

Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy • Hazardous Waste from Fossil Fuels • Investment in Fossil Fuels Industries

Further Reading

- Beér, J. M. (1988). "H. C. Hottel Plenary Lecture," *Symp. (Int.) Combust.*, 22nd, pp. 1–16. The Combustion Institute, Pittsburgh, Pennsylvania.
- Beér, J. M. (1996). Low NO_x burners for boilers, furnaces and gas turbines. *Combust. Sci. Technol.* **121**, 169–191.
- Blum, R., and Hald, J. (2000). "ELSAM." Skaerbaek, Denmark.
- Boyce, M. P. (2002). "Handbook for Cogeneration and Combined Cycle Power Plants." ASME Press, New York.
- Electric Power Research Institute (EPRI). (2003). "Electricity Technology Roadmap." [www.epri.com]. EPRI, Palo Alto, California.
- Energy Information Administration. (2000). "Annual Energy Outlook 2001 with Projections to 2020." U.S. Department of Energy, Washington, D.C.
- Folsom, B., Hong, C., Sommer, T., and Pratapas, J. M. (1993). Report TR103265-V2. *Proc. Joint Symp. Stationary Combust. NO_x Control*, 1993, pp. 7A; 15–34. Electric Power Research Institute (EPRI), Palo Alto, California.
- "Fuel Cell Handbook," 5th ed. (2000). DOE National Energy Technology Laboratory (NETL), Morgantown, WV.
- Haupt, G., and Karg, J. (1998). *Proc. Conf. Electric Power Supply and Industry (CEPSI)*, 12th. Pattaya/Thailand.
- Herzog, H. J., and Drake, E. M. (1993). "IEA Greenhouse Gas R&D Program, IEA/93/0E6." International Energy Agency, Cheltenham, U.K.
- Robertson, A., Garland, R., Newby, R., Rehmat, A., and Rebow, L. (1989). "Second Generation Pressurized Fluidized Bed Combustion Plant". Foster Wheeler Dev. Corp. Rep. FWC/FWDC-TR 89/11 to the U.S. DOE, DE-AC21-86MC21023.
- Ruth, L. A. (1998). "U.S. Department of Energy Vision 21 Workshop." DOE Federal Energy Technology Center, Pittsburgh, Pennsylvania.
- Sarofim, A. F., and Beer, J. M. (1978). *Symp. (Int.) Combust.*, 17th, pp. 189–204. The Combustion Institute, Pittsburgh, Pennsylvania.
- Scott, D. H., and Nilsson, P. A (1999). "Competitiveness of Future Coal-Fired Units in Different Countries." IEA Coal Research, CCC/14, London.
- Soud, H. N., and Fukusawa, K. (1996). "Developments in NO_x Abatement and Control." IEA Coal Research, London.
- Steinbach, C., Goudeaux, D., Troger, C., and Moore, R. (2000). "Operation Experience with Alstom Power's GT10 Gas Turbine and Ultra Low NO_x Combustion System." PowerGen Conference, Orlando, FL.
- U.S. National Coal Council Report to the Secretary of Energy. (2003). "Coal Related Greenhouse Gas Management Issues." Washington, DC.
- Wendt, J. O., Sternling, C. V., and Mafovich, M. A. (1973). *Symp. (Int.) Combust.*, 14th, pp. 897–904. The Combustion Institute, Pittsburgh, Pennsylvania.
- Wiswanathan, R., and Bakker, W. T. (2000). "Materials for Ultra Supercritical Fossil Fuel Plants." EPRI Report TR-114750.

Electric Power Generation: Valuation of Environmental Costs



THOMAS SUNDQVIST

Swedish Energy Agency
Eskilstuna, Sweden

PATRIK SÖDERHOLM

Luleå University of Technology
Luleå, Sweden

ANDREW STIRLING

University of Sussex
Brighton, United Kingdom

-
1. Introduction
 2. The Valuation of Electricity Externalities
 3. Electricity Externality Studies
 4. Critical Issues in Externality Assessment
 5. Conclusions

Glossary

abatement cost The implicit cost associated with controlling or mitigating damage.

damage cost The direct cost associated with damage.

economic valuation Measuring individual preferences for or against change, as expressed by changes in welfare.

environmental cost Externality arising from environmental impacts.

externality An unpriced and uncompensated side effect of one agent's actions that directly affects the welfare of another agent.

social cost The total costs to society of a production or consumption activity; the sum of private and external costs.

neoclassical welfare economics A branch of economics that studies efficiency and the overall well-being of society based on alternative allocations of scarce resources.

utility The welfare or satisfaction obtained from the use or consumption of goods and services.

willingness to accept Monetary measure for a change in welfare; the monetary amount that someone is willing to receive or accept to give up a good or service.

willingness to pay Monetary measure for a change in welfare; the monetary amount that someone is willing to give up or pay to acquire a good or service.

The availability of reliable electric power at low cost is a necessary condition for sustained global economic welfare. However, at the same time, electric power generation is the source of a number of social and environmental problems. Therefore, a wide variety of regulations and economic incentives exist worldwide to promote the introduction of more environmentally benign power generation technologies. When implementing such policy measures, two questions typically arise. First, which technologies are the more relatively environmentally benign? Second, how does one find a proper overall balance between the benefits of electricity production and the costs of environmental degradation?

1. INTRODUCTION

Over recent decades policymakers have shown an increased interest in the market-oriented measures recommended by neoclassical economics. According to this body of analysis, normal market processes are assumed to approximate optimal balances between those private costs and benefits that are reflected in market prices. For those costs and benefits that are "external" to market prices, the neoclassical prescription is to apply nonmarket valuation techniques

to determine the magnitudes of these externalities and, thereby, to enable regulatory or fiscal measures to explicitly include these values in market prices. Under this view, normal processes of market optimization will then be based on total social costs (private costs plus external costs) and, thereby, will reflect a proper balance between the benefits and impacts in their widest sense. Formally, then, an externality (i.e., an external cost or benefit) may be defined as an unpriced and uncompensated side effect of one agent's actions (in our case electric utilities) that directly affects the welfare of another agent. Therefore, the challenge lies in finding techniques by which to derive complete and reliable values for the many different types of external cost associated with electricity production.

In response to increasing demands from policymakers during the 1980s and 1990s, this challenge received intense attention from economists and environmental scientists worldwide. A number of large-scale international studies embarked on the task of quantifying and valuing the externalities associated with electric power production. Examples include the ExternE project in the European Community and the External Costs and Benefits of Fuel Cycles study in the United States. The ultimate goal of these studies has been to guide policymaking by providing a robust basis for new regulations and taxes and, thereby, to aid decision makers in their choices among fuels and technologies.

The results obtained by several of these externality studies have already begun to exert some influence on regulatory decision making and wider policy debates. For instance, certain external cost values have been featured prominently in legislative initiatives for "environmental cost adders" in a few U.S. states, and some other results (e.g., those associated with global warming) have from time to time become the subject of high-profile public policy debates. However, the results obtained in the accumulated extensive body of literature on the external environmental costs of electricity production so far have generally exerted only a limited impact on real policy decisions. Some authors argue that this is because electricity environmental impact studies may have raised more questions than they have answered and because there are important limits to their usefulness in deriving robust recommendations for policymakers.

This article reviews the economic theory behind externality valuation as well as the various methods that can be used to assess external costs in practice. It also surveys some results from the vast externality

valuation literature and discusses some important practical issues that bear on efforts to improve the environmental performance of electricity systems by adopting these particular values as a basis for regulatory and policy measures.

2. THE VALUATION OF ELECTRICITY EXTERNALITIES

2.1 The Background to Externality Valuation

The presence of an externality indicates a situation in which the first theorem of welfare economics fails to apply. In other words, externalities cause market failures, which then lead to a failure to accomplish Pareto efficiency. The market failures caused by externalities can arise in both production and consumption activities. Either way, they give rise to resource allocations that are nonoptimal from the point of view of society as a whole. Fig. 1 shows the relationship among a number of key concepts associated with market failure through the existence of externalities. The private costs facing a producer represent the best alternative uses of resources available as reflected by the market prices of the specific resources used by the producer. However, the social costs of production equal private costs plus external costs and represent the best alternative use of resources available to society as a whole. Therefore, a negative externality is equivalent to the difference between the private and social costs of an activity such as electricity production. Because by definition there exists no market for an external impact, a profit-maximizing utility has no incentive to integrate this effect into the decision-making process. Thus, private costs are lower than social costs. However, the existence of a difference between private and external costs is not inevitable. If the external costs can—by means of various fiscal or regulatory instruments—be "internalized" in electricity prices (i.e., be made private), producers and consumers in electricity markets will have an

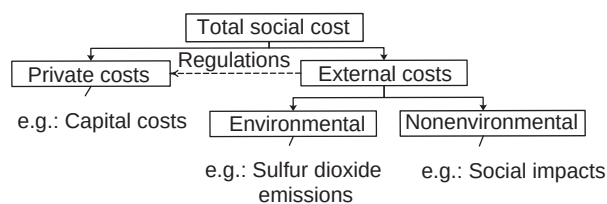


FIGURE 1 Total costs to society of a productive activity.

incentive to undertake actions that help to mitigate the negative environmental impacts of electricity use.

In his seminal work, Coase argued that bargaining between the polluter and the affected agents can, under certain circumstances (e.g., low transaction costs and full information), internalize externalities and achieve an efficient outcome. However, in most cases, because of the large number of parties involved, this calls for some kind of government intervention. One way of correcting the inefficiency of an external cost is the use of so-called Pigovian taxes as originally suggested by Pigou. This implies setting a tax equal to the value of the marginal external cost (at the optimal level of the activity in question) so that the private decision maker is provided with an incentive to economize not only on the traditional input factors but also on unpriced goods and services such as those provided by the natural environment. However, this solution to the externality problem requires that the tax authority is able to identify the external cost function. This returns us to the challenge of how to go about assessing the size of this function and, hence, the value of the damage caused by a negative externality (or the benefits incurred by a positive one).

2.2 Externality Valuation in Theory

The theoretical basis for the economic valuation of externalities (and thus environmental impacts) is outlined in the welfare economics literature. This strand of research recognizes that the economic value of a resource or service is ultimately a function of individual preferences and, therefore, that the tool for analyzing welfare changes is utility theory. For example, consider the welfare of individual i and assume that her utility (u_i) depends on a vector of private consumption possibilities ($\mathbf{x}$) and on a vector of environmental and resource flows ($\mathbf{q}$) given to the individual (i.e., environmental quality). Hence, individual i 's utility function,

$$u_i = u_i(\mathbf{x}, \mathbf{q}), \quad (1)$$

shows the various combinations of $\mathbf{x}$ and $\mathbf{q}$ from which the individual derives utility. Furthermore, assume that some of the elements in $\mathbf{q}$ have positive prices, represented by the vector $\mathbf{r}$. The individual maximizes her utility subject to the budget constraint,

$$M_i = \mathbf{p}\mathbf{x} + \mathbf{r}\mathbf{q}, \quad (2)$$

where M_i is individual i 's income and $\mathbf{p}$ is a vector of prices for $\mathbf{x}$. The first-order condition for this

maximization problem implies that individual i 's conditional demand function (i.e., conditioned on the imposed $\mathbf{q}$), x_i , for marketed goods can be written as

$$x_i = x_i(\mathbf{p}, M_i - \mathbf{r}\mathbf{q}, \mathbf{q}). \quad (3)$$

Inserting Eq. (3) into Eq. (1) gives the conditional indirect utility function, v , for individual i :

$$v_i = v_i(\mathbf{p}, M_i - \mathbf{r}\mathbf{q}, \mathbf{q}). \quad (4)$$

The use of indirect utility functions enables us to express welfare measures in money metric terms. Assuming that $\mathbf{q}$ consists of only one element q , that $\mathbf{x}$ is the numeraire, and that q is an environmental "good" (i.e., *ceteris paribus*, individual i would choose more of q if given the option), consider a project that causes the environmental quality to improve from q^0 to q^1 (i.e., $q^1 > q^0$). The project causes a positive change in individual i 's indirect utility (or welfare) so that

$$\Delta v_i = v_i(\mathbf{p}, M_i - \mathbf{r}q^1, q^1) - v_i(\mathbf{p}, M_i - \mathbf{r}q^0, q^0) > 0. \quad (5)$$

Thus, given the presence of the project, individual i is made better off. However, because utility is not directly observable and because environmental "goods" are not part of market transactions, we need to find alternative ways of assessing the value of this welfare change. Theoretically, two standard monetary measures of quality-based welfare changes are the compensating surplus and the equivalent surplus. The first measure, compensating surplus (CS), is given by the solution to

$$v_i(\mathbf{p}, M_i - \mathbf{r}q^0, q^0) = v_i(\mathbf{p}, M_i - \mathbf{r}q^1 - \text{CS}, q^1). \quad (6)$$

In this case, CS is equal to the maximum willingness to pay (WTP) for the environmental improvement. Second, consider the equivalent surplus (ES) measure, which is given by the solution to

$$v_i(\mathbf{p}, M_i - \mathbf{r}q^0 + \text{ES}, q^0) = v_i(\mathbf{p}, M_i - \mathbf{r}q^1, q^1) \quad (7)$$

Here, ES equals the minimum willingness to accept (WTA) compensation for the environmental improvement not occurring. The interpretation of the CS and ES measures is reversed in the case of an environmental quality deterioration, where CS is equal to the minimum WTA and ES is equal to the maximum WTP.

In empirical studies, it is generally the case that WTA measures tend to be substantially higher than WTP measures for the same change. Thus, the choice of WTP or WTA as a measure of economic value may significantly affect the size of the resulting externality estimate. Even though there is no firm theoretical basis for choosing one approach rather

than another, *WTP* is generally advocated as the most appropriate measure of changes in welfare. This is partly because *WTA* is not constrained by income and, therefore, is seen to introduce an incentive problem. On the other hand, the sensitivity of *WTP* to income introduces certain equity issues.

In sum, the economic valuation of externalities, and thus of many environmental impacts, builds on the assumption that people seek to satisfy their preferences, which are exogenously determined, complete, continuous, and ethically unchallengeable (subjective). The environment is essentially treated as any other private commodity, and people are willing to consider trade-offs in relation to the quantity or quality of environmental “goods.” According to the welfare economics literature, the appropriate role of policy in the field of energy externalities is to aggregate the monetary estimates of individual preferences and to weigh them against other (more tangible) economic benefits and costs. Thus, the economics of nonmarket valuation builds on (1) clear but relatively restrictive behavioral assumptions (i.e., utility maximization), (2) a sense of society as the sum of the preferences of its individual members, and (3) a view of the task of public policy involving the internalization of external impacts and with utilitarianism as the ethical principle guiding social choice.

2.3 Externality Valuation in Practice

In practice, there are two basic methodological approaches used for the valuation of external costs in the energy sector: the abatement cost approach and the damage cost approach.

The abatement cost approach uses the costs of controlling or mitigating damage or the costs of meeting legislated regulations as an implicit value of the damage avoided. The rationale behind this approach is that legislatures are assumed to have considered the willingness of the public to pay for alleviation of the damage in setting the standard, thereby providing a revealed preference damage estimate no less reliable than the more direct valuation methods. One of the serious caveats with this approach is that it relies on the rather strong assumption that decision makers always make optimal decisions, that is, that they know the true abatement and damage costs. Also, a necessary condition for social optimality is that the abatement costs used are derived from the pollution control strategy that provides the least cost of control. If not, the estimates cannot reflect damage costs adequately.

Another limitation of the abatement cost approach is that society’s preferences change over time as information, analysis, values, and policies change. Hence, past revealed preferences might bear little relation to actual impacts today and their current value to society. For instance, the implicit value of carbon dioxide (CO_2) emissions indicated by a revealed preference analysis would, in many cases, be very low because regulations targeted toward this problem are relatively few in number and low in scale. This built-in “tautology” of the approach means that estimates need to be constantly revised as regulations and policies change. Perhaps more important, given that policy is by definition optimal, the analysis provides no room for relevant policy implications; therefore, one can question why the analysis is needed in the first place.

The damage cost approach is aimed at measuring the net economic damage arising from negative externalities by focusing more or less directly on explicitly expressed preferences. This approach can be subdivided into two main categories: top-down and bottom-up. Top-down approaches make use of highly aggregated data to estimate the external costs of, say, particular pollutants. Top-down studies are typically carried out at the national or regional level, using estimates of total quantities of pollutants and estimates of total damage caused by the pollutants. Specifically some estimate of national damage is divided by total pollutant depositions to obtain a measure of physical damage per unit of pollutant (Fig. 2). These physical damages are then attributed to power plants and converted to damage costs using available monetary estimates of the damages arising from the pollutants under study. The main critique against the top-down approach is that it “generically” cannot take into account the site specificity of many types of impacts or the various stages of the fuel cycle. Another argument that has been raised

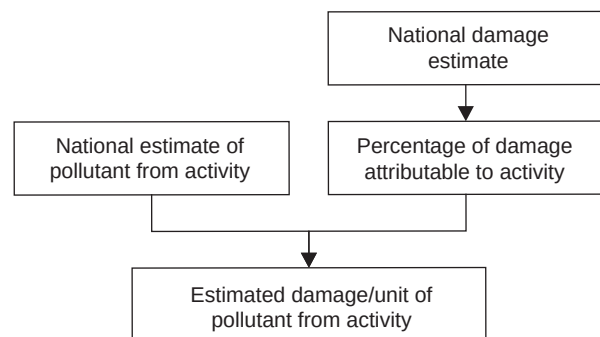


FIGURE 2 The top-down approach.

against the approach is that it is derivative because it depends mostly on previous estimates and approximations.

In the bottom-up approach, damages from a single source are typically traced, quantified, and monetized through damage functions/impact pathways (Fig. 3). This method makes use of technology-specific data combined with dispersion models, information on receptors, and dose-response functions to calculate the impacts of specific externalities. The bottom-up approach has been criticized because

applications of the method have unveiled a tendency for only a subset of impacts to be included in assessments, focusing on areas where data are readily available and rendering impact pathways easier to establish. That is, in bottom-up applications important impacts may be left out, especially in cases where data are not readily available. Furthermore, the bottom-up approach relies on models that might not adequately account for complexities in “the real world” given that there may be synergistic effects between pollutants and environmental stresses and

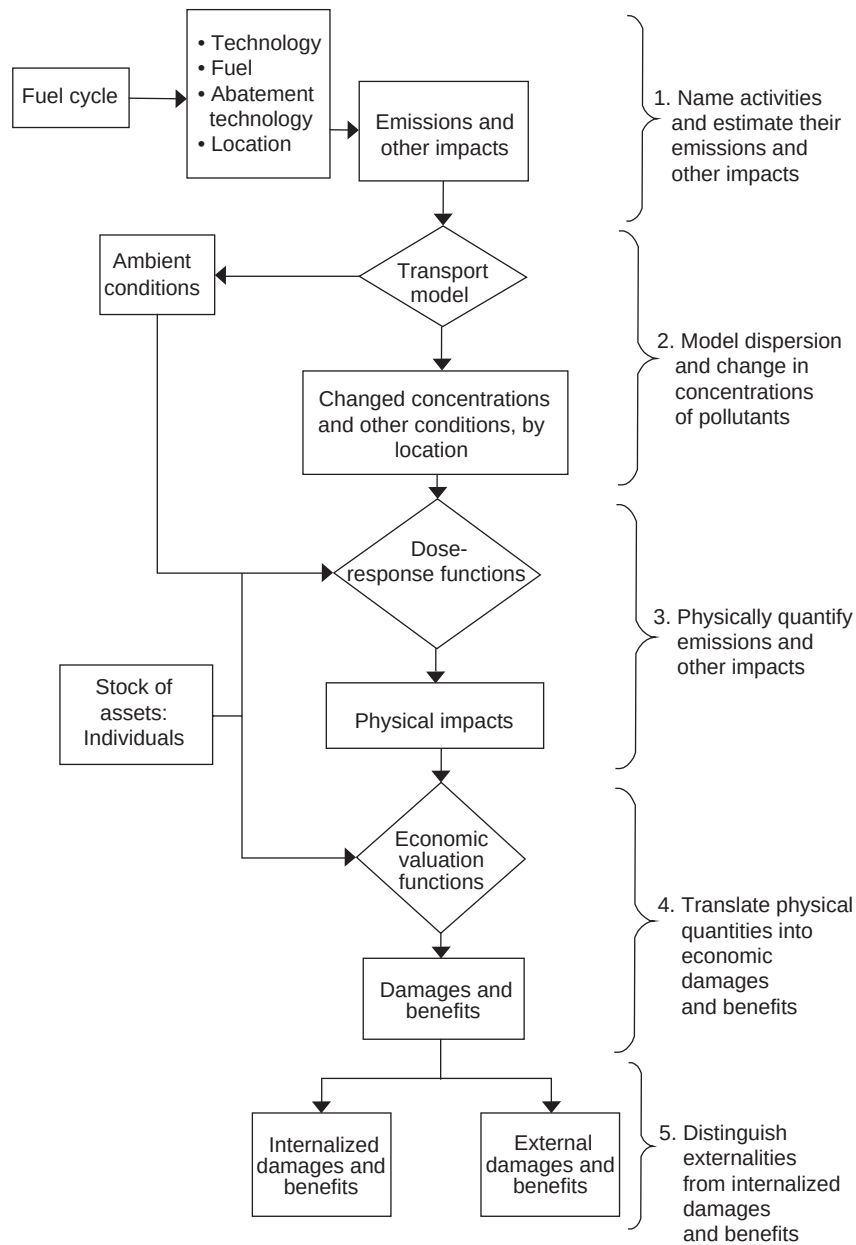


FIGURE 3 The impact pathway (bottom-up approach).

that there may be problems in establishing the timing of effects (e.g., between exposure and impact). Therefore, the argument is that bottom-up approaches might not be sufficiently transparent. Still, this is the approach that, due to its focus on explicit estimates of economic welfare (rather than on implicit ones such as in the abatement cost approach), appears to be most in line with economic theory. As is evident by the methodological choices of recent externality studies, it is also the most preferred approach to the empirical assessment of externalities in the electricity sector (see Section 3.1).

There are several ways of monetizing the identified externalities. The first two approaches discussed previously, abatement cost and top-down damage cost, directly provide a monetary estimate of the damages associated with the externalities. However, in the third approach, bottom-up damage cost, one needs to translate the physically quantified impacts into monetary terms. By definition, externalities are external to the market in question. However, certain costs (e.g., pollution damage to property or agriculture) may be external to electricity markets but subject to pricing in other markets. In such cases, the relevant market prices are used as a direct basis for valuation. Impacts from externalities that are not reflected in prices in any market must then rely on nonmarket valuation methods, which can be subdivided into direct methods and indirect methods.

The direct methods attempt to create a hypothetical market for the environmental good. These methods are direct in the sense that they are based on direct questions to households about willingness to pay. The direct methods possess the advantage that they can assess total economic values, that is, both the use and nonuse values (i.e., existence values) associated with the good. Well-known examples of such techniques include contingent valuation and choice experiments. The indirect methods take their basis in the actual (rather than the hypothetical) behavior of individuals. In terms of willingness to pay, the welfare effects show up as changes in costs or revenues either in observable markets or in markets closely related to the resource that is affected by the externality. Thus, the damage is indirectly valued using an existing relation between the externality and some good that is traded in a market. Examples of indirect methods are hedonic pricing (e.g., looking at property values) and travel costs (e.g., looking at amenity values).

3. ELECTRICITY EXTERNALITY STUDIES

This section gives an overview of the results from the multitude of externality studies carried out during the 1980s and 1990s. This is followed by a brief look at the results from one particularly detailed study, the European ExternE project.

3.1 Overview of Results

A considerable number of externality studies were carried out during the 1980s and 1990s. Figure 4 presents the results from 63 externality studies, with the numbers in square brackets showing the total number of observations for each fuel source (note the use of logarithmic scale). The ranges overlap, rendering ambiguous the ranking of various fuels with respect to externality impacts. Indeed, it is possible, using these overlapping ranges, to argue for virtually any ranking order one likes across the options in question. However, when the preponderance of findings are taken into account, the results do suggest that fossil fuel-fired power (in particular coal and oil) gives rise to the highest external costs, whereas some of the renewable energy sources (solar, wind, and hydropower) tend to have the lowest.

For a specific fuel source, the difference between low and high values is substantial, and this is also true if one looks at single studies. The ranges reported can often vary from a tiny fraction of electricity market prices and the private costs of producing power to a number that is significantly above private cost levels. However, the wider sensitivities are not always fully conveyed or discussed in individual studies, where results are often expressed with a fine degree of precision and are routinely referred to as “real,” “full,” or “true” values. When attention extends to the literature as a whole, the scale of the discrepancies between the results obtained by various studies for similar fuels does raise concerns about the validity and reliability of an approach to regulatory appraisal based purely on environmental valuation.

This said, it is of course entirely reasonable that there should be some degree of variability among studies. This may be due, for instance, to factors such as (1) a focus on different technologies (e.g., implying separate emission factors); (2) different site-specific characteristics (e.g., population density, income, transport distances); (3) differences in scope (e.g., different selections of externalities may be included, different taxonomies may be used to organize effects,

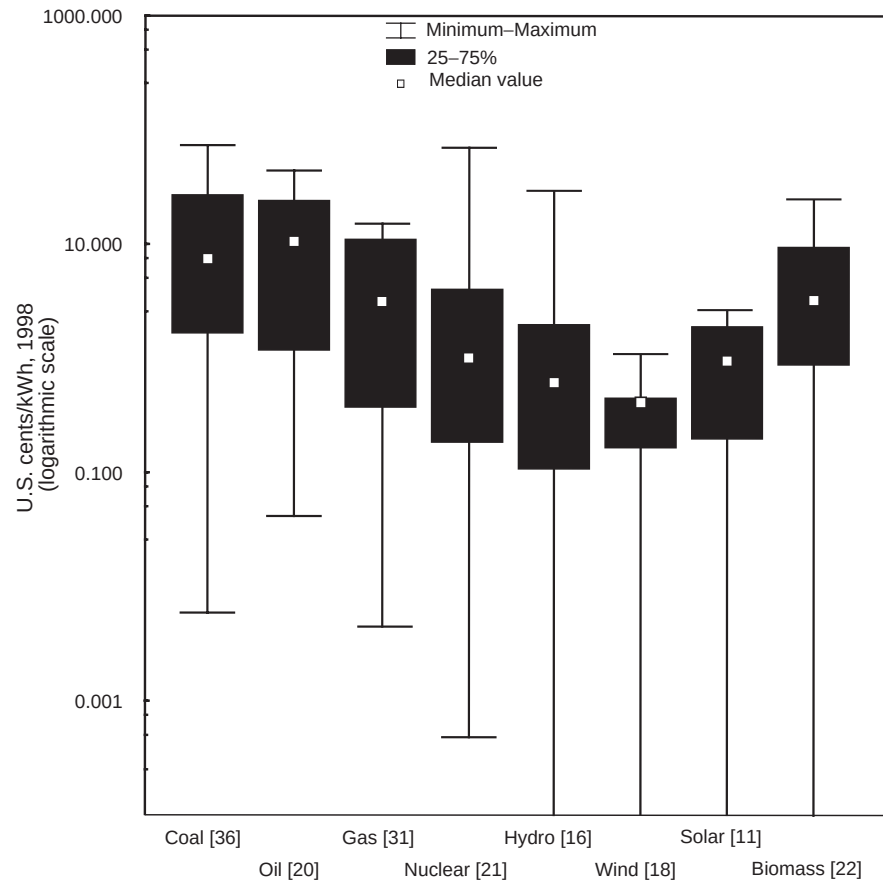


FIGURE 4 Range of external cost estimates for various fuel sources. From Sundqvist, T., and Söderholm, P. (2002). Valuing the environmental impacts of electricity generation: A critical survey. *J. Energy Lit.* 2, 3–41. Reprinted with the permission of the Oxford Institute of Energy Studies.

different stages in fuel cycles or resource chains may be evaluated); (4) different trade-offs may be implicit among the various dimensions of environmental effects (e.g., injuries and disease, immediate or latent effects, concentrated or dispersed effects); (5) different treatments of complexity, uncertainty, or gaps in knowledge (e.g., *ex ante* rather than *ex post* data; likely technological trajectories; indirect, contingent, or synergistic effects). Some of these issues are further discussed in the following section. For the moment, note that it is difficult to determine whether such factors justify such large variability in results, especially given that there is no objective metric against which to compare the “truth” of the contending empirical estimates.

Sundqvist and Söderholm provided an overview of roughly 40 externality studies and showed that, in addition to the apparent disparity of results, several other conceptual issues are of importance. First, most of the fuel sources available for power generation,

including coal, oil, natural gas, nuclear, hydro, wind, solar, and biomass as well as (in a few cases) lignite, waste incineration, geothermal, peat, and orimulsion, have been addressed in previous valuation efforts. However, most studies focus on the traditional fuels such as coal and nuclear. Thus, there is a tendency for many studies to focus on existing technologies rather than on those technologies that are generally accepted as being likely to play a significant role in the future (e.g., wind, biomass). In many cases, this is understandable given that empirical data clearly are more available for existing technologies than for emerging ones. Nevertheless, an important goal of externality valuation in the power sector has been to “level the playing field” in the selection between traditional and new generating technologies. This would probably require a stronger focus on promising but as yet uncommercialized technologies.

Second, a majority of the studies have been carried out for the developed world (mostly for Western

Europe and the United States). Thus, only in some rare cases has the focus been on developing countries, where the pressure for additional power capacity is by far the greatest. There are also reasons to believe that externality estimates should differ substantially between developing and developed countries. In developing countries, incomes are lower, sociocultural priorities are likely to differ, and the environmental effects of power production may be fundamentally different. An important example of the latter is the environmental externalities stemming from hydropower. Hydroelectric development in tropical climates may give rise to global warming impacts due to decay of vegetation left in the reservoir, whereas hydroelectric development in cooler climates will not. This raises serious concerns about transferring values from studies conducted in, say, Western Europe for use in a developing country context.

Third, examining the methodologies used over time reveals that the bottom-up damage cost approach has become the dominant paradigm, whereas the abatement cost and top-down approaches were used predominantly during the 1980s and early 1990s (Fig. 5). An important reason for this development is that the national implementation phase of the ExternE project relies solely on damage cost bottom-up models, and these studies together represent a large share of the total number of projects conducted during the latter part of the 1990s. However, this also indicates that the bottom-up model has been accepted as the most appropriate method with which to assess power generation externalities. The ExternE project has largely served as a vehicle in the methodological development of

externality valuation. The sophistication and scope of the ExternE work and the detailed attention to methodology have been widely respected at the international level, and many followers rely heavily on the numbers and methods presented. However, this development raises the question of whether the choice of methodological approach (between abatement costs and damage costs) matters for the results. We return to this question in more detail later and suggest that this choice might very well matter. This, in turn, raises important concerns about the reliability of external cost valuation exercises in the power-generating sector.

3.2 An Example: The ExternE Core Project

As noted earlier, one of the most influential externality studies is the European ExternE project. This project aimed at evaluating the external costs of the various power generation fuel cycles in the European Union using the bottom-up approach. Because the ExternE project represents one of the most ambitious and internationally respected attempts to yield a “true” external cost estimate for the various power technologies, it serves well as an example. Tables I and II present the various power generation externalities quantified and priced within the ExternE core project. The ExternE core project was followed up by the so-called national implementation project, whose aim has been to develop a European Union-wide set of external cost data for the various fuel cycles and countries as well as methods for aggregation of results to make them useful for policymakers.

When inspecting Tables I and II, we first observe that most of the fuel cycles involve significant health and mortality impacts on humans (public and occupational). In this project, considerable attention was devoted to the evaluation of public health impacts and much was learned, especially about the importance of fine particle emissions. Overall, the results support the rather unexceptional conclusion that fossil fuels, especially coal, present more serious environmental impacts than do renewables such as hydro and wind. It is notable that the most favorable results are attributed to nuclear power. However, most of the externality damages priced in the ExternE core project are reported to reflect an uncertainty range characterized as “medium to high.” This raises a number of salient critical issues bearing on the robustness and meaning of these results and their associated prescriptive rankings.

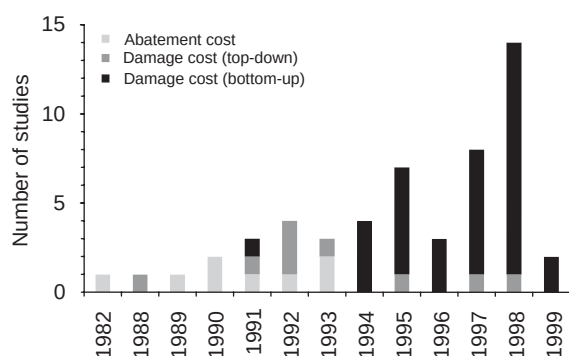


FIGURE 5 Methodological choice over time. From Sundqvist, T., and Söderholm, P. (2002). Valuing the environmental impacts of electricity generation: A critical survey. *J. Energy Lit.* 2, 3–41. Reprinted with the permission of the Oxford Institute of Energy Studies.

TABLE I

Externalities Priced within the ExternE Core Project: Coal, Oil, and Gas

Externality	Coal	Oil	Gas
Public health	Particulate matter, ozone, and accidents: mortality, morbidity, and transport impacts	Particulate matter and ozone: mortality, morbidity, and transport impacts	Particulate matter: mortality, morbidity, and transport impacts
Occupational health	Diseases from mining and accidents during mining, transport, construction, and dismantling	Accidents: death and injury impacts	Accidents: death and injury impacts
Agriculture	Sulfur, acidification, and ozone: crop and soil impacts	Sulfur, acidification, and ozone: crop and soil impacts	
Forests	Sulfur, acidification, and ozone damages	Sulfur, acidification, and ozone damages	
Marine	Acidification impacts	Accidents with oil tankers	Fishery: extraction impacts
Materials	Sulfur and acidification damages on surfaces	Sulfur and acidification damages on surfaces	Sulfur and acidification damages on surfaces
Amenity	Noise: operational road and rail traffic impacts		Noise: operational impacts
Global warming	CO ₂ , CH ₄ , and N ₂ O damages	CO ₂ , CH ₄ , and N ₂ O damages	CO ₂ , CH ₄ , and N ₂ O damages
Total estimate ^a (U.S. cents/kWh)	2.8–4.1	2.7–2.9	1.7

^aThe global warming impacts constitute roughly half of the reported external cost estimates for coal-, oil-, and gas-fired power.

TABLE II

Externalities Priced within the ExternE Core Project: Nuclear, Hydro, and Wind

Externality	Nuclear	Hydro	Wind
Public health	Radiation and nonradiation: mortality and transport impacts from operations and accidents		Accidents: travel to and from work
Occupational health	Radiation and nonradiation: mortality and transport impacts from operations and accidents	Accidents during construction and operation	Accidents during manufacturing, construction, and operation of turbine
Agriculture		Loss of grazing land	Acidification: damage on crops
Forests		Forest production loss due to flooding and land use	Acidification damages
Marine		Water supply and ferry traffic	Acidification damages
Materials			Acidification damages
Amenity		Visual amenity loss	Noise and visual amenity loss: Operational impacts
Global warming			CO ₂ , CH ₄ , and N ₂ O damages
Recreation		Fishing and hunting	
Cultural objects		Objects of cultural and archaeological interest	
Biodiversity		Terrestrial and aquatic ecosystems	
Total estimate (U.S. cents/kWh)	0.0003–0.01	0.3	0.1–0.3

4. CRITICAL ISSUES IN EXTERNALITY ASSESSMENT

This section discusses a number of critical issues concerning nonmarket valuation of energy externalities and focuses in particular on the role of externality valuation in policymaking.

4.1 Methodological Issues

As already noted, we need not necessarily be concerned about the variability in external cost estimates if we can identify rational reasons for the discrepancies. However, one possible reason for such discrepancies that would itself raise concerns is that the choice of methodology may have a significant impact on the results. By using the results from more than 40 electricity externality studies across eight fuel sources, Sundqvist provided some econometric evidence in support of the conclusion that methodological choice does matter for the results. He reported that the probability of obtaining a “low” externality value is, *ceteris paribus*, generally lower when the abatement cost approach or the top-down damage cost approach is used, whereas the opposite is true when the bottom-up damage cost approach is used. Abatement costs will (theoretically) be representative of damage costs if and only if they are derived from the least cost strategy, but normally most studies employ the most commonly used (or mandated) abatement technology when assessing pollution abatement costs. As a consequence, it can be expected that the results for the abatement cost applications would deviate. The reason why the top-down approach also tends to yield relatively high results is that practical problems may arise in attributing the “exact” damage to each individual source, and this may force researchers to rationalize and use standardized rules for the attribution process. These rules may fail to ascribe the aggregated damage to each and every individual source, especially smaller sources, thereby producing estimates for specific power plants that are positively biased, given that these plants normally are easily identifiable as well as significant sources of pollution.

A fundamental question raised by these results concerns the extent to which the three broad approaches to externality valuation are comparable. So far, we have suggested that the differences reported may be due primarily to a number of technical factors in analysis such as identifying the least cost control strategy and addressing all relevant

end points. If these issues can be resolved, it might be hoped (at least in principle) that the methods could begin to generate more similar results. In what follows, we suggest a number of reasons why this is not necessarily the case. However, before turning to this, it may be worth reviewing one final observation that has been made concerning the pattern evident in the picture of environmental valuation results provided in Fig. 4.

Stirling observed that the scope for multiple divergent, but equally legitimate, “framing assumptions” is a feature held in common between environmental valuation and many other forms of regulatory appraisal, including risk assessment and decision analysis. However, one feature that is unique to neoclassical environmental valuation is that what is conveyed to policymaking is not simply the relative performance orderings for different options but also the particular numerical values themselves. These are expressed as particular values for monetary externalities, suitable for adoption as a basis for Pigovian taxes. Stirling argued that this may make environmental valuation peculiarly vulnerable to a phenomenon he referred to as a “price imperative.” This describes a situation where, when faced with widely differing but equally valid possible assumptions, the choices made in analysis—implicitly or explicitly—lead results to fall into a domain that is most likely to be practically usable in policymaking. A Pigovian tax set at much less than one-tenth of market prices would have minimal impact on producer or consumer behavior. A Pigovian tax set at 10 times market prices would probably be politically unfeasible. This may be a factor in explaining why, despite ranges of variation spanning more than six orders of magnitude in Fig. 4, the bulk of valuation results lie well within one order of magnitude of existing wholesale electricity prices.

4.2 Externality Classification

The neoclassical welfare economics literature provides a relatively straightforward definition of the concept of externality. However, in practice, the choice of “relevant” externalities tends to differ among various valuation studies. In other words, some studies seem to differ considerably in their definition of what constitutes an externality. As an example, the inclusion of employment benefits (following a power generation investment) as an external benefit strains the definition of what an externality is. For employment effects to be considered external, the local and regional labor markets

must function poorly (i.e., market failures must be present) and workers must be immobile. However, some authors make a strong case for treating employment impacts as external benefits. For instance, to the extent that people obtain disutility (e.g., less self-confidence) from unemployment as such, this notion is valid. However, in such a case, the employment benefits should be compared across all fuel sources and not only for renewable energy projects (as is often the case).

For economists, the existence of externalities motivates regulatory action and the use of taxes and subsidies to improve efficiency. However, lay people and politicians often hold the view that regulations and other policy measures should be used to promote the “good” and discourage the “bad” in a much broader sense. This suggests that economists have to be careful in explaining what the results stemming from externality appraisals show and what they do not show. The preceding also raises the more philosophical question of what should be the ultimate end of policy: economic efficiency based on individual preferences or other value judgments formed in public deliberations (or any mix of these policy principles)? If the latter path is chosen, externality valuation may still provide an important input into the policy decision process, but the notion of some “total” (or “true”) cost of electricity production appears less valid. We return to this question later.

4.3 Scope

Even when all of the external costs and benefits of a given power generation source have been identified, there remains the issue of choosing the appropriate level of scope for the valuation effort. At least two choices have to be made. First, which of the externalities are important enough to include in the valuation? Second, should the study address externalities across the entire fuel cycle (including downstream and upstream impacts) or focus only on the generation stage? Studies conducted in the past have made different choices in these respects, and this is likely to explain some of the disparity in the estimates presented. Let us first look at how this question has been dealt with in previous studies.

As a part of the process of deciding what externalities to include, many studies begin with a comprehensive screening of the relevant economic and scientific literature to determine which pathways and end points are likely to be important in the analysis. Thus, some of the externalities are *de facto* assigned a zero value on the basis of their presumed

insignificance, whereas others are included in the analysis because the initial screening could not rule out the possibility of significant effects. It should be clear that this approach may lead to total externality values that are significantly downward biased. However, what is perhaps of more importance is those cases where externalities are left out of the analysis because there is insufficient scientific information to establish defensible monetary damage values. The most important example is global warming and CO₂ emissions. During the first phase of the ExternE project, it was noted that the environmental damage cost estimates for greenhouse gas emissions presented in the literature spanned a range of several orders of magnitude, and the main report concluded that “all attempts to value these impacts require important normative judgements, and therefore the potential for synthesis or consensus is remote.”

After additional research efforts within the ExternE project, these general conclusions appear to remain valid. However, it is generally agreed that the external costs of CO₂ emissions are substantial and, therefore, may constitute a large share of the total value. Thus, the environmental economist faces a dilemma: is it better to leave out potentially important external damages from the valuation and present biased estimates, or should one make use of rough proxy estimates (e.g., mitigation costs) to provide (or at least approach) some kind of “full-cost” estimate? In the end, the ExternE project chose the latter path and recommended the use of “minimum,” “central,” and “maximum” estimates.

This raises a number of important issues. The choice of which externalities to include in the assessment cannot be entirely objective but is rather a matter of judgment. The judgment that has to be made is essentially whether the externality under consideration is “mature” enough to “undergo” economic valuation. However, this is not only a question of whether the scientific knowledge is more or less established; it also involves the issue of whether the public is sufficiently informed and, hence, able to form a valid opinion about the issue at hand. Again, economic valuation is ultimately about measuring people’s given preferences toward goods, but if no relevant preference structure for a particular good exists (as with global warming), the valuation effort may be seen as arbitrary. We also noted previously that the use of abatement cost estimates (i.e., regulatory-revealed preferences) provides poor substitutes given that such estimates rely on the notion that all relevant preferences

have already been perfectly integrated into policy decisions.

This raises questions about the decision in the ExternE project to present rough estimates of global warming impacts. Moreover, global warming also entails important ethical and justice-related issues such as the question of whether human lives lost in developing countries should count less than corresponding lives in the developed world and to what extent impacts affecting future generations should be discounted (if at all). In such a case, the initial challenge of policy might not lie in “measuring” and aggregating individual preferences but rather might lie in specifying the conditions for public discourse over common ways of understanding the meanings and relative priorities attached to the various pertinent issues.

The preceding also suggests that any notion of “full,” “total,” or “true” cost of electricity has to be understood as hypothetical at best. The fact that externality studies leave out potentially important externalities might not be a problem per se. On the contrary, nonmarket valuation builds on specific basic assumptions, primarily about the behavior and preferences of the public, and if these are not well articulated, it may be better to refrain from monetary valuation. These concerns are likely to be particularly valid for “goods” with which people have relatively little past experience and that are “complex” and involve far-reaching and unknown consequences. Global warming, like the risk profile for nuclear accidents and radioactive waste, provides examples.

Finally, the choice of what parts of the fuel cycle to focus on also complicates the assessment of the full cost of electricity. Where should one draw the appropriate analytical boundaries and, thus, the limits of the external cost assessment? This choice is perhaps somewhat more technical than that of what externalities to include, but it is no less important. It is also complicated by the fact that the values of many upstream and downstream impacts tend to be highly site specific and depend, for instance, on the mode of transportation and the location of the place where the fuel is extracted. One problem is how one is to define the “entire” fuel supply chain. For instance, if we include in the analysis the externalities incurred by transporting fuels from the mine or well to the power station, should we then not also include the corresponding transport-related externalities involved in bringing building materials and workers to the power plant construction site? The choice of a relevant level of analysis is not easy, and this also suggests that the

notion of any “total” or “true” cost of power generation is flawed. Ideally, this choice should be guided by the need for relevant policy information.

4.4 Parameter Input Assumptions

There are essentially two different categories of parameter input assumptions made in electricity externality studies: technical assumptions (e.g., energy efficiency, dose–response functions, emission factors) and economic assumptions (e.g., monetary values, discount rates). Arguably, the assumptions regarding, for example, CO₂ emissions are major drivers of the results. Another striking example concerns the assumptions made about the monetary values used to value mortality impacts due to air pollution. Many studies use the value of a statistical life (VOSL) to assess these types of impacts. For instance, the ExternE core study relies on a VOSL value of U.S. \$2.6 million for Europe. In the national implementation part of the ExternE project, the decision was made to introduce an alternative measure on which to base the valuation of mortality impacts due to air pollution. This is the so-called years of life lost (YOLL) approach, which essentially assigns a WTP value to the risk of reducing life expectancy rather than to the risk of death. The YOLL values attributed to the mortality impacts are, as is evident from Fig. 6, reduced by up to two orders of magnitude as compared with the values based on the VOSL method. In addition, the core project that relied on the VOSL approach did not include values for chronic mortality impacts due to air pollution, something that the national implementation studies do.

Thus, it is clear that the assumptions underlying the valuation of human health and mortality impacts, global warming impacts, and dose–response functions are major drivers of external cost estimates. Clearly, this also provides a major explanation as to why the total externality estimates from various studies often differ considerably (even for the same fuel source).

So far in this subsection, we have noted that the input assumptions made in externality assessments play a significant role in affecting the overall external damage estimates and, thus, in guiding policy. To some extent, this is how it should be, of course; different sites and different technologies (depending, e.g., on vintage) incur different emission impacts and, hence, different damages. Still, the large sensitivity in results due to parameter input assumptions also creates problems for policymakers. First, policymakers often wish to gain some notion as to which

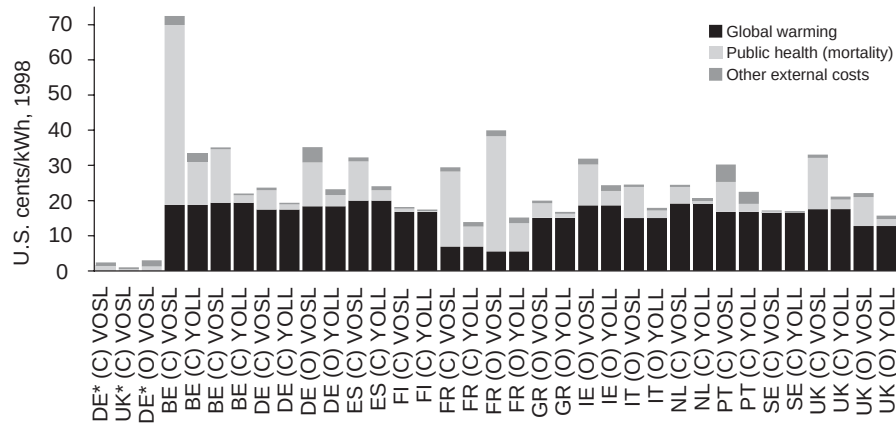


FIGURE 6 External cost estimates in the ExternE core (*) and national implementation projects: coal (C) and oil (O) fuel cycle. From Sundqvist, T., and Söderholm, P. (2002). Valuing the environmental impacts of electricity generation: A critical survey. *J. Energy Lit.* 2, 3–41. Reprinted with the permission of the Oxford Institute of Energy Studies.

are the most important external costs (i.e., environmental damages) for each power generation source. This enables them to target policy measures, such as subsidies, research and development (R&D) support, and regulations, toward the most important impacts. However, previous research efforts have provided only limited guidance on this particular point. Also, in many studies, a single category of external damages seems to dominate the total external cost estimates. For example, in the initial ExternE coal study, human health-related impacts dominate the aggregate damage (some 76–95% of the total), whereas global warming impacts constitute an important part of the estimated damages from the coal fuel cycle in the national implementation part of ExternE (Fig. 6). Second, in deregulated electricity markets, one of the most important uses of external cost estimates are as inputs in the development of environmental taxes and tradable permit schemes. However, for this reason, policymakers would need relatively “safe bets” about the general impacts involved, but so far previous studies have provided only wide ranges of estimates. Again, the impact of global warming is a good illustration of this.

This discussion raises a fundamental issue in nonmarket valuation that is rarely touched on in the electricity externality debate. Most environmental economists would agree that environmental valuation requires that a relevant “project” has been defined and that it involves the choice between two or more relevant alternatives. In the case of electricity externalities, these “projects” are normally the investments in various power plants (few companies or governments undertake investments

in entire fuel cycles). However, in externality studies, these investment projects are often hypothetical; that is, they do not represent an existing (real) situation, and valuation estimates are transferred from other studies. The problem with this is that according to the literature on nonmarket valuation (and indeed that on market valuation as well), economic values are context dependent and project specific. In other words, it might not make much sense to talk about a universal *WTP* for avoiding 1 ton of SO_2 being emitted, even though that is just what many policymakers would like to know about because they often prefer more or less harmonized standards and taxes across different regions.

4.5 Public versus Private Preferences

Critics of environmental valuation often stress that the methods used rely on overly restrictive assumptions. This implies that such studies often produce poor descriptions of the environmental values that people actually hold as well as underlying processes of preference formation. Specifically, externality valuations rest on an assumption that individuals aim at maximizing personal utility and that they possess well-articulated, ethically unchallengeable, and exogenous preferences for any environmental good. They also effectively assume that these individual preferences may be combined to yield an aggregate preference function for society as a whole. Both types of assumption are quite seriously problematic.

This second assumption is problematic even when approached within the confines of neoclassical welfare economics on which environmental valuation is

built. It glosses over the profound intractabilities involved in seeking definitively to compare the intensities of subjective preference displayed by various social agents. Indeed, even where social choices are addressed simply in ordinal terms as a matter of relative option rankings, economist Kenneth Arrow went a long way toward earning his Nobel Prize by demonstrating formally that it is impossible, under basic axioms of rational choice in neoclassical welfare economics, to guarantee any definitive aggregation of preference orderings in a plural society. In other words, the business of aggregating individual preferences, lying at the heart of neoclassical approaches to environmental valuation, violates its own well-established theoretical principles.

Beyond this, the first assumption raises a number of further concerns arising outside the narrow field of rational choice in neoclassical welfare economics. For instance, rather than being taken simply as an expression of personal utility maximization, environmental values may instead be seen to hold a broad ethical content. Because ethics is a matter for cultural and political discourse, environmental valuation should, it is argued, be endogenous to the political process and rely on social agreements. In other words, the initial challenge of environmental policy might lie not in “discovering” private preferences but rather in specifying the conditions for public discourse over what is worth valuing and for what reason.

Sagoff claimed that individuals essentially have at least two different preference orderings. Their private preferences reflect only their own well-being as consumers of private goods, whereas the public preferences reflect moral values about what persons, as citizens, believe is right and just for society as a whole. In their roles as citizens, people may express a rights-based (or deontological) belief system, which denies the principle of utilitarianism (and trade-offs) and instead recognizes the priority of the right over the good. Such a moral position is likely to be particularly prevalent in the case of environmental goods. The environment is often the subject of ethical concern and involves many cross-cutting dimensions that cannot be causally separated. In this way, the “market analogy” and “commodity fiction” of environmental valuation may break down.

There is empirical evidence that people express public, rather than private, preferences when considering complex environmental issues and when confronted with *WTP* questions in contingent valuation surveys. It is also the case that the

consumer–citizen distinction provides one explanation of some puzzling phenomena frequently observed in contingent valuation studies. For example, a large number of studies have found very large discrepancies between *WTP* and *WTA* measures for environmental goods, with the latter often being several times larger than the former. This may be explained by the notion that in the *WTA* case, if the environmental good is morally considerable, the acceptable level of compensation will be extremely high or even undefined.

The distinction between private and public preferences is highly relevant when considering the external costs of power generation. Most important, many power generation externalities are either relatively novel (e.g., the risk profiles of nuclear power) or complex (e.g., ecosystem changes due to hydropower developments). For these types of externalities, most people have yet to find an overall ethical position; consequently, there are no “given” preferences on which economic valuation may rely. In addition, most power generation fuel cycles involve significant impacts on the health and deaths of humans. These impacts raise a moral dilemma: to what extent should we treat humans as means to an end (utility) or as ends in themselves? Such a question should ideally be resolved within the realms of public discourse. Thus, the social choice problem with respect to many energy and environmental issues is first, about advancing common ways of understanding the pertinent issues. This implies that environmental research in the social science field must increasingly address the instruments and content of political and moral debate and not simply the technicalities of established valuation methods. In other words, the process may count every bit as much as the outcome.

The preceding discussion also adds a new perspective to the observed differences in reported externality estimates between the abatement cost approach and the damage cost approach. The two methods involve different ethical bases. For example, the damage estimates developed within the ExternE project are considered *ex ante* (i.e., the damages themselves determine whether one power source is “better” than another), whereas the estimates derived using the abatement cost approach are *ex post* (i.e., the price is an outcome of a political process and does not play a direct role in the decision). Therefore, the damages developed in “advance” (as in the ExternE project) might not be directly comparable to “implicit” estimates that are based on the cost of abating as revealed by decision makers given that

they reflect different reasoning processes. Policy-makers, in their formulation of regulations, may very well rely on ethical foundations other than economic welfare theory.

5. CONCLUSIONS

This discussion has considered the theoretical framework, reviewed the principal methodologies, and surveyed the detailed results associated with the valuation of external environmental costs of electricity production. Given the magnitude and pressing nature of the environmental imperatives in this area, it is clear that serious careful study of the kind that is well represented in this field has a crucial role to play in informing policy. However, it supports neither the cause of rigorous analysis nor that of environmental protection to neglect the series of profound difficulties and intractable challenges that attend neoclassical environmental valuation. In particular, it is important to note that the theoretical foundations are contestable, the methods are problematic, and the practical results are highly variable and sensitive to divergent “framing assumptions”.

Overall, the externality studies carried out have provided a welcome contribution to increasing the knowledge about the economic effects associated with the environmental problems in the electricity sector. However, more research is likely needed before these and future studies will be of practical use in shaping policymaking. Thus, perhaps the main positive conclusion that may be drawn from this critical survey is that the real value of environmental externality studies lies not so much in the particular numerical results they yield but more in the opportunity they provide to focus on the key sources of variability and sensitivity. In this regard, there may be considerable potential to combine these methods with techniques such as sensitivity and scenario analysis to systematically explore the implications of various possible frameworks and assumptions. By embedding such broader based approaches in inclusive deliberative procedures involving different industry and public stakeholders, the emphasis might usefully shift away from the numbers themselves and toward the contending perspectives and assumptions that “frame” these numbers. In this way, environmental valuation might contribute toward establishing a truly well-informed and robust basis for future environmental decision making in the electricity sector.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electricity, Environmental Impacts of • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy

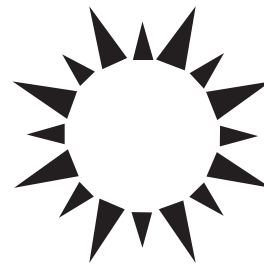
Further Reading

- Arrow, K. (1963). “Social Choice and Individual Values.” Yale University Press, New Haven, CT.
- Arrow, K., Solow, R., Portney, P. R., Leamer, E. E., Radner, R., and Schuman, H. (1993). Report of the NOAA Panel on Contingent Valuation. *Federal Register* 58, 4601–4614.
- Baumol, W. J., and Oates, W. E. (1988). “The Theory of Environmental Policy.” 2nd ed. Cambridge University Press, Cambridge, UK.
- Coase, R. H. (1960). The problem of social cost. *J. Law Econ.* 3, 1–44.
- European Commission. (1995). “ExternE: Externalities of Energy,” vols. 1–6. Office for Official Publications of the European Communities, Luxembourg.
- European Commission. (1999). “ExternE: Externalities of Energy,” vols. 7–10. Office for Official Publications of the European Communities, Luxembourg.
- Garrod, G. D., and Willis, K. G. (1999). “Economic Valuation and the Environment, Methods and Case Studies.” Edward Elgar, Cheltenham, UK.
- Kelly, J. (1978). “Arrow Impossibility Theorems.” Academic Press, New York.
- Krewitt, W. (2002). External costs of energy: Do the answers match the questions? Looking back at 10 years of ExternE. *Energy Policy* 30, 839–848.
- Oak Ridge National Laboratory and Resources for the Future. (1994–1998). “External Costs and Benefits of Fuel Cycles,” report 2-8. McGraw-Hill Utility Data Institute, Washington, DC.
- Pigou, A. C. (1924). “The Economics of Welfare.” Macmillan, London.
- Sagoff, M. (1988). “The Economy of the Earth.” Cambridge University Press, Cambridge, UK.
- Sen, A. K. (1970). “Collective Choice and Social Welfare.” Elsevier North-Holland, New York.
- Spash, C. L. (1997). Ethics and environmental attitudes with implications for economic valuation. *J. Environ. Mgmt.* 50, 403–416.
- Stirling, A. (1997). Limits to the value of external costs. *Energy Policy* 25, 517–540.
- Sundqvist, T. (2003). What causes the disparity of electricity externality estimates? *Energy Policy*, in press.
- Sundqvist, T., and Söderholm, P. (2002). Valuing the environmental impacts of electricity generation: A critical survey. *J. Energy Lit.* 2, 3–41.

Electric Power Measurements and Variables

MARIESA L. CROW and NIRUP SHETTY

University of Missouri, Rolla
Rolla, Missouri, United States



-
1. A Brief History of Electrification
 2. Electric Power
 3. Power Transformers
 4. Transmission System Components
 5. Substations
 6. Distribution Systems
 7. Power System Protection

Glossary

active power The part of the instantaneous power that can do real work. Real work is done when the resistive component of a load consumes electrical energy. Active power is measured in watts (W).

apparent power The total power (the complex vector sum of active and reactive power), measured in volt-amperes (VA).

bundling Transmission lines are bundled when there is more than one conductor per phase in extra high-voltage lines.

distribution The process of delivering electricity from substations to customers. The distribution system “steps down” power from higher voltage transmission lines to a level that can be used in homes and businesses.

grid The transmission network through which electricity moves from suppliers to customers.

load A device that uses the power supplied from the source for doing work.

reactive power The part of the instantaneous power that does no real work. It is absorbed by the inductive and/or capacitive components of the power system. It is measured in volt-amperes reactive (VAR).

short circuit A fault that occurs in an electric circuit or apparatus usually due to imperfect insulation causing a large current to flow through the system to ground. A short circuit has the potential of inflicting significant damage to the system.

substation The set of apparatus in a power system network that is used to change some characteristic, such as frequency or voltage, of power supplies.

transformer A device that is used to convert high alternating voltage to low alternating voltage and vice versa.

transmission The process that transports electric energy in bulk form, usually at high voltages, from a source of supply to a distribution system.

The modern electric power system is comprised of three primary entities: the generating system, which produces the electricity; the transmission system, which transfers the electricity over high-voltage transmission lines to regional substations; and the distribution system, which distributes the electricity to local customers. The transmission system consists of transmission lines and structures, substations, and transformers to convert the electricity and one voltage level to another. The substations also house a variety of equipment to protect against lightening, short circuits, and other faults that may occur in the system that adversely affect the distribution of the electricity to the customers. Electric power system customers range from residential home and small business owners to commercial customers and large industrial customers. Each type of customer impacts the transmission and distribution of electricity in different ways during daily and seasonal electric power variations.

1. A BRIEF HISTORY OF ELECTRIFICATION

The electrical utility industry is arguably the largest and most complex man-made system in the world and has been hailed as one of the greatest engineering achievements of the 20th century. The invention of the electric incandescent lamp by Thomas A. Edison

in 1878 marked the first step toward electrification. Edison is also credited with first proposing the idea of a centrally located power station for distributed lighting and developed the first electric network in the United States providing direct current (DC) power for lighting the lower Manhattan area of New York City. After the invention of the transformer, alternating current (AC) power became more popular than DC, with the first single-phase AC line being constructed in Oregon in 1889. The first three-phase system was installed by Southern California Edison at 2300 V in 1893. The development of electric power generation, transmission, and distribution has been increasing steadily ever since. From the construction of the first 220-V DC grids by Thomas Edison to the recent construction of AC grids with transmission voltages of 765,000 V, the drive to provide more reliable and efficient power has led to many significant developments in the field.

2. ELECTRIC POWER

The sinusoidal waveforms that comprise the voltages in most three-phase power transmission system typically have the form $v(t) = V \cos(\omega t)$ and the current has the form $i(t) = I \cos(\omega t + \phi)$, where ω is the system frequency (typically either 50 or 60 Hertz) and ϕ is the power factor angle. The angle ϕ is introduced due to the inherent inductive or capacitive nature of the transmission system and the system load, and it is the angle difference between the current and voltage waveforms. In an inductive circuit, the current lags the voltage, whereas in a capacitive circuit the current leads the voltage. In a purely inductive system, the current will lag the voltage by exactly 90° , whereas in a purely capacitive system the current will lead the voltage by exactly 90° . A purely resistive system will have the current and voltage in phase with each other. Most power systems are inductive due to the large reactance of the transmission lines. Figure 1 shows the phasor diagram and voltage and current waveforms of an inductive circuit in which the current lags the voltage by 30° .

For a circuit that has a combination of resistances and inductances, the product of the voltage (V) and the component of the current I that is in phase with the voltage ($I \cos \phi$) is the actual power that is used by the load. This quantity is known as the active power, P . The out-of-phase component ($VI \sin \phi$) is not used by the load directly but rather is responsible for the energization and deenergization of the reactive components in the circuit. This component is called

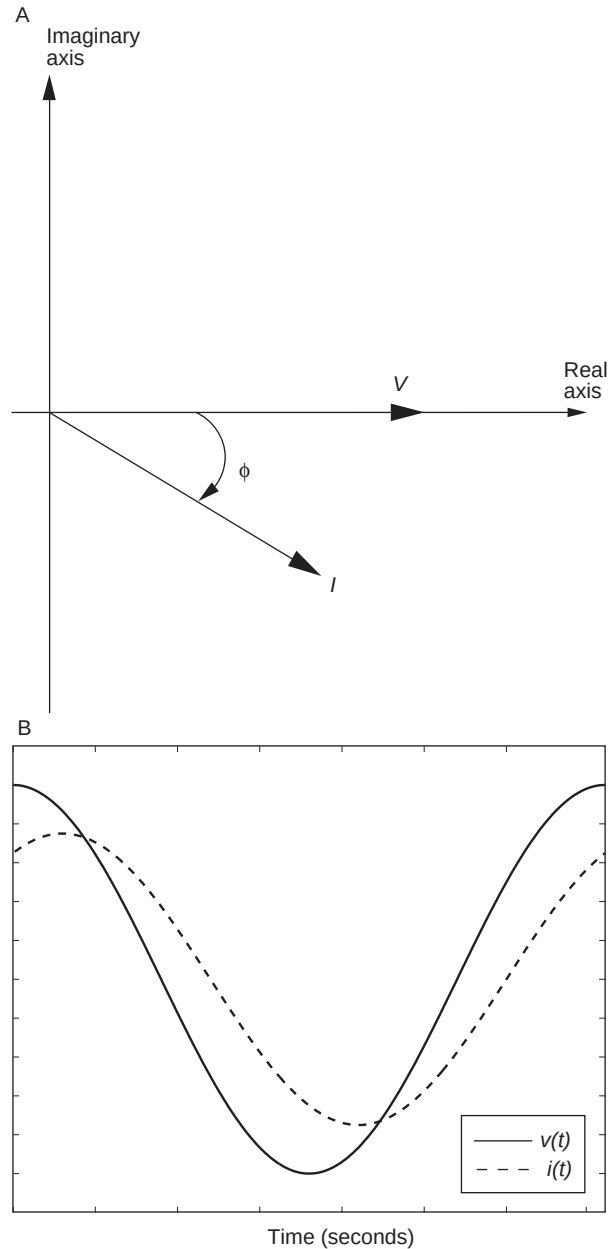


FIGURE 1 Inductive circuit (A) phasor diagram and (B) waveforms.

the reactive power, Q . Given the unique relationship between P and Q , the power in the AC circuit can be represented by $S = P + jQ$, where this complex sum S is called the apparent power with units of volt-amperes (VA). These relationships exist only in AC circuits. In a DC circuit there is no phase difference between the voltage and current; therefore, there is no reactive power and all the power generated is useful active power P . An inductive circuit will absorb both active and reactive power, whereas a

capacitive circuit will absorb active power but will inject reactive power.

2.1 Three-Phase Power

With the invention of the power transformer, AC systems became more popular than DC systems since the electric power could be transmitted across the transmission lines at higher voltage levels. Since power is the product of voltage and current, higher voltage levels require less current flow to transmit power. The reduced current subsequently reduces the active power losses in the line since the losses are given as I^2R , where R is the transmission line resistance.

Most electric power is transmitted in three phases, a–c. The three sets of sinusoidal voltages (and currents) are equal in magnitude but differ in phase angle of $\pm 120^\circ$ from each other. The voltages shown in Fig. 2 have three phases. By convention, the a phase is usually taken as the reference voltage. From these waveforms, it can be seen that the combination of three waveforms at any instant in time is identically zero. The same is true for the a–c phase currents. However, the product of these voltage and current waveforms yields the instantaneous electric power. If the three phases of electric power are combined, the resulting power is always constant at every instant under steady-state conditions even though the voltages and current are varying sinusoidally. This property of constant power is the primary reason that three-phase power is used. The same constant power result cannot be achieved by two-phase or four-phase power; thus, three-phase power is unique.

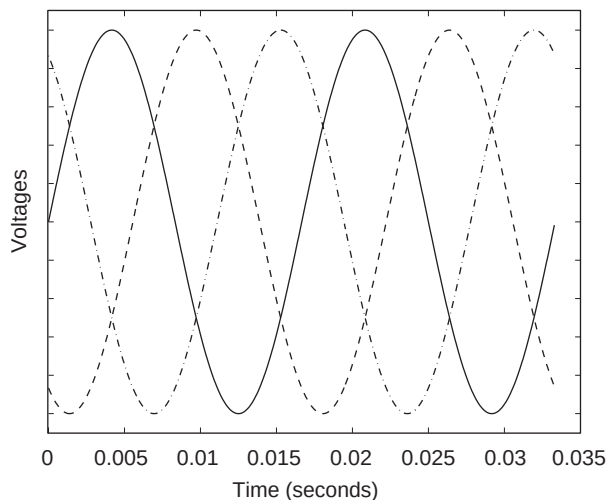


FIGURE 2 Power system voltages. Solid line, a phase; dashed line, b phase; and dashed–dotted line, c phase.

The second reason for using three-phase power is the ease of generation. A single-phase generator always experiences pulsating torque because the generated magnetic field does not rotate. The magnetic field simply changes direction at synchronous speed (i.e., it changes direction every $1/60$ of a second). Conversely, the magnetic field generated by three-phase windings smoothly rotates at synchronous speed and does not produce pulsating torque. Similarly, motor loads require three-phase power for smooth and efficient operation. The three-phase system is also capable of eliminating the third harmonic currents that are often present in the distribution system by the use of delta-wye transformers, which are unique to the three-phase system. This idea can be extended to a six-phase system; however, such a system would not be economical because of the cost of additional cables and insulation. Hence, the three-phase system is considered to be optimal for electric power transmission.

2.2 Power System Voltage Levels

The three-phase power is typically generated at a particular voltage and is then “stepped up” to a higher voltage using a three-phase transformer. The higher voltage is used for transmission (to reduce currents and thereby losses) and is then “stepped down” to lower values at the distribution substation. The voltage ratings of large generators employed at generating stations range from 13.8 to 24 kV. This is stepped up to transmission-level voltages of 69–765 kV using transformers. The evolution of voltages from 138 to 765 kV in the United States was a result of the developing transmission technology. These voltage increases met the needs of the larger and more economical generators begin developed. This evolution is shown in Fig. 3. Electricity is delivered from the transmission system to the customer’s home through a primary distribution system. This system is composed of a network of power lines delivering electricity at lower voltage. The voltage is stepped down again to the distribution system by a transformer. The voltages on a distribution level range from 220 V to 5 kV depending on the load connected. A typical power system configuration is shown in Fig. 4.

3. POWER TRANSFORMERS

A power transformer is used in a substation to step up or step down the incoming and outgoing voltages.

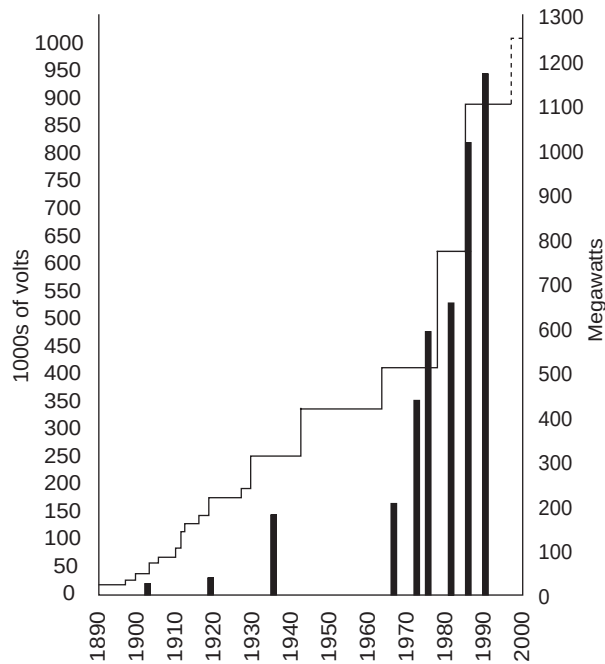


FIGURE 3 Evolution of maximum generating units and maximum transmission voltages.

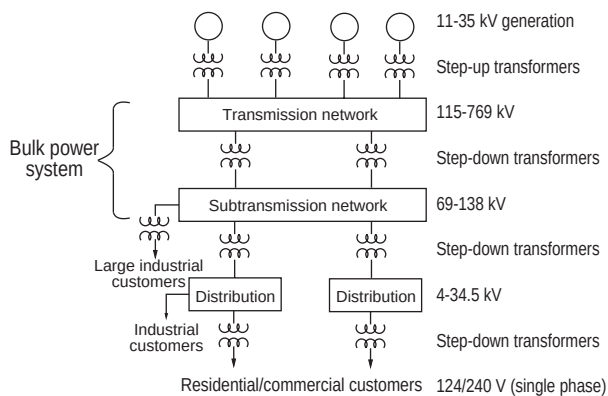


FIGURE 4 Power system configuration.

The transformers at distribution substations always step down the incoming transmission voltages to lower voltages used at the distribution level. Distribution lines emanate from substations at voltages of 4–25 kV. The power is then transmitted to a distribution transformer, where the voltage is further stepped down to the utilization voltage that ranges from 110 to 480 V for direct consumer usage.

When the transformer was invented by William Stanley in 1886, the early restrictions on the range and power levels of power systems were eliminated. In the DC system designed by Thomas Edison, the transmitted power required exceptionally high cur-

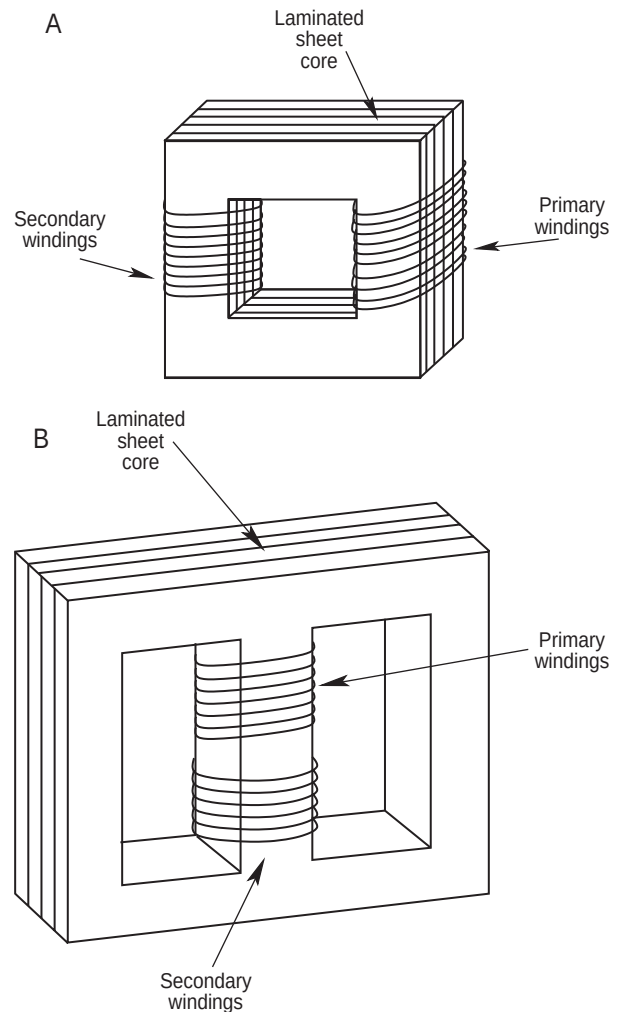


FIGURE 5 Single phase transformers: (A) core type and (B) shell type.

rent levels because the voltage was low. These large current levels caused large voltage drops and power losses in the transmission lines. The use of transformers greatly reduced voltage drops and power losses. The basic premise of an ideal transformer is that if the system voltage is increased, the current levels decrease proportionally to maintain a constant power output.

A transformer is typically constructed in one of two ways. The following basic configurations are shown in Fig. 5:

- Core type: This consists of a simple rectangular laminated piece of steel with the transformer winding wrapped on both sides. The winding connected to the source is called the primary winding, and that which is connected to the impedance or load is known as secondary winding.

- **Shell type:** This type of construction consists of a three-legged core with the windings wrapped around the center leg.

In both types of transformers, the core is constructed of thin laminations electrically isolated from each other in order to minimize eddy currents or those currents that circulate within the core material and cause heating losses.

The transformer works on the principle of magnetic induction. The time-varying voltage, V_p , that is applied to the input (primary) side produces a time-varying flux in the core. This changing flux couples with the secondary winding and induces a secondary voltage, V_s , in proportion to the ratio of the windings. The relationship between the voltages at both the sides of an ideal transformer is given by

$$V_p/V_s = N_p/N_s,$$

where N_p and N_s are the number of turns of the primary and secondary windings, respectively. The relationship between the current flowing in both sides of the ideal transformer is given by

$$N_p I_p = N_s I_s,$$

where I_p and I_s are the primary and secondary currents, respectively. If the transformer is lossless, then the input power is equal to the output power and the voltages and currents are related by $V_p/V_s = I_s/I_p$. Thus, as the voltage increases from secondary to primary, the current decreases; therefore, the current is lower at higher voltages.

In modern power systems, transformers can be found everywhere. Transformers have a variety of names, depending on the purpose for which they are used in a power system. A transformer connected to the output of the generator and used to step up its voltage to transmission levels is called a unit transformer. The transformers that bring down the transmission voltage levels to distribution levels are known as substation transformers. Lastly, the transformers that bring down the distribution-level voltage to the voltage level made available to the customers are known as distribution transformers.

Transformers are typically placed in a variety of locations. Many transformers are overhead transformers hung on poles with direct access to the wires. Underground transformers are typically placed in underground vaults to protect them from public access and environmental disturbances. The most common type of transformer is the pad-mounted transformer, which typically sits on a concrete pad at ground level. Pad mounts are designed for contact with the general public and therefore must adhere to

strict American National Standards Institute (ANSI) standards. The last type of transformer is the interior transformer, which is designed to be placed in a building, usually at ground level. Due to concerns about the oil used for coolant, most interior transformers are cooled with a less flammable coolant such as gas and are frequently referred to as dry-type transformers.

A unique set of transformers are the under-load-tap-changing (ULTC) transformers. Since the voltage along a distribution line typically decreases with an increase in load being drawn by customers, it is desirable to regulate the voltage magnitude at the substation. The most economical method of voltage regulation is with an ULTC transformer. The ULTC transformer is built with several tap settings that allow the number of secondary windings to be adjusted based on the desired voltage level. As the ratio of primary to secondary windings varies, the voltage ratio on the primary and secondary side also changes. This changing of the tap setting enables the secondary voltage to be changed in response to changes in load. These taps are fully automatic and operate in response to voltage on the secondary side of the transformer.

4. TRANSMISSION SYSTEM COMPONENTS

An overhead transmission line consists of conductors, insulators, support structures, and, in most cases, shielding wires.

4.1 Power Cables

Copper has been replaced by aluminum for construction of overhead lines due to the availability and cost factor. The most commonly used conductor is aluminum conductor-steel reinforced (ACSR), which has a central core of galvanized steel wire. Surrounding this central core are strands of aluminum, which have good conductivity. The steel strands provide strength and the aluminum conducts the current. This type of conductor is shown in Fig. 6. The other conductors in common use are all-aluminum conductor, all-aluminum alloy, aluminum-clad steel conductor, aluminum-clad steel conductor (Alumoweld), and extra high-strength steel conductor. Fillers such as fiber or paper can be added between the aluminum and steel strands to increase the conductor spacing to reduce the electric field at the conductor surface.

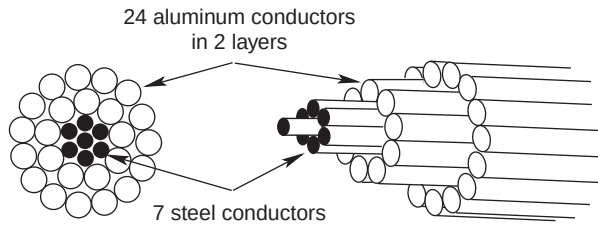


FIGURE 6 ACSR 24/7 conductor.

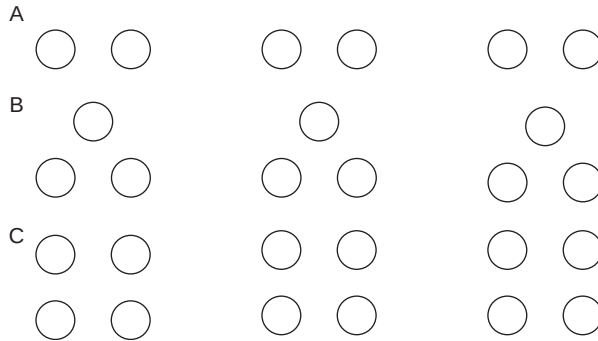


FIGURE 7 Typical bundle configurations.

Extra high-voltage lines often have more than one conductor per phase. This configuration is commonly referred as conductor bundling. Line bundling allows transmission of large amounts of power over long distances. A bundle consists of two–four conductors. Bundling increases the effective radius of the conductor and reduces the electric field strength near the conductors, which reduces corona power loss, audible noise, and radio interference. Line bundling also reduces line reactance, allowing greater current carrying capacity. Figure 7 shows typical bundle configurations.

4.2 Shielding Wires

Shielding wires are located over the load-carrying conductors on transmission lines to prevent a direct lightning strike. Almost all lines above 34.5 kV employ shielding wires. The shield wire provides a 30° protection on either side of the wires. The number and location of the shielding wires are selected such that a lightning strike will terminate on the shielding wires and not on the phase wires. Shielding wires are essentially grounded neutral wires located above the phase wires. They bypass a lightning strike harmlessly to the ground, provided that the tower impedance and tower footing resistance are small.

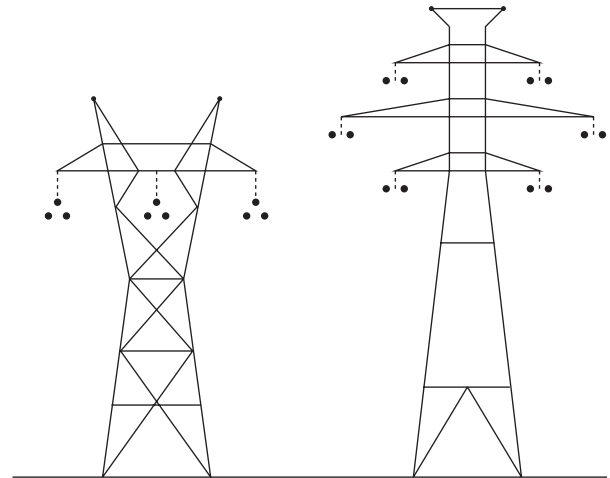


FIGURE 8 Typical lattice structures for 345-kV transmission lines.

4.3 Transmission Line Structures

The primary function of a transmission line structure is to support the transmission lines. The design of the transmission line structure must consider environmental conditions, area requirements, voltage levels, and right-of-way (ROW) requirements. The ROW is the amount of clearance required on either side of the transmission line for safety purposes. Double-circuit 345-kV lines usually have self-supporting steel towers with the phases arranged in a triangular configuration to reduce tower height or in a vertical configuration to reduce tower width. For transmission of voltages of 345 kV and less, wood-frame configurations are commonly used. Typical lattice-type structures for a 345-kV transmission line are shown in Fig. 8. Single-pole, tubular-steel structures are also popular for 345-kV transmission lines. These structures are preferred over the less expensive lattice-type structures when the ROW is restricted. Tubular-steel structures are typically more expensive than lattice-type structures, but lower maintenance requirements may make them competitive in situations in which access is difficult.

4.4 Insulators

Overhead line conductors must be supported on poles or towers such that currents from the conductors do not flow to the ground through the supports. This requires proper insulation of the lines from the supports, which is achieved by securing the line conductors to the supports by insulators. Insulators for transmission lines higher than 69 kV are suspension-type insulators that consist of a string

of discs, typically constructed of porcelain. All the discs are connected in series by metal links in the form of a string. The conductor is suspended to the bottom end of this string. The other end of the string is secured to the cross-arm of the tower. Each disc can only withstand a voltage of approximately 16 kV; therefore, the number of discs per insulator is proportional to the voltage rating of the line. A 120-kV insulator has 6–8 units, whereas a 500-kV insulator is constructed of 26–32 units.

Another type of insulator in operation is the post-type insulator, which is housed in substations to support conductors, busbars, and other equipment. This type of insulator is used for medium- and high-voltage transmission lines in which the insulator replaces the cross-arms of the transmission tower.

Composite insulators are becoming more common as an alternative to porcelain. Composite insulators are constructed around a fiberglass rod, which gives it high mechanical strength. A forged-steel or iron fitting is then compressed to the rod to provide a slippage-free end fitting.

5. SUBSTATIONS

Electrical power is produced at power-generating stations, which are located at favorable but typically geographically remote locations. The power is delivered to the customers at homes or offices through a large network of interconnecting transmission and distribution lines. At many places in the power system, it is necessary to alter some of the electric power characteristics, such as voltage or current. For example, the transmission-level voltages are reduced to distribution-level voltages for the purpose of supplying it to the customers. The place where this conversion of transmission to distribution takes place is known as a substation. Similarly, the generating voltages are also stepped up to transmission levels at substations near the generating unit. A typical distribution substation contains equipment discussed in the following sections.

5.1 Busbar

The busbar, or bus, is used to distribute the power in different directions. Many feeders come out of the substation because they have to cater to different customers in various geographical locations. The bus is a node in the power system circuit. In a power system, when two or more lines are attached to a common point, that point is called a bus. Physically,

busbars are copper rods or thin-walled tubes at a common voltage in a substation to which two or more lines (feeders) are connected. If a number of generators or feeders at the same voltage are directly connected at a point, busbars are used as the common electrical component.

5.2 Isolating Switches and Circuit Breakers

In substations, it is often desired to disconnect a part of the system for general maintenance and repairs. This is accomplished with the help of isolators or isolating switches. A circuit breaker is a device that can open or close a circuit under normal as well as fault conditions. It is designed so that it can be operated manually under normal conditions and automatically under fault conditions.

5.3 Converter Stations

Converter stations are present only in high-voltage DC substations. The components are power electronic-based inverters or rectifiers, which are needed to convert the incoming AC components to DC and vice versa. If a DC link exists between two substations, then the voltage has to be converted at the first substation and back to DC by inverters at the second substation.

6. DISTRIBUTION SYSTEMS

The distribution system is the primary means of distributing the electric power from the substation to the system customers. A typical distribution substation serves 1–10 feeder circuits. Each feeder circuit may serve a variety of load types.

6.1 Power System Loads

A load is any device that draws electrical energy from a power system. A load can be resistive, inductive, capacitive, or a combination of these characteristics. Loads in a typical power system are classified as residential, commercial, and industrial. Of these, residential loads are mostly single-phase loads, whereas the other two types are generally three-phase loads. Residential loads have lower voltage levels and include lighting, air-conditioning, heating, and other household appliances. They are usually rated below 2 kW. On the other hand, industrial

loads have the highest voltage levels. Industrial loads include high-power motors used for drilling, pumping, grinding, or other manufacturing purposes and furnaces. Industrial loads are generally rated from 25 kW for small-scale industry to 500 kW or higher for large-scale industry. Lighting, heating, or cooling of large commercial areas such as shopping malls are commercial loads. Figure 9 shows the utility sales by sector for the United States from 1990 to 2001.

The magnitude of a load may vary continuously throughout the day, week, season, and year. Therefore, loads may also be differentiated in terms of time as opposed to end-use sector:

- **Base load:** The minimum load that has to be served throughout the day. This load is typically served by the largest and/or most inexpensive generators.
- **Peak load:** The maximum load that has to be served by the utility during the time period of interest. For example, industrial loads usually peak once during each day, but the summer peak load occurs just one day per year.
- **Off-peak load:** Any load other than peak or base loads is referred to as an off-peak load.

7. POWER SYSTEM PROTECTION

A fail-free power system is not technically feasible since faults occur in a power system at regular intervals for various reasons. Short circuits occur in power systems when equipment insulation fails due to system overvoltages caused by lightning or

switching surges. Faults can occur in any power system component, including generators, transformers, buses, or lines. The most prone to faults are the transmission lines since they are the component most exposed to the environment. Typical transmission line faults that can occur are insulation degradation and short circuits caused by contact with tree branches, lightning, and grounding failures. These faults can cause sustained or momentary interruption of power to the customer and may cause considerable economic losses. Approximately 80% of the faults that occur in the power system are temporary in nature. Proper protection schemes can ensure that reliable power can still be supplied to the customer.

ANSI standards require that faults should be cleared within 3 cycles in extra high-voltage lines and in 5–20 cycles in lower voltage equipment. Faults that occur on a power system are cleared by placing protection equipment throughout the line.

7.1 Circuit Breakers

Circuit breakers are a self-controlled device for automatically interrupting and reclosing an AC circuit with a preset sequence of opening and reclosures. Circuit breakers do not have built-in controls, and they open or reclose based on relay operation. Relays sense the current or voltage and energize the trip coil of the circuit breaker, which in turn causes the poles to separate. As the poles separate, an electric arc is induced across the intervening gap. Different quenching mediums are used to interrupt the arc, thereby opening the circuit:

- **Air-break circuit breaker:** This type of circuit breaker is generally the indoor type and is installed on vertical panels. They are used for DC circuits up to 12 kV and for medium- and low-voltage AC circuits, usually up to 6.6 kV. In this case, the arc is lengthened in the breaker by a magnetic field and is extinguished by arc splitters.
- **Oil circuit breaker:** Oil is employed as the quenching medium in this family of breakers. Early bulk-oil breakers used a large oil tank for quenching the arc. However, modern practice is to use minimum-oil circuit breakers in which the arc is extinguished in a small arc control device filled with oil. The voltage rating for this type of breaker is 3.6–245 kV. The disadvantage of this type of breaker is that since oil is hygroscopic, it must be sealed in an air-tight chamber.
- **Vacuum circuit breaker:** The principle for this type of circuit breaker is that a vacuum is a dielectric

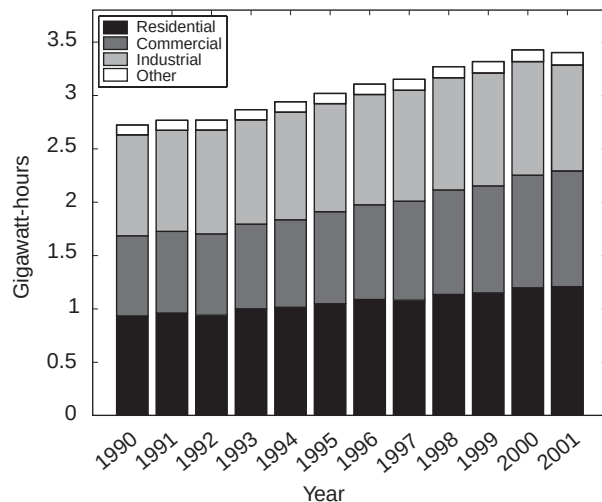


FIGURE 9 U.S. electric utility retail sales of electricity by sector, 1990–2001.

medium and an arc cannot persist in an ideal vacuum. The arc is generally extinguished in half a cycle. Vacuum circuit breakers are generally employed up to 72.5 kV and are smaller than oil circuit breakers.

- **Air-blast circuit breaker:** To extinguish the arc in an air-blast circuit breaker, a blast of air is blown at the current zero-crossing. These circuit breakers can be employed up to a voltage of 275 kV.

- **SF-6 circuit breaker:** SF-6 is a heavy, chemically inert, nontoxic, nonflammable gas. Breakers using SF-6 are popular and are used for voltages of 6.6–765 kV.

7.2 Relays

Relays are devices whose primary function is to detect faults and take proper control action to clear the faults. Whenever the relay senses a faulty condition in the power system, it energizes the trip coil of the circuit breaker to clear the fault. Relays and circuit breakers comprise the protection system of the power system. The following are the different types of relays:

- **Overcurrent relays:** This type of relay operates when the current through it exceeds the pickup value of the relay. For currents lower than this pickup value, the circuit breaker is not tripped. For system-wide coordination of overcurrent relays, it is essential that the relays conform to an inverse-time characteristic in which the operating time (time to trip for a given current) decreases as the relay current increases. Such relays are known as inverse definite minimum time lag relays.

- **Directional relays:** Directional relays are designed to operate for fault currents flowing in only one direction. When two or more sources are present at different locations, it is frequently difficult to coordinate overcurrent relays. In such cases, directional relays can be used.

- **Impedance relays:** Impedance relays operate on the basis of the voltage-to-current, or impedance, ratio. The impedance relay is also called a distance relay since the line impedance is proportional to distance. The relay trips when the measured impedance $|Z|$ falls below $|Z_r|$, where $|Z_r|$ is an adjustable relay setting and $|Z|$ is the voltage-to-current ratio at the relay location.

- **Differential relays:** A differential relay is a suitably connected overcurrent relay operating on the difference between the currents at the two ends of the protected element. It does not operate for faults

outside of the equipment. Only for an internal fault will a difference of currents occur, causing the trip coil to be energized. Differential relays are commonly used to protect generators and transformers.

Relays were introduced in the early 1900s and were electromechanical in nature. They are relatively maintenance free and are still found in operation today. Solid-state relays using analog circuits and logic gates provided the same reliability as electromechanical relays but at a more competitive price. The latest technological devices are digital relays that allow the relay characteristics to be changed from a remote location in real time. The principles of digital relays are the same as those of previous electromechanical relay, but these relays have the advantage that a central computer can coordinate between the different local digital relays.

7.3 Fuses

The simplest overcurrent protective element in a power system network is a fuse. Fuses are generally used in large numbers by utility companies because they are inexpensive and maintenance free. The fuse consists of a metal “fusible” link or links encapsulated in a tube, packed in a filler material such as sand. The metal link is generally silver, and it is connected to the contact terminals. If the current flowing through the fuse is normal, the fuse acts as a conductor. However, when abnormal current levels are reached, the link melts. Most of the fuses are current limiting in nature (i.e., the fuse will melt in less than half a cycle for high fault current). In the event of a fault, only the fuse closest to the fault should interrupt and open the circuit immediately downstream. The fuse with the highest continuous current rating is placed closest to the source in the case of a radial feeder. The primary disadvantage of fuses is that they operate only once and cannot operate again without manually replacing the fuse link.

7.4 Circuit Recloser

An automatic circuit recloser is an overcurrent protective device that automatically trips and recloses a preset number of times to clear temporary faults or to isolate permanent faults. It also has provisions for manual opening and reclosing. A recloser unit has a built-in relay and a circuit breaker. Most temporary faults are cleared in one or two recloser operations. Since 80% of the faults that

occur on the power system are temporary in nature, the presence of a recloser helps eliminate the unnecessary tripping of loads for temporary faults. The recloser and the fuses must be coordinated so that only the faulted section is affected.

7.5 Sectionalizer

A sectionalizer is a self-opening device that is used in conjunction with the source-side recloser or circuit breaker to automatically isolate faults. When the source-side protective device operates to deenergize the circuit, the sectionalizer will count the number of operations. After a set number of operations, the sectionalizer permanently opens, thus limiting the outage to the faulted section only. Sectionalizers isolate permanent faults to smaller sections of the line, which in turn improves the reliability of the system. Another advantage is that the sectionalizer does not have a time-current curve and therefore does not have to be coordinated. It can be placed between existing reclosers and fuses without changing its setting.

7.6 Lightning Arrestors

A lightning arrestor, or surge arrestor, is a protective device that conducts the high-voltage surges on the power system to the ground. One end of the arrestor is attached to the terminal of the equipment and the other end is effectively grounded. It consists of a spark gap in series with a nonlinear resistor. The length of the gap is set such that normal line voltage

is not enough to cause an arc across the gap, but a dangerously high voltage will break down the air insulation and form an arc. The property of the nonlinear resistance is such that its resistance decreases as the voltage or current increases and vice versa.

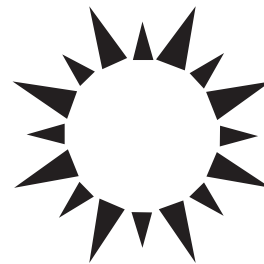
SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

- Blackburn, J. (1997). "Protective Relaying: Principles and Applications." Dekker, New York.
- Grainger, J., and Stevenson, W. (1994). "Power System Analysis." McGraw-Hill, New York.
- Grigsby, L. (2000). "The Electric Power Engineering Handbook." CRC Press, Boca Raton, FL.
- Kersting, W. (2001). "Distribution System Modeling and Analysis." CRC Press, Boca Raton, FL.
- Winders, J. (2002). "Power Transformers: Principles and Applications." Dekker, New York.

Electric Power Reform: Social and Environmental Issues



NAVROZ K. DUBASH

World Resources Institute

Washington, D.C., United States

-
1. Introduction
 2. Reasons for Reform of the Electric Power Sector
 3. Social Concerns in Electric Power Reform
 4. Environmental Concerns in Electric Power Reform
 5. Governance of Electric Power Reform

Glossary

cross-subsidy A subsidy to one set of consumers of a service to reduce the cost of providing access or the cost of consumption, paid for by charging other consumers more.

electricity service concession A contract to a private sector firm to provide electricity in a defined area under a set of rules and incentives established by the government.

end-use energy efficiency programs Efforts at increasing the efficiency of the energy system by focusing on changing use patterns of and technologies deployed by the final consumer.

environmental dispatch The dispatch of electricity from electricity generators based on a pollution index that allows adjustment of both the order of dispatch and the price to internalize pollution costs.

integrated resource planning A planning framework that allows for comparison of the costs and benefits of alternative electricity supply mechanisms—including demand-side savings and nontraditional electricity sources such as renewable sources—across an electricity system. This framework should include consideration of environmental and social externalities.

lifeline rates Reduced rates, often financed by cross-subsidies, which allow poor households that use minimal amounts of a service to pay less than wealthier households that use more of those services.

universal service obligation A commitment by service providers to provide electricity services to those who request it within their service area, often accompanied by a commitment to provide this service at an affordable price.

The issue of equity and environment in electric power reform is rooted in a larger debate over the nature of the electricity sector. Is electricity a commodity like any other? Should the production, transmission, and distribution of electricity be managed largely by the private sector? Or is electricity an instrument of policy aimed at sustainable development, over which governments should retain policy control, if not ownership? In furthering the debate, a position that has found resonance is that although provision of electricity services may indeed be benefited by some commercial role, electricity provision also has strong positive and negative externalities that call for proactive government involvement.

1. INTRODUCTION

Electric power reform commonly refers to changes in ownership and structure of an electricity system, with the purpose of introducing private ownership and competition. Discussions over how to implement these reforms have focused on improving the technological and financial performance of the sector. Scholars and policy makers have also recognized that power sector reform can have implications for social and environmental outcomes, and that these concerns are not captured by a focus on technical and economic concerns alone. Prominent among relevant social outcomes are the extent to which members of a society have access to electricity, the affordability of electricity, the impact on labor employed in the sector, and the reliability of electricity supply. Relevant among environmental outcomes are the impacts of sector reforms on global, regional, and local pollution caused by production of electricity. Important factors shaping environmental outcomes include the impact of

reforms on consumption levels, technology change, efficiency of supply and demand, and regulatory practice. Finally, an important factor in determining equity and environmental outcomes of electric power reform is the quality of governance in the sector.

2. REASONS FOR REFORM OF THE ELECTRIC POWER SECTOR

2.1 Electric Power Reform Defined

As described in a widely used formulation by authors Sally Hunt and Graham Shuttleworth, electric power reform captures changes in the electricity sector along two dimensions: ownership/control and market structure. Changes in control and ownership can be further broken into three categories:

1. Commercialization. A government surrenders control over a state-owned enterprise in order to promote operation along commercial lines. However, ownership remains in government hands, with no change in organizational form.

2. Corporatization. A government formally and legally relinquishes control and management over a public enterprise to create a corporation. Management is independent of the government, although ownership remains in government hands, and the corporation may be subject to regulatory oversight.

3. Privatization. A government sells a corporation to private owners. The new owners' primary accountability is to investors, although the corporation may be subject to regulatory oversight.

Changes in market structure to promote competition can be subdivided into four steps:

1. Monopoly. No competition at any level. A single entity, usually government owned and operated, handles generation and transmission of electricity. This company provides electricity to distribution companies, which have a monopoly relationship with the final consumer.

3. Single buyer. Competition in generation. Public or private power generators sell power to a single purchasing agency. The purchasing agency transmits power to distribution companies, which have a monopoly relationship with the final consumer.

3. Wholesale competition. Competing generating companies sell directly to distribution companies, and also to very large consumers. Generators have open access to the transmission grid in order to deliver power to their consumers. With the exception

of large consumers, distribution companies have a monopoly over final consumers.

4. Retail competition. Generating companies sell directly to distributing companies and final consumers. Generators have open access to both transmission and distribution grids. Final consumers can choose their supplier.

This simple framework conceals considerable discussion and debate over the details of creating markets, and the appropriate sequencing of these various stages.

2.2 Drivers of Electric Power Reform

Historically, electric power systems have been publicly owned and operated because of their natural monopoly characteristics. Governments were better able to exploit the economies of scale in power generation, raise the large amounts of capital required, bear the long time periods for returns on investment, and address the networked nature of the system. Over time, electric power became part of a set of essential services, provision of which was associated with governments. In much of the industrialized world, this approach succeeded in providing reliable and affordable power to citizens, although in the developing world success was far more modest. However, it was in industrialized countries that this approach to power provision was first questioned. At least three factors spurred the questioning, albeit with different emphases in different countries. First, demand for electricity began to slow, undermining the predictable sources of growing income on which utility financing had been based. Second, new generation technologies, particularly gas turbines, reversed the scale economies in the sector. Third, costs and risks increased in the sector due to a rising environmental consciousness, a corresponding increase in regulations, and burdensome investments in high-capital-cost nuclear power plants.

Countries began to act on these changed realities. In the 1970s, the United States allowed independent power producers to sell electricity to utilities, undermining the idea that electricity was best run as a natural monopoly. In the 1980s, the United Kingdom and Chile remade their electricity sectors around the objective of promoting competition. In particular, they unbundled the component parts of the electricity system—generation, transmission, and distribution—and offered components for sale to the private sector. Around the same time, Norway also introduced competition, but without an

accompanying process of privatization. The rapid growth and declining costs of information technology facilitated these changes by enabling the development of new control techniques consistent with greater decentralization of the grid.

Advocates of market-based reforms suggest that reforms bring greater incentives for cost-saving and productivity enhancement. Specifically, it is argued that market reforms provide appropriate price signals to consumers (which spurs more efficient economy-wide resources allocation), reward the efficient use of inputs with profit, and force cost cutting through competitive pressures. Critics point to several difficulties with creating a well-functioning market in electricity. These include the difficulties of defending against market power, the higher cost of capital faced by the private sector, which can translate to higher average costs in a capital-intensive sector such as electricity, and the difficulty of planning, particularly for transmission, in a decentralized system when information is hidden and incomplete. In addition, critics challenge the emphasis on reforms aimed at facilitating private participation, and question whether a change in ownership is a critical element in solving the sector's problems.

It is beyond the scope here to settle these debates conclusively, but the empirical record provides a few lessons of relevance to this discussion. Designing and implementing a market for electricity is a complicated affair, and it is certainly possible to make an electricity system worse off through reforms, as the disastrous California experience showed only too well. Where electricity reforms have resulted in some efficiency gains, competition is at least as important as privatization in achieving these gains. Where efficiency gains have led to lower electricity prices, such as in the United Kingdom and Europe, benefits have gone disproportionately to industrial users. Finally, controlling price volatility has been a considerable challenge, in some cases even threatening the reliability of the system. Despite these continued uncertainties, electricity reform has proceeded apace worldwide, including in developing countries. Reforms in developing countries have been driven by financial and economic concerns such as the poor performance of public utilities, a shortage of capital for investment, a broader fiscal crisis due in part to electricity subsidies, and a short-term need to raise capital through the sale of electricity assets. In many cases, countries have embarked on a reform trajectory at the urging of international financial institutions, and as part of a larger package of economic reforms.

The application of a uniform reform approach as just described to countries in the developing world has led to criticism. One prominent critic, A. K. N. Reddy, points out that the approach followed in the developing world was developed for quite a different purpose—to exploit opportunities for efficiency gains in mature electricity systems in industrialized countries. Existing problems—technological and financial performance crises, the failure to provide access to electricity to large portions of the population, and the lack of attention to the environmental trajectory of the sector—find little explicit attention in the reform approach that has been carried out to date. This observation is an important reason for seeking a broader and more context-specific set of design parameters for electric power reform in the developing world.

3. SOCIAL CONCERNS IN ELECTRIC POWER REFORM

Electricity services enable many positive economic and social benefits. Electricity is a key infrastructural element for industrialized growth and for irrigation. At a household level, electricity can have positive impacts on education, due to better lighting; on time and effort savings, by substituting for traditional fuels used for cooking and heating; on health, by reducing indoor air pollution and through use of refrigeration for foods; and on entrepreneurial activity, through better information and digital connectivity. To achieve these gains, citizens must have access to electricity, and this access must be affordable and reliable. In addition, the electricity sector is also a large employer in many countries, which has provoked concerns among labor groups about the impacts of reforms. The discussion here focuses on access, affordability, and labor concerns. Although reliability is also an important concern, it is already an important and explicit goal of the reform agenda, and so is not treated further here.

3.1 Access to Electricity

Worldwide, rates of access to electricity remain low. Two billion people around the world are commonly estimated to lack electricity. The World Energy Assessment states that in rural areas the situation is particularly bad; 56% of the world's rural population has to manage without electricity. The situation is particularly dire in Sub-Saharan Africa, where the World Bank found that in 1990, 62% of the urban

population and 92% of the rural population rely entirely on energy sources other than electricity.

Historically, many public utilities have undertaken grid extension by setting targets, and financing investment through levies or cross-subsidies. For example, South Africa's public utility, Eskom, exceeded its target of 350,000 new connections a year through much of the 1990s, financed by a cross-subsidy from existing users. South Africa, however, is an exception, and the experience with target setting for public utilities has been disappointing in much of the world. There are several reasons for this failure to provide access to electricity. The rural poor are particularly costly to serve because of remote locations, low population density, high transmission line losses, low consumption levels (which raise average costs), poor credit and minimal collateral, and a lack of purchasing and political power. Many of these factors, other than geographically determined factors, also apply to poor urban and periurban populations. These problematic conditions have been compounded by institutional failure at public utilities. Uncollected levies, the failure to apply least-cost electrification approaches, and the use of electricity as an instrument of political patronage have further stymied efforts at electrification.

The proposed new model of private participation in a market-oriented sector may be no better placed to solve this problem than the public utility approach. In a market model driven by profitability, interest in serving the rural or urban poor is likely to be muted in the absence of specific incentives provided by public policy. Research conducted at the World Resources Institute suggests that through the 1990s, access to electricity was rarely a priority in developing country reform efforts. Early indications suggest that the outcome is indeed minimal progress toward increased access to electricity. For example, a review of reforms in Bolivia concluded that expansion of the grid will not take place as a consequence of privatization and restructuring. In addition, the private sector may choose to disconnect nonpaying customers that the public utility had continued to supply as an implicit form of social welfare. For example, in Argentina, reforms led to a measure of deelectrification of the poor as the privatized utility clamped down on nonpayment and informal connections.

As it has become clear that additional policy incentives are necessary if increasing access to electricity is to remain a priority, more attention has been paid to ways of integrating access into reform approaches. In a market-based electricity sector,

two possible entry points for access concerns are (1) the conditions of privatization and (2) concession agreements with the private sector for provision of electricity.

3.1.1 Institutional Mechanisms for Promoting Enhanced Electricity Access

Governments may place obligations on private providers of electricity, as a condition of participation in the market. Such an obligation can take various forms. One approach is to specify an "obligatory service" commitment, which requires a company to provide access to its services to all who wish to avail themselves of services at the prevailing tariff. However, many consumers may not be able to afford consumption at the prevailing tariff. A "universal service obligation" addresses this issue by requiring a company to provide a specified level of service to all customers at an affordable tariff. Because this tariff is necessarily below the price the company would otherwise seek to charge, establishment of a universal service obligation also entails finding financing to support this obligation. This can be arranged through three approaches: a fund collected through a special charge on consumers, which is then paid to the service-providing company; by allowing cross-subsidies; or by defining the service area in such a way that profitable and unprofitable areas are bundled together.

For populations beyond the scope of the existing grid, however, more explicit and targeted programs of access are necessary. Electricity concessions provide one such approach that is potentially compatible with a postreform electricity sector. The central idea is that the private sector is awarded a contract to provide electricity in a defined area under a set of rules and incentives, including financial incentives, established by the government. The desired end is access to electricity, and not necessarily grid extension. Because generation closer to the point of consumption will obviate the need for investment in transmission lines, the concession approach provides an incentive to seek out the lowest cost technology for providing access. Consequently, distributed electricity using technologies such as small hydropower, wind, solar, and biomass may prove to be cheaper and more practical options for providing electricity, and may be adopted by concessionaires.

There are several variants on this approach, which vary in their method of subsidy allocation and degree of private participation. For example, in Argentina, the government has sought to maximize cost-effectiveness by awarding concessions to the company

that requires the lowest subsidy to provide service to a concession area. In Chile, private distribution companies, acting in partnership with municipalities or rural communities, submit project proposals to receive subsidies to cover investment, but not operating, costs. Only projects that have a negative private return but a positive social return are eligible. The state agency evaluates these projects on the basis of cost-benefit analysis, financial contribution by communities and the company, and the degree of social impact. In another example, electrification in Morocco is promoted by a combination of a public and private approach. In some cases, the national electric utility maintains full control over installation, maintenance, and collection of fees; in other cases, it contracts private retailers to install distributed power systems in return for a share of fees. In addition, the utility requires municipalities and households to cofinance 45% of the investment costs of electrification.

The discussion here so far has focused on institutional mechanisms for involving the private sector in electricity provision. However, the same mechanisms can work effectively for nonprofit entities, such as community groups or cooperatives. There are precedents for electrification organized around cooperatives, most notably in the United States. Starting in 1935, electrification in the United States was built around rural electricity cooperatives financed by low-interest public funding. These cooperatives applied a principle of universal coverage and drew on standardized engineering across the country to reduce costs. Cooperatives and smaller scale entities also benefit from greater knowledge of local circumstances, which can enable them to better adapt technologies and more effectively fulfill difficult tasks such as bill collection.

3.1.2 *Financing Access*

It is significant that the approaches described in the preceding sections do not lift from the state the burden of raising funds, but rather provide a new way of channeling those funds, and potentially leveraging additional funds. Historically, electrification has been financed through cross-subsidies—charging some consumers more to subsidize investment in access for others. In the drive to introduce market reforms, cross-subsidies are increasingly frowned on as introducing distorted price signals leading to suboptimal patterns of use. More transparent, but also distorting, are explicit levies on some consumers to be used for electrification. The use of levies requires sufficient state capacity to effectively admin-

ister such a fund. A third approach, and one most favored by market-oriented reform advocates, is an explicit allocation from the central budget for electricity access. However, aspirations to transparent and explicit subsidy allocation ignore the political reality in many countries where access is a concern. Where budget allocation is determined by a complex of national and subnational political interests, it is far from clear that forcing subsidy allocation to pass through a national governmental process will yield a satisfactory allocation of funds for promoting electricity access. In general, the choice of financing mechanism should be determined not only by considerations of economic efficiency, but also by considerations of government fiscal and administrative capacity and the likely impact of local politics.

3.1.3 *Local-Level Competition for Electricity Provision*

The approaches described so far are all built around monopoly provision of electricity at some local scale, even if the selection of the monopoly provider is on a competitive basis. An alternative approach would allow free entry to the market for electricity services. Advocates of competition at the local level suggest that the result will be increased efficiency of service provision and incentives for innovation, with the result that the need for subsidies may be reduced. To realize this vision, arrangement for exclusive provision will have to be opened to all providers, and regulations will have to be designed minimally in order to reduce entry costs. However, open entry and competition for provision of services bring potential problems. First, there is a strong link between market structure and financing of electrification. Where there is an existing service provider, competition could allow an entrant to “cherry pick” rich customers or rich areas, leaving only the poorer customers. This problem is particularly pernicious where services to poor customers are financed through cross-subsidies, or where the incumbent provider has a universal service obligation. In areas where there are no existing providers and where subsidies are necessary to lure entrants, the burden of administering subsidies increases considerably with multiple players. This is particularly true if provision of subsidies is dependent on performance criteria, such as number of customers served. Second, in the absence of adequate regulation, there is scope for formation of cartels and price gouging of customers. Finally, there are basic issues of safety in electricity provision that are left unaddressed in an unregulated or lightly regulated market for electricity services.

3.2 Price

In both industrialized and developing countries, electricity reform leads to price changes, although for different reasons and sometimes in opposite directions. In industrialized countries, reforms are expected to lead to efficiency gains and lower prices for consumers. In practice, where these gains have been realized, they have benefited shareholders and rich consumers more than they have benefited poor consumers. The results can be highly regressive. In Argentina, for example, prices went down postreform for all consumer classes, but the price for high-consumption (and likely wealthy) users dropped by 70% whereas that for low-consumption (and poor) users fell only marginally. The net effect was that, in absolute terms, the rich paid 40% less than the poor paid per unit of electricity. In the United Kingdom, electricity companies reaped the greatest share of benefits, government gained from increased sales revenue and higher taxes on profits, but power purchasers paid higher prices than they would have under public ownership.

In many developing countries, prices for consumers are likely to rise, at least in the short to medium term, as cross-subsidies and price restrictions are removed. In some instances, these price increases have been considerable. In Ghana, the tariff increased by about 300% in the course of implementing reforms. Higher prices disproportionately affect the poor, who tend to spend a larger proportion of their income on energy services than do wealthy consumers. Although the higher prices may be necessary to put the sector on a sound commercial footing, they also bring social costs. Consequently, to ensure that social policy goals of providing effective access to electricity are met may require a measure of public support in the form of subsidies.

The question of subsidizing the price of electricity is closely tied to the discussion on access to electricity (Section 3.1), because a given subsidy allocation can be used either for the cost of facilitating access to electricity or for reducing the per-unit cost for the consumer. Both benefit the consumer, but in different ways. There are good arguments for subsidizing access rather than consumption. Populations without access to electricity are likely to be among the poorest in a country. By focusing on access, subsidies are automatically targeted toward the neediest. In addition, subsidizing use can skew consumption patterns in detrimental ways. For example, electricity subsidies for farmers in India have encouraged heavy use of groundwater, including the introduction of

water-intensive crops in areas of water shortage. One means of avoiding this problem is to subsidize the fixed costs associated with consumption, such as connection fees. The result can be substantial savings for low-income groups, even while eliminating the potential for price distortions.

Although subsidizing access is preferable to subsidizing consumption, subsidizing a minimal level of consumption can afford low-income groups access to basic electricity services, with considerable social gains. Such subsidies are normally structured as “lifeline rates,” which are cross-subsidies that allow poor households that use basic services to pay less, compared to wealthier households that use more electricity services. Although challenged under a reform approach that seeks to remove market distortions, these lifeline rates continue to be an effective way to balance social objectives with efficiency gains.

Several factors bear consideration in the design of price subsidies. First, subsidies should be well targeted to a population. In many cases, subsidies have become mechanisms for appeasing politically important constituencies rather than an effective way of supporting a social objective. Second, the amount of subsidy should be consistent with the goal of the subsidy program. For example, a lifeline rate should be limited to basic electricity services, and should not be set at a consumption level that encompasses groups beyond the target group. Third, the subsidy should be designed so as to minimize distorting consumer decisions.

Another means of minimizing social pain due to price increases is to increase the efficiency of electricity use. Doing so would enable consumers to extract greater services out of each unit of electricity consumed. In this sense, programs to promote end-use energy efficiency, typically conceived of as environmental programs, also have social benefits.

3.3 Employment

In many countries, the electricity sector is a significant employer. Reform of the sector is often accompanied by job cuts, as privatized utilities seek to cut costs and improve efficiency. In the European Union, for example, the research unit of Public Services International, a labor union, found that between 1990 and 1995, 212,000 jobs were lost in the energy sector, equivalent to a 17% reduction. The United Kingdom, where 42% of jobs in the sector were cut, accounted for more than half of these losses.

In some countries, reform of the sector to restore efficient functioning and financial viability may indeed require some measure of employee reduction. For example, in many African countries, the number of customers per employee hovers in the range of 20–80, as against a nominal international target of around 125. However, these numbers must be read with caution, because they do not account for differences across companies that only generate power and those that also distribute power. Moreover, the relative costs of labor and capital vary across countries, which suggests that the appropriate ratio of customers to employees need not be the same across the world. Nonetheless, it is safe to assume that there are many cases in which effective reform requires a measure of labor retrenchment. In these cases, a socially progressive approach to reform would ensure safeguards for laid-off workers, such as retraining programs and assistance with reemployment. Even more wide-reaching are plans to include workers at large in the capitalization of electric utilities. In Bolivia, for example, labor unions, political parties, business groups, and others forged a “capitalization plan” whereby a portion of the shares of state-owned enterprises, including electricity, were distributed among the broader working-age population. Specifically, half the shares were sold to strategic investors and the other half were credited to workers’ individual retirement accounts, with a portion distributed as annuities. The general point is that although employee reductions may be necessary in some cases, the associated social pain can be reduced by appropriate mitigation, and by undertaking privatization in a manner that distributes benefits across the population.

4. ENVIRONMENTAL CONCERNS IN ELECTRIC POWER REFORM

The electricity sector has a considerable environmental footprint. Many of the environmental impacts associated with electricity are related to the heavy reliance on fossil fuels for electricity generation in much of the world. As the World Energy Assessment summarizes, combustion of fossil fuels releases local pollutants such as particulates and lead, regional pollutants such as sulfur dioxide, and global pollutants such as carbon dioxide. These pollutants have consequences for human health and the well-being of ecosystems. In addition, there are upstream environmental impacts such as those caused by mining and fuel transport, storage, and

processing, and downstream impacts such as ash disposal. Nonfossil sources of electricity also cause environmental disturbance. For example, hydropower (in particular, large hydropower) causes several land-use-related impacts, such as submergence, displacement of populations, greenhouse gas emissions caused by decaying biomass, and disruption of fisheries. The discussion in this section, however, focuses largely on air pollution impacts.

What will be the net effect of electric power reform on the environment? There is no single answer to this question. Instead, the net effect of the several changes that collectively constitute reform will depend at minimum on the time frame (short versus long term), the specifics of how policies are implemented, and the initial conditions in the reforming country. However, systematic tendencies can be associated with particular reform steps. And, in addition, countries may choose to take proactive measures to address some of these tendencies. The discussions that follow outline these environmental tendencies and spell out possible proactive steps to increase the likelihood of environmentally positive outcomes.

4.1 Environmental Implications of Institutional Change for Electricity Reform

There are several overlapping ways to analyze the environmental implications of institutional change in the electric power sector. One approach is to examine the various components of electricity reform (as described in Section 2.1), with an eye to assessing the environmental implications of each component. This is the approach followed in this discussion. Another would focus on pathways through which reforms would have an environmental impact, such as on choice of generation technologies. These are not exclusive approaches, and there is necessarily some overlap in the categories under each approach. To complement an analysis based on the reform component, Table I, which is intended to be illustrative rather than comprehensive, summarizes a pathway approach.

The remainder of this section follows a framework laid out by Keith Kozloff in studies for the U.S. Agency for International Development and for the Renewable Energy Policy Project. The discussion is organized around five dimensions of reform: commercialization, privatization, unbundling, wholesale competition, and retail competition. The first two dimensions are related to ownership, and

TABLE I
Pathways through Which Electricity Reform Affects Environmental Outcomes

Pathway	Relevant questions	Intermediate variables
Consumption: price response and output substitution	How will electricity price change?	Price elasticity of demand
	Will output substitution occur toward or away from electricity?	Existing generation mix
	Will a market for electricity services be created?	Existing nonelectricity fuel choices
Generation technology: scope for input substitution	Which technologies will be favored under reform, and which will be disadvantaged?	Fixed versus operating cost of technology
		Transmission investment or cross-border trade as a substitute for generation investment
		Extent to which generation facilities using older technologies have been amortized
Efficiency of supply and demand	Will incentives for more efficient generation, transmission, and distribution be introduced?	Commercial pressure to cut costs
	Will end-use energy efficiency become easier or harder to promote?	Transactions costs for end-use efficiency
Regulation: environmental and economic	To what extent will system-wide planning be carried out in a competitive market?	Legal framework for regulation
	How will economic and environmental regulations be coordinated?	Capacity of economic regulators to address environmental incentives and vice versa Degree of enforcement capability

the last three, to structure. These categories often blur at the edges, and links listed under one could often feasibly be listed under several other categories.

4.1.1 Commercialization

Commercialization in the power sector is likely to have a range of positive impacts on the environment. First, the principle of cost recovery introduced through commercialization could increase incentives for better management and operation of existing capacity and more efficient transmission and distribution. Cost recovery will also provide an incentive to utilities to implement end-use efficiency programs, especially to reduce losses in rural areas, where the revenue generated often does not meet the marginal cost of supplying electricity.

Second, commercialization is often accompanied by tariff reform that releases controls on electricity rates. Although in industrialized countries reform has led to greater efficiency and often to lower prices, in developing countries reforms are likely to lead to higher prices as government-mandated price caps are removed. Higher prices will provide consumers greater incentive to implement end-use efficiency programs, and may provide market opportunities for

firms that provide energy efficiency products and services. However, price effects may be muted because the price elasticity of electricity is considerably lower—perhaps as much as 50% lower—in developing countries as compared to industrialized countries. On the negative side, they will also partially undermine the benefits utilities gain from loss-reducing efficiency programs in rural areas.

Third, the combination of cost recovery and more accurate price signals suggests that a commercialized utility has a strong incentive to identify least-cost ways of extending services to rural areas. This will likely provide an incentive to distributed renewable energy technologies.

4.1.2 Privatization

Privatized utilities have little or no commitment to achieving social or environmental goals, unless they are explicitly required to or are given incentives to do so through regulatory structures. However, on the positive side of the environmental ledger, because a private utility not only seeks to recover costs but also to turn a profit, the incentives for greater efficiency in supply-side efficiency improvements in operation and maintenance also introduced through

commercialization are strengthened. Similarly, a private utility may have an incentive to invest in low-cost technologies, which will promote a switch to cleaner burning natural gas.

On the negative side, the private sector typically has a higher cost of capital and a higher discount rate than does the public sector, both of which provide disincentives to environmentally sensible investments that yield benefits over time. For example, because demand-side management (DSM) measures yield benefits over a number of years, they will yield a lower return under private ownership and will be less attractive. Similarly, higher capital costs and discount rates are unfavorable to renewable energy sources that require large up-front capital investments, and are more friendly to technologies such as gas turbines, which have low capital costs but high fuel costs. In addition, privatization could attract greater investment to the sector, which would increase generation capacity and result in a larger environmental footprint. However, to the extent that new capacity replaces small-scale diesel self-generation by industry, this effect would be mitigated.

If requirements are placed on the private sector to reverse some of the negative incentives just discussed, then it is important that these be put in place prior to privatization. Once existing assets are sold off, it will be politically difficult to change the rules, and sends future investors a signal of inconsistent policy.

4.1.3 Unbundling

As a result of unbundling, the costs of providing generation, transmission, distribution, and retailing services are accounted for separately. As a result, unbundling can provide disincentives to activities that promote energy efficiency. In an unbundled sector, generation and transmission companies do not have direct contact with the consumer in order to undertake end-use efficiency programs. Distribution companies, which are in a position to undertake these programs, are unable to capture the full benefits of greater efficiency because the benefits of reduced generation and transmission requirements are shared with other distribution companies. This separation also limits access to useful information—those on the supply side are cut off from information on consumption patterns, and distribution companies may not have access to billing data from retailing companies. In addition, unbundling makes it more difficult for a single entity to identify and capture the full upstream savings in generation and transmission costs provided by distributed renewable

sources. Consequently, there is diminished incentive to invest in distributed renewable power.

Beyond these general observations, the environmental impact of the unbundling process depends on the details of implementation. First, in an unbundled structure, the terms and conditions of access to transmission have environmental implications. Open transmission access could stimulate investments in grid-connected renewable sources. However, this advantage will be negated if renewable power developers are charged higher prices, as is possible under some pricing schemes. In addition, transmission access opens opportunities for interregional electricity exchange that may affect the overall fuel mix. If access to broader markets increases use of low-cost coal plants, which operate under their operating potential, then the result will be negative environmental outcomes. This may be less of a problem in developing countries, where such excess idle capacity typically does not exist. Finally, transmission access could also promote cogeneration. However, in order for this to work, access prices must be low enough for cogeneration to compete with other supply sources.

Second, in the single-buyer model (many generators but one monopoly buyer), the environmental consequences are influenced by the details of power purchase agreements (PPAs). Existing PPAs are structured around low-capital-cost technologies such as fossil fuel plants, which may be unfavorable to renewable energy technologies with equivalent life-cycle costs but higher up-front costs. Important considerations are payment schedules and the treatment of risk in fuel price. In addition, because renewable power projects tend to be smaller but transactions costs tend to be fixed, the transactions costs per unit tend to be higher for renewables. In order to overcome these problems, PPAs need to be tailored to renewable sources and standardized. Moreover, the long-term nature of PPAs is potentially damaging because it locks in technologies that might rapidly become obsolete. A spot market is more likely to take advantage of new technological options as they develop.

4.1.4 Wholesale Competition

Wholesale competition may promote more efficient operation of thermal power plants, but has disincentive effects on energy efficiency. First, when wholesale competition is organized around spot markets in electricity, plant operators have an incentive to run their plants as much as possible in order to maximize returns. Second, price signals in

spot markets are based on short-term costs rather than on life-cycle costs. Because the benefits of energy efficiency are spread over a long period, these benefits tend not to be factored into spot markets. Third, spot markets potentially increase the expected variability in prices, which leads to more uncertain savings from energy efficiency, which are then discounted compared to more certain savings derived from life-cycle costs.

Wholesale competition will likely also negatively affect adoption of renewable energy technologies, because these technologies are more difficult to finance in short-term electricity markets. Spot markets imply a more uncertain revenue stream than do long-term PPAs, thus lenders are less willing to finance the high capital costs of renewable power projects in a short-term market. Moreover, the high debt burden of renewable power projects leaves the projects susceptible to a slump in spot market prices. In addition, some “intermittent” renewable energy sources, such as wind and solar, are at a disadvantage in spot markets because they are unable to provide power on demand in peak periods. This disadvantage needs to be addressed explicitly in the formulation of rules governing the spot market.

4.1.5 Retail Competition

Retail competition, if appropriately structured, may introduce some features that mitigate the negative environmental effects of wholesale competition, whereas others worsen the problem. There are three possible positive effects. First, retail competition requires access to distribution lines. The costs of access require that area-specific marginal costs are assessed. If these costs are charged to customers, their incentive to manage their load is enhanced. Second, a generator who is also a retailer could invest in conservation measures within its customer base, and use the saved energy to compete in other markets. This mechanism is enhanced where generators face an “obligation to serve” in rural areas at an unprofitable tariff. Efficiency measures allow the utility to reduce losses in these unprofitable areas and sell the saved power in profit-making areas. Third, retail competition allows renewable energy suppliers to distinguish themselves by the source of their energy, by, for example, promoting themselves as “green” power suppliers. However, this strategy may be less successful in developing countries, in which electricity prices account for a larger share of budgets, as compared to industrialized countries.

On the other side of the ledger, retail competition will likely lead to an electricity price decline, further

diminishing the incentives to undertake efficiency-improving measures. Second, because retail competition allows a customer to switch between suppliers at any time, a retailer is not guaranteed of capturing the full stream of benefits of end-use energy efficiency investments, providing a strong disincentive to make those investments.

4.2 Measures to Improve Environmental Outcomes Following Reform

As the preceding discussion suggests, the environmental impact of electricity reform in a particular place will depend on the details of implementation. In this section, three important elements of implementation are summarized, each with examples of environmentally positive policy steps.

4.2.1 System Planning and Portfolio Management

Effective system planning and portfolio management require that the costs and benefits of alternative electricity supply mechanisms are comparable across a system. This comparison should include consideration of demand-side savings and non-traditional sources such as renewable energy, should internalize environmental and social considerations, and should provide a mechanism for public input. Integrated resource planning (IRP) was the tool devised to meet these requirements under an integrated system. IRP provides a means to minimize the “total resources cost” involved in meeting electricity service needs, which includes not only the costs to the utility and its customers, but also the indirect social and environmental costs.

Centralized IRP at the system level is incompatible with a market structure. However, the essential planning and portfolio management functions remain important, and should be performed through other institutional mechanisms. For example, under wholesale competition, distribution utilities can still conduct, or can be required by law to conduct, an integrated analysis of alternative portfolios, and can draw up a plan based on that analysis. Instead of directly acting on the plan, however, the distribution utility must then solicit bids from the marketplace to meet the plan. Over time, the demand pattern would force a more balanced portfolio of resource investment.

Another approach to internalizing environmental externalities would shift the focus from procurement to dispatch. An “environmental dispatch” approach would use a pollution index to adjust both the order in which electricity is dispatched from generators and the price the generators are paid. Environmentally

clean generators are dispatched preferentially and receive a higher price.

4.2.2 Environmental Implications of Regulatory Decisions: Transactions Costs and Prices

Many decisions routinely made by electricity regulators for economic reasons often have environmental implications. Regulatory decisions can unintentionally increase the transactions costs for more environmentally sound forms of electricity. In other cases, opportunities to reduce these transactions costs may be missed. In addition, rules on pricing decisions may systematically underplay the economic benefits that energy efficiency and renewable energy can bring. For example, a regulatory decision that links utility profits to the amount of electricity sold works as an obstacle to energy efficiency. Specifically, a price cap does just this, whereas a revenue cap decouples sales from revenues. In addition, unbundling raises the transactions costs to investing in efficiency—price signals are less clear, distribution utilities have an incentive to “free ride,” and generators have inadequate access to consumers. Regulators can play a role in fixing these problems. If retail bills reflect actual service costs, customers receive signals that allow them to tailor demands to reduce costs. On the other hand, if these costs are represented as fixed charges, then effective demand management becomes harder.

With respect to renewable energy, a simple approach to minimizing transaction cost is to offer a standardized contract for power purchase to small-scale renewable energy suppliers. Without such a standard contract, the fixed cost of negotiating a power purchase agreement may be prohibitive for small suppliers. This approach has worked with some success for small hydropower developers in Sri Lanka. In addition, open access to the transmission grid is a prerequisite for renewable energy suppliers to compete in a market. These policies are credited with stimulating a renewable energy industry in India.

The absence of facilities for net metering (a bidirectional meter that registers flow in both directions) introduces a transactions cost that discourages investment in very small-scale renewable energy. The existence of net metering would allow embedded small-scale generation to realize its full potential by allowing the sale of extra power generated on site to the grid. Finally, the bidding rules that govern power pools may be written based on the characteristics of large generation plants, and may systematically disadvantage renewable energy, small generating units, and energy efficiency.

Similarly, there is a set of regulatory decisions that operate through the mechanism by which prices are set. For example, because distribution costs can vary widely across a service area, an average cost-based price conceals the true gains from distributed energy generation and from energy efficiency. By contrast, a deaveraged distribution price would send consumers the appropriate price signals. In another example, subsidies for line extensions also artificially lower the cost of on-grid generation technologies, undermining the competitiveness of off-grid technologies. Finally, transmission prices and terms of access will need to reflect the special circumstances of renewable energy and small generating units if they are to be competitive. For example, distance-based pricing, pricing based on maximum rated capacity rather than actual generation, and a penalty for intermittent transmission could all work against renewable power developers.

4.2.3 Proactive Policy Measures to Promote Energy Efficiency and Renewable Energy

In addition to ensuring that economic regulation does not disadvantage energy efficiency and renewable energy, there are a number of proactive measures consistent with competition available to promote these approaches. A number of countries have adopted price-driven, or supply-push, renewable energy incentives programs. In the United States, utilities were required to purchase all the renewable energy offered to them at a price equivalent to the “avoided cost” of generating that power. The United Kingdom has imposed a tax on fossil fuel generation; under the Non Fossil Fuel Obligation program, the tax is used to support nonfossil energy sources. Germany and Denmark have implemented a “feed-in” law, which sets a guaranteed price for the purchase of renewable energy. India has allowed generous depreciation allowances to promote wind energy.

Quantity-driven incentives such as a renewable portfolio standard (RPS) offer another approach. An RPS requires all retail electricity service providers to purchase a specific proportion of their electricity from renewable energy sources. This approach has been adopted in several European countries and in several states in the United States.

A systems benefits charge provides a more general way to support environmentally (and socially) desirable investment requirements by means of a general surcharge across the system. For example, Brazil has adopted a modified version of a system benefits charge, requiring distribution utilities to spend 1% of revenues on energy conservation, of which 0.25%

must be spent on end-use efficiency. More generally, the funds gathered through a systems benefits charge may be spent on increasing access to electricity for disadvantaged populations, on energy efficiency, or on research and development of sustainable energy technologies.

5. GOVERNANCE OF ELECTRIC POWER REFORM

It is increasingly clear that although provision of electricity services would benefit from the application of commercial principles, application of these principles should not exclude a strong government role to maintain essential public benefits (social and environmental) in the electricity sector. As the preceding discussion suggests, there are various mechanisms through which to promote social and environmental interests.

A technical understanding of the links between social and environmental concerns and electric power reform is a necessary, but not sufficient, condition for these interests to be addressed. If decisions about electricity reform are made without advocates for concerns such as access to electricity, price, and environmental outcomes, present and past experience suggests that these issues are likely to be crowded out by more narrow economic considerations. Equitable and environmentally sound reform will require that decision-making processes be open to concerns from rural populations, civil society, environmentalists, and others who typically articulate a public benefits agenda. Hence, equity and environment in electricity reform are equally about an effective system of sectoral governance and a technical or policy understanding of social and environmental concerns in the sector.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation:

Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy • Environmental Injustices of Energy Facilities • Equity and Distribution in Energy Policy

Further Reading

- Brook, P. J., Besant-Jones, J. (2000). Reaching the poor in an age of energy reform. In "Energy Services for the World's Poor." Energy Sector Management Assistance Program, Washington, D.C.
- Burtraw, D., Palmer, K., and Heintzelman, M. (2000). "Electricity Restructuring: Consequences and Opportunities for the Environment." Resources for the Future, Washington, D.C.
- Clark, A. (2001). Making provision for energy efficiency investment in changing electricity markets: An international review. *Energy Sustain. Dev.* V(2), 26–39.
- Dubash, N. K. (ed.). (2002). "Power Politics: Equity and Environment in Electricity Reform." World Resources Institute, Washington, D.C.
- Goldemberg, J., and Mielnik, O. (1997). Electricity system reform and sustainable development in Latin America. *Energy Sustain. Dev.* III(6), 17–26.
- Hamrin, J. (2002). Policy options for building environmental protection into the restructuring process. In "Energy Market Restructuring and the Environment" (M. Harriss, Ed.), pp. 197–220. University Press of America, Lanham, Maryland.
- Hunt, S., and Shuttleworth, G. (1997). "Competition and Choice in Electricity." John Wiley & Sons, Chichester.
- Karakezi, S., and Kimani, J. (2002). Status of power sector reform in Africa: Impact on the poor. *Energy Policy* 30(11–12), 923–946.
- Kozloff, K. (1998). "Electricity Sector Reform in Developing Countries: Implications for Renewable Energy." 2. Renewable Energy Power Project, Washington, D.C.
- Martinot, E. (2000). "Power Sector Reform and Environment: A Role for the GEF?" Global Environment Facility, Washington, D.C.
- Reddy, A. (2001). Indian power sector reform for sustainable development: The public benefits imperative. *Energy Sustain. Dev.* V(2), 74–81.
- U.S. Agency for International Development. (1998). "The Environmental Implications of Power Sector Reform in Developing Countries." 98-05 USAID, Washington, D.C.
- Wamukonya, N. (ed.). (2003). "Electricity Reform: Social and Environmental Challenges." United Nations Environment Programme, Nairobi, Kenya.

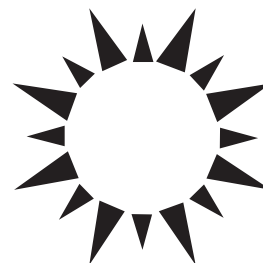
Electric Power Systems Engineering

MARIJA D. ILIC

Carnegie Mellon University
Pittsburgh, Pennsylvania, United States

JOHN ZABORSZKY

Washington University
St. Louis, Missouri, United States



-
1. Introduction
 2. Basic Objectives of EPSE
 3. EPSE Tools for System Operation under Normal Conditions
 4. EPSE for Emergency Operation
 5. Challenges to EPSE in the Changing Industry
 6. Objectives of EPSE as a Function of Changing Industry Structures
 7. A Glimpse into the Future: More Smooth and Secure on a Building Block Structure

Glossary

forced outage or blackout Loss of electrical power due to the shutdown of a major component, such as a power plant, transformer, or transmission line.

gains Increase in signal power, voltage, or current by amplification.

governor A control device on a machine or engine that is used to maintain a steady speed, pressure, or temperature automatically (as by controlling the fuel supply).

load A device or destination to which electrical power is delivered; also, output of a generator or power plant.

primary controller An electrical instrument or set of instruments that serves as the main regulating or guiding device for a mechanical or electrical system.

transformer An electrical device used to transfer electric energy from one circuit to another, especially equipment that produces such a transfer with a change in voltage.

utility A company that performs a public service (such as sewage disposal or supply of gas for heating) subject to government regulation; in this context, an electricity generation/supply company.

This article is devoted to the electric power systems engineering field, focusing on systems aspects of

modeling, analysis, and decision-making tools for operating and planning complex electric power grids. It first describes basic objectives of electric power systems engineering. The primary emphasis of the article is on IT tools and the key challenges that must be overcome to further benefit from using electric power systems engineering (EPSE) tools in support of the electric power industry under restructuring. The material in this article is of direct relevance for managing the electric power systems in a reliable and efficient way without major widespread loss of service to customers.

1. INTRODUCTION

Electric power systems engineering (EPSE) is a subfield of electric power engineering, with the prime concern being performance of an interconnected electric power system. The EPSE subfield has evolved over the past 40 years in response to the industry needs to manage the system (1) securely under unexpected equipment outages and unusually wide variations in load demand patterns and (2) as economically as possible in response to normal load-demand variations. Its growth has been made possible through tremendous progress in the general fields of computers, communications, and control engineering for complex, large-scale systems. Nevertheless, the general tools had to be formulated and made useful in the context of the specific applications to the electric power systems.

This article is being completed the eve after the major blackout in North America that took place on August 14, 2003. Society has once more been caught

by surprise in a widespread industry failure to provide electricity to over 60 million customers for over 24 h. Questions about why such an event spread over very large geographical areas are being asked. A typical belief that an electric power grid has automated protection against cascading failures despite an unexpected failure of a major piece of equipment, such as a transmission line or a power plant, has once more been proven wrong. A related category of questions concerns prevention of massive power failures by isolating only parts of the grid that had to be disconnected. Still other questions focus on what should be done to modernize the electric power grid in order to prevent future cascading failures.

This unplanned coincidence between the blackout and the completion of this article has forced the authors to revisit its earlier version. To the authors, the 2003 blackout represents yet another example of challenges to the operating practices and the information technology (IT)-based methods for facilitating reliable and efficient operation of complex electric power grids. It is, unfortunately, one of many in the long list of hard-to-predict, high-impact, cascading events of this type that have taken place since the early 1960s worldwide. It is hoped that the reader will see several connections between the material in this article and the recent events.

The following quotation from *Dynamics and Control of Large Electric Power Systems*, by Ilic and Zaborszky, briefly summarizes the gradual evolution of the U.S. interconnection and points out the rationale for its present complexity:

The U.S. structure of power-system operation was developing generally during the twentieth century without much awareness or attention to the implications of the evolution of power-system structure. With the gradual infiltration of the computer into on-line operation in the 1950s, a surge of interconnections developed for utilizing the potential for improved efficiency and economy. As is typically the case in human activities, the development took the form of designs for individual cases of construction based only on the local situation of the case and only during normal, stationary system operation. The vital questions of the effects of off-nominal conditions were essentially ignored. A striking example is the omission in new construction of self-starting house units and the total dependence for starting large power plants on power drawn from transmission lines to neighboring stations. This general attitude was the root cause of the famous Northeast Blackout, a cascading system collapse that disabled ten states for approximately 1 week in 1964 followed by a few other large events, such as that occurring in Manhattan in 1967.

After this rude awakening, consciousness of the importance of the structure of the complete system (the

interconnected U.S. electric power system) started to gradually creep into the awareness within the field.

—Ilic and Zaborszky, 2000, p. 642

Also, recent electric power restructuring has created very new challenges to the existing EPSE methods. One of the qualitative changes related to the organizational changes in this industry involves the shift from a top-down coordinated monitoring and decision-making to more distributed information managing and decision-making. This change is opening doors to new technical and pricing paradigms for operating complex power grids in a self-healing manner by the system end users adjusting to the changing conditions in a bottom-up manner.

The most recent blackout, the above quote summarizing the beginnings of EPSE, and the recent industry reorganization all concern on-line information essential for determining the system-wide status beyond single components and single utilities. Since the specifics of particular physical and policy/regulatory phenomena vary with the nature of triggering events and system designs, the complexity of data and difficulty of extracting critical information quickly are issues at the heart of any blackout. The critical physical information may concern frequency-, voltage-, or transmission-related physical phenomena; the importance of these factors is well documented in several post mortem analyses of major blackouts. The critical policy/regulatory information may affect capacity available for managing the system as a whole under forced outage and its valuation or it may concern state versus federal responsibility for building stronger transmission lines for exchanging power under outages.

This article is devoted to the electric power systems engineering subfield of electric power engineering, focusing on systems aspects of modeling, analysis, and decision-making tools for operating and planning complex electric power grids. It first describes basic objectives of electric power systems engineering. The primary emphasis of the article is on IT tools and the key challenges that must be overcome to further benefit from using EPSE for operating and planning electric power systems of the future. The final section describes fundamental challenges to developing EPSE tools in support of the electric power industry under restructuring.

Electric power systems engineering must evolve as the industry needs evolve, both as the industry reorganization takes place and as the new hardware technologies become more available and cost-effective. Much could be improved by supporting the

power grid's on-line operations via carefully designed minimal coordination at its various layers. Doing this correctly remains a tremendous challenge in pursuit of increased system-wide efficiency and ability to prevent huge failures.

2. BASIC OBJECTIVES OF EPSE

The main objectives of EPSE are monitoring, analysis, and decision-making tools for facilitating operations and planning of an electric power system as a whole. The EPSE methods are available for monitoring, analyzing, and controlling electric power grids in a semiautomated manner under normal operating conditions, i.e., when the status of the equipment is as planned. The EPSE approaches to managing the system under forced outages are not as mature in comparison with the tools available for normal operation; instead, these methods are generally utility-specific and based on human expert knowledge about the system. The events on August 14, 2003 confirmed once more that industry practices are not fully capable of preventing widespread power failures under forced equipment outages. Moreover, as the electric power industry restructures and new technologies become more cost-effective, the objectives of electric power systems engineering evolve, requiring novel problem formulations and computer-aided tools for relating often conflicting engineering and economic objectives.

The early stages of electric power engineering were primarily concerned with the design and operation of individual pieces of equipment, power plants, transformers, transmission lines, and the like. However, as the electric power grid has expanded over time, it has become essential to understand the interactions among various system components and the effect of these interactions on the overall quality of electricity service. The quality and cost of power service to customers have greatly improved as a result of optimizing utilization of system-wide resources under normal conditions. IT-based power systems engineering tools for scheduling the electric power generation required to meet the time-varying load demand have been basic to efficient use of the least operating cost power generation. In particular, under normal conditions, the cheapest generation is scheduled and delivered across the power system to supply both forecasted demand and demand fluctuations that are hard to predict. Ideally, under unexpected forced outages of major components (power plants, transformers, or transmission lines), power should be delivered from other parts

of the system so that the load in the affected area remains served. To achieve this, power systems engineering tools have been developed for scheduling short-term generation reserve to supply customers in case major components of power equipment fail as a result of hard-to-predict forced outages.

The software tools for balancing supply and demand under normal conditions are based on implicit assumptions about the electric power system structure. Naturally this structure allows simplification of very complex electric power system dynamics by the separation of operations into subprocesses that evolve at different temporal rates and at various system levels. It is important to use the IT-based tools to assess the validity of the assumptions and underlying algorithms used, as system conditions vary over time. In particular, the useful structure and unique features of complex power system dynamics are typically violated under forced outages. Moreover, industry restructuring may inadvertently lead to qualitatively new power system architectures, which, in turn, will require careful assessment of the software tools used for operating them.

Some of the fundamental policy- and technology-related changes are discussed in Sections 5 and 6 below. A typical power systems architecture characteristic of normal operation and regulated industry is described in Section 2 in preparation for reviewing the IT tools for operating regulated utilities under normal conditions. The most difficult challenges concern trade-offs between efficient operation under normal conditions and the stand-by reserve to be used in case of forced outages. Since the critical events are unknown and are very hard to predict in a timely manner, conventional probabilistic tools are difficult to use for efficient on-line decision-making.

2.1 Basic Structure of Electric Power Systems

A typical organization of today's electric power systems is shown in Fig. 1. The generation and distribution of electric power geographically are accomplished over one single electric network in the United States. Traditionally in the regulated industry, generation is provided in very large quantities at power plants. Under restructuring, small-quantity distributed generation also enters with a naturally limited scope of providing power locally. Electricity consumption ranges from very large levels (factories) to very small levels (residential).

This immense interconnected network inescapably has an internal structure. Generation in very large

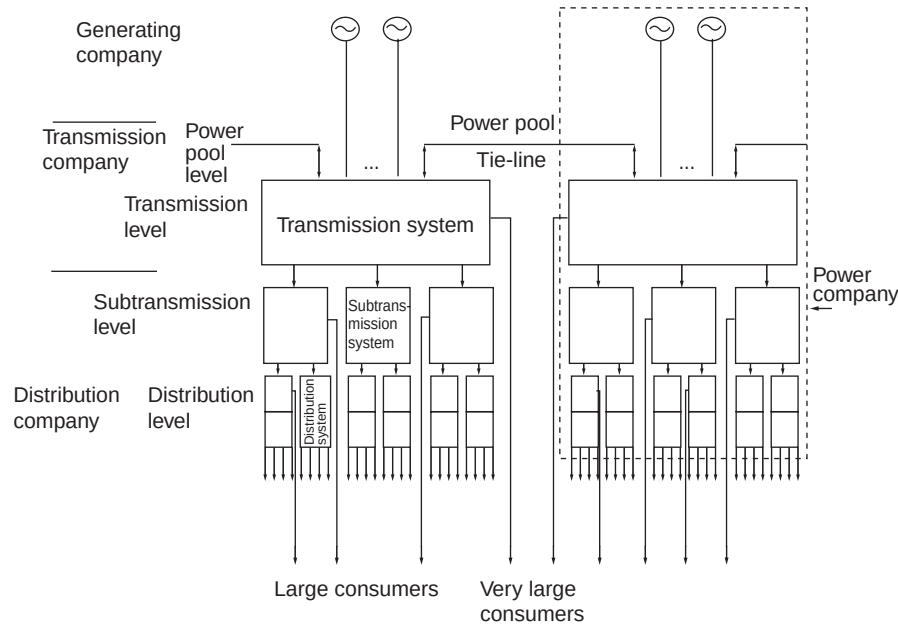


FIGURE 1 Typical organization of existing electric power systems.

quantities in power plants demands high- to very-high-voltage (ranging from 20,000 to 1 million V) transmission line networks. Such networks make up the pool or bulk transmission level network that covers the United States but is also geographically subdivided into large blocks according to ownership, as indicated in Fig. 1. The power distribution process within each geographic bulk transmission block is typically subdivided into geographic subtransmission areas, which historically do not connect to individual generating facilities; instead, only large loads are supplied from the subtransmission networks along with the various distribution sections within each subtransmission area. Finally, the lowest voltage distribution level also splits into two voltage sublevels, ending at the residential households.

The network consists of transmission lines that interconnect nodes, known as buses, at high voltage levels. Different levels of voltage within this structure are interconnected via transformers. Buses are the sole connection points to the transmission system of all equipment:

- (1) the sources, that is, generators;
- (2) the loads, either large and concentrated like an aluminum plant or distributed like the feeding point of subtransmission networks supplying a composition of smaller loads and eventually leading to distribution lines on residential streets;
- (3) reactive-power storage devices, such as capacitor banks or reactors;

- (4) increasingly numerous electronically switched power devices (thyristors), such as static Var compensators, the coupling equipment between the high-voltage (HV) alternating current (AC) bus and the high-voltage direct current (HV-DC) transmission lines, and flexible AC transmission systems (FACTS) devices that are in development;

- (5) series capacitor compensation; and
- (6) lightning arrestors and other equipment connected to a bus by links other than transmission lines.

Buses form large islands of complex equipment connected to the outside world exclusively through the high-voltage lines (AC or DC) of the transmission network.

The business organization of this gigantic system in the traditional regulated sense was single ownership of each vertical column; see Fig. 1. This division preserves the coast-to-coast nature of the total network, but the sale and purchase of power are carried out between companies owning the individual vertical blocks within Fig. 1 under very strict government regulations. In the new industry environment, individual blocks in Fig. 1 are separately owned, and free market trading proceeds for a fee between individually owned transmission blocks under relatively mild government regulation. The above-described strict separation by size is also fading; any amount of power can be bought and sold at any power flow level.

The above-described organizational structure in the regulated industry maps into various power flow levels. The bulk power is transmitted from a very few large power plants to the load centers within a region or from a power pool composed of several subsystems; these subsystems most often are defined as the boundaries of the individual electric power utilities, loosely coupled with the neighboring utilities. The delivery of power between utilities within an electric power pool or within similar horizontal structures occurs by means of an extra-high-voltage (EHV) sparse electric transmission grid, at typical voltages of 500 kV and higher; the pool (or region) level is referred to as a tertiary level. Within each subsystem, the transmission grid is less sparse and at high voltage, typically above 200 kV, known as the secondary level. Finally, major pieces of equipment, such as power plants, transmission lines, transformers, and substations (load centers), are connected to an even lower level voltage network, known as subtransmission; they are referred to as the primary (component) level. The substations are further connected through a low-voltage distribution system to different loads; larger industrial loads are connected at voltage levels similar to the substation voltage level and the smaller loads are progressively connected to lower and lower voltage levels within a distribution system, terminating at the lowest level residential end users.

2.2 Primary Control of an Electric Power System

An electric power system characterized by the above structure serves the basic purpose of producing and distributing electric power to the customers. The system as a whole is operated in anticipation of and in response to the temporal variations in load demand. Electric power is produced by the turbine-generator-governor sets at the power plant locations. The mechanical power is produced by the prime movers. The key control of how much power goes to the network is the governor, which is a local (primary) controller of each modern power plant. Like any other automated controller, a governor has a setpoint and reacts to the deviations from this setpoint automatically. This controller reacts to the frequency deviations from the frequency setpoint value. The faster the turbine turns, the more power is decreased by controlling fuel income to the prime mover by means of a governor (Watt regulator). Standard operating practice has been to adjust the setpoints on governors of generators in each area so that the

anticipated area load is fully supplied and the system frequency is nominal. The mechanical power is then converted to the electric power in generators and distributed via the electric power transmission network to the load centers and further distributed to the individual end users via the local electric distribution networks. It is generally hard to predict load exactly. This results in an imbalance between mechanical power produced and power consumed. In this case, according to Kirchhoff laws, the electric power into each node of the electric power network still equals electric power out of the node instantaneously. If there is not enough mechanical power to supply electric power consumed, the generators slow down, resulting in reduced system frequency.

Most of the electric power transmission today is AC, requiring all power generators to operate at the same nominal frequency. In order to maintain frequency very close to its nominal value, each system has several power plants that participate in so-called automatic generation control (AGC). These plants automatically adjust their electric power output by resetting the setpoints on their governors in response to frequency deviations in each control area. Similarly, power plants have so-called automatic voltage regulators (AVRs), which automatically control reactive power produced by the power plant in order to maintain voltage magnitudes at the location of power plants within acceptable voltage deviations. The voltage regulators are also local (primary) controllers. The setpoints of these controllers are adjusted in anticipation of reactive power consumption on the system to ensure that voltage magnitudes at the consumers' side stay within the acceptable limits. Given a setpoint, the primary AVR regulates the reactive power output of a generator so that the voltage magnitude remains within the threshold of the setpoint. This is very similar to the way the governor responds to frequency deviations by producing real power in order to maintain frequency close to the nominal frequency. Some power plants also have power system stabilizers, which represent more advanced local controllers that respond simultaneously to the deviations in frequency and voltage and/or acceleration and have a faster response to the error than the governors and AVRs. If the power system stabilizers are tuned correctly and are implemented on the key power plants, the electric power system as a whole responds in a much more robust way to major disturbances than if the power plants do not have these controllers.

The response of all primary controllers is tuned for the assumed range of operating conditions. The

tuning is achieved by modeling the rest of the complex system as a simple equivalent circuit. However, it is critical to observe that system operating conditions could deviate significantly from the assumed conditions if either (1) load deviates significantly from the assumed or (2) a piece of major equipment is out of operation. In either case, the primary controllers may not perform their function most effectively because their gains are not adjusted to respond effectively to the new conditions. Re-adjusting these gains in anticipation of or in response to the unusual conditions forms the basis for some form of adaptive control to enable the most effective use of system resources over the wide range of changes in its conditions.

Advanced logic using local measurements in typical primary controllers (e.g., using acceleration measurement in addition to frequency) could contribute significantly to the power plant's ability to adjust better and faster to the changing conditions—such measurements might, for example, result in the development of a controller that would return the system back to its normal frequency and voltage by adapting its power output faster and more effectively. Also, much could be gained from at least partial communication among the primary controllers. This is particularly critical for detecting qualitative changes in system conditions, in order to trigger the change in the logic of the primary controllers determining more adequate settings for the new system conditions. There exist representative examples of the potential gains from wide area measurement systems for stabilizing system conditions under forced outages that would be otherwise uncontrollable without coordination among several controllers; on-load tap changing transformers for regulating voltage at the receiving end of the line have been known to be critical in several voltage problem-related blackouts in France and Belgium; however, the change in their logic when the operating conditions are qualitatively different could potentially prevent these blackouts from occurring.

These seemingly small improvements in primary control of power system equipment could be shown to enhance the ability of the system as a whole to withstand unpredictable difficult conditions without collapsing. The major challenge here is the design of control logic and supporting communications to implement advanced primary control in a robust way—that is, if the controller and/or its communication fail, the system as a whole will still not be affected in a detrimental way. Since this is hard to do with provable robustness, the more common way is

to enhance system performance by adding major redundant hardware for stand-by reserves. Establishing a balance between flexibility and over-design has always been the challenge in EPSE.

2.3 System Protection

Under some major forced outages, the system may not be able to return to the nominal frequency after a fault is cleared. In particular, if a fault is not cleared within the so-called critical clearing time, the basic integrity of the operation is endangered. Computing critical clearing time for complex electric power systems still presents an enormous technical challenge.

At each time-step, the proper commands are generated for correcting the abnormality or protecting the system from its consequences. If no abnormality is observed, then the normal operation proceeds for the next sample interval. If the frequency deviations are large and sustained, the under- and over-frequency protective relays automatically disconnect power plants at which the unacceptable deviations occur. Similarly, if a large transmission line goes out of operation, this could lead to current overload in the neighboring transmission lines. A typical scenario here is that a line becomes shorted as a tree touches it, resulting in an unacceptably large current passing through it, and the over-current protective relay of that line disconnects the line to prevent it from further damage. This further activates the over-current protective relays of other lines and they in turn become disconnected.

Protective relaying plays a tremendous role in preventing particular pieces of equipment from becoming damaged under unusual operating conditions. Recall that protective relaying has evolved around the objective of ensuring the safety of individual pieces of equipment under abnormal conditions. For example, transmission lines are allowed to carry only a certain maximum current, blades on turbines can turn only at the speeds specified by the design specifications, and appliances in households operate without being damaged if the voltage applied to them is within the prespecified thresholds of deviations from the nominal voltage at which they are intended to operate. The principles of setting thresholds on relays have evolved over many years and are often system-specific. The relay settings are not changed dynamically to adapt to the changing operating conditions.

For these reasons, relays are key to the cascading events causing major blackouts. The relay settings and their coordination are critical to the efficient use

of the system as a whole. While it is often difficult to differentiate the role of relays in protecting individual pieces of equipment from their role in the system as a whole, it is known that the malfunctioning of protective relaying has played a critical role in several major blackouts. A typical scenario of a blackout of this type starts by a major piece of equipment undergoing forced outage, which, in turn, causes unacceptable over-currents, over- or under-voltages, or unacceptable frequency deviations somewhere else in the system. This situation then triggers the disconnection of the protective relays in the pieces of the affected equipment for safety reasons, which creates further imbalances in the remaining portion of the system, where the protection disconnects the remaining pieces of the equipment. This ultimately results in the disconnection of all pieces of equipment, known as the system blackout or collapse. An even more difficult scenario is related to the false tripping of a relay, which disconnects a piece of equipment without a forced outage actually causing it. Once this piece of equipment is disconnected, the same cascading process of disconnecting the affected pieces of equipment throughout the systems begins. This scenario is known as a “hidden failure.” The need to rethink the overall concept of protective relaying in light of system needs and progress in fast computer/communications technologies was recognized some time ago; it has been only recently that this direction is being revived.

Because of this critical role that protective relays play in affecting the interactions among various pieces of the system, it is important to improve their logic and make them both error-free and more adaptive to the changing conditions. Similarly as for primary controllers, this is possible either through a redundant design or through more communications among relays associated with the individual pieces of equipment as these interact to perform system-wide objectives. With the major breakthroughs in the communications technologies, significant progress is being made toward adaptive relaying. Whereas protective relaying could be viewed as the extreme case of primary control, the two problems have not been viewed this way in the past.

2.4 Structure-Based Hierarchical Operation under Normal Conditions

A typical electric power system of the structure described above has millions of nonuniform components and local primary controllers and relays. One of the major goals of EPSE is to design power system

control for achieving a desired prespecified performance. Despite many years of research and development, it has remained difficult to provide a reliable, flexible methodology for automated tuning and switching of a variety of control equipment spread throughout the system. In addition to the problem of high dimensionality, the development of such general tools is particularly challenging for this class of systems because of the nonuniformity of control components, reflected in their uneven spatial distribution throughout the system as well as in a large spread of response times.

Instead of developing the most general computer tools for any possible condition, current operating practice is much simpler and different under normal conditions than under the forced outages. Under normal conditions, the power systems are operated hierarchically by various control centers and their on-line interactions. Typically, decisions are made by individual control centers about the status of the area (utility, control area) under often strongly implied or explicitly understood assumptions about the status of the rest of the interconnection and their interactions with the neighboring areas. As explained later in more detail, the entire hierarchical control of today’s electric power systems is based on the assumption that any system change would primarily affect the system only locally and that the effects on the rest of the system are secondary. This underlies the decentralized logic of the primary controllers and relays; it further underlies decentralized decision-making by individual control areas within the overall system interconnection. These simplifying assumptions, generally valid under normal operation, lead to the hierarchical control with spatially and temporally separable sub-objectives within an otherwise very complex system.

The critical assumptions under which the interconnection is planned and operated are generally violated under some abnormal conditions associated with forced outages and/or large deviations in load demand from its forecast. When the system enters such conditions, many of the software tools routinely used under normal operation are disabled by the system operators and the main decision maker becomes a system operator. In addition to adequacy problems, i.e., the failure of power plants to supply power, the other major aspect of abnormalities concerns unusual power flow patterns through the delivery system that are very different from the conditions assumed when primary controllers and protective relays are set. It is, therefore, at least in concept, clear why these “malfunction” under

emergency conditions. Under forced outages, system survivability is the prime objective and the cost is secondary. This means that in regulated industry different control areas cooperate by sharing stand-by reserves needed for emergency conditions. The system interconnection as a whole is considered to be viable under contingencies when there are no widespread failures affecting service to a large number of customers and the system grid remains intact; i.e., conditions that may trigger disconnecting pieces of equipment automatically for safety are avoided. The entire concept assumes no active participation by the end users in emergencies.

Because an electric power system is operated very differently in these two modes, one of the largest challenges to effective power systems engineering tools has been to classify the system status according to how close it may be to either its normal condition or the abnormal states. This challenge was conceptualized for the first time by DyLiacco, who introduced the classification of power system operation shown in Fig. 2. There are many degrees and many timescales of abnormal operation that are shown in Fig. 2. Conditions are generally assessed as normal when they are stationary and as expected. All equipment is working that is supposed to work; loads, fuel supplies, water, and weather conditions are within the expected ranges. This classification was later enhanced, but its overall philosophy has remained as initially proposed. Using this classification of system status, a system operator is in a position to schedule power for the forecasted demand and leave the rest to automation at the secondary and primary levels or override the automation and rely on an expert system knowledge about the specifics of the area of interest and act in real time to prevent major system-wide failures. It is worth observing, however, that the boundaries between normal and abnormal conditions are not straightforward to establish, making it necessary to use this classification always

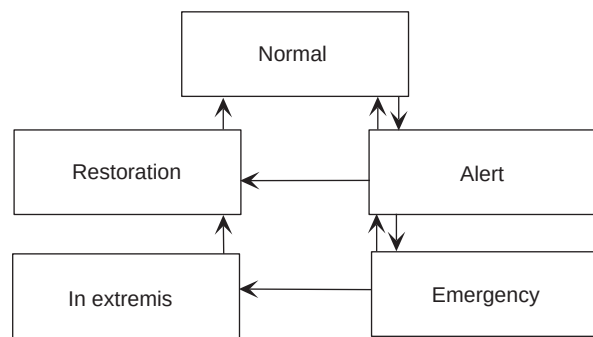


FIGURE 2 Classification of power system operation.

with caution. Requiring much margin between normal and abnormal conditions generally results in subefficient use of the existing resources; also, pushing the system to the boundaries of its secure operation for efficiency reasons could result in less robust and less reliable operation. Balancing these two conflicting objectives is generally very difficult.

2.5 Operating the System by Decision and Control

To actually operate the system within any of the basic conditions defined in Fig. 2 and during intentional or spontaneous transition between them inherently needs to be done in time-steps where the system state is identified at each time and a decision is made regarding control action to be made at each time-step. After these control decisions are set up at each bus system-wide, then the system responds during the interstate intervals, monitored appropriately to provide the basis for decision at the next sampling time.

The system-wide flow of information and action for this process is of immense size and complexity. Its basic structure is summed up in Fig. 3. Note the two columns for decision and for control at each sampling time. The information base—the system-wide measurement data—is indicated as entering at the bottom of the decision column, where the first task is the classification of the system state into basic operating modes of Fig. 2 or the transition between them. Identifying the position of the current system condition within this column represents the decision regarding the phase of the operation. This information is then passed over to the appropriate block into the control column of Fig. 3 for appropriate control action, which then is passed to the appropriate equipment within the system as indicated in Fig. 3. The equipment then carries out the action called for in system-wide control.

Figures 1 through 3 give a summary of the structure and functioning of a nationwide electric power system. The details and their complexity and volume of the contents of this gross frame of the national power system and economics are immense.

3. EPSE TOOLS FOR SYSTEM OPERATION UNDER NORMAL CONDITIONS

EPSE tools have been developed to accommodate the hierarchical structure of typical electric power

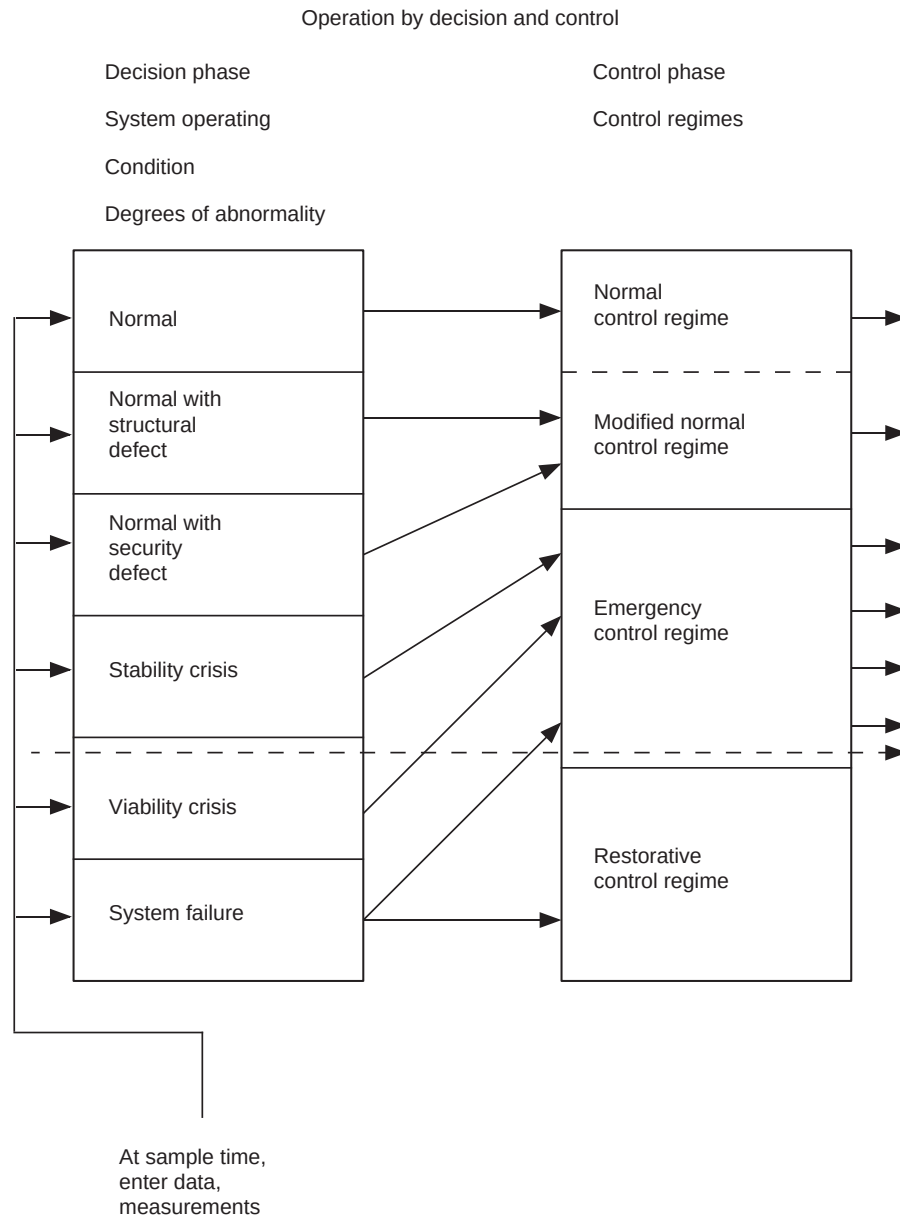


FIGURE 3 Basic structure of system-wide flow of information and action. The dashed line indicates an example of a signal path.

interconnection shown in Fig. 1, horizontally structured into utilities with their own sub-objectives and loosely coupled with the neighboring utilities for sharing generation reserves in case of major equipment outages within the entire interconnection. System monitoring, control, and protection methods are based on these hierarchical structures.

Principal computer-based algorithms for this condition in a decision phase are as follows:

- monitoring and estimation of load and generation;

- static state estimation;
- monitoring of the system loading conditions and security.

Some principal algorithms used for controlling systems under normal operation include the following:

- unit commitment (UC);
- economic dispatch (ED) or optimal power flow (OPF);
- automatic generation control (AGC) to match area real power supply and demand automatically as real power load varies slowly;

- automatic voltage control (AVC) to match area reactive power supply and load and voltage profiles (in some parts of the world only);
- load management.

3.1 System Monitoring

Modern electric power systems are monitored in normal operation to estimate load variations, the status of generators (functional, out for maintenance, or under forced outage), and the status of transmission lines (connected to the rest of the system or shorted out). This information is used as an input to the decision and control phase to balance supply and demand as the system status and levels of load demand vary. In a specific utility area, load variations have typical periodic patterns: annual, seasonal, monthly, weekly, daily, and hourly. Different computer-based methods are used to forecast these system load variations; the long-term forecast (annual and longer) is the hardest and least accurate and the shorter term forecasts are more accurate.

Much progress has been made since the time static state estimation was first introduced to the area of electric power systems. However, this type of estimation is still in use. The basic principle is to use on-line redundant transmission line flow measurements, which estimate false data on the system. The most challenging false data are caused by the incorrectly assumed status of equipment. After many years of researching and developing state estimators, many control centers have them in on-line operations.

The system is also monitored for its proximity to its safe operating limits. The system is generally considered to be in a secure state if none of the operating conditions are violated and the status of the equipment is as expected. The System Control and Data Acquisition System (SCADA) is the basic system in support of area monitoring and decision-making in each control center.

3.2 System Decision and Control

As the loading conditions, power plants' availability status, and the status of transmission lines vary over time, system operators rely heavily on computer-aided software for adjusting overall system resources as needed to supply customers, while maintaining currents/voltage/frequency within the prespecified acceptable threshold of variations. The basic objective of system decision and control in normal

operation is to cover the consumers' demand for power using generating sources available throughout the system.

3.2.1 Objectives under Regulated Industry

- The total operating cost of the full system including transmission losses is minimized.
- The price the consumers pay for power is subject to government regulation and control in such a way that the total cost of estimated power production and delivery is covered with adequate profit margins.
- There is enough power available in the system to cover current load estimated statistically using the past year's records plus adequate reserves to cover deviations of the current load from the estimated values and for emergencies of a first-contingency class.
- The power quality is adequate for consumer satisfaction; that is, voltage and frequency are kept within an adequate viability band and kept sufficiently steady. Note that synchronous time frequency is a courtesy to customers. Much less precision would be enough for engineering purposes.
- The system has security that minimizes outages.

3.2.2 Hierarchies in Regulated Generation Control

The only major changes considered in normal operation are the result of load dynamics. Therefore, power systems dynamics under normal conditions are driven primarily by changes in demand. Typical demand curves exhibit several rates of response, ranging from very fast random variations (order of seconds), through hourly, daily, weekly, and seasonal patterns of larger deviations. In today's hierarchical organization, generation-based scheduling is open-loop for balancing anticipated load on weekly, daily, and hourly bases; the secondary level of control (AGC and/or AVC) is closed-loop, automated, and responsive to the voltage and frequency deviations on a minute-by-minute basis, as the unexpected slow and small load deviations in real and reactive power occur. This power balancing on a minute-by-minute basis is performed by each control area within an electric interconnection for the assumed transmission line flows to the other neighboring areas.

There is very little or no coordination of control areas to adjust their line flow exchanges in near real time; instead, control areas agree how much to exchange for the following day and jointly adjust

their internal resources to implement these agreements. However, since the transmission line flows are not directly controllable in typical AC systems, variations from pre-agreed upon schedules take place in real time. This has not been considered to be a big problem, and utilities have established ways to pay back this inadvertent energy exchange approximately once a month, without any major disturbances to the overall system operations, since the system is operated according to the preventive security criteria for the given regional design. As the line flow patterns between control areas deviate for economic transfers—creating considerably different patterns than originally anticipated when the regional system was planned—mainly for implementing energy trades across control areas, the need for on-line minimal coordination of exchanges between control areas at the regional (tertiary) level and even among the regions may become necessary. Several blackouts, including the August 2003 blackout in the Northeast, clearly indicate the need for such monitoring.

Variations of the hierarchies and their automation can be found throughout the world; for example, the system in the United Kingdom does not have automated frequency control, and, moreover, since the entire system is effectively a single control area, tertiary and secondary levels merge into a single level. On the other hand, the European Interconnection is similar in structure to the U.S. horizontal structure, with the control areas corresponding to the countries, instead of particular utilities, respectively. As the electrically interconnected systems open increasingly to the free energy trades across the traditional control area boundaries, the need for advanced EPSE coordinating tools is becoming more pronounced. The blackout of the Italian electric power grid in September 2003, much in the same way as the U.S. August 2003 blackout, was also related to the failure of pieces of equipment outside the affected network; in this case, it was triggered by the failure of an EHV transmission line interconnecting Italian, French, and Swiss subsystems.

3.2.3 Temporal Separation of Basic Control Functions under Normal Operation

The normal-state standard system control functions are summarized as follows:

- Time frame function;
- 2–3 s Inertia, loads, excitation systems;
- 7–10 s Governors;

- 5–10 min AGC;
- 5–10 min AVC;
- 5–30 min ED/OPF;
- 1–10 h UC, restarting and shutting off units.

3.3 System Control

During normal operation, real power and voltage control functions are separable within a certain decoupling error. The operating practices described here are developed assuming this separation.

3.3.1 Real Power Unit Commitment and Economic Dispatch

The unit commitment and economic dispatch are basic computer-aided tools for optimizing total variable operating costs in industry. This is accomplished by using the least expensive units to supply base load and more expensive units for load following and (load) frequency control. At a control center, the most cost-effective units are scheduled on and off to optimize generation use over daily or weekly time intervals in accordance with coordinated unit commitment. The function is open-loop, performed only for anticipated, not actual, system conditions. The longer term economic efficiency is determined by how power plants are turned on and off and cost optimization is achieved by performing economic dispatch as the information about system loads becomes available through a SCADA system.

3.3.2 Automatic Real Power Generation Control

Of course, the actual current load will virtually never match its statistically estimated value for which generation dispatch is done, and the system frequency is also subject to small fluctuations. Consequently, some form of on-line control will be needed to keep the system frequency close to nominal. The specific control generally depends on the horizontal structure of the given electric power system and it could range from a manual frequency control of an isolated system, as in the United Kingdom, through a hierarchical generation control in a multi-control-area setup, as in the United States.

An important observation concerning frequency control is that even the most advanced hierarchical frequency control is strikingly simple with respect to its on-line information and coordination requirements. The area control error (ACE) is the basic measurement to which AGC reacts to balance supply and demand in each control area in near real time. It is defined as the sum of the weighted deviation of the

control area frequency at each particular sampling instant from the nominal system frequency and the deviation of the net power flow exchange with the neighboring areas measured at the same time from the pre-agreed upon net tie-line flow exchange. This entirely decentralized balancing of supply and demand in the entire electric interconnection comprising several control areas is based on the fact that when the ACE is zero for each area, the system is running exactly on schedule. Therefore, if each area independently controls its generation economically to drive its own ACE to zero, then:

- The actual load for the system is covered;
- The system frequency is as scheduled;
- All net tie-line exchanges are at their scheduled values area by area, thus satisfying the scheduled trades;
- If an area is unable to supply its scheduled net tie-line exchange and drive its own ACE to zero, then system frequency will drop and net tie-line exchanges at other areas will rise to cover the deficiency in the load of the deficient area.

3.3.3 Basic Hierarchical System for Real Power Control

- The top (tertiary, regional, interconnection) level simply adjusts the setpoint frequency to drive the actual system frequency and synchronous time back to normal. This one piece of information, the setpoint frequency, needs to be communicated to all areas. This minimal central action will rearrange the area net tie-line exchanges to balance the entire load. This is known as time-error correction in the regulated industry structure.

- Each area individually acts on the change of frequency setpoint by keeping its ACE to zero. No communication is needed between the areas.

- The system is fully decoupled; the areas do not need to communicate with one another or the center. The center needs only to watch the synchronous time to generate its single control command.

- Primary controllers (governors and power system stabilizers) stabilize very fast transients as the power supply/demand balancing at each control area level is performed through AGC.

It is important to recognize the role of accurate short-term load forecasting: the more accurately and more frequently the load is projected, the more efficient the use of available generation will be. The ED and UC are an integral part of a single complex function whose economic performance and technical

performance are the result of the actual interdependencies of UC and ED, on one side, and the AGC, on the other. For example, it was estimated that the main inefficiency when performing UC comes from leaving unused generation on some regulating units for frequency control. One could envision, with sufficient computing power, that generation control based solely on dynamic economic dispatch and stochastic unit commitment would be sufficient to balance supply and demand in real time without violating frequency quality specifications.

3.3.4 Reactive Power Economic Dispatch

For an interconnected power system to remain viable, the reactive power must balance at each bus. The problem of reactive power dispatch in the regulated industry is generally bundled with the problem of maintaining (load) voltages within the prespecified limits. The (generator) voltage setpoint values are optimized with respect to certain performance criteria subject to the reactive power balance constraints, the load-voltage acceptable limits, and the limits on generator voltages. Although the voltage control of an interconnected large-scale power system is widely recognized as a very important problem, its basic formulations and solutions are often utility-specific. In the U.S. interconnection, voltage control at the secondary (control area) and tertiary (regional) levels is viewed strictly as a static optimization problem. The most common tool for solving this optimization problem is an algorithm based on OPF.

Compared with generation-based frequency control, generation-based reactive power or voltage control is less standardized and less automated. Only a handful of countries in the world have semiautomated reactive power dispatch of generators.

The generation-based reactive power dispatch falls under the category of the OPF. Depending on the actual choice of the optimization criterion, two formulations are of interest:

1. The cost function is the total fuel cost required to supply a given demand.
2. The cost function is the total transmission loss that occurs as power is delivered from the power plants to customers.

An increased severity of voltage-related operating problems has resulted in an effort to make OPF-based formulations more directly related to these problems than the general OPF methods. Among many developments, the most notable are as follows: (1) an improved formulation of the decoupled OPF;

(2) OPF-based optimization of reactive power margins; and (3) an OPF targeted on improving energy transfer over far electrical distances.

3.3.5 A Hierarchical System for Reactive Power/Voltage Control

A qualitatively different approach to the voltage-control design is hierarchical and is based on temporal separation of reactive power fluctuations and corresponding control levels. Similar to the hierarchical real power-frequency control schemes, the approach is based on decomposing a large system into control areas equipped with decentralized closed-loop secondary-level voltage controllers. A close look into this approach indicates that the objectives of a hierarchical voltage control scheme are analogous to the objectives of frequency control; the main objective at the secondary (control area) level of voltage control is, for example, to update setpoints of generator exciters to maintain voltages at critical (“pilot”) points within the prespecified deviations as the reactive power demand and generation deviate from their scheduled (anticipated) values. The hierarchical control for the regulated industry, based on the systematic implementations in several European countries, is often referred to as AVC.

3.3.6 Load Management

It has been well understood for quite some time that having an active response to the changing system conditions seen in frequency and voltage deviations from their nominal values by the end users could significantly contribute to a more robust performance of the system as a whole. In basically the same way that the control area supply/demand imbalance (seen in its ACE) becomes reduced by increasing generation, one could decrease control area load and drive ACE back to zero. Responsive load management is particularly important at times of unexpected shortage in generation and transmission capacity, as well as for using less costly generation reserves instead of peaking units under normal operation.

Except for some early experiments known as demand-side management programs, customers have largely remained passive power consumers. As the industry restructures, the need for more system conditions and price-responsive load will increase. The early vision of homeostatic control, in which all end users react to local deviations in frequency and demand in an automated way, could go a very long way toward more efficient use of resources under

normal conditions and reduced reserves for emergency conditions. Sections 5 and 6 state that price-responsive load is the key to the well-functioning electricity markets; the difficult question is providing on-line incentives to the end users to respond at the value as seen by themselves.

4. EPSE FOR EMERGENCY OPERATION

The ultimate challenge to electric power systems engineering at its present stage is reflected in the lack of model-based analytic tools and algorithms for detecting abnormalities in system operation and for deciding what actions to take as the abnormal situation evolves. The classification of potential abnormal modes of operation according to the famous DyLiacco diagram shown in Fig. 2 and its later enhancements suggests that the following operating modes could occur: alert mode and emergency mode.

This section briefly summarizes the tools for monitoring and responding under these abnormal conditions.

4.1 System Monitoring

4.1.1 Alert Operating Mode

SCADA-based monitoring could detect a situation in which conditions are normal; the system is viable and within the acceptable operating constraints. However, the structure has been altered from what was expected by an earlier event—a transmission line is missing, for example. If this is detected, the operating mode is recognized to be under an alert condition.

4.1.2 Emergency Operating Mode

The emergency conditions could be reflected in several different system-wide problems, ranging from stability and viability crisis through system failure. This division, however, is not clear-cut. The conditions overlap, but normally one dominates. Take in sequence:

4.1.2.1 Stability Crisis The system is in a momentary dynamic state, typically caused by a fault, which is sufficiently violent to endanger the integrity of the system.

4.1.2.2 Viability Crisis The system in its present condition is incapable of operating within voltage,

current, frequency, etc., operating constraints, given the available generation and transmission capability for the existing load demand. This condition could result from a fault and is then often preceded by a stability crisis. Nonviable conditions can, however, come on quietly, for instance, if major generation or other equipment does not become available when it is scheduled to come on-line for peak demand.

4.1.2.3 Integrity Crisis The integrity of the system is violated—for instance, load was dropped or the portion of the system is dropped off. Principally, this is a severe and extensive accumulation of stability and viability crises or the damage left by them with islanding and blackout patches in the system. During the crisis phase, events are still occurring and a struggle is still going on to avoid as much disintegration as possible.

4.1.2.4 Restoration This is an extensive and collective control (generally interpreted) effort to remedy the damage left by any combination of the three crises.

4.2 System Decision and Control

4.2.1 Alert Operating Mode

The principal operating algorithms are as for the normal operating mode, but, in addition, future consequences of the structural change must be evaluated and, if necessary, remedied. For example, the structural change may violate security constraints while approaching peak load. Industry practice is preventive in this operating mode, in the sense that the system is operated so that it meets normal operating constraints when any one of the most likely and most severe unexpected structural changes (contingencies) takes place. It is often cost-inefficient because it does not allow for least-cost use of resources without considering future events. To circumvent this problem, the operating limits under structural changes are allowed to be less restrictive in the alert mode than in the normal mode. For example, typically accepted voltage variations under normal operation are within 2% or 0.002 per unit, and in the alert mode they are relaxed to 5%. When a more severe structural change occurs, it may bring the system very close to the limits of acceptable operating constraints. For such conditions, economic criteria are overruled by the security criteria. Different ancillary algorithms for readjusting network flows, bringing on new, more expensive

generation, modifying load dispatch, etc., are used to assist the operator under these conditions.

Moreover, since the system operation is still normal under the alert state, the control tools and actions of the normal state are used. The special alert-state control actions consist primarily of equipment switching and changes in control setpoints and schedules to improve security.

4.2.2 Emergency Operating Mode

The decision and control actions in emergency mode are often left to the human operator and are often dependent on the specifics of a particular control area or region. At best, off-line computer-based analysis is performed for scenarios of interest and particular procedures are in place in case the particular scenario occurs. Model-based simulations of such scenarios are generally very time-consuming and are therefore not feasible as on-line means of facilitating decision-making and control in emergency situations, when time is critical. One of the typical decisions concerns the so-called critical clearing time, i.e., the longest time of fault duration, so that when the fault is removed, the system can still return to its normal operation. The following are representative decision and control actions characteristic of an emergency operating mode.

4.2.2.1 Stability Crisis Some of the principal control tools can be:

- Local structural control or protective relaying to eliminate faulty components and local control actions to preserve system stability.
- Stability augmentation utilizing FACTS, technologies, breaking resistors, load skipping, etc.
- Load dropping or major structural changes to separate the system in the least objectionable manner when integrity cannot be maintained. Note that this action pushes the system into the integrity crisis or an *in extremis* condition.

4.2.2.2 Viability Crisis Principal algorithms in an approximate order of increasing undesirability include:

- Frequency reduction of the AGC type;
- Use of spinning reserves or cold reserves;
- Special measures, such as fast turbine run-back;
- Help from neighboring areas;
- Exploiting time-limited overload capability of the equipment;
- Special measures, such as voltage reduction;

- First-stage structural control, such as starting a new generation, returning equipment from maintenance, or load dropping (for minutes or hours);
- Second-stage structural control-islanding involving separation of a power network into parts, some of which are not supplied with power.

4.2.2.3 Integrity Crisis The tools are mostly the same as for the stability and viability crisis.

4.2.2.4 Restoration The main processes taking place during restoration are reconnection of the islands and restoration of loads. This is, of course, an operation by itself and it has been researched extensively over time.

5. CHALLENGES TO EPSE IN THE CHANGING INDUSTRY

Over the past decades, the electric power industry has been undergoing major organizational changes. It is critical to recognize that an electric power structure, along with the objectives and performance metrics used in various EPSE methods for monitoring, decision-making, and on-line operations of a complex electric power system, must be posed in the context of each specific industry structure. The interplay between technical, economic, and policy signals over various time horizons and among various industry layers is fundamental to the overall system performance as a whole. One of the more difficult tasks of EPSE even for fully regulated industry has been how to establish the right balance between system robustness to unexpected events, such as forced outages, and the efficient use of the existing equipment under normal conditions. In other words, the tradeoffs between system security/reliability and the economics of providing power to the end users have been generally difficult to mediate. The performance metrics of the system as a whole have been affected by the utility-dependent performance metrics within the system. Basic security criteria adopted by the regulated industry are discussed above; all software tools for monitoring and decision-making in control centers are based on these well-established objectives.

Electric power restructuring has brought about many fundamental changes to the way EPSE might be performed in the future. To start with, a once vertically integrated utility with a single objective of

supplying its native customers has become unbundled into power supply, transmission, and distribution businesses and/or functions. Delivery and consumption are becoming businesses with their own objectives, typically not aligned with the objectives of the power suppliers. In addition, the service is provided through some forms of markets, with marketers managing both the physical, forced outage-related risks and the financial risks with the objectives of making their own profits. Design of markets for various services (delivery, voltage, frequency control) in addition to the basic power market is at rudimentary stages. Providing electric power service competitively asks for complex new economic and policy designs, with a full understanding of their overall effects on the system and the unbundled market participants themselves.

The EPSE tools in support of electric power provision in these evolving industry structures are still rudimentary. One thing is certain, although not widely recognized: it is going to be impossible to continue using the same EPSE models, analysis, and decision-making tools developed for the regulated industry to manage effectively the power system under restructuring. This lack of understanding of the gaps between the engineering tools and market objectives has already caused unexpected outcomes. The remainder of this section briefly summarizes evolving EPSE concepts. The last part of this section provides a possible systems view for bridging this gap between regulated and restructured industry models. The section explains how potentially new paradigms for the electric power system may be based on distributed decision-making by all market participants actively pursuing their sub-objectives under incomplete information. Minimal coordination at various industry layers in order to ensure well-understood performance metrics is the ultimate goal. This challenge is briefly elaborated at the end.

6. OBJECTIVES OF EPSE AS A FUNCTION OF CHANGING INDUSTRY STRUCTURES

6.1 EPSE Tools for Vertically Unbundling Industry

6.1.1 Objectives in Deregulated Industry

- Suppliers can obtain the maximum price for their product competitively established in free

commodities market-type economy restricted only in the sense of the rules of exchange.

- Consumers of electric power can pay the lowest price for their purchase competitively established in a free commodities market-type economy restricted only in the sense of the rules of exchange. Note that this may work fine for large customers who can effectively trade at the commodities market to their advantage. Small customers cannot effectively do that, which raises the difficult question of how to establish the price of the power at individual homes, with the potential outcome to retain some kind of government regulation in this area.
- There is enough power available in the system to cover the current load of the system for emergencies of a first contingency class.
- The power supply quality is adequate for consumer satisfaction; that is, voltage and frequency are kept within an adequate viability band and sufficiently steady.
- The system has security to minimize outages.

6.1.2 Evolution from the Hierarchical Operation into Open Access Operation

In the changing power industry, balancing supply and demand in real time is generally market-based. Three qualitatively different market mechanisms are of interest: (1) the primary electricity market(s) for supplying anticipated demand; (2) various “ancillary” markets for ensuring that system conditions remain within prespecified limits as demand deviates in real time from its anticipated pattern; and (3) transmission markets for delivering power requested by (1) and (2).

These markets are necessary for ensuring that the system conditions remain within their technically acceptable limits as power is provided competitively. Here, only some fundamental changes needed in EPSE tools for serving the new industry are summarized.

In addition to the temporal decomposition-based hierarchies present in load dynamics, further temporal decompositions are emerging via the establishment of electricity markets whose on-line decision-making evolves at daily, hourly, and even near-real-time rates. Two observations are relevant here for effective on-line use of generation in deregulated industry: (1) the distinction between predicted and real time is no longer as pronounced as when unit commitment is done assuming demand to be known for the entire next day and (2) it is, therefore, conceptually difficult to “unbundle” the economic

generation scheduling process, assuming the predicted load from the on-line control function for responding to unexpected load fluctuations.

Furthermore, some customers are willing to accept power interruptions in exchange for a lower price of electricity—that is, generation is beginning to be provided by some customers on a nonfirm basis. This is a qualitative change from the utility’s unconditional obligation to serve and charge its connected customer the same price of electricity over time. This raises a fundamental question about the inherent temporal decomposition of load dynamics and the conceptual applicability of hierarchical power systems operation under competition.

Finally, a transmission grid is managed as a regulated monopoly in a multiownership environment, but is expected to serve markets (1) and (2), which are competitive. This asymmetry of regulatory structure within the single electric interconnection intended to be used in an open access manner, without recognizing ownership boundaries, is creating tremendous challenges for operating the transmission grid consistently. Because of this, systematic policy/regulatory frameworks for moving toward value-based transmission service need to be introduced.

6.2 EPSE for Operating the System under Open Access

Operating under open access implies, at least in principle, the ability for power suppliers to competitively serve customers across utility boundaries within an electric power interconnection. The hierarchical operation in regulated industry must be modified to facilitate delivering power across the boundaries of former control areas established in the old industry without endangering the reliability of the interconnection as a whole. As power is traded across the control areas, each control area must monitor these exchanges in order to balance the remaining needs for power to the customers that it serves as a provider of last resort. The interconnection is being used for power flow patterns different than the flow patterns expected under the originally intended system control and protection design. These market-driven exchanges could result in much stronger couplings among the control areas and, therefore, could require some degree of on-line coordination of flow exchanges between the control areas in order to operate the interconnection as a whole reliably. EPSE tools for coordinating control

areas while enabling power trading across control areas do not yet exist.

One possible approach to operating the interconnection under open access would be a market-based approach to (1) ensuring enough capacity, both generation and transmission, for reliable service and (2) giving incentives to the end users to adjust in an on-line way to the changing technical and/or economic conditions. An early vision of the 20th century electric power operation was introduced by Fred Schweppe: all end users react to variations in local frequency and voltage and the system remains in homeostasis, or balance. The success of this concept, borrowed from biology, must obey two key rules: (1) all end users must respond and (2) the right automation must be implemented. The ongoing industry restructuring has not yet provided a regulatory/policy framework in support of this technological vision.

Furthermore, although the technology for automated participation by the end users is available, it is not cost-effective when implemented by each end user, independent of its size and characteristics. Instead, what are available are cost-effective technologies for managing very large consumers or groups of smaller end-users. Although in concept there should not be any difference between one large end user of X megawatts and several smaller users whose total use is X megawatts, there is a very big difference between two end users of the same average X-megawatt energy consumption, where one end user is adaptive to the system conditions (by “shaving” its peak consumption by redistributing the needs to a different time of the day and/or by reducing its nonessential consumption during extreme system conditions) and the second end user is completely nonresponsive to the system conditions and requires an unconditional supply without considerations for the overall system conditions.

To implement price-responsive energy consumption on-line, communications begin to play a very different new role than in the hierarchical operation under regulation. For an open access system to work well, it becomes necessary to have on-line communication of important technical and/or economic information (1) between the aggregator of various end users and (2) between the aggregators and energy providers, transmission providers, and system operators as well as among several system operators within the interconnection. Note that, at least in principle, a bundled production could be provided by the sellers of power with the delivery service included or the

aggregator at the end user's side would look for the best combination of subservices as needed by the end users represented by this aggregator.

A highly effective paradigm of this type has not been actively pursued due to the lack of regulatory policy incentives to differentiate quality of service and value of these differences. In today's industry under restructuring, there is hardly any incentive to the end users to respond to the system conditions, since the service is cost-based with the costs more or less equally distributed among all end users according to their average X-megawatt consumption. The final part of this section provides a possible vision of futuristic EPSE tools that, given the right policy incentives, could lead to the value-based provision of quality of service under open access.

6.3 EPSE for Risk Management under Open Access

As the industry reorganizes, it has become necessary to differentiate between the physical and the financial risks. Major physical risks concern system-wide failures like that experienced during the 2003 black-out, reflected in a fundamental inability of the system to operate as intended and to deliver according to the prespecified performance metrics. This could take place for a variety of reasons, such as (1) under-design of system hardware for the perceived needs; (2) poor on-line IT-based monitoring, control, and protection coordination of various individual pieces of equipment; (3) human error; and (4) nonresponsive power consumption. The society must take it as given that some very-low-probability, potentially high-impact triggering events will happen. It is, however, not obvious at all how eliminating any of the reasons (1)–(5) would fully guard against widespread system failures. The individual relays and breakers very rarely fail to disconnect pieces of equipment that they are responsible for. However, in order to minimize the effects of triggering events on the system as a whole, the tradeoffs between designing an “optimal grid” for the anticipated use under normal and abnormal conditions and developing novel EPSE tools for managing these physical risks under open access must be understood and managed. No matter how robust the hardware design in place, its system-wide use is not going to be as effective as possible without good software tools.

The role of EPSE software tools in managing physical risks in complex electric power grids has been grossly underestimated. Consequently, the tools are

very rudimentary and very little technology transfer has been attempted. In particular, feed-forward software tools for scheduling available resources for the predicted load are well established for the regulated industry, as reviewed in Section 3. The software and hardware options for managing resources under open access are much more complex than in the regulated industry. EPSE tools for distributed decisions by market participants and by the system operators, in particular with the load being an active decision maker, must be developed. The old industry has been characterized by being truly risk-averse. As the levels of risk-taking become more diverse under open access, EPSE tools must be developed as a means of decision-making by those taking risks.

6.4 Toward EPSE-Supported Protocols under Open Access

Given a strong movement toward the industry restructuring along with an emergence of cost-effective technologies for producing, delivering, and consuming power in qualitatively different ways than in the past, it is essential to develop EPSE tools for flexible provision of service to the end users according to their choice. Instead of pursuing the rigid hierarchical control of the past and/or over-designing the system, an IT framework would provide flexibility in delivery and consumption.

The underlying technological weaknesses that led to the major blackouts of 1966 (Northeast), 1975 (Northeast), 1996 (California), and 2003 (Northeast) are best corrected with information and control technologies. The first weakness is simplistic equipment protection and control. When system conditions evolve to threaten individual pieces of equipment, such as a line, transformer, or generator, protection devices respond by removing that equipment from service. This increases the stress on other parts of the system, possibly causing their protection to act, and so on. The result can be a large-scale cascading blackout. A better alternative would be to shift resources to aid the equipment first threatened, avoiding a blackout altogether. This requires that local controllers of power plants, transmission lines, and consumer consumption act quickly in a coordinated manner to relieve system stress before protective action is required.

Unfortunately, controllers cannot identify emergency conditions and are not flexible enough to act in a coordinated manner, and the protection devices lack the information and decision-making abilities

required to postpone their actions. The second weakness is the inability to quickly detect that the system is not operating as desired, because of equipment failure or unusual power demand somewhere within the very large and complex electric power grid. In particular, if individual operators could quickly recognize that their region is under stress, then they could adjust their own resources fast enough to keep operation under control or at least minimize the extent of a blackout. Today's operators do not have the information to do so because there is no on-line monitoring at the regional level or higher. Taken together, these two weaknesses result in a lack of information, flexibility, and coordination that will inevitably lead to future large-scale blackouts as demand increases. With economically minimal investment in information and control technology, this need not be the case.

To improve system operation, the protection and control of equipment, such as power plants, transformers, transmission lines, and loads, must first be improved. This is a challenge because the system response to protection and control actions under stress may be different than under the conditions for which the protection and controls were initially built and tuned. Typical assumptions under normal operation are that (1) the local effects of protection and control are much stronger than their effects elsewhere on the system and (2) producing more power when frequency and voltage drop always helps. These two assumptions are frequently violated when the system is under stress, requiring more on-line information about the rest of the system and more advanced logic to decide how to act.

Next, an on-line coordinated exchange of power between utilities and larger regions of the electric power grid in the United States and Canada must be implemented. This will allow the system operators to better manage large-scale system operation. Managing the entire interconnected grid on-line as a single system is not possible for a variety of technical and business reasons. Instead, a multilayer IT-based framework is required for the system to support the delivery of power between arbitrary generators and consumers. This framework must (1) allow all system users to adjust to changing conditions; (2) allow all consumers to communicate their short- and long-term demands for power and its delivery; (3) have clearly defined technical and contractual responsibilities for the performance of all users; and (4) monitor and control power exchanges between all utilities and regions. Such a framework would facilitate the evolution of the rigid grid of today to an

Internet-like grid in which the users decide when and how to use the system. Not only will this improve reliability, but it will permit value-based reliability in which users pay more (or less) for more (or less) reliable power.

The same IT-based framework for power system management can also support long-term forward markets for reliable service by providing current and future system needs for that service. The money paid in advance for reliable service could be invested in the information and control technology required to implement it. Having forward reliability markets would avoid unpopular regulatory decisions to increase the price of electricity without a good understanding of customer needs for reliable service and their willingness to pay for it. It would also promote the value-based evolution of the grid according to customer needs.

An adequately reliable power system will most economically come from investment in information and control technology. Improvements in system protection and control, built on an IT framework for on-line reliable service, and the forward markets for reliable service complement each other. One without the other does not work and both are long overdue.

7. A GLIMPSE INTO THE FUTURE: MORE SMOOTH AND SECURE ON A BUILDING BLOCK STRUCTURE

7.1 The Current Bulk Power Transmission Network

The electric power system in its present state is a dynamic system of enormous size coast to coast, operating at an equilibrium state that is stable under the slight perturbations imposed by normal load variations. However, faults or mistakes must always be expected even in the best of equipment—and even in humans. An abnormal sudden change like the loss of a large power-generating plant or a large, vital transmission line abruptly changes the system equilibria, including the current operating state, potentially drastically. While the system is trying to move automatically (too fast for human interference) to a new equilibrium, the huge difference in energy present in the network before and after the change between the two operating equilibria is left to swing freely over the system in giant energy waves, the postfault transient oscillation, which challenge the current, voltage, or power-carrying capacity, i.e., the backbone strength of equipment such as transmission

lines, generators, and transformers. It is the system operators' duty to maintain a loading condition by their dispatch activity at all times for the existing loading of the system without system breakup for all fault contingencies. Through existing advanced technology including power electronics, computer, and communication, this effort can be successful most of the time, but not always, and one should not forget the new emergence of terrorist possibilities.

7.2 Reorganizing Bulk Power Transmission Network to Building Block Structures Individually Controlled by a Computer Control Center

Drastic dangers call for drastic remedies. The authors' group, led by John Zaborszky, has been developing such a drastic remedy by mostly retaining its currently used power equipment, but rearranging the structure of the interplay of the power system and its controlling computer and communication system for maximum effectiveness.

Recent events served as a reminder that the current power system is vulnerable to widespread failure caused by the failure of even one major component. This vulnerability derives from the fact that after such a failure the system becomes a substantially different one and the operating equilibria cease to be equilibria in the new system. However, the states in the new state space will generally lie in the region of attraction of one of the new equilibria, which in turn will lie quite close to their existing ones except in the disturbed area, so they should move to their new position through a dynamic transient unless the system is becoming unstable. The latter results from the fact that each of the components (building blocks) carries a large amount of energy at an equilibrium and in AC this energy is swinging across any connection at 120 cycles.

So, at either of the equilibria (before or after the event), each building block will carry two components of energy:

- Type 1: the steady state energy content of the building block in the new system after the event.
- Type 2: the difference between the energy content of the building block at the time of the event and the type 1 energy content.

Each of these should be available in the computer control center as it is planning the postevent operation of the system after the event.

7.3 The Outline of the Management of the Envisioned Approach

The proposed new, building block (for example, power plants or generators, or major loads, such as aluminum plant, capacitor banks) structure eliminates any direct control interaction between individual building blocks even if they are neighbors but not their physical dynamic interconnection through the bulk power network system. The only information and control command exchange exists between the control centers and individual building blocks before and after the disturbance.

Each of the individual building blocks in their current physical state contains an amount of energy of type 1. This energy will differ from the energy at the state assigned by the control center to the blocks. In effect, the distribution of the type 1 energy defines the operating state of the system.

The excess energy (type 2) becomes the free swinging power in the system postevent transient oscillation, unless it is eliminated by draining it out or in locally at each of the bulk transmission buses into storage capacitors and then fed back smoothly into the system by routine dispatch (such storage already exists sporadically in the power system). While the excess energy is thus drained out or in (if available in storage) at individual buses, the system voltages and current are sliding on smoothly under dispatch control without wild swinging with only a perturbation ripple, provided enough generation is available. If not, i.e., in an emergency, the central computer control will order dropping of loads in priority order and then they will be reconnected when possible.

In fact, the system may already be broken up into disjoint pieces (e.g., in a disaster situation). If so, the building block structure allows the control centers to make the separate pieces by themselves viable by load dropping and set them up like “individual power companies” to run independently of one another while still connected to the control center until their regular connection can be restored. This approach outlined above can become an effective defense in emergencies of the 9/11 or the New York 2003 type.

7.4 Summary of the Process for Building Block Use Operated and Controlled by the Computer Control Center

In summary, the envisioned building block structure of the power system has the following distinguishing features:

- The system is restructured as a mosaic of building blocks, such as generators, power plant buses, and transmission lines.
- Building blocks are connected only to their neighbors.
- Each block has only one single control command connection going to the system’s computer control centers. There is no control command connection, i.e., input, directly between two building blocks, even if they are direct neighbors physically.
- The one single communication between each individual building block and the control center carries full report of the dynamic state of that particular building block to the control center and control command instruction to the block.
- The control center sends to each building block a command to adjust one particular system state variable (located inside in that block) to an assigned value, using a single-variable feedback loop control producing a single control signal input generated independently within the building block for setting the state variable of the block.
- On the basis of block-by-block information received by the control center about components of the system state in each block and the information the center has about the demands on the system, the control center computes the type 1 energy content required for stable operation of the system in each of its blocks.
- Note that this computation is simply the steady-state stable “load flow” for the system.
- Block by block, this information is deducted from the present energy contents of the block.
- This difference is drained in or out (depending on whether it carries a plus or minus sign) of the storage equipment at each particular block.
- The result is that the system has slid over into its desired steady state value.
- No violent transients exist to cause new failures.
- Note that on an all-DC bulk power system, this result derives automatically but it applies as described on a conventional mixed AC and DC bulk power system—with the addition of some more power electronic equipment.

In either case (all HV-DC or conventional bulk transmission), the dynamics of the system trajectories in the state space will change instantly and smoothly to that of the postevent system composition without transient waves but the energy controls of the energy storage devices (typically capacitor banks) will shift virtually unilaterally, to its post event form.

Note the following:

- The above assumes the system to be linear around its operating equilibria and that its response is instantaneous, assumptions that are true only approximately. It needs to be proven that on the actual systems this would cause only a tolerable perturbation around the operating point in case of faults.

- This approach exploits the full potential of power electronics to achieve maximum possible smoothness and security of the system operation without violent transient swings that can cause the spreading of major faults (e.g., East Coast 1965, 2003).

- This, however, requires installation of some power electronics, computer, and storage equipment, which need to be properly analyzed.

- The economy of such full restructuring of the current bulk power system needs to be established to fully utilize the current technology of power system electronics, computers, and communications, even if it requires only the addition of a moderate amount of equipment but major restructuring of their operation into a building block base.

In conclusion, as briefly summarized here, this approach assumes linearity and instant response (both of which can only be approximated, leaving the question: how well?) and assumes that it would be possible to drain out the type 2 energy at each building block so that the system would find itself in the new equilibrium position with no transient. Without any such draining process, as is happening today, the type 2 energy differences create a violent transient, which is proven to cause new faults since it creates highly abnormal voltages, currents, etc., on the equipment until it is dissipated through the line losses and radiation.

With what is proposed here, the system—that is, the bulk power transmission network—would possess a smoothly varying equilibrium state without any over-voltages or currents and hence without the spreading of the disturbances and system breakup following a major fault as in New York 2003 or a 9/11-type attack.

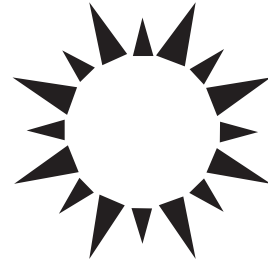
SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

- Amin, M. (Ed.). (2002). Complex networks. IEEE Control Systems Magazine, February 2002.
- Chen, J., Thorp, J. S., and Parashar, M. (2001). "Analysis of Electric Power System Disturbance Data," Hawaii International Conference on System Sciences (HICSS-34), Maui, Hawaii, January 2001.
- Cohn, N. (1977). "Research Opportunities in the Control of Bulk Power and Energy Transfers on Interconnected Systems," Special Report on Power System Planning and Operations: Future Problems and Research Needs, EPRI E-377-SR.
- Fink, L., and Carlsen, K. (1978). Operating under stress and strain. *IEEE Spectrum* March, 48–53.
- Ilic, M., Galiana, F. D., and Fink, L. (Eds.). (1998). "Electric Power Systems Restructuring: Engineering and Economics." Kluwer Academic, Dordrecht, the Netherlands/Norwell, MA.
- Ilic, M. D., and Liu, S. X. (1996). "Hierarchical Power Systems Control: Its Value in a Changing Electric Power Industry, Advances in Industrial Control." Springer-Verlag, London.
- Ilic, M. D., and Zaborszky, J. (2000). "Dynamics and Control of Large Electric Power Systems." Wiley-Interscience, New York.
- Interconnected Operations Services. (1996). NERC WG Report.
- Joskow, P., and Schmalensee, R. (1983). "Markets of Power: An Analysis of Electric Utility Deregulation." MIT Press, Cambridge, MA.
- Schweppe, M., Caramanis, F., Tabors, R., and Bohn, R. (1988). "Spot Pricing of Electricity." Kluwer Academic, Dordrecht, the Netherlands/Norwell, MA.
- Zaborszky, J., and Ilic, M. (2001). Exploring the potential of the all DC (ADC) bulk power systems. In "Proceedings of the Bulk Power System Dynamics and Control V," August 26–31, 2001, Onomichi, Japan." [Plenary paper]
- Zaborszky, J., and Ilic, M. (1999). "General Structure of the Regulated, Deregulated and Other Generation Dispatch." IEEE Power Tech '99 Conference, Budapest, Hungary, Paper No. BPT-440-21.

Electric Power: Traditional Monopoly Franchise Regulation and Rate Making



KENNETH ROSE

Institute of Public Utilities

East Lansing, Michigan, United States

-
1. History of Traditional Regulation
 2. The Utility Franchise
 3. The Regulatory Process
 4. The Revenue Requirement Formula
 5. Rate Base
 6. Rate of Return
 7. Expenses
 8. Cost Allocation and Rate Design
 9. Limitations of the Traditional Approach
 10. Summary and Conclusions

Glossary

cost allocation The process of allocating commission-approved utility cost of service among customer classes based on demand characteristics or other similarities.

cost of capital Percentage return that investors require for the use of their money.

cost of service Total amount of dollars needed to supply the utility's customers, including operating and maintenance expenses, taxes, depreciation expenses, and a fair return on investment.

franchise Exclusive territorial rights granted by state or local authority.

prudent investment test A regulatory standard that requires that costs recovered from customers by a regulated public utility must have been reasonable given what was known or knowable at the time the costs were incurred.

rate base A measure of a utility's total property value, including plant, equipment, materials, supplies, and cash; excluded are accumulated depreciation, amortization, deferred taxes, and customer-provided capital.

rate case Formal process used by state and federal commissions to determine and set rates.

rate design The rate structure or tariff schedule devised for each customer class or subclass; it is intended to collect

the cost of service allocated to the customer class or subclass.

rate of return The return on investment received by the utility or its investors; the allowed rate of return is the commission-approved percentage return on the value of the rate base the utility is given an opportunity to earn on its investments that are used and useful for public service.

revenue requirement The total expenses, or the cost of service, determined by a commission to be needed to provide service to the customers.

traditional regulation A form of economic or price regulation conducted by a governmental agency or commission that estimates the regulated firm's cost of doing business and determines a "just and reasonable" price for consumers based on approved costs, including a fair return.

used and useful test Regulatory standard for utility investments and expenditures that must be met to be included in the rate base or recovered as expenses from customers that specifies that investments and expenditures must be used and useful for public service.

Traditional regulation as applied to electric utilities is a form of economic or price regulation that evolved and often changed form over a period of more than 100 years. Also called rate base, cost of service, cost-based regulation, or rate-of-return regulation, this form of regulation is a method that attempts to estimate the regulated firm's cost of doing business and derive a "fair" price for consumers. The concept of public utility regulation was developed from the common law prohibition of unfair competition. It was applied to many industries, including the railroad industry beginning in the mid-19th century and later urban transit (streetcars and buses), telephone, water transportation, manufactured and

natural gas, and water utilities. An important goal of this type of regulation is to balance the interests of the firm while protecting customers from monopoly prices and other abuses. The term “traditional regulation” is used here with the caveat that many industries and state and federal regulators have abandoned or significantly modified this basic form of regulation during recent years for one or more sectors. Charles Phillips, Jr., noted that utility regulation is an economic, legislative, and legal concept. These themes are repeated often when regulatory questions arise. Economic concepts are used for rate design, determination of rate of return, and analysis of company incentives for cost control and behavior. Legislative involvement from the enabling statutes, statutory changes made in response to economic conditions and technology changes as well as through the political process, and legal concepts form the basis of regulation and are grounded in constitutional law and set the boundaries for commission authority and utility actions. Other disciplines are also involved in influencing regulation, especially accounting through accounting standards and reporting requirements and engineering through the prominent technical complexities of power generation and delivery.

1. HISTORY OF TRADITIONAL REGULATION

There is some difference of opinion about the origin of traditional utility regulation in the electric industry. Some have argued that it was due to the efforts of Progressive Era reformers who were crusading against monopoly abuses in many industries during the late 19th century. Others maintain that it was the industry itself that decided that regulation was better than either “ruinous” competition or the possibility of government acquisition. Because there is strong evidence to support both views, it is likely that its origin is one of those points in history where opposing interests converge to support a similar solution. However, once a common frame of reference was reached, the actual implementation of the regulatory model was often less congenial.

The regulation of the electric industry generally followed the changing structure of the industry. At the beginning of the industry’s history, there was little regulation and it was often thought to be unnecessary. Electric power was seen as a means to further economic development, much as railroads were seen

a few decades earlier. The main use for electric power during the late 19th century was for industrial purposes and a few wealthy households. Franchises were granted by municipalities, and rate limits were sometimes set by local governments. Early in the 20th century, most of the power was still generated for industrial use and was produced at the industrial site that used the power. However, central station power, or utility production, became more prevalent and eventually became the leading form of production by the 1920s.

Along with the rise of central station power production and the consolidation of the industry into larger regional companies due to economies of scale in the production and delivery of electricity, utility borders expanded beyond municipalities. As a result, regulation shifted to the state level, beginning in 1907 in New York and Wisconsin and spreading to nearly every state by the 1920s. Large regional and multistate holding companies formed, and investor and consumer abuses were rampant, peaking in 1929 with the financial collapse of the Insull companies. The congressional and regulatory response was to force the breakup of the large holding companies and to place restrictions on the operations and management structures of the remaining regional companies (through the Public Utility Holding Company Act of 1935). This marked the beginning of federal involvement in the industry that continues to the current day. However, state commissions continue to be the primary regulator of retail rates for investor-owned and regulated utilities in the United States.

2. THE UTILITY FRANCHISE

During the late 19th century, it was generally decided that efficient competition was not feasible or practical for the developing electric industry. Rather than allow redundant systems and parallel distribution lines (which occurred in some locations where competing electric companies had their own lines), local municipalities and (later) state governments granted exclusive territorial rights to utilities. This was and is done through an agreement with state or local authority by franchise territorial exclusivity laws, a certificate of public convenience and necessity, or both. All 50 states had laws setting up exclusive retail marketing areas for investor-owned utilities, and many states provided specific service area assignments under territorial exclusivity statutes. These territorial exclusivity statutes provide a

utility with the exclusive right and obligation to serve in an identifiable service area. Many states make service area assignments through the commission granting the utility a certificate of public convenience and necessity.

Because this sanctioned monopoly was by design the only supplier in the area, the utility was required to serve all customers in its service area. In general, utilities have historically had an obligation to serve all retail customers requesting service and to make the investments necessary to provide this service on demand (with limits on this obligation for new customers requiring new lines). To meet this obligation to serve, utilities made investments and entered into legal obligations, such as construction and purchased power contracts, and usually have rights of eminent domain for the construction of new power lines and other facilities. The obligation to serve derives from the utility being granted an exclusive public utility franchise. The requirement to serve is in exchange for an exclusive but revocable right or privilege to serve an area.

In addition to the obligation to serve all who apply for service within their franchise or service areas, public utilities have three other main responsibilities. First, utilities must provide safe and reliable service. Second, they may not engage in undue price discrimination. The prohibition against undue price discrimination requires that all similarly situated customers receiving identical service must be served on the same terms and conditions and for the same price (usually enforced through a required standard tariff filing). Third, utilities may charge only just and reasonable rates and may not earn monopoly profits with rates determined by a commission. What constitutes just and reasonable rates is determined by the regulator, whose decisions are subject to judicial review. These requirements are designed for the express purpose of limiting the potential for monopoly abuse. In return for undertaking these obligations, utilities are granted an opportunity (but not a guarantee) to earn a reasonable return on their prudent investment and to recover their prudently incurred expenses.

The granting of a certificate to serve, along with the accompanying obligations to serve and other legal interpretations and evolving reinterpretations embodied in state and local statutes and case law, constitutes what has become known as the regulatory compact. The regulatory compact does not and never did exist in a written form; rather, it was formed through a series of court cases and commission actions and practices. The regulatory compact

is, as Justice Oliver W. Holmes described it in 1912, a balancing of a utility's constitutional rights and its responsibilities.

3. THE REGULATORY PROCESS

The general procedure used to set rates under traditional regulation is through a formal rate case. The full rate case is a quasi-judicial procedure similar to a court proceeding and follows procedures proscribed in the public utility commission's state or federal enabling statutes and the commission's own administrative procedures. Rate cases can last up to a year or longer but are sometimes time limited by statute. Typically, testimony is presented by the participants' witnesses, and cross-examination may be done. A full rate case for a large electric utility may involve thousands of pages of testimony from expert witnesses, hearing transcripts, cross-examination, reply and rebuttal briefs, re-cross examination, and commission statements and decisions. A decision of the commission can be appealed to state and federal courts, which can take several years to resolve the issue. Although deference is often given to the regulatory authority on findings of fact, decisions are sometimes overturned on procedural grounds or due to questions of statutory or constitutional interpretation.

To avoid the time and expense of the full proceedings, settlements may be reached among the parties to the case. Once rates are established, automatic adjustments or rate modification may also occur without the formal rate case procedure. During periods of escalating costs and plant and equipment construction, rate cases may be common and may even overlap or run end to end. However, during periods of stable input costs, low inflation, and limited building activity, regulated firms might not request a rate adjustment for many years.

Rate cases can be initiated by the regulated firm or the commission, and in some states and under certain circumstances, others can initiate rate cases as well. Besides the commissioners, commission staff, and the regulated company, other participants in a rate case may include advocates for large and small consumers, environmental groups, and other "intervenor." Commissions conduct other types of hearings and investigations besides rate cases to consider commission procedures or possible rule changes, certificate applications, utility asset transfers, and other matters.

4. THE REVENUE REQUIREMENT FORMULA

From a purely functional and operational viewpoint, at the heart of traditional regulation is a deceptively simple formula—the revenue requirement formula. The revenue requirement is an approximation of the regulated company's total expenses to provide service to its customers and can be written as follows:

$$\begin{aligned} &\text{Revenue Requirement} \\ &= \text{Operating and Maintenance Expenses} \\ &+ \text{Depreciation and Amortization Expenses} \\ &+ \text{Taxes} \\ &+ [\text{Gross Property Value} \\ &\quad - \text{Accrued Depreciation}] \\ &\times \text{Allowed Rate of Return.} \end{aligned}$$

The part of the formula in brackets, gross property value minus accrued depreciation, is referred to as the rate base. Both the rate base and the rate of return are discussed in more detail in the next two sections. Operating expenses include wages, salaries, fuel, advertising, and regulatory costs. Also included are maintenance expenses, depreciation and amortization expenses, and federal, state, and local taxes. Expenses are also discussed later. The revenue requirement formula is designed to provide the regulated firm with an opportunity to recover its investments and expenses to serve its customers. Unfortunately, applying the formula is not just a matter of plugging in the numbers and getting the “correct” answer. Although the formula is straightforward, determining the value for each of the components has been shown, through decades of experience, to be a complex, contentious, and time-consuming task.

Before determining the value of each component, a “test year” or base period must be established. If a historic test year is used, as is most often the case, a 12-month period prior to the filing of the rate case is chosen. Alternatively, a future or forecasted test year, or a “rolling” test year, is used. In any case, the aim is to establish a representative sample of the utility's sales revenues as well as its expenses, taxes, and other costs to serve its customers.

5. RATE BASE

The rate base measures the total value of the firm's property, including plant, equipment, material, sup-

plies, and cash, and may include fuel inventories. Excluded are accumulated depreciation, amortization, deferred taxes, and customer-provided capital. An average rate base for the test year is typically used rather than a value at the end of the test year.

Although all the components of the revenue requirement formula are often contested and argued during rate-making proceedings, the valuation method and what investments are to be allowed in the rate base have been the most contentious and litigious part of the formula over the years. This is because the determination of the size of the rate base often involves large investments that have a significant impact on rates and because a certain amount of subjective opinion and discretion is used to determine it.

The methods and procedures developed over the years to determine rate base value exemplify the fact that traditional regulation is based largely on case law and legal precedent. Before 1900, regulation was price based and set by legislatures or by franchise bargaining by municipalities with companies more often than it was based on the value of companies' properties. In an early and important case in regulatory law, *Munn v. Illinois* (1877), the Supreme Court suggested that although private property cannot be taken away without due process, legislatures can change the law. If there are “abuses by legislatures,” the solution was for “the people [to] resort to the polls, not to the courts.” This left the matter of regulation to be decided by legislatures and not the courts. The *Munn* case marked the beginning of the regulation of private property held by businesses that are “affected with a public interest,” a phrase cited by the Supreme Court with an origin back in 17th-century British common law. Although many of the businesses that were regulated during the 19th and early 20th centuries are no longer regulated or have since faded away (e.g., bakers, innkeepers, milk and ice delivery services), the basis of electric utility regulation still used today by most states has its origins back in the case law that formed the public utility concept.

The judicial noninvolvement of *Munn* was clearly set aside in 1898 by the Supreme Court's decision in *Smyth v. Ames*. In the *Smyth* case, the court laid out its idea as to how utility property should be valued as follows: “The basis of all calculations as to the reasonableness of rates to be charged by a corporation ... must be the fair value of the property being used by it for the convenience of the public.” Originally, the court had in mind six specific measures of value and allowed for others as well.

This marked the beginning of the “fair value” doctrine. This case was used as a means to determine the rate base for nearly a half-century. Subsequent cases clarified what was meant by fair value. What was upheld by the courts was that fair value was to be defined as either original cost (i.e., the amount actually paid for installing the plant and equipment less accumulated depreciation plus improvements) or reproduction cost (i.e., the cost of plant and equipment estimated at market price levels prevailing at the date of valuation less depreciation).

For the railroad company at the time of *Smyth*, the original cost was higher than the reproduction cost because the investments were made during and just after the Civil War when costs were higher than during the 1890s. The state of Nebraska, represented by William Jennings Bryan, advocated the lower reproduction value. The Supreme Court generally also favored reproduction cost valuation around the turn of the 20th century because of the difficulty in determining original cost (due to different accounting practices and possible fraud). Later, during periods of inflation, positions changed whereby companies supported a reproduction cost standard while regulators and consumer representatives advocated original cost. The principle of reproduction cost remained firmly in place with a majority in the Supreme Court through the late 1920s.

According to Emery Troxel, “the inexact, variable, and sometime inexplicable meaning of fair property value” made determining rate base value extremely difficult. The subjective nature of the calculations inevitably led to the various interested parties coming up with values that differed widely. In an important judicial turning point, Justice Louis D. Brandeis wrote a separate concurring opinion (together with Justice Holmes) in *Missouri ex rel. Southwestern Bell Telephone Co. v. Missouri Public Service Commission* in 1923 that began the move away from reproduction cost. Brandeis stated that he believed that *Smyth v. Ames* was “legally and economically unsound” and that using this method was a “laborious and baffling task.”

But it was not until 1944 that the Supreme Court deemphasized the importance of finding a rate base value and the valuing methodology and placed more emphasis on the “end result” of a commission’s decision. In *Federal Power Commission v. Hope Natural Gas Co.*, the Supreme Court stated,

It is not theory but the impact of the rate order which counts. If the total effect of the rate order cannot be said to be unjust and unreasonable, judicial inquiry under the [Natural Gas] Act is at an end. The fact that the method

employed to reach that result may contain infirmities is not then important. ... Rates which enable the company to operate successfully, to maintain its financial integrity, to attract capital, and to compensate its investors for the risks assumed certainly cannot be condemned as invalid, even though they might produce only a meager return on the so-called “fair value” rate base.

In other words, the focus should be on the “end result,” not on the specific property of the utility or on a rate base valuation formula. This allowed state and federal regulatory authorities the discretion to also use an original cost standard based on “prudent investment,” that is, original cost minus fraudulent, unwise, or extravagant expenditures. The *Hope Natural Gas* case had the effect of changing the emphasis from rate base valuation methodology to return on investment.

To be included in the rate base, the investment must be both “prudent” and “used and useful” to the customers being served. The prudent investment test or standard establishes that costs that were reasonable at the time they were incurred, and given the circumstances and what was known or knowable at the time, are to be included in the firm’s rates. This provides the utility with an opportunity to recover these costs. Capital costs are recoverable along with a reasonable return on equity. Reasonably incurred operating costs are also included in the revenue requirement. In general, investments and expenditures have a “rebuttable presumption” that they are prudent when incurred; that is, they are presumed to be prudent unless it can be demonstrated otherwise. Regulators can consider cases of “hidden” imprudence, even if the investments or expenditures are already in the rate base, but are required to consider what was known or knowable at the time the decision was made; in other words, there should be no “Monday morning quarterbacking.”

The used and useful test is a much older test established in law. This test holds that utility investments and expenditures must be used and useful to the customers the company serves to be included in the firm’s rate base. Also, investments that are no longer used and useful to the public because they are either technologically or economically obsolete can be excluded from the rate base. The used and useful test is intended to prevent a regulated company from deliberately overinvesting in plant and equipment to inflate the rate base and to discourage a utility’s investment in assets that will not provide a tangible useful service to its customers. Investments that were once determined to be not

used or useful and were excluded from the rate base may later be allowed in the rate base if conditions change. An example would be if a utility's sales have increased since the last rate case such that additional capacity is needed to serve the customers. In such a case, it may be determined that a previous investment that was excluded from the rate base is now used and useful and should be included in the recalculated rate base.

An additional controversial issue in the determination of the company's rate base is the treatment of plant under construction or "construction work in progress" (CWIP). An issue is how much, if any, of CWIP should be included in the rate base calculation. This treatment of plant under construction, which for electric utilities can be considerable relative to the size of the firm, varies by jurisdiction. Factors that regulatory authorities use to decide whether to add CWIP to the rate base include whether the completion date is near for the work under construction, the project is of a short duration, the company's investment is relatively significant so that excluding it could impair the financial health of the company and its ability to raise capital, the construction project is for a certain type of investment (e.g., pollution control equipment), and the interest charges would be a large portion of the company's earnings. Some jurisdictions will not allow CWIP in the company's rate base at all because, consistent with the used and useful standard, they believe that current rate payers should pay only for facilities that are currently in use for public service. A related and historically older mechanism that state commissions have often used is the "allowance for funds used during construction" (AFUDC), which provides for recovery of the cost of capital for plant under construction once that plant is in service. Because AFUDC was not recovered from customers until the plant was in service, it has been generally regarded as less controversial than CWIP, particularly when the construction costs were viewed as not being excessively burdensome.

In summary, *Munn v. Illinois* laid the constitutional foundation for the regulation of businesses that are "affected with a public interest" such as electric utilities. *Smyth v. Ames* began more judicial involvement and efforts to establish reasonable rates based on fair value. Finally, *Federal Power Commission v. Hope Natural Gas Co.* switched the emphasis from finding a rate base value and the valuing methodology to the importance of obtaining a fair and reasonable end result of the regulatory process.

6. RATE OF RETURN

Alfred Kahn points out that every part of the regulatory process determines the level of allowed earnings for the regulated company. The regulated company's earnings are affected by what investments are included in the rate base, the allowed depreciation methodology, what expenses are allowed to be recovered, and the timing of the recovery, among other factors. But when the commission determines the allowed rate of return, it is explicitly determining the return on the company's property used to serve the public. Kahn noted, as observed by the Supreme Court in the *Hope Natural Gas* case, that there is no scientifically correct allowed rate of return. Commissions must exercise their judgment to determine a "zone of reasonableness," again applying the just and reasonable standard and balancing utility shareholder rights with protecting the public from monopoly abuses.

When a regulatory authority sets the company's allowed rate of return, the authority is determining the return that the utility will be allowed to earn on the rate base, beyond the operating costs and expenses that the company is also allowed to recover. This provides the utility with an opportunity to earn the allowed return, but it does not guarantee such a return. Because conditions (e.g., utility sales) may be different from what was assumed for the test year when the rate of return was set by regulators, the actual return earned by the utility may be different. The utility is not protected from losses due to mismanagement or adverse business conditions. The allowed rate of return refers to the percentage return on the rate base the regulator approves, which may be different from the rate of return received by investors or the utility from the total investment of the utility. Return refers to the total dollars that the utility receives or that investors receive in interest and earnings. In general, there is no provision for retroactive rate making, that is, making up for returns that were allowed but were not received by the utility.

The case law history again played a significant role in determining the method of setting the rate of return. An early and influential case was *Bluefield Water Works & Improvement Co. v. Public Service Commission of West Virginia* from 1923, when the Supreme Court outlined the major elements that regulators have considered over the years when determining rate of return. These elements include that (1) the return should be "reasonably sufficient" to maintain the financial soundness of the utility; (2)

the return should be adequate to maintain the company's credit sufficiently to allow it to raise money necessary for public service, assuming the company is operated efficiently; and (3) the utility should be allowed to earn a return on the value of the property dedicated to public service comparable to that being made on similar investments in unregulated businesses with similar risks, but it is not entitled to earn profits in excess of that. On this last point, the court stated that a public utility "has no constitutional right to profits such as are realized or anticipated in highly profitable enterprises or speculative ventures." The court also noted that a "rate of return may be reasonable at one time and become too high or too low" due to changing financial and business conditions.

In summary, the legal standard is that rate of return should not be confiscatory; rather, it should justly, but not excessively, compensate the utility for use of the shareholders' property. The company is obligated to operate efficiently, invest prudently, and control expenses reasonably. Again, when considering these factors, the regulatory authority attempts to balance the interests of the company's investors with those of the rate payers.

In terms of economic theory, the company should be allowed to earn a "normal" profit, that is, what it could earn on its investments if it were a business with similar risk and capital structure in a competitive industry. This should allow sufficient return on investment so that the utility can compensate its investors for the use of their money, that is, the investors' "opportunity cost."

A utility's actual return to its investors and its cost of capital are determined by financial markets. An investor's return is the sum of the yield of a security (stocks or bonds) and the growth (or loss) of that security. Both yield and growth are dependent on the security's price, which may change each regular business day in the financial markets. If a utility is perceived to be more risky because, for example, investors believe that the firm is less likely to maintain its dividend, the price of its stock will decline. In such a case, investors who hold the utility's stocks or bonds lose part of their investment (on paper at least). The expected return on such an investment increases, and new investors now require an additional risk premium to become or remain investors in the firm. Overall, investments of similar risk will be priced in financial markets to have equivalent expected returns to investors. For example, if something occurs that lowers risk, a security's price will rise, thereby decreasing its yield.

Conversely, if risk increases, the price will fall and the yield will increase.

Traditionally, from an investor's standpoint, electric utility stocks and bonds have been considered "quasi-fixed income" securities. Their traditional monopoly status and relatively inelastic demand meant that they were a relatively safe and a steady income source. Growth potential was not the main reason why investors purchased these investments. In general, it was viewed that there was limited opportunity for growth either within or outside the utility's service territory but that the risk of a loss was low as well. This view has changed as the industry is being transformed into a more competitive structure. However, when and where traditional regulation is applied, investor service and rating agencies observe and analyze carefully the results of the regulatory process and how the regulated companies are treated. A key factor they watch is where the allowed rate of return is set.

Leonard Hyman traced historical rates of return and found that between 1915 and 1929 they were in the range of 7 to 8%. During the Depression years of the 1930s, they dropped to the range of 5.5 to 6.0% and remained there until the mid-1960s. After the mid-1960s, because of rising interest rates and inflation, rates of return rose and peaked during the late 1970s and early 1980s. It is not always clear how commissions determined rates of return early on; perhaps they were using prevailing business and interest rates of the time. During the latter half of the 20th century, it became more common for commissions to base the rate of return on the utility's cost of capital.

Table I illustrates how the cost of capital can be determined. Assuming a capital structure of 50% debt and 50% equity (weighted for preferred and common equity), where the utility raises \$500,000 from borrowing at 4.0% and another \$500,000 at 6.0% from equity, the weighted cost of capital would be 5.0%. In practice, determining the cost of debt and preferred stock is relatively easy because the interest rate on debt and preferred dividends are fixed. However, determining the return on common equity is not so straightforward and has had the affect of increasing the complexity of determining an allowed rate of return considerably. Because common equity is usually a significant part of an electric utility's capital structure, it is important to gauge the return on common equity as accurately as possible.

Hyman described four methods used to estimate return on the common stock equity. First is the comparable earnings method, which is based on

TABLE I

Cost of Capital Example

Capital component	Amount (U.S. dollars)	Cost (percentage)	Weighted cost (U.S. dollars)
Debt	500,000	4.0	20,000
Equity	500,000	6.0	30,000
Total	1,000,000		50,000
Cost of capital = 5.0%			

Source. Based on example from Hyman (1988).

returns on equity of comparable businesses in terms of risk and capital structure. Among the limitations to this approach are that it is difficult to reach an agreement on a sample distribution of comparable companies and that the results are dependent on the selection of companies picked by the interested party in the case. The second approach is the discounted cash flow technique, which is an attempt to objectively estimate a market cost of capital. This is based on the premise that the price of a common share equals the present value of the sum of all future income to be received from the share. This method also has limitations that stem from assumptions used in the estimation. A third approach is the capital asset pricing model. This method divides the return on investment into a risk-free return and an added return for risk and attempts to estimate investors' expectations of a future return from the past performance of stocks with similar characteristics. The fourth method is a group of three financial integrity approaches: (1) set a return to produce a necessary interest coverage ratio, (2) provide equity with a return set at a fixed amount above current debt costs, and (3) provide a return on equity designed to keep the common stock price at or slightly above book value.

These approaches used to determine the rate of return are complex and vary widely by jurisdiction. In actual practice, there are wide gaps among the theory of how rate of return should be determined, opinions on how it should be done, and what in practice is actually done. It is widely recognized that although regulatory authorities may use information from the four methods or other approaches to help inform them of the cost of capital, they still must rely on discretion and judgment, guided by the extensive legal history, to decide on a just and reasonable rate of return within a zone of reasonableness. This explains why the process of determining a rate of return has been described as more like informed conjecturing than like exact science.

7. EXPENSES

As described previously, the total revenue requirements calculation also includes operating and maintenance expenses, annual depreciation expense accrual, income taxes, and other taxes. For electric utilities, fuel and purchased power are typically the largest expenses. A problem with establishing a representative or historic test year is that it is likely that conditions will change, particularly during periods of inflation or fuel cost changes. Of course, changes can either increase or decrease the utility's cost of service. When the utility's earnings decline and it is unable to earn the allowed return on investment, this is referred to as attrition. To remedy this, states may use attrition allowances or use a partially forecasted test year to adjust the rate base or expenses.

Some expenses, such as fuel costs, are nearly certain to change and are very difficult to forecast with any accuracy. This became especially apparent during the rapidly increasing fuel price increases of the 1970s. As a result, most states use automatic adjustment mechanisms that periodically adjust companies' fuel expenses and rates. This allows a utility to recover increased fuel costs without having to go through the time-consuming process of a full rate case. Other expenses have been added to the adjustment mechanism over time, for example, for recovery of purchased power costs.

8. COST ALLOCATION AND RATE DESIGN

Once the total revenue requirement has been determined, the rates charged for each customer class are derived through a two-part process of cost allocation and rate design. The first part allocates the cost of service for each customer class, usually through a cost-of-service study. Customers with

similar patterns of use are grouped into customer classes. For electric service, the general customer classes are residential, business (commercial and industrial), lighting, and miscellaneous. Each general class is often also divided into subclasses. Examples include standard residential service, residential with electric space heating, and small and large industrial and commercial classes. Electric utilities in the United States typically have one to two dozen or more customer tariffs for each subclass with optional service categories such as interruptible, time of use, conservation, or other special rates. A guiding principle of cost allocation is that the rates customers pay should correspond with the costs of serving those customers.

The next step is to design the individual rates for each customer class to provide the utility with an opportunity to collect the authorized revenue requirement. Electric utilities typically have three main components of the customer's rate (often called a rate schedule) for each customer category. First are customer charges that are for costs incurred to serve the customer independent of the amount used such as metering equipment, meter reading, billing, and distribution line. Customer costs vary not by the amount for power used but by the type of customer; these charges are usually expressed in dollars per meter per month. Second are demand charges that are for the fixed costs of plant and equipment capacity needed to meet the customers' maximum or peak demand, usually calculated in terms of kilowatts per month. Third are energy charges that are intended to recover the cost of operation and that vary with the amount of power used such as fuel, labor, and materials. Energy charges are usually in cents per kilowatt-hour and are often seasonal, that is, higher during the summer when demand is higher and power plants with higher operating costs are used to meet demand. Additional charges may be added for an automatic fuel adjustment, conservation or renewable program cost, federal- or state-mandated environmental equipment cost, and/or commission-approved or -mandated programs for low-income household assistance.

The rate schedule will vary by customer class and specifies rates by season, volume use, any discounts, and any additional charges. Seasonal rates are usually used to allow higher prices during the summer (for summer-peaking utilities, e.g., June through September). A uniform rate applies a constant energy charge no matter how much is consumed. If the per unit energy charge falls as quantity consumed increases, this is a declining- or

decreasing-block rate (e.g., the first 600 kilowatt-hours per month are 6 cents per hour and each additional kilowatt-hour is 5 cents per hour). Finally, with an increasing-block rate structure, the per unit rate increases as quantity consumed increases.

James C. Bonbright outlined eight "criteria of a sound rate structure" that identify desirable characteristics that regulators have used, or at least should use, to develop rate structures. These criteria were described by Charles Phillips as follows:

1. The related "practical" attributes of simplicity, understandability, public acceptability, and feasibility of application
2. Freedom from controversies as to proper interpretation
3. Effectiveness in yielding total revenue requirements under the fair return standard
4. Revenue stability from year to year
5. Stability of the rates themselves, with a minimum of unexpected changes seriously adverse to existing customers
6. Fairness of the specific rates in the apportionment of total costs of service among the various consumers
7. Avoidance of "undue discrimination" in rate relationships
8. Efficiency of the rate classes and rate blocks in discouraging wasteful use of service while promoting all justified types and amounts of use (a) in the control of the total amounts of service supplied by the company and (b) in the control of the relative uses of alternative types of service (e.g., on-peak vs off-peak electricity)

These principles are widely accepted and used by regulators as guidelines for developing a rate structure. Phillips noted that these principles are broad and ambiguous but are of value because, as Bonbright noted, they are useful for reminding regulators of considerations that might be neglected and for suggesting reasons why practical rate design does not "readily yield to 'scientific' principles of optimum pricing." As with determining the allowed rate of return and other decisions, regulators must again rely on their informed judgment when designing the rate structure.

9. LIMITATIONS OF THE TRADITIONAL APPROACH

Many observers would argue that traditional or cost-based regulation has performed satisfactorily under

many different conditions in terms of cost to consumers, reliability of service, and determining a “fair” outcome. This form of regulation was shown to be adaptable and able to deal with a wide range of crises and challenges placed before it over nearly a century of experience. However, it has come under intense criticism from many observers. One of the more common and persistent limitations pointed out by critics is the lack of incentive a utility has to minimize and control its cost under a cost-based system of regulation. Regulators can attempt to oversee and scrutinize the utility’s management and operation of its system, but they cannot possibly impose the same level of cost control that competitive pressures exert on a competitive firm. Other criticisms include the relatively high administrative cost (from being overly cumbersome, legalistic, and bureaucratic), the lack of incentive for a regulated utility to use innovative technologies or rate structures, and the tendency for the utility to attempt to overcapitalize in rate base investments under certain conditions.

In response to these criticisms, regulators have been introducing more incentive- or performance-based mechanisms to ameliorate the reduced incentive to control costs under traditional regulation. The idea of introducing stronger incentives into the regulatory process is not new. For example, the used and useful and prudent investment standards were designed to induce more cost control in utility investments. During more recent years, regulators have used commission-ordered company audits, performance incentives to induce increased plant efficiency, industry benchmarking standards to compare a utility’s performance with that of other utilities, price cap regulation that sets rates for several years (with annual productivity and inflation adjustments) and allows the utility to retain any gain (or loss) from improved cost efficiency, and profit-sharing arrangements based on performance.

Much of the electric supply industry’s attention recently has been placed on the restructuring of the industry to increase competition in wholesale and (in some states) retail markets. However, traditional regulation continues to be an important regulatory tool. Currently, most states continue to use traditional or cost-based regulation for all electric utility services and will likely continue to do so for the foreseeable future. States that have changed to retail access, allowing retail customers to choose their suppliers, have abandoned cost-based regulation for generation service but will likely continue to use cost-based methods for the distribution services offered by their local distribution utilities.

10. SUMMARY AND CONCLUSIONS

It was recognized that under traditional monopoly franchise regulation and rate making, electric utilities and other regulated companies were subject to a regulatory compact or, as Justice Holmes referred to it in 1912, a “bargain.” Based on historical case law and more than a century of development and change, this compact can be summarized as the careful balance between compensatory rates and confiscation of utility property that provides a utility with an opportunity to earn a reasonable return on its investment in exchange for providing safe and reliable power at reasonable cost to all customers who request service. This opportunity is held in check by the used and useful and prudent investment tests as well as by competition from government ownership, fuel substitutes, self-generation, and (more recently) incentive- or performance-based rates and competition from other suppliers. This compact has been rebalanced occasionally to accommodate changing conditions in the industry from economic impacts and technological change. Over decades, state and federal commissions have developed procedures to implement the laws passed and have refined the regulatory mechanisms to determine what investment and operating costs should be recovered, set a return that the regulated company should be allowed to earn on its property, and determine rates for consumers. State and federal commissions have used incentive- and performance-based mechanisms to improve operating efficiency and investment decisions. The recent move to competition for generation nationally and by some states notwithstanding, many aspects of traditional regulation are still in use by most states and will continue to be used for the noncompetitive segments of the industry, such as transmission and distribution services, for the foreseeable future.

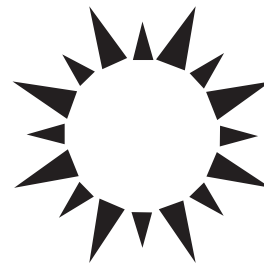
SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Transmission and Generation Reliability and Adequacy

Further Reading

- Anderson, D. D. (1981). "Regulatory Politics and Electric Utilities: A Case Study in Political Economy." Auburn House, Boston.
- Beecher, J. A. (2003). "Economic Regulation of Public Utilities: An Introduction and Overview." Institute of Public Utilities, East Lansing, MI.
- Bonbright, J. C. (1961). "Principles of Public Utility Rates." Columbia University Press, New York.
- Bonbright, J. C., Danielsen, A. L., and Kamerschen, D. R. (1988). "Principles of Public Utility Rates." Public Utilities Reports, Arlington, VA.
- Bryant, J. M., and Herrmann, R. R. (1940). "Elements of Utility Rate Determination." McGraw-Hill, New York.
- Burns, R. E., *et al.* (1984). "The Prudent Investment Test in the 1980s." National Regulatory Research Institute, Columbus, OH.
- Clemens, E. W. (1950). "Economics and Public Utilities." Appleton-Century-Crofts, New York.
- Edison Electric Institute and American Gas Association (2001). "Introduction to Public Utility Accounting." Edison Electric Institute and American Gas Association, Washington, DC.
- Fainsod, M., Gordon, L., and Palamountain, J. C. (1959). "Government and the American Economy." Norton, New York.
- Hyman, L. S. (1988). "America's Electric Utilities: Past, Present, and Future." Public Utilities Reports, Arlington, VA.
- Kahn, A. E. (1988). "The Economics of Regulation: Principles and Institutions." MIT Press, Cambridge, MA.
- Phillips, C. F. (1984). "The Regulation of Public Utilities: Theory and Practice." Public Utilities Reports, Arlington, VA.
- Posner, R. A. (1999). "Natural Monopoly and Its Regulation." Cato Institute, Washington, DC.
- Troxel, E. (1947). "Economics of Public Utilities." Rinehart, New York.

Electric Power: Transmission and Generation Reliability and Adequacy



FRANCISCO F. GARCÉS

Universidad Nacional de San Juan

San Juan, Argentina

-
1. Introduction
 2. Generation System Reliability Evaluation
 3. Composite Generation and Transmission Reliability
 4. Electric Power Industry Restructuring and Reliability

Glossary

adequacy The ability of the electric systems to supply the aggregate electrical demand and energy requirements of their customers at all times, taking into account scheduled and reasonably expected unscheduled outages of system elements.

bulk electric system The part of the system that encompasses the electrical generation resources, transmission lines, interconnections with neighboring systems, and associated equipment, generally operated at voltages of 110 kV or higher.

distribution system The part of the power system that functions to move electric energy from the bulk supply points to individual consumers.

embedded generation The generation sources that feed into middle- or low-voltage distribution networks.

generation system The part of the power system that functions to convert different forms of primary energy in electric energy.

reliability The ability of a technical system to perform its function properly, for the period of time intended, under the operating conditions intended.

restructuring The transformation process of the electric power industry, from vertically integrated monopoly utilities to market-oriented structures.

security The ability of electric systems to withstand sudden disturbances, such as electric short-circuits or unanticipated loss of system elements.

transmission system The part of the power system that functions to transport electric energy from generation points to bulk supply points.

Power system reliability is the field of engineering devoted to assessing the adequacy and security of power systems. Adequacy is defined as the ability to meet the aggregate electrical demand and energy requirements of customers at all times, taking into account scheduled and reasonably expected unscheduled outages of system elements. Security is the ability of electric systems to withstand sudden disturbances. It is generally understood that adequacy assessment involves steady-state consideration of the systems, whereas security assessment must include analysis of the dynamic system behavior following disturbances. This article deals with the problem of bulk power system adequacy assessment, involving both transmission and generation systems.

1. INTRODUCTION

The behavior of technical systems is conditioned by internal and external factors, many of them of stochastic nature, that can lead to system failure. In order to evaluate the reliability of a system, possible system states should be identified, the ability of the system to perform its function in each of these states should be analyzed, and probabilistic measures about the occurrence of system states should be calculated. The function of a power system is to supply consumers with electric energy economically, at any time and location, observing reliability and quality standards. The practical economic impossibility to store significant amounts of electric energy implies that a balance between generation and load, including network losses, must be maintained at all points in time.

According to the North American Electric Reliability Council, the reliability of the interconnected bulk

electric systems is defined through adequacy and security. Both adequacy and security are affected by stochastic events. It is generally understood that adequacy assessment involves steady-state consideration of systems, and that security assessment must include the treatment of the dynamic system behavior following disturbances. Tools for adequacy assessment, based on probabilistic procedures combined with methods for steady-state power system analysis, involve the solution of large nonlinear equation systems. The analysis of power system security requires that sets of differential equations describing the dynamic system behavior be solved. Even considering the powerful computers available today, the combination of dynamic analysis with probabilistic procedures leads to very time-consuming calculations. Generally, security analysis of power systems is carried out within a very short time frame after a disturbance, disregarding the possibility of multiple independent events.

The typical structure of a power system includes the subsystem generation, transmission, and distribution components. Generation and transmission affect bulk power system reliability and are the focus here. The most important types of studies of bulk power system reliability involve generation system reliability and composite-generation and transmission-power system reliability.

During the last 20 years of the 20th century, the industry of electric power supply underwent a great structural transformation, going from a regulated, vertically integrated monopoly to a more competitive industry that is thought to be better suited to provide service more efficiently. The ongoing restructuring process, however, adopting different structures in different regions, generally involves privatization and unbundling of generation, transmission, distribution, and retail activities. The change of paradigm in the electric industry, from minimizing total system costs to maximizing system-participant profits, can strongly affect system reliability. Because most of the tools for reliability assessment currently in use have been adapted from those developed for vertically integrated systems, it is relevant to question the impact of change on reliability.

2. GENERATION SYSTEM RELIABILITY EVALUATION

The aim of generation system reliability calculations is to find the generation capacity needed to achieve a desired (often called optimal) reliability level. Because electricity cannot be stored economically in

large quantities, uncertainties affect the amount of load to be supplied, and components are subject to unscheduled outages, it is necessary to have generation reserve for a reliable service.

A first approach in generation reliability evaluation assumes that network components are substantially more reliable, compared to generation units, and that those networks have enough redundancy. In this context, a fundamental simplification can be made, assuming that power can be moved unconstrained from suppliers to consumers.

2.1 Reliability Criteria and Indexes

The optimal reliability level can be obtained, minimizing the total cost, including reserve and outage costs. With increasing reserve, the utility costs will increase due to the investments needed and the outage costs will decrease, and vice versa. The actual determination of the optimal reliability level requires knowledge of outage costs. This is problematic because it involves both objective measurable quantities and subjective perceptions of customers. Generally, the reliability level to be maintained is determined on the basis of estimation of outage costs by regulatory agencies, which set planning and operation standards to guarantee quality, reliability, and security of the service.

Historically, deterministic criteria, such as a percentage of the peak load or the capacity of the largest unit dispatched, were used to define generation reserve requirements. Then, probabilistic criteria were developed and implemented to take into account the stochastic nature of factors affecting reliability. One of the most widely used reliability indexes for generation systems is loss-of-load probability (LOLP), which is the probability of the customer load being greater than the available generation capacity. Other well-known reliability indexes are loss-of-load expectation (LOLE), generally given in number of days per year on which the peak load is expected to exceed the available generating capacity, and expected energy not served (EENS), which gives the expected megawatt-hours not supplied due to capacity shortages in a given time unit (a year, for example).

2.2 Reliability Calculation

The first step in reliability evaluation of generation systems is the modeling of the stochastic behavior of generation units. Many different generator models for reliability studies have been developed, varying in

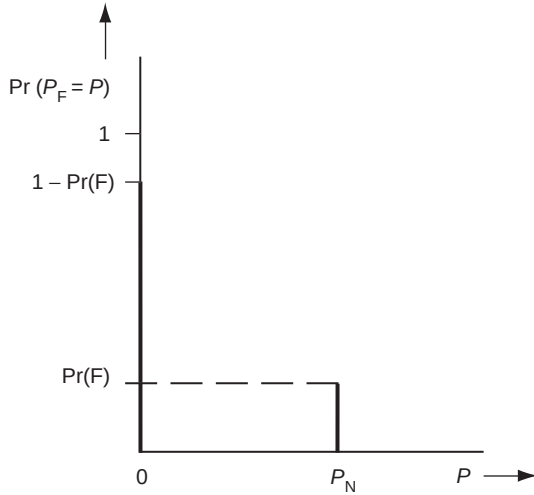


FIGURE 1 Simplified two-state generator model.

the degree of sophistication to describe the actual behavior of generating units. From the system viewpoint, it is important to identify those generator states in which reserve is needed because the generator is not able to supply load. These states can be combined to form the failure state, obtaining a two-state generator model such as the one depicted in Fig. 1. Given the failure probability function $\Pr_i(F)$ of each generator i in the system and assuming independent behavior of generators, the cumulative probability function of unavailable capacity, $F_F(P) = \Pr(P_F \geq P)$, can be calculated through simple convolution of the probability functions of all generators in the system. Figure 2 shows the qualitative form of the function $F_F(P)$.

The reserve capacity P_R of a generation system is defined as the difference between installed generation capacity and maximum load to be supplied. For a

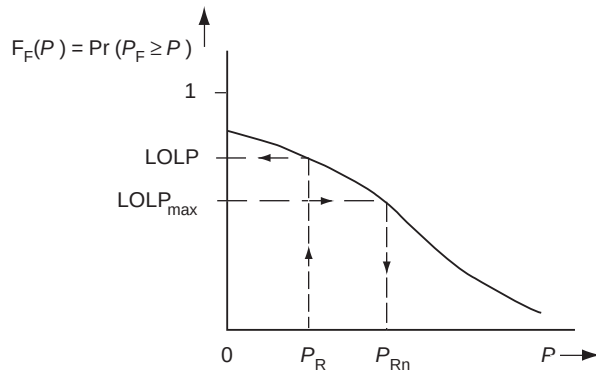


FIGURE 2 Loss-of-load probability and reserve requirement calculation.

given value of P_R , LOLP can be calculated as the value of $F_F(P)$ for $P = P_R$ (Fig. 2). According to its definition, LOLP is the probability that the system cannot fully supply the customer load. This happens when the outage capacity exceeds the available reserve. Because the load is known only as a forecast value, it should be modeled through a probability distribution. Usually a normal distribution with mean equal to the forecast value is used. Assuming independence of generation and load, the probability function describing the stochastic load behavior can be included in the convolution calculation to evaluate the function $F_F(P)$. On the other hand, if a maximum acceptable value of LOLP, called here $\text{LOLP}_{\max}$, is given, the generation reserve, P_{Rn} , needed to achieve the reliability level defined through $\text{LOLP}_{\max}$ can be obtained as indicated in Fig. 2. With the exception of the simpler, deterministic approaches, this is probably the most common methodology used by system operators to calculate generation reserve requirements.

LOLE is determined as the annual average value of LOLP calculated for each daily peak load in the year, multiplied by the number of days in the year. The weakness of the indexes LOLP and LOLE is that they do not give any information about the severity of a failure. The index EENS overcomes this problem, encompassing probability and severity of the failure, and can be calculated using the yearly load duration curve shown in Fig. 3. For each value of available generation capacity $P_A = P_i$, with probability $\Pr(P_A = P_i)$, there is an energy deficit value, E_{di} , such that

$$\text{EENS} = \sum_i E_{di} \cdot r(P_A = P_i).$$

2.3 Embedded Generation

An increasing concern about environmental issues has led many governments to establish incentives for

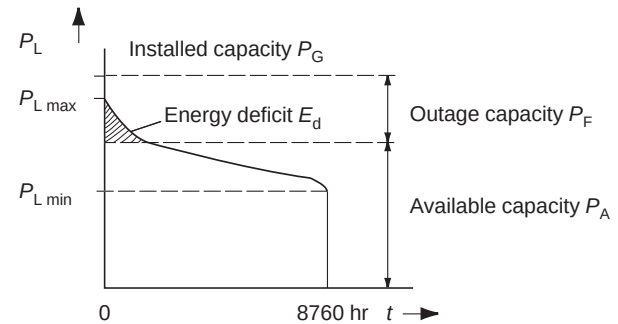


FIGURE 3 Calculation of expected energy not supplied.

clean generation using renewable energy sources such as sun or wind. Unlike large conventional generation units, small solar or wind generators often feed into the distribution networks, creating a completely new set of problems. However, reliability studies of systems with sizable participation of embedded generation should consider this type of generation. Even when network constraints are neglected, typical in reliability evaluation of generation systems, some new problems arise: (1) prediction of embedded generation output is no easy task because often embedded generation is not dispatched by the system operator; (2) primary sources, especially wind, are highly variable and difficult to forecast; (3) there are dependencies in the behavior of generators, and loads are correlated with the availability of primary energy sources, notably the sun; and (4) frequently, distributed generators are not permitted to generate if the distribution network they are feeding into is disconnected from the bulk power system. Although some efforts have been made in order to overcome these problems, further work is needed to develop models that properly consider the contribution of embedded generation in reliability studies of generation systems.

3. COMPOSITE GENERATION AND TRANSMISSION RELIABILITY

Generation systems without the constraints that originate in transmission networks connecting generation with load points were considered in the previous section. Because actual transmission systems are not 100% reliable and have limited transport capacity, reliability evaluation of bulk power supply requires consideration of both generation systems and transmission networks. The distribution of power flows in the transport network follows Kirchhoff's laws, which, together with network impedances, generator outputs, and bulk supply point loads, determine the loading of each transmission line in the system and the nodal voltages. A load flow model, one of the most powerful tools for power systems analysis, allows the calculation of voltages and load flows in transmission networks.

The function of the composite system is to generate enough energy to satisfy the load and to transport it to the supply points. As already stated, reliability analysis includes adequacy and security evaluation. Security evaluation requires analysis of the dynamic behavior of a system when disturbances

lead to changes of the system state. Although the subject of intense research, probabilistic assessment of the dynamic behavior of power systems has not yet reached the maturity needed for successful implementation in real power systems. Adequacy studies analyze power system states from a steady-state viewpoint using a load flow model. A system state will be termed adequate whenever it has enough generation available to supply the customer load fully and has the ability to transport the energy from the generating plants to the supply points, observing voltage and thermal limits. Dynamic constraints are often taken into account indirectly through reduction of transmission line transport capacity limits on the basis of stability considerations.

3.1 The $(n - 1)$ Criterion

The most frequently used approach to assess adequacy of transmission and composite systems is the so-called $(n-1)$ criterion, which states that a system must be able to withstand any single contingency without affecting functionality. Using load flow calculations, all system states, including one major component (generator, line, or transformer) forced out of service, are examined. If the system load is fully supplied, satisfying all constraints (including active and reactive generation limits, nodal voltages, and thermal or stability capacity limits), the considered state is deemed adequate. Otherwise, the state is considered inadequate. Then, reinforcements are proposed until all $(n-1)$ states fulfill the adequacy criterion.

The appeal of the $(n-1)$ criterion derives from its simplicity. The $(n-1)$ criterion is also the most frequently used approach to composite systems security analysis. Replacing the steady-state load flow analysis of each $(n-1)$ state by a dynamic behavior analysis following each possible single failure of the system, the system states can be classified as secure, when all constraints are satisfied, or insecure otherwise.

The weakness of the $(n-1)$ approach to composite system adequacy assessment is that it considers neither the probability of the analyzed states nor the magnitude of the problem in case of inadequate states. It must be pointed out that the $(n-1)$ criterion neglects all multiple contingencies regardless of their probabilities. An extension of the approach to consider multiple contingency states and the introduction of load point-related, individual component-related, and system wide reliability indexes help to overcome these problems. Regarding the

severity of the problems, the $(n-1)$ criterion can be extended to consider a more detailed classification of the system states: (1) a system state with fully supplied load, no constraint violations, and fulfilling the $(n-1)$ criterion is termed the normal state; (2) when the load is fully supplied, no constraint is violated but the $(n-1)$ criterion is not fulfilled, the state is called the alert state; (3) if the load is fully supplied but there are constraint violations, the system state is designated as an emergency state; and (4) any state with load curtailment is called an extreme emergency state. Probabilistic indexes can be calculated for each of these states.

3.2 Reliability Indexes

System-wide reliability indexes reflect the average adequacy of the composite system but do not give any information about reliability as observed at individual bulk supply points. Therefore, it is important to use a combination of both system-wide and supply point-related indexes to describe the reliability of composite power systems. A variety of different indexes have been proposed, including probability, frequency, and duration of failures, load not supplied, energy curtailed, etc. Only a few of them are presented in Table I as examples of often used reliability indexes. For each specific reliability study, the more appropriate reliability indexes should be selected.

Probability indexes can be calculated for a given load state, as in the case of the index LOLP in the previous section, or as average values for a given time period, as in the case of LOLE. It is obvious that energy-not-supplied indexes require the consideration

of a certain time period. Because the load state of a power system changes with time, leading to changes in the generation units scheduled to supply load and in the network state, any calculation of average reliability indexes requires definition and analysis of representative load states for the time period of interest.

3.3 Reliability Calculation

There are six essential steps in the procedure for reliability evaluation of composite power systems.

3.3.1 Definition of Representative Load States To Be Considered

The question about the load states to be considered depends on the aim and scope of the study. If, for example, the probability of the system not being able to supply the load fully at a given point in time (usually the peak load) is desired, the load at that time point is to be considered. When the expected value of energy not supplied in a given time period is required, many load states representing the possible load conditions in that period must be studied.

3.3.2 Active and Reactive Power Dispatch

Once the load states are defined, it is necessary to schedule the operation of the generation system to supply the load. This has been done in the vertically integrated industry using well-known and generally accepted optimization methods. In the new competitive environment, generators submit to the Independent System Operator (ISO) bid prices at which they are willing to sell energy, and the ISO allocates the resources economically.

3.3.3 Load Flow Analysis

Given the loads and the generation scheduling, a load flow analysis is performed in order to verify compliance of technical constraints, such as transport capacity limits or supply point voltages.

3.3.4 Contingency Simulation

Contingency simulation is the heart of composite systems reliability evaluation and is based on a load flow model. It aims to find out the system states in which there are load curtailments, constraint violations, or both. The selection of the system states to be analyzed can be made using simulative or analytical techniques. Simulative techniques, usually called Monte Carlo methods, are based on the use of computer-generated random numbers to determine the state of each system component. Outages of

TABLE I
Reliability Indexes

System-wide indexes	Supply-point or component-related indexes
Probability of failing to fully supply the system load satisfying all technical constraints	Probability of failing to fully supply the load at a given bulk supply point
Probability of any capacity limit violation	Probability of capacity limit violation for a given component
Probability of any voltage limit violation	Probability of voltage limits violation at a given network node
Expected energy not supplied for the whole system	Expected energy not supplied for a given load point

components with high failure rates are likely to be simulated more frequently, compared to outages of more reliable components. The simulation process produces a sample of the possible behavior of the system. Reliability indexes are calculated through statistical evaluation of this sample. The number of simulations required to obtain a representative sample can be calculated as a function of the desired accuracy of results.

In the analytical approach, the system states are enumerated, generally in an increasing order of contingency level, stopping either after a given contingency order or when the state probability falls below a given value. Power system components are very reliable, implying that for increasing contingency order, state probabilities rapidly become very small. For each system state, selected either through simulation or enumeration, probabilistic measures regarding its occurrence are calculated and a load flow analysis is performed to verify if any constraint is violated. In case of constraint violations, the procedure must mimic the operator behavior, determining and simulating the introduction of remedial actions. The impact of these corrective actions is then studied through a new load flow calculation. After all corrective possibilities for a given contingency state have been used, the remaining constraint violations determine that the analyzed state is a failure state.

3.3.5 Corrective Actions

The simulated corrective actions must reflect with reasonable accuracy the actual possibilities of the operator in a real system. The possibilities of the system operator include modification of active and reactive power dispatch and change of the network topology through connection or disconnection of transmission facilities. Optimization tools with the objective of operating cost minimization can easily include the possibility of changes in generation dispatch to satisfy technical constraints. It is not so easy to consider topology modifications because of the discrete nature of this problem. The most popular tool used by power engineers to obtain an economic dispatch considering network constraints is optimal power flow (OPF). OPF is an extension of the well-known simple economic dispatch (ED). The objective function of ED is minimization of generator operating costs subject to the constraints that generator output must be within technical limits for each generator and the total output of the dispatched generators must equal the system load, including transmission losses. OPF adds to the economic dispatch a set of alternating current (AC) power

flow equations as equality constraints. The solution of the OPF problem is an operating plan with minimal costs and a complete load flow solution. A further development of OPF is the so-called security-constrained OPF, which introduces inequality constraints on power flows and voltages.

Security-constrained OPF allows the calculation of optimal corrections of active and reactive generation aimed to eliminate or mitigate violations of power flow or voltage constraints. Attempts are being made by researchers to develop accurate and robust procedures to determine corrective actions, including all possibilities available to system operators using different modeling approaches, such as expert systems or other artificial intelligence tools. The reliability indexes must reflect the constraint violations remaining after exhausting the possibilities regarding corrective actions.

3.3.6 Calculation of Reliability Indexes

Once the preceding five steps have been carried out, the system states selected through enumeration or simulation can be classified either in normal operative states or in problem states because the load is not fully satisfied or constraints are violated. For any specific class of adequacy deficiency, insufficient generation, or network violations, system-wide, component-, or node-related probabilistic measures such as probability, frequency, or duration can be calculated. The load flow analysis provides the results necessary to link these reliability parameters to the severity of the deficiency. For example, for a bulk supply point, probabilities can be calculated for states with load curtailment exceeding prespecified values.

4. ELECTRIC POWER INDUSTRY RESTRUCTURING AND RELIABILITY

Before successful restructuring in other industry branches led to significant changes in the electric power industry organization and structure, beginning in Chile in the early 1980s and then in Britain in 1988, the most common organizational forms of electricity supply were the state-owned monopoly and the monopoly franchise system with vertically integrated utilities. Reliability planning in these systems was done in similar ways. In state-owned monopoly systems, the supply duty is generally established by law, requiring that all customers are reliably supplied, with costs as low as possible. Because the whole system is owned by the state,

central planning, including generation and transmission, is the natural way to achieve high reliability while minimizing total costs. All necessary data are available to the power company and the problem is easily formulated as an optimization problem with total cost minimization as the objective function, subject to technical and reliability constraints. The reliability constraint is formulated as an inequality, requiring that a reliability index, typically loss-of-load probability or expected energy not supplied, does not exceed a prespecified limit.

A vertically integrated utility does its planning much in the same way as the state monopoly. In its geographically delimited concession region, the vertically integrated utility has to economically supply the customer demand subject to the same constraints as the state-owned monopoly. Reliability standards are set by the state through regulatory agencies and electricity prices to end consumers are regulated on the basis of audited costs and a guaranteed rate of return.

New generation technologies, notably high-efficiency gas-fueled combined cycle thermal plants, broke the notion that scale economies made electricity production a natural monopoly. This, combined with the fact that a regulated monopoly has little incentive to reduce its costs because any reduction is passed on to the consumers through adjustment of the regulated prices, and the success of deregulation in other industry branches, such as telecommunications and airlines, led to the movement toward restructuring in the power industry. On the other hand, there is a feeling that the high reliability standards used in the vertically integrated industry, state or private owned, have sometimes led to overcapacities that probably increase costs beyond the values consumers would be willing to pay for reliability.

In the new deregulated environment, there is no more central planning. Regarding generation investment, instead of central planning it is expected that economic signals will play a fundamental role in defining location, capacity, and time of investments. Concerning short-term operation, it is necessary to have a set of market rules such that reliability-related products can be commercialized in a more flexible and competitive way. It is expected that the adequacy of generation systems is a problem that can be properly addressed in a framework of competitive market rules, in both long-term planning and short-term operation.

Transmission systems are essentially different from generation systems due to the factual impossibility of direct competition. Because of the large investments required and the problems in finding

new sites for high-voltage lines, transmission systems will continue to be seen as natural monopolies. How to assure reliability in the long term under the new industry environment is yet an open problem. Even though there is no consensus about the solution to this problem, currently there is an ongoing discussion about considering transmission as a sector that can be contemplated as a performance-based industry rather than a guaranteed rate-of-return industry. With this change in regulation plus the implementation of new transmission technologies such as flexible AC transmission systems (FACTS) and the acceptance of distributed generation, competition for some of the transmission reliability-associated products can be expected.

On the other hand, in the short term, for a given transmission network, the real questions are reserve requirement calculation and reserve allocation procedures. Adequacy problems derived from transmission bottlenecks or transmission failures lead to undesired supply interruptions and highly volatile prices. This problem is much more severe in cases in which demand plays a passive role. To be consistent with the aim of competition in power systems, procedures for reserve allocation should be able to consider explicitly reliability requirements submitted by the demand. Composite systems reliability analysis should be further developed on the basis of the fundamentals discussed here, with the aim to provide tools for proper reliability-related risk management for both system operator and market participants. The ISO needs reliability analysis tools to define the reserve requirements on the basis of the desired reliability level, and market participants need tools to be able to manage their exposure to dispatch risks for any kind of generation capacity. On the basis of this analysis, a market participant can optimize bids for energy and reserve on a portfolio basis.

SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electric Power: Critical Infrastructure Protection • Electric Power Generation: Fossil Fuel • Electric Power Generation: Valuation of Environmental Costs • Electric Power Measurements and Variables • Electric Power Reform: Social and Environmental Issues • Electric Power Systems Engineering • Electric Power: Traditional Monopoly Franchise Regulation and Rate Making

Further Reading

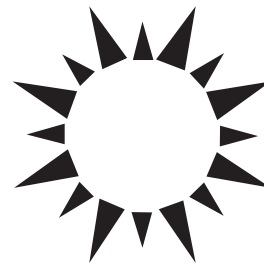
- Allan, R., and Billinton, R. (2000). Probabilistic assessment of power systems. *Proc. IEEE*, **88**, 140–162.
- Alvarado, F. (2000). The relationship between reliability and market design: Where are we? “NSF/DOE/EPRI Workshop: Future Research Directions for Complex Interactive Electric Networks,” Nov. 16–17. Dept. of Energy, Washington, D.C.
- Billinton, R., and Allan, R. N. (1996). “Reliability Evaluation of Power Systems.” 2nd Ed. Plenum, New York.
- CIGRE Task Force 38.03.12. (1997). Power system security assessment: A position paper. *Electra* **175**, 49–78.
- North American Electric Reliability Council. (1997). “Operating Policies and Planning Standards.” Available on the Internet at <http://www.nerc.com>.

Electromagnetic Fields, Health Impacts of

DAVID L. MCCORMICK

IIT Research Institute

Chicago, Illinois, United States



-
1. Introduction
 2. General Toxicity Studies of EMF
 3. Developmental and Reproductive Toxicity Studies of EMF
 4. Immunotoxicity Studies of EMF
 5. Genetic Toxicity Studies of EMF
 6. Oncogenicity Studies of EMF
 7. Conclusion

Glossary

clastogenic Capable of inducing DNA strand breaks.

mutagenic Capable of inducing mutations in DNA.

oncogenic Capable of inducing a tumor or other neoplasm.

terata Birth defects.

teratogenesis Abnormal processes of fetal development leading to birth defects.

Efforts to identify and quantitate possible health hazards of exposure to power frequency (50/60 Hz) magnetic fields include both epidemiology studies of human populations and hazard identification and mechanistic studies performed in experimental model systems. The results of these studies demonstrate that magnetic field exposure induces no systemic toxicity in humans or in experimental animals. Similarly, magnetic field exposure has no adverse effects on reproduction or fetal development in humans or in animal models that are commonly used for hazard identification. Magnetic fields do not induce mutations or chromosome damage in standard genetic toxicology test systems, and extensive studies in a wide range of animal models provide no clear evidence that magnetic field exposure is oncogenic in any test system. Epidemiology studies

suggest a small increase (relative risk ≤ 2.0) in the risk of leukemia in children exposed to magnetic fields; this increase is not seen in other sites in children, or in any organ site in adults. Because no compelling biological mechanism has been identified through which magnetic fields may induce neoplasia or other adverse health effects, further studies are required to determine whether magnetic fields are causally linked to childhood leukemia.

1. INTRODUCTION

Over the past two decades, the possible relationship between human exposure to power frequency (50 and 60 Hz) electromagnetic fields (EMF) and adverse health outcomes has generated significant interest in both the scientific community and the general population. In the late 1970s and early 1980s, several epidemiology studies were published whose data suggested that adverse health effects may be seen in human populations receiving either occupational or environmental (i.e., residential) EMF exposures. Health effects associated with EMF in these studies included reproductive dysfunction, increased risk of birth defects and other developmental anomalies, and increased risk of cancer in several sites. These suggestive epidemiology data stimulated a large number of additional epidemiologic and experimental investigations, whose goals are to (1) confirm or refute the existence of specific health hazards of EMF exposure and (2) where such hazards are demonstrated to exist, identify EMF flux densities at which adverse health effects are observed, in order to quantify the risks of human exposure to EMF.

On the basis of the type of system studied, scientific investigations into the possible health

effects of EMF can be divided into three general categories:

1. Epidemiology studies, in which the risk of a disease is compared in human populations with different levels of EMF exposure
2. *In vivo* experimental studies (animal bioassays), in which the risk of disease is compared in groups of experimental animals with different levels of EMF exposure
3. Mechanistic studies, in which the effects of EMF exposure on biochemical, molecular, or other disease-associated end points are evaluated using experimental animals or *in vitro* model systems

Each category has important strengths and limitations that should be considered in the overall interpretation of EMF bioeffects data.

1.1 Strengths and Limitations of Epidemiology Studies of EMF

Clearly, the most important strength of epidemiology studies is that they are investigations of human responses to EMF exposure. As such, no interspecies extrapolations are required, and the EMF flux densities being investigated in epidemiology studies are, by definition, “real world.” For these reasons, any observation of a strong positive association between EMF exposure and risk of a specific disease would provide compelling evidence of a relationship between the exposure and that adverse health effect. Conversely, however, epidemiology studies are limited in that they can only demonstrate associations between an exposure and disease risk. Epidemiology cannot provide conclusive demonstrations of causality: spurious associations may be identified in epidemiology studies, and weak causal associations may be missed due to limited statistical power.

A critical factor that must be assessed in the interpretation of EMF epidemiology studies is the accuracy and relevance of the exposure metric on which comparisons are based: Is the estimate of EMF exposure developed using (in order of decreasing desirability)

- Measurements of EMF flux densities to which an individual is actually exposed, as obtained by a personal dosimeter?
- Measurements of EMF flux densities present at a relevant location in the home or work site?
- Measurements of EMF flux densities made at a constant point outside of the home (“front door” measurements)?

- An imputed or calculated level of EMF exposure that is based on job classification, residential wire code, residential distance from a power line or substation, or other metric?

It has been demonstrated that EMF exposures that are measured or imputed using different strategies can vary substantially; more specifically, important criticisms have been raised concerning the accuracy of residential wire codes that have been used widely as a surrogate for residential EMF exposure levels and job title classifications that have been used as an estimator of occupational EMF exposures.

A parallel consideration is the temporal variability of EMF measurements: Does today’s measurement of an EMF flux density in a house or other location provide an accurate estimate of the magnetic field environment that was present in the same location 5 years ago, 10 years ago, or at some other earlier time that is relevant to the etiology of the disease in question? Because EMF flux densities in any specific location can demonstrate considerable temporal variability, the ability of such measurements to support retrospective epidemiology studies remains an important issue.

1.2 Strengths and Limitations of Experimental Studies of EMF

The strengths and weaknesses of studies of the effects of EMF on disease induction in experimental animals are essentially opposite of those seen in epidemiology studies. Because animal bioassays are performed in a nonhuman test system, interspecies extrapolation of study data is required. In addition to possible species differences in response, field strengths studied in animal bioassays often greatly exceed those seen in human exposures and may thereby induce artifactual responses by overwhelming a physiologic system. These two factors provide the most important limitations to the use of experimental animal data to assess the potential risk of human exposure to EMF or any other exogenous agent.

Conversely, however, several factors provide important support for the use of animal models in EMF hazard assessment. A number of animal model systems that have been used in studies of EMF toxicology are also used widely in the safety assessment of pharmaceuticals, environmental chemicals, and other agents to which humans receive broad population exposure. These model systems are well-characterized, have extensive historical databases, and have been shown to be generally

predictive of human responses. Animal studies also provide the opportunity to conduct bioassays under precisely controlled and monitored exposure conditions and to incorporate statistical considerations into study designs. In consideration of the difficulties associated with EMF exposure assessment in epidemiology studies, the ability to conduct animal studies under rigorously controlled exposure conditions is a major strength of this approach to hazard identification. Finally, through the use of genetically engineered animal models, possible effects of EMF on susceptible subpopulations can be investigated. Studies in such susceptible subpopulations may identify hazards that cannot be identified in normal animals or humans, but that may substantially increase the risk of disease in individuals with a genetic predisposition, pre-existing condition, or other factor that increases disease risk.

1.3 Role of Mechanistic Studies in EMF Hazard Assessment

The results of mechanistic investigations performed in experimental animals or *in vitro* model systems provide critical data that can establish the biological framework in which observed health effects of EMF can be interpreted. Although identification of a mechanism of action for a toxic agent is not essential for hazard identification, elucidation of one or more biochemical, molecular, or cellular pathways through which an agent exerts biological effects will clearly strengthen hazard assessments developed on the basis of epidemiology and animal bioassays.

Conversely, should the results of epidemiology studies and animal bioassays be negative or conflicting, the identification of a clearly definable mechanism of action can both (1) provide a rational biological context for hazard identification and (2) support the design of optimized hazard identification studies. However, should the results of epidemiology studies and animal bioassays be negative, the lack of an identified biological mechanism of action further weakens the assertion that a substantive human health hazard exists.

2. GENERAL TOXICITY STUDIES OF EMF

In consideration of the absence of documented case reports of systemic EMF toxicity, it is generally

agreed that acute, subchronic, or chronic exposure to EMF causes little or no systemic or generalized toxicity in humans. Because certain occupational environments contain "high field" areas in which EMF flux densities exceed those of residential environmental by tens to hundreds of fold, it is logical to expect that such high field environments would be the primary sites of any observed EMF toxicities. However, no case reports of EMF-associated generalized toxicity or malaise in any exposed worker have been published.

The lack of systemic toxicity of EMF exposure is further supported by the results of acute and subchronic animal bioassays conducted in rodent models. Although occasional instances of statistically significant differences from sham controls have been reported, no clear pattern of EMF toxicity identifiable by effects on animal survival, clinical observation, body weight, food consumption, hematology, clinical chemistry, or histopathology has been reported.

As such, no evidence of generalized toxicity associated with EMF exposure has been identified in studies of either humans or experimental animals. On this basis, it can be concluded that EMF exposure presents little or no hazard of systemic acute, subchronic, or chronic toxicity.

3. DEVELOPMENTAL AND REPRODUCTIVE TOXICITY STUDIES OF EMF

Investigations into the possible developmental and reproductive toxicity of EMF were stimulated by early reports of teratogenesis in chick embryos exposed to high EMF flux densities. Although the processes of prenatal development differ substantially in chickens and in humans, the induction of terata in early chick embryos has been interpreted by some investigators as supporting the hypothesis that EMF is a teratogen. By contrast, however, teratologists who use animal models to develop human hazard assessments for chemical and physical agents have questioned the value of the chick embryo model for human hazard identification.

The early teratology data generated in chicken models were extended by reports of small increases in the incidence of birth defects and spontaneous abortions in women who used electric blankets and electric waterbed heaters during pregnancy. These early findings were hypothesis generating, but were limited in that they were retrospective studies that

included no measurements of EMF exposure. Furthermore, the design of these studies essentially precluded the separation of possible EMF effects from possible thermal effects associated with the use of an electric blanket or bed heater.

Since the publication of these early findings, a large number of epidemiology studies have been performed to investigate the hypothesis that EMF exposure is a risk factor for teratogenesis. These studies have examined the incidence of birth defects, growth retardation, and premature delivery, among other end points, in the offspring of women who used electric blankets or electric bed heaters during pregnancy. Other studies have examined rates of these abnormalities in pregnant women with different levels of residential EMF exposure and in pregnant women with different measured or imputed levels of occupational EMF exposure. In general, although occasional EMF-associated differences have been reported, the results of these studies failed to confirm the hypothesis that EMF exposure during pregnancy is a significant risk to the developing fetus. In most studies, no statistically significant relationships were found between EMF exposure and the risk of birth defects or other developmental abnormalities, regardless of whether the increased EMF exposure resulted from the use of an electric blanket or bed heater, from a higher level of EMF in the residential environment, or from EMF exposures that occurred as a result of employment.

Supporting these epidemiologic data are the results of more than a dozen teratology studies that have been conducted in animal model systems. In this regard, it is essential to note that the rodent models and study designs used in the EMF teratology studies are used routinely to evaluate the possible teratogenicity of pharmaceuticals and environmental chemicals; these models have been validated extensively and have been shown to be generally predictive of human responses. The results of these studies, which have been conducted by several different teams of investigators in several countries, have been uniformly negative. When the negative results generated in these experimental evaluations of EMF teratogenesis are considered together with the negative results of most epidemiology studies, the overwhelming majority of available teratology data do not support the hypothesis that EMF exposure during pregnancy constitutes an important hazard to the developing fetus.

As was the case with teratology studies, no body of evidence exists to implicate EMF as a significant risk for fertility. Although the total number of available studies is small, no evidence of reduced fertility

has been reported in either men or women receiving occupational exposure to high levels of EMF. Although the number of experimental studies of EMF and reproductive function is similarly small, animal studies of male and female fertility are similarly negative. These studies, which commonly involve exposure of several generations of rodents to the test agent, have failed to identify any decreases in fertility in either male or female animals exposed to EMF.

4. IMMUNOTOXICITY STUDIES OF EMF

Evaluation of the possible effects of EMF exposure on immune function is important to support our understanding of the possible role of EMF in carcinogenesis and as a modifier of host resistance to infection. Any reproducible down-regulation of immune function in EMF-exposed animals could have important consequences in terms of the host's ability to resist infectious disease challenge or to provide effective immune surveillance against neoplastic cells.

The body of studies in which the influence of EMF on immune function parameters has been evaluated is small and is almost entirely limited to studies conducted using rodent model systems. In these studies, a battery of immune function assays (lymphoid organ weight, lymphoid organ cellularity, lymphoid organ histology, B- and T-cell mediated immune responses, NK cell function, lymphocyte subset analysis, host resistance assays) was performed on mice and/or rats exposed to magnetic fields for period ranging from 4 weeks to 3 months. In these studies, no alterations were found in lymphoid organ weights, lymphoid organ cellularity, or lymphoid organ histology; similarly, no deficits in T-cell function or B-cell function were identified, and lymphocyte subset analysis did not reveal any effects of EMF exposure. Interestingly, however, a statistically significant suppression of NK cell function was seen in two studies performed in female mice; this alteration was not seen in male mice of the same strain, nor in either sex of rat evaluated in a parallel study.

Although the authors conclude that this suppression of NK cell function is real, it appears to have little or no biological significance. The primary role of NK cell appears to be in immune surveillance against cancer cells. However, completely negative results were obtained in a 2-year carcinogenesis bioassay of EMF performed in the same strain of mice that demonstrated the reduction in NK cell

function. Because the deficit in NK cell function was not associated with an increase in cancer risk, it is concluded that either (1) the quantitative reduction ($\sim 30\%$) in NK cell function induced by EMF is too small to be biologically significant or (2) other arms of the host immune system were able to perform required immune surveillance in these animals.

5. GENETIC TOXICITY STUDIES OF EMF

The possible activity of EMF exposure in the induction of genetic mutations, DNA breaks, and other types of DNA damage is a key mechanistic consideration that underlies the interpretation of data relevant to the possible activity of EMF as a carcinogen. Although nonmutational mechanisms may also be responsible for carcinogenic activity, mutation and DNA damage are a common mechanism of action for carcinogens. As such, any evidence of genetic toxicity for EMF would have substantial implications for its potential activity as a risk factor for human cancer.

Numerous studies have been performed over the past 20 years to investigate the hypothesis that EMF exposure induces mutations or other genetic damage. Studies to evaluate the possible genetic toxicity of EMF have been performed using a broad range of experimental model systems, including *in vitro* bacterial mutagenesis assay systems (such as the Ames test), *in vitro* mutagenesis and DNA damage test systems using mammalian cells, and *in vivo* assays of DNA damage. A substantial number of the *in vitro* genetic toxicology studies of EMF have been performed using test systems that are used widely in the safety assessment of pharmaceuticals and chemical agents. As such, a large database exists to support the utility of these assays to predict *in vivo* responses.

The results of the overwhelming majority of genetic toxicology studies published in the peer-reviewed literature have been negative; as such, a broad assessment of the genetic toxicology database does not support the hypothesis that EMF is genotoxic in either bacteria or in mammalian cells. Virtually all genetic toxicology studies that have been performed using what are considered to be “standard” genetic toxicology test systems (bacterial cell mutagenesis [Ames test], mammalian cell mutagenesis, micronucleus assays) have generated negative data; the quantity of negative genetic toxicology data for EMF in these test systems would ordinarily be

sufficient for a federal regulatory agency to consider EMF to be nonmutagenic.

It should be noted that not all genetic toxicology studies of EMF have been negative. Positive evidence of genetic toxicity has been obtained under certain very limited experimental conditions (i.e., electric shocks and spark discharges) associated with power frequency electric fields; however, these specialized situations are not encountered in most residential or occupational environments and therefore appear to be a little relevance to overall EMF hazard assessment. Most other positive reports have involved studies in nonstandard experimental systems whose relevance to human hazard identification is unclear. A few positive results have been obtained in studies using test systems in relatively wide use in the genetic toxicology community; however, efforts to replicate the positive data have failed.

Aside from specialized situations that are not encountered in routine residential or occupational environments, there is no consistent body of experimental evidence that demonstrates that either power frequency magnetic fields or a combination of power frequency magnetic + electric fields (nonshocking and nonsparking) induce mutations, DNA breaks, or other DNA damage in any standard genetic toxicology test system. On this basis, it is generally accepted within the scientific community that power frequency EMF is not genotoxic.

6. ONCOGENICITY STUDIES OF EMF

Although other possible health effects of residential or occupational EMF exposure are clearly of concern, the primary focus of most EMF hazard assessment studies has been to determine whether exposure to EMF increases the risk of cancer or other neoplasms. The majority of epidemiologic studies of EMF published by the end of the 20th century have been focused on cancer end points. Similarly, a large fraction of experimental resources dedicated to studies of EMF action have been directed toward the question of cancer hazard assessment.

6.1 Epidemiology Studies of EMF and Cancer

Although the results of a number of individual epidemiology studies have provided suggestive evidence of the potential oncogenicity of EMF, the

total body of epidemiologic data linking EMF exposure and cancer risk is by no means conclusive. The results of several EMF epidemiology studies published in the late 1970s and early 1980s suggested that EMF exposure may be a risk factor for cancer of the brain, breast, and hematopoietic system. However, other epidemiologic studies of similar populations failed to confirm these findings, and no clear picture has emerged from this body of literature.

Since 1990, nearly 100 epidemiology studies and meta-analyses designed to investigate the possible association between occupational or residential exposure to magnetic fields and cancer risk have been published. The experimental methods used in these studies are often substantially improved in comparison to methods used in earlier investigations; specific areas of improvement include the use of larger sample sizes and substantial improvements in EMF exposure assessment. However, despite these improvements in epidemiologic methods, EMF cancer epidemiology studies continue to generate both positive and negative results. When considered together, the results of these studies are insufficient to either clearly support or refute the hypothesis that exposure to EMF is a significant risk factor for human cancer in any organ system.

The most comprehensive evaluation of the possible carcinogenicity of power frequency magnetic fields was that published by a working group assembled by the International Agency for Research on Cancer (IARC). The conclusion of this group, which reviewed essentially every epidemiology study of EMF and cancer that has been published in a peer-reviewed journal, was that limited evidence exists to support the hypothesis that EMF exposure is a risk factor for childhood leukemia. The increased risk of childhood leukemia associated with EMF exposure is small (relative risk of <2.0), but is a generally consistent finding across a relatively large number of epidemiology studies. By contrast, however, the conclusion of this group was that the sum of the epidemiology data is inadequate to support the hypothesis that EMF exposure is a significant risk factor for leukemia or any other type of neoplasm in adults and is inadequate to support the hypothesis that EMF exposure is a significant risk factor for any type of malignancy other than leukemia in children. As such, the IARC panel concluded that suggestive evidence of EMF oncogenicity exists only for childhood leukemia and not for any other neoplasm in either children or adults.

Because the increased risk of childhood leukemia in EMF-exposed groups is small, epidemiology studies are continuing in order to quantify more precisely the risks involved, and to confirm that this observed elevation in leukemia risk does, in fact, result from EMF exposure and is not the result of some confounding factor.

6.2 Experimental Studies of EMF and Cancer

In situations where epidemiology does not support the conclusive identification and quantitation of the potential risks associated with exposure to an environmental agent, laboratory studies conducted in appropriate experimental model systems increase in importance. Well-designed and controlled animal studies permit evaluation of biological effects *in vivo* under tightly controlled exposure and environmental conditions, as well as in the absence of potential confounding variables. In consideration of the conflicting results of EMF epidemiology studies and difficulties associated with exposure assessment in such studies, animal studies may provide the best opportunity to identify effects of EMF exposure that could translate into human health hazards.

6.2.1 Assessment of the Possible Oncogenic Activity of EMF as a Single Agent

Several teams of investigators have published the results of studies that are designed to investigate the possible oncogenic activity of chronic exposure to EMF alone (as opposed to EMF exposure in combination with simultaneous or sequential exposure to other agents). These studies have generally been conducted using the chronic rodent bioassay design, a standardized study design in which animals are exposed to the test agent for the majority of their natural life span (2 years), followed by a complete microscopic evaluation of approximately 50 tissues from every animal to identify induced cancers and other lesions. This study design has been used widely to evaluate the potential carcinogenicity of new drugs, chemicals, and environmental agents; the IARC and the United States National Toxicology Program (NTP) have both performed statistical analyses demonstrating the utility of the chronic rodent bioassay as a predictor of human carcinogenicity. As a result, it is generally accepted among toxicologists that the chronic bioassay in rodents currently provides the best available experimental

approach to identify agents that may be carcinogenic in humans.

In all studies conducted using this study design, chronic exposure to power frequency EMF had no effect on the incidence of neoplasia in any of the three putative targets for EMF action (brain, breast, and hematopoietic system) that had been identified by epidemiology. Furthermore, chronic exposure to EMF did not induce any pattern of biologically significant increases in cancer incidence in other tissues. The very close comparability of the results of chronic rodent oncogenicity bioassays conducted independently in three laboratories strengthens the conclusion that exposure to magnetic fields does not induce cancer in rodents. On the basis of the demonstrated predictive nature of the rodent 2-year bioassay, these findings do not support the hypothesis that EMF is a significant risk factor in the etiology of human neoplasia.

6.2.2 Site-Specific Rodent Oncogenicity Studies

It is generally accepted that exposure to power frequency EMF does not induce mutations or other genetic damage that is identifiable using standard genetic toxicology model systems. Because EMF is apparently not genotoxic, it is logical to conclude that any oncogenic effects of exposure are mediated through co-carcinogenic or tumor promotion mechanisms. To address this possibility, a relatively large number of *in vivo* studies have been conducted to identify possible co-carcinogenic or tumor promoting effects of EMF in various target tissues. The design of these studies involves simultaneous or sequential exposure to EMF in combination with another agent (generally a chemical carcinogen or ionizing radiation) and is focused on neoplastic development in a specific organ. Most multistage tumorigenesis studies of EMF have been conducted using animal models that are designed as research tools to study site-specific carcinogenesis. As a result, the utility of these assays to predict human oncogenicity is less completely understood than are the results of the chronic oncogenicity bioassay.

Studies of the effects of EMF on cancer induction in multistage models for cancer of the brain, hematopoietic system, and liver have been uniformly negative: in all cases, exposure to an inducing agent + EMF induced no more neoplasms than were induced by exposure to the inducing agent alone. Conflicting data have been presented for studies of the effects of EMF exposure on the induction of skin and breast cancers in multistage animal models: in

each case, positive data have been presented by one laboratory, but independent studies conducted in one or more additional laboratories have failed to confirm the original finding.

Overall, the peer-reviewed literature provides little compelling evidence that exposure to EMF has co-carcinogenic or tumor promoting activity in any organ site in experimental animals. Completely negative results were found in studies conducted using models for cancer in two of the three target organs (brain and hematopoietic system) identified in epidemiology studies as possible targets for EMF action; data in the third putative target tissue (breast) are conflicting. These data generally support the conclusions from the 2-year oncogenicity bioassays and do not support the hypothesis that EMF exposure is a key risk factor for cancer induction in these target tissues.

6.2.3 Oncogenicity Studies in Genetically Engineered Animals

Insights into the molecular mechanisms of disease have provided the scientific basis for the development and use of genetically modified animals to study disease processes and the application of these models in hazard identification. The science supporting the use of genetically engineered animals in safety assessment stems from initial observations that the incidence of malignancy is increased or its latency is decreased in transgenic animal strains into whose germ line a tumor-associated gene (oncogene) has been inserted and in knockout animal strains from whose germ line a tumor suppressor gene has been deleted. Genetically engineered animals demonstrate a predisposition to specific diseases and thereby provide an approach to hazard identification that may identify effects that occur in sensitive subpopulations but rarely or not at all in the general population. As a result, genetically modified models may demonstrate sensitivity to EMF bioeffects that do not occur in standard animal strains.

Although the total number of completed oncogenicity evaluations of EMF in genetically engineered animals is small, the results of published studies are consistent. In three studies conducted in the United States and in Australia, EMF exposure had no effect on the induction of hematopoietic neoplasia in one transgenic mouse strain (two independent studies) and in one knockout mouse strain. Because EMF exposure did not increase the incidence of hematopoietic neoplasia in any experimental group, the data from these genetically engineered mouse models can be considered to confirm and extend the negative

oncogenicity data generated in the 2-year rodent oncogenicity bioassays of EMF.

6.3 Possible Mechanisms of EMF Action in Tumorigenesis

Over the past two decades, researchers have proposed a number of possible mechanisms through which exposure to EMF may act to induce or stimulate carcinogenesis; these proposed mechanisms include a wide range of potential effects at the molecular, cellular, and whole organism levels. Review of the published scientific literature provides no clear support for any specific mechanism of EMF action in carcinogenesis. Furthermore, no reproducible body of scientific evidence exists to link EMF exposure to alterations in any biological end point that is mechanistically related to neoplastic development in any tissue.

The majority of hypothesized mechanisms of EMF action in carcinogenesis can be grouped into five general types of effects:

1. Genetic toxicity (DNA damage, clastogenesis or mutagenesis)
2. Alterations in gene expression
3. Biochemical changes associated with regulation of cell proliferation
4. Alterations in immune function
5. Alterations in melatonin levels or action

These general classifications are necessarily arbitrary and were developed to simplify the purposes of this review. It is clear that several of these mechanisms overlap and that not all possible mechanisms of EMF action are included in these five groups. However, these groups do encompass the vast majority of mechanisms proposed for EMF bioeffects and do include mechanisms that have received the most extensive study in order to identify possible linkages between EMF exposure and carcinogenesis.

6.3.1 Genetic Toxicity

The possible genotoxicity of EMF exposure was reviewed in Section 5. As indicated in that section, numerous well-designed, complete, reproducible, and apparently well-conducted studies have failed to demonstrate any genotoxic effects of EMF.

It is generally agreed that ELF magnetic fields do not possess sufficient energy to damage DNA via direct interaction. Although it is possible that power frequency EMF could damage DNA through an indirect mechanism (i.e., via generation of free

radicals), an overwhelming majority of genetic toxicology studies conducted in standard safety assessment models have failed to identify any significant mutagenic or clastogenic effects of EMF exposure. As such, a consensus has emerged in the scientific community that exposure to power frequency EMF presents little to no risk of genetic toxicity.

6.3.2 Alterations in Gene Expression

The possible influence of EMF exposure on the expression of cancer-related genes is one of the most controversial areas of research in bioelectromagnetics. Over the past decade, several investigators have reported increases in the expression of one or more protooncogenes (generally *c-myc* or *c-fos*) in cell cultures exposed to 60 Hz EMF. However, replication studies conducted by numerous other investigators have not confirmed these findings, and other, more broadly based screening efforts to identify genes that are up- or down-regulated by EMF exposure have generally yielded negative results. On this basis, the robustness of the reports of altered gene expression, as well as the generality of those findings, must be questioned.

Given the huge universe of genetic targets in the mammalian cell, it is considered likely that exposure to 60 Hz EMF may alter the expression of one or more genes. It is important to note, however, that a finding of enhanced expression of an oncogene or any other gene is not, by itself, sufficient evidence to suggest physiologic or pathophysiologic activity. Because many cellular pathways exhibit biological redundancy, alterations in one pathway may be compensated for by an antagonistic change in another pathway. For this reason, although findings of altered gene expression may be considered to suggest a biologically important effect, demonstration of that effect is necessary to establish the biological relevance of the observed alteration in gene expression. Lacking the demonstration of any physiological/pathophysiological effect, an isolated finding of altered gene expression cannot be construed to be biologically significant. Restated with emphasis on the process of carcinogenesis, a finding of altered expression of an oncogene that is associated with EMF exposure can by no means be equated with any activity of EMF as a carcinogen or tumor promoter.

6.3.3 Biochemical Changes Associated with Regulation of Cell Proliferation

Changes in the expression or activity of enzymes and protein products involved in cell proliferation, cell

cycle regulation, and apoptosis would be expected if EMF exposure was significantly associated with either complete carcinogenesis or tumor promotion. Several reports of this type of EMF bioeffect (e.g., alterations in the activity of ornithine decarboxylase) have been presented in the literature. However, these reports generally demonstrate little consistency across investigators or experimental model systems; as such, the results of individual studies do not appear to be supportable as a general mechanism for EMF activity.

As discussed earlier with regard to possible alterations in gene expression, an isolated finding of changes in enzyme activity or other pathway in cells that are exposed to EMF can be considered to be no more than suggestive of a possible adverse outcome. Linkage of these biochemical or cell-based effects to the physiological or pathophysiological effect is essential to demonstrate their biological significance.

6.3.4 Alterations in Immune Function

Down-regulation of the host immune response by EMF is a tempting hypothesis through which EMF exposure could be associated with risk of carcinogenesis. However, as reviewed in Section 4, this hypothesis is supported by little experimental evidence: with the exception of an apparently biologically insignificant effect on NK cell function, the results of all studies of EMF effects using standardized immunotoxicology test systems and end points have failed to demonstrate effects of EMF on any arm of the host immune response. Although the hypothesis that EMF exposure stimulates tumorigenesis through an indirect mechanism involving suppression of host immunity is intellectually appealing, it is unsupported by any substantial body of experimental evidence.

6.3.5 Alterations in Melatonin Levels or Action

The “melatonin hypothesis” is perhaps the most widely accepted potential mechanism for the biological effects of EMF exposure. A simplified version of this hypothesis states that EMF bioeffects are mediated through alterations in the function of the pineal gland; these changes in pineal function are then responsible for alterations in patterns of melatonin synthesis, kinetics, or action. Ultimately, reductions in melatonin synthesis may have a number of potential adverse effects on the host, including reduced activity as an endogenous free radical scavenger, and influences on the hypothalamic-pituitary-gonadal axis.

A considerable body of evidence to support the melatonin hypothesis was generated in bioelectromagnetics studies performed during the 1980s and early 1990s; in these studies, reductions in circulating melatonin levels or time shifts in the circadian rhythm of melatonin biosynthesis were reported in both experimental animals and in humans. However, several leading investigators who previously reported significant effects of EMF exposure on pineal function have been unable to replicate their own data. This failure of replication is often associated with the acquisition of new, more sophisticated EMF exposure and monitoring equipment. Furthermore, the majority of papers from other laboratories have failed to demonstrate any consistent pattern of effects of EMF exposure on melatonin synthesis or kinetics in humans or in rodent model systems.

One notable exception to this trend is a series of studies that have been conducted using the Siberian hamster as a model system; EMF effects on melatonin secretion in this species appear to be a reproducible finding. However, this animal demonstrates a seasonal circadian rhythm that is considerably different from that seen in either humans or rodent species that are commonly used for safety assessments; as such, a relationship between this finding and possible human health effects is difficult to establish.

Concomitant with the weakening of the data set relating EMF exposure to alterations in pineal function, the melatonin hypothesis has been revised to suggest that EMF influences on melatonin may be effected through alteration in melatonin effects at the level of the target cell, rather than through effects on its synthesis or secretion. The database supporting this hypothesis is very limited and is considered insufficient to provide strong support for this potential mechanism of EMF action.

7. CONCLUSION

With one possibly notable exception, the large body of epidemiologic and experimental data generated over the past 20-plus years generally fails to support the hypothesis that exposure to power frequency EMF is associated with adverse health effects in either humans or experimental animals. No substantive evidence of generalized toxicity resulting from acute, subchronic, or chronic exposure to EMF has been reported in either humans or animal model systems. EMF exposure appears to have no biologically significant effects on host immunity and does not impair fertility in either sex. Although some

conflicting data have been generated, EMF exposure does not appear to be a significant risk factor for the developing fetus.

Those data notwithstanding, however, the primary consideration in most hazard assessments is whether the agent being evaluated causes cancer. EMF exposure is not genotoxic, and the results of oncogenicity evaluations conducted using standard animal models, multistage animal models, and genetically engineered animal models are overwhelmingly negative. Similarly, the results of epidemiology studies provide no clear evidence of EMF oncogenicity in any organ site in adults. However, a small, but consistent increase in the risk of childhood leukemia has been observed in a substantial number of EMF epidemiology studies; the consistent nature of this finding was considered to be sufficient for the IARC Working Group to conclude that “limited evidence” of activity exists for EMF in the induction of childhood leukemia.

In summary, epidemiologic evidence provides limited evidence that EMF exposure is associated with an increased risk of childhood leukemia; by contrast, epidemiology studies have not identified increases in cancer risk for any other site in children or for any tissue in adults. No compelling body of epidemiologic or experimental evidence exists to demonstrate any other toxicologic effect of exposure to power frequency EMF, and no substantive biological mechanism has been identified for EMF

action. Further epidemiology studies will be required to characterize the relationship between EMF and childhood leukemia more completely.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution, Health Effects of • Climate Change and Public Health: Emerging Infectious Diseases • Electricity, Environmental Impacts of • Electric Power Generation: Valuation of Environmental Costs • Electromagnetism • Radiation, Risks and Health Impacts of

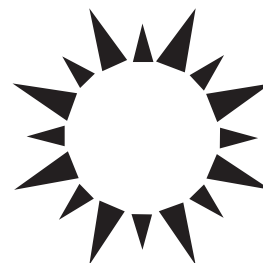
Further Reading

- International Agency for Research on Cancer (2003). IARC Monographs on the Evaluation of Carcinogenic Risk to Humans, Volume 80. Non-Ionizing Radiation, Part 1: Static and Extremely Low-Frequency (ELF) Electric and Magnetic Fields. Lyon, France.
- McCann, J., Dietrich, F., and Rafferty, C. (1998). The genotoxic potential of electric and magnetic fields: An update. *Mutation Research* 411, 45–86.
- McCann, J., Dietrich, F., Rafferty, C., and Martin, A. O. (1993). A critical review of the genotoxic potential of electric and magnetic fields. *Mutation Research* 297, 61–95.
- Portier, C. J., and Wolfe, M. S. (eds.). (1998). “Assessment of Health Effects from Exposure to Power-Line Frequency Electric and Magnetic Fields.” National Institute of Environmental Health Sciences, Research Triangle Park, NC.

Electromagnetism

DANIEL R. STUMP

Michigan State University
East Lansing, Michigan, United States



-
1. Introduction
 2. Electric Charge and Current
 3. The Field Concept
 4. Electromagnetic Induction
 5. Applications of Electromagnetic Induction
 6. The Maxwell Equations of the Electromagnetic Field
 7. Electromagnetic Waves
 8. Summary

Glossary

electric charge A property of the elementary particles of matter that gives rise to electric and magnetic interactions; electrons have negative charge and protons have positive charge.

electric current Net flow of electric charge due to the motion of charged particles in a conducting material.

electric field Lines of force produced by an electric charge and exerting a force on other electric charges.

electromagnetic induction Phenomena in which a change of magnetic field induces an electric field or electromotive force (EMF); the basis of the electric generator.

magnetic field Lines of force produced by an electric current or ferromagnet and exerting a force on other electric currents or on the poles of a magnet.

Maxwell's equations A set of four equations that describe the electric and magnetic fields as functions of position and time.

Electromagnetism is the science of electric and magnetic fields and of the interactions of these fields with electric charges and currents, and as such it is one of the fundamental interactions of nature. Its physical origin lies in a property possessed by elementary particles of matter—electrons and protons—called electric charge. The electromagnetic interaction also governs light and other forms of radiation. Electricity and magnetism appear in many natural phenomena and are applied in many inventions of technology and

everyday life. This article describes the basic science and applications of electromagnetism.

1. INTRODUCTION

The earliest discoveries of electric and magnetic forces were made by philosophers of ancient Greece. They observed that when amber is rubbed with animal fur, it acquires the ability to attract small bits of reed or feathers. This small effect was the first observation of static electricity. (The Greek word for amber, *ελεκτρον*, is the origin of our word “electric.”) They also observed that a lodestone exerts a force on iron—the first example of magnetism. (The Greek province of Magnesia, where the iron ore magnetite occurs naturally, is the origin of our word “magnetic.”) These discoveries of weak and mysterious forces were the first steps toward our scientific understanding of electromagnetism. Today science has dispelled much of the mystery of electricity and magnetism, and created technological power beyond the dreams of the ancient philosophers.

2. ELECTRIC CHARGE AND CURRENT

2.1 Electrostatics and Electric Charge

Electric charge is a property of matter. At the most basic level, the constituents of atoms are charged particles—electrons with negative charge ($-e$) and protons with positive charge ($+e$). An atom has equal numbers of electrons and protons so it is electrically neutral. However, a sample of matter becomes electrically charged if the balance between electrons and protons is broken. For example, when amber is rubbed with fur, electrons are transferred from the fur to the amber; the amber then has net negative charge.

Like charges repel and unlike charges attract. That is, if two samples of matter have net positive charge (or both have net negative charge), they exert equal but opposite repulsive forces on one another. If the samples have unlike charges, one positive and the other negative, then each exerts an attractive force on the other. The strength F of the electric force $\mathbf{F}$ was measured accurately by Charles Augustin de Coulomb (Vectors are indicated in boldface, scalars in plain face). The force is proportional to the product of the charges, q_1 and q_2 and inversely proportional to the square of the distance of separation r ,

$$F = \frac{kq_1q_2}{r^2} \quad \text{where} \quad k = 8.99 \times 10^9 \text{ Nm}^2/\text{C}^2.$$

This simple mathematical formula has been tested to extreme precision. The dimensional units for the constant k are Nm^2/C^2 , where N=newton=unit of force, m=meter=unit of length, and C=coulomb=unit of charge. The proton charge is $e = 1.602 \times 10^{-19}$ C. It forms part of the foundation for the theory of electromagnetism.

2.2 Magnetostatics and Electric Current

Everyone has observed magnets and their forces, which can be quite strong even in small magnets. Every magnet has polarity—north and south poles. Two magnets repel if their north poles approach one another and repel if the south poles approach; but they attract if north approaches south. Magnets have a special attraction to iron; if either pole is placed near an iron object, a magnetic force pulls the magnet and the object toward one another. Science has identified the origins of magnetic forces. Technologies based on magnetic forces are used every day throughout the world.

Magnetism is very closely connected to the electric charge of subatomic particles. The most familiar example of a magnet is a ferromagnet—a piece of magnetized iron. However, a ferromagnet is not the only source of magnetism nor even the most basic. Electric currents also produce magnetic forces, and in some ways the magnetic field associated with electric current is the most basic form of magnetism. The term magnetic field used in this section refers to any magnetic effect. The more technical meaning of the term is explained in Section 3.

2.2.1 Electric Current as a Source of Magnetic Field

An electric current is a stream of electrically charged particles (of one sign) moving in the same direction.

The current may be constant in time (DC, or direct current), oscillating in time with a constant frequency (AC, or alternating current), or varying in time. Currents can exist in metals and in several other forms of conducting matter. In a metal, some electrons occupy states that extend over large distances (i.e., not bound to a single atomic core). If an electric force is applied, then these conduction electrons will move in response, creating an electric current. (Ohm's law, $V=IR$ where V =potential difference in volts, I =current in amps, and R =resistance in ohms, expresses quantitatively the relation between the electric force and the current.) In metals, the positively charged atomic nuclei are fixed in a crystalline lattice, so the electric current is due to motion of electrons.

The first observation of the magnetic field associated with an electric current was an accidental discovery by Hans Christian Oersted during a public lecture in 1819. The current existed in a metal wire connected across a battery. Oersted noticed that a nearby compass needle deflected while the current was flowing. The strength of the magnetic field at a distance of 1 centimeter from a 1 ampere current is 2×10^{-5} tesla, comparable to the earth's magnetic field of approximately 5×10^{-5} tesla.

Oersted's discovery was studied in detail by Jean Marie Ampère. The magnetic field can be determined quantitatively by measuring the force on a pole of a magnet or the torque on a compass needle. A compass needle in equilibrium points in the direction of the field vector at the position of the compass. Therefore, a compass can be used to map the field directions. Ampère found that the field direction "curls around" the current. Figure 1 shows a segment of current-carrying wire and the associated magnetic field encircling the wire. The field directions follow the right-hand rule: with the thumb of your right hand pointing along the wire in the direction of the current, the fingers naturally curl in the direction of the magnetic field around the wire. The theory of the magnetic field created by an electric current is called Ampère's law.

Electric currents can also exist in materials other than metals, such as plasmas and ionic solutions. The associated magnetic fields may be measured for scientific purposes. For example, the current in a neuron in the brain creates a magnetic field that is measured in magnetoencephalography. Or a lightning strike (current in an ionized path through the atmosphere) creates a magnetic field that may be measured to study the properties of lightning. The magnetic field of the earth is another natural example

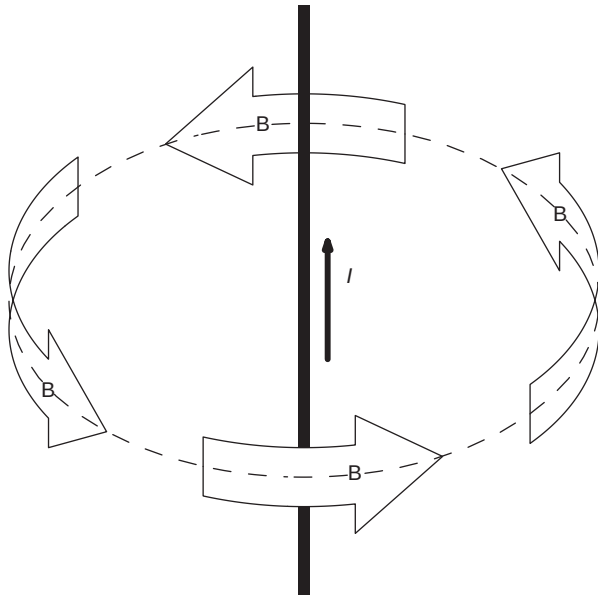


FIGURE 1 The magnetic field B curls around the current I . The dashed curve indicates an imaginary circle around the wire segment. A compass needle placed near the wire will point in the direction of the field.

of a field produced by an electric current. The core of the earth is highly metallic and at high temperature and pressure. Electric currents in this metal core create the earth's magnetism, which we observe at the surface of the earth.

2.2.1.1 Electromagnets Ampère's law is applied in electromagnets. The magnetic field due to a straight length of current-carrying wire is weak. However, if the wire is wound around a cylinder, making a coil with many turns as illustrated in Fig. 2, then the field inside and near the ends of the cylinder can be strong for a practical current. The cylindrical coil is called a solenoid. The field strength can be increased by putting an iron core in the cylinder. Unlike a permanent ferromagnet, the field of an electromagnet can be turned on and off by an electric switch. Electromagnets are commonly used in electric motors and relay switches.

2.2.2 The Magnetic Force on an Electric Current

A magnetic field, whether produced by a ferromagnet or by an electric current, exerts a force on any electric current placed in the field. Indeed, there exists a magnetic force on any moving charged particle that moves across the magnetic field vectors. So the interaction between electric charges in motion and a magnetic field has two aspects: (1) an electric

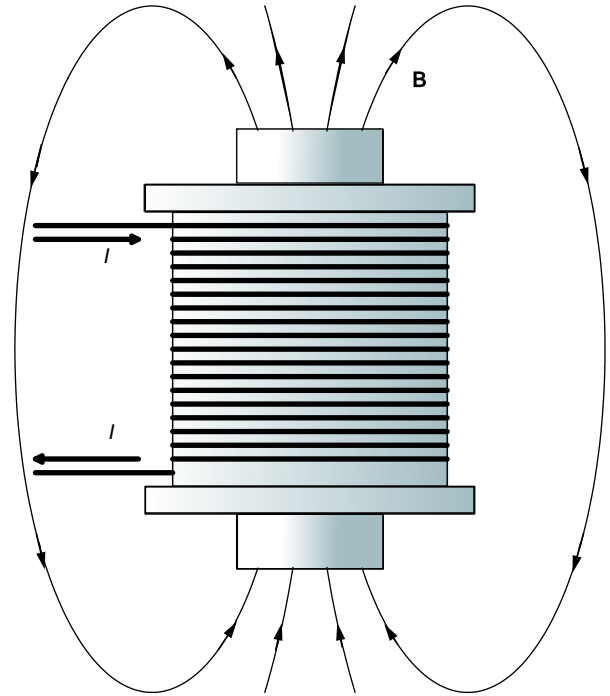


FIGURE 2 An electromagnet.

current creates a magnetic field (Ampère's law); (2) a magnetic field exerts a force on an electric current.

2.2.2.1 The Electric Motor The magnetic force on an electric current is the basis for all electric motors. The magnetic field may be produced by a permanent magnet (in small DC motors) or by an electromagnet (in large DC and AC motors). The current flows in a rotating coil of wire and may be produced by a battery or some other source of electromotive force. Many practical designs have been invented, with the common feature that a magnetic force acts on the current-carrying coil, in opposite directions on opposite sides of the coil, creating a torque that drives the rotation of the coil. A co-rotating shaft attached to the coil is then connected to a mechanical system to do useful work.

3. THE FIELD CONCEPT

In field theory, electric and magnetic forces are described as the effects of electric and magnetic fields. The theory of electromagnetic phenomena is entirely based on the field concept.

Any electrically charged particle q creates an associated electric field E . The field extends throughout the space around q and varies with the position

$\mathbf{x}$ in space. If the particle is at rest (in some frame of reference), then its field is independent of time. In a system with many charges, the full electric field is the sum of the electric fields of the individual charges. Thus, the electric field in an electrostatic system, denoted mathematically by $\mathbf{E}(\mathbf{x})$, is a function of position $\mathbf{x}$ and depends on the locations and charge strengths of all the charged particles.

The electric field extends throughout a volume of space. It exerts a force on any charge in the space. To consider the field concept, suppose two charged particles, q_1 and q_2 , are located at positions $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively (see Fig. 3). The electric field at an arbitrary point $\mathbf{x}$ is

$$\mathbf{E}(\mathbf{x}) = \frac{kq_1}{r_1^2} \mathbf{e}_1 + \frac{kq_2}{r_2^2} \mathbf{e}_2,$$

where r_1 is the distance from q_1 to $\mathbf{x}$, $\mathbf{e}_1$ is the unit vector in the direction from q_1 to $\mathbf{x}$, and q_2 and $\mathbf{e}_2$ are the analogous quantities for q_2 . A small test charge q placed at $\mathbf{x}$ will experience a force $\mathbf{F} = q\mathbf{E}(\mathbf{x})$. Since $\mathbf{E}(\mathbf{x})$ is the sum of the fields due to q_1 and q_2 , the force $\mathbf{F}$ on the test charge is the sum of the two forces exerted on q . The charges q_1 and q_2 also experience forces due to the presence of each other. For example, the force on q_2 due to q_1 is $q_2\mathbf{E}_1(\mathbf{x}_2)$, where $\mathbf{E}_1$ is the field due to q_1 alone. The field due to a charged particle is inversely proportional to the square of the distance from the particle, so the force between charges obeys the inverse-square law observed by Coulomb.

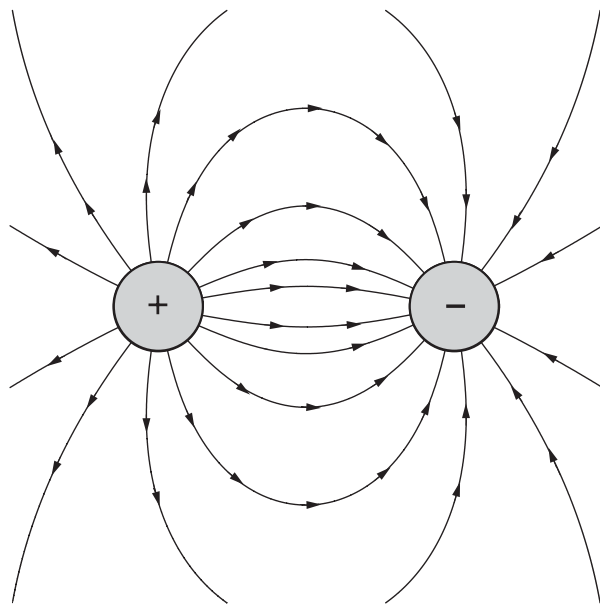


FIGURE 3 The electric field lines of a system of two charges with equal but opposite charge.

In field theory, the force on a charged particle q is attributed to the field $\mathbf{E}(\mathbf{x})$ created by the other charges. Thus, the field concept is significantly different from “action at a distance.” The force on the particle, equal to $q\mathbf{E}(\mathbf{x})$, is exerted by the field at the position of q , rather than by direct actions of the distant charges. In other words, an electrostatic system consists of two physical entities: a set of charged particles and an electric field $\mathbf{E}(\mathbf{x})$. The field is just as real as the particles.

Figure 3 illustrates the electric field for a system of two charged particles with equal but opposite charges. The curves in Fig. 3, called the electric field lines, represent the field. The electric field is a vector at each point throughout the space around the charges. A positive test charge q in the field would experience a force in the direction of the field vector at its position. We visualize the field by drawing the field lines, which are curves that are everywhere tangent to the field vector directions.

Electric charges at rest create an electric field $\mathbf{E}(\mathbf{x})$. Ferromagnets and electric currents create another field—the magnetic field $\mathbf{B}(\mathbf{x})$. Figure 4 illustrates the magnetic fields for two elementary current sources: (1) a small current loop and (2) a long current-carrying wire. In a ferromagnet the atoms behave as small current loops with a uniform orientation. The combined atomic fields make up the field of the magnet.

Both $\mathbf{E}$ and $\mathbf{B}$ are force fields, which extend throughout a volume of space. But they exert distinct and different forces. The electric field $\mathbf{E}$ exerts a force on a charge q in the direction of the field vector. The magnetic field $\mathbf{B}$ exerts a force on moving charges or current-carrying wires. The direction of the magnetic force is perpendicular to both the current and the field, as illustrated in Fig. 5. The magnetic field also exerts forces on the poles of a ferromagnet. The direction of the force is parallel (for a north pole) or antiparallel (for a south pole) to the vector $\mathbf{B}$. A compass needle aligns with the magnetic field because the equal but opposite forces on the two poles of the needle compose a torque that twists the needle toward alignment with the field. The magnetic forces on the current elements around a small current loop also compose a torque; in equilibrium, the plane of the loop is oriented perpendicular to the $\mathbf{B}$ vector. The combined torques on the atoms in a ferromagnet make up the net torque on the magnet.

The field concept was first stated by Michael Faraday. From years of experimental studies on electricity and magnetism, Faraday had formed the idea that a physical entity extends throughout

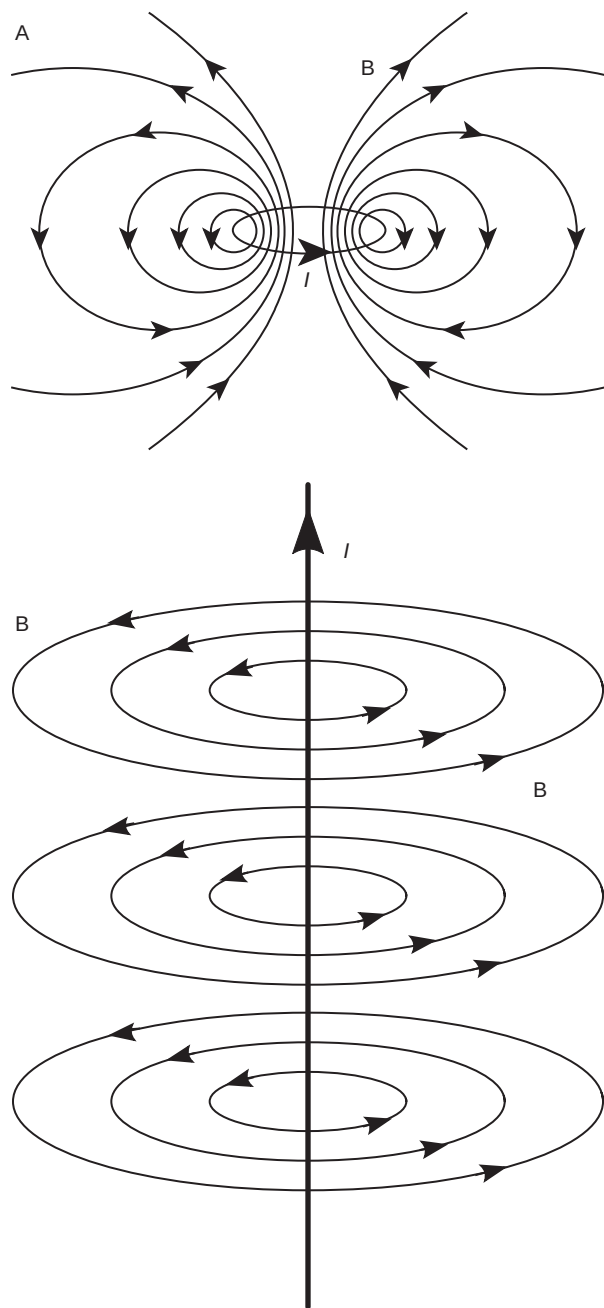


FIGURE 4 Magnetic field of (A) a small current loop and (B) a long straight wire segment.

the space outside charges or magnets and exerts forces on other charges or magnets in the space. He referred to this extended entity as the “lines of force.” The term “electromagnetic field” was coined by James Clerk Maxwell, the renowned theoretical physicist who developed a mathematical theory of electromagnetism based on the field concept.

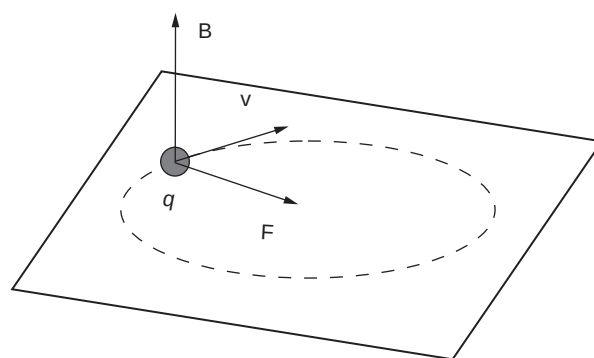


FIGURE 5 The magnetic force on a moving charged particle (or current segment). The direction of the force is perpendicular to both the particle velocity $\mathbf{v}$ (or current) and the magnetic field: $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$. A charged particle moving on a plane perpendicular to a uniform magnetic field moves along a circle.

4. ELECTROMAGNETIC INDUCTION

The previous sections were concerned with static systems of electric charge or current—that is, in which the fields $\mathbf{E}(\mathbf{x})$ and $\mathbf{B}(\mathbf{x})$ do not change in time. Some relationships between electricity and magnetism were discussed: a steady electric current produces a magnetic field (Ampère’s law); a magnetic field exerts a force on any electric charge moving across the field lines. However, for static fields electric and magnetic phenomena appear to be rather distinct.

Time-dependent fields will be described next. The electric field $\mathbf{E}(\mathbf{x}, t)$ and magnetic field $\mathbf{B}(\mathbf{x}, t)$ are functions of time t as well as position $\mathbf{x}$. In dynamic systems, the two fields affect each other significantly. Therefore electric and magnetic phenomena are connected and must be described by a unified theory. Electricity and magnetism are then combined into electromagnetism.

The phenomenon of electromagnetic induction was discovered in 1831 by Michael Faraday in England and independently by Joseph Henry in the United States. The effect is that when a magnetic field changes in time, an electric field is induced in directions that curl around the change of the magnetic field. This phenomenon has important technological applications.

Electromagnetic induction may be observed directly in simple physical demonstrations. Figure 6 shows schematically a coil of conducting wire C connected to a galvanometer G . The galvanometer acts as a current indicator: When no current flows around C the galvanometer needle points toward the

central (zero current) position; when current flows, the needle is deflected. No current source (such as a battery) is connected to the wire coil. In Fig. 6, M is a magnet that can be moved toward or away from the coil C . When the magnet is at rest, no current flows in C and the galvanometer needle points in the central direction.

If the magnet in Fig. 6 moves toward the coil, the galvanometer needle will be deflected in one direction, indicating that a current is flowing in C . The current exists while M is moving. When the motion of M ceases, the needle will return to the central position indicating that the current has stopped. If the demonstration is repeated with the magnet moving away from the coil, the galvanometer needle will be deflected in the opposite direction while M moves. These demonstrations show directly that a change in the magnetic field through the coil induces a current around the coil.

The magnetic field in the demonstration might be varied in a different way. Suppose the bar magnet M in Fig. 6 is replaced by an electromagnet that does not move. Electric current $I(t)$ in the electromagnet solenoid produces a magnetic field $B(\mathbf{x}, t)$ according to Ampère's law. If the solenoid current is constant, then B is constant in time and there is no current in the coil C . But if the solenoid current changes, the magnetic field changes. A deflection of the galvanometer needle will be observed while the magnetic field is varying in time. A current around the sensing coil C is again induced when the magnetic field through the coil is changing.

These demonstrations show that when a magnetic field $B(\mathbf{x}, t)$ changes in time, a current is induced in a conductor that is present in the field. However, the induced current is actually a secondary effect. The current in C is created by an induced electric field $E(\mathbf{x}, t)$, and E is the primary effect. Electromagnetic induction is fundamentally a phenomenon of the fields, $B(\mathbf{x}, t)$ and $E(\mathbf{x}, t)$. A magnetic field that

changes in time induces an electric field in directions curling around the change of the magnetic field. If there happens to be a conducting coil present—as, for example, C in Fig. 6—then the induced electric field drives an electric current around C . But the primary effect is induction of the electric field. The induced electric field $E(\mathbf{x}, t)$ exists while the magnetic field $B(\mathbf{x}, t)$ is varying in time.

The apparatus shown in Fig. 6 is only schematic. The induced current in C would be very small for an ordinary bar magnet M moving at reasonable velocities. A practical demonstration would require a sensing coil with many turns of wire and a sensitive galvanometer. The effect might be increased by putting an iron core inside the coil to enhance the magnetic field.

Faraday and Henry performed laboratory experiments similar to the demonstrations illustrated in Fig. 6 in their discoveries of electromagnetic induction. Faraday described the results of his detailed studies in terms of the lines of force—his concept of a magnetic field filling space around a magnet. In modern language, a statement summarizing his observations is

Faraday's law: When the flux of magnetic field through a loop changes in time, an electromotive force (EMF) is induced around the loop.

This statement is Faraday's law of electromagnetic induction. In equation form, $d\Phi/dt = -\varepsilon$, where Φ is the magnetic flux through the loop, $d\Phi/dt$ is the rate of change of the flux, and ε is the electromotive force around the loop. The “flux” is a quantitative measure of the field lines passing through the loop, defined by the product of the component of the magnetic field vector perpendicular to a surface bounded by the loop and the area of the surface.

The “loop” in Faraday's law is a closed curve (e.g., a circle). The loop may be a conducting loop, in which case the induced EMF drives a current, or it may just be an imaginary curve in space. In either case, the EMF is an induced electric field curling around the change of the magnetic field.

Another, related demonstration may be carried out with the simple apparatus of Fig. 6. Instead of moving the magnet M and holding the coil C fixed, suppose the coil is moved while the magnet is held fixed. Again, a current will be observed around C . The phenomenon in this case is called “motional EMF”; an electromotive force is induced in a conductor that moves relative to a nonuniform magnetic field. In the language of Faraday, when a conducting wire moves through magnetic lines of

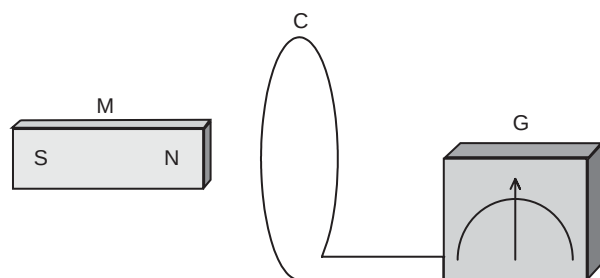


FIGURE 6 Schematic apparatus for demonstrations of electromagnetic induction.

force, an induced current flows in the wire. Evidently, any change of the magnetic flux through a conducting loop will induce a current in the loop.

Yet another way to change the magnetic flux through the coil *C* in Fig. 6 is to change the orientation of the coil. In Fig. 6 the plane of the coil is shown perpendicular to the bar magnet. In this orientation the magnetic field lines pass straight through *C*; the flux is the product of the magnetic field strength *B* and the area *A* of the loop, $\Phi = BA$. Now suppose the coil rotates about a vertical axis, with *M* fixed. Then the flux of magnetic field through the coil changes as the plane of the loop is at a varying angle to the field vector. When the plane of the loop is parallel to the field lines, the flux is zero because no field lines pass through the coil. While the coil is rotating, a deflection of the galvanometer needle will be observed, consistent with Faraday's law, because of the changing flux. This is another example of motional EMF: The conducting wire moves through the magnetic field lines, and there is an induced current in the wire.

5. APPLICATIONS OF ELECTROMAGNETIC INDUCTION

A number of important inventions are based on the phenomenon of electromagnetic induction. Two that have great technological significance will be described here.

5.1 The Electric Generator

An electric generator (or dynamo) is a machine that generates an electromotive force (EMF) by electromagnetic induction. One or more coils of wire rotate in a magnetic field. The magnetic flux through a coil changes as the coil rotates, inducing an EMF around the coil by Faraday's law, $\varepsilon = -d\Phi/dt$.

Figure 7 shows schematically the design principle of an electric generator. In the model, the square coil rotates about the vertical axis that bisects the coil. The model field *B* is the constant field between north and south poles of ferromagnets. As the coil rotates, the flux through it varies sinusoidally as a function of time. Then there is an induced alternating EMF around the coil.

The model generator in Fig. 7 has the ends of the coil in contact with two slip rings. The voltage difference between the rings is equal to the EMF ε

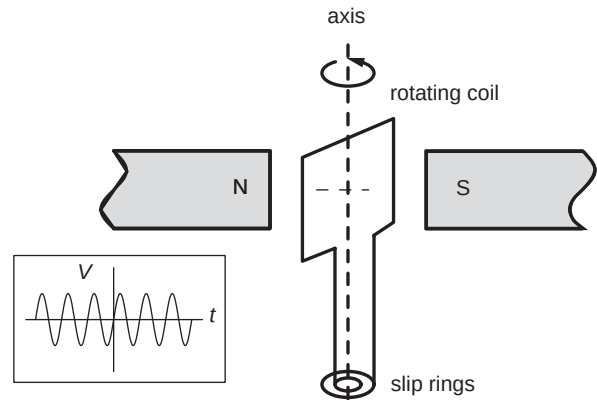


FIGURE 7 The design principle of an electric generator. The magnetic flux through the coil varies as the coil rotates and an EMF is induced in accord with Faraday's law. The voltage difference *V* between the slip rings varies sinusoidally in time (inset graph).

around the coil. Thus the device generates an alternating voltage *V* (shown in the inset graph) with constant frequency equal to the rotational frequency of the coil. If the rings are connected by a wire with resistance *R*, an alternating current (AC) *I* will occur in the wire with $V = IR$ by Ohm's law.

An external torque must be applied to the coil of the generator to maintain the constant rotation. If the slip rings are connected to an electric appliance, then energy is supplied by the generator because the induced EMF drives an alternating current in the appliance. But energy cannot be created—only converted from one form to another. If no external torque is applied, the generator coil will slow down and stop as its rotational kinetic energy is converted to the work done by the appliance. The slowing of the rotation is an example of magnetic braking: a conductor moving in a magnetic field experiences a force opposite to the motion. To maintain a constant rotation of the generator coil (and hence AC power to the appliance), the coil must be driven by a torque exerted by an external device. In a portable generator, the external torque is produced by a gasoline-powered engine. In an electric power plant, the external torque is produced by a turbine driven by a flow of hot steam; the steam is produced in a boiler (e.g., from burning coal in a coal-fired power plant or from the heat of nuclear reactions in a nuclear reactor).

The electric generators producing electric power around the world are large and complex machines; but the fundamental physics in these devices is simple: Faraday's law and the design principle in Fig. 7.

5.2 The Transformer

A transformer takes an input alternating EMF V_1 (e.g., from an electric generator) and makes an output EMF V_2 that might be supplied to an electric appliance (see Fig. 8). The frequencies of V_1 and V_2 are equal, but the amplitudes are different. A step-down transformer has $V_2 < V_1$; for example, a transformer might take in 110 volts (alternating) from a wall socket and put out 10 volts to a small electric motor in an appliance. A step-up transformer has $V_2 > V_1$; high-voltage applications require such a transformer.

The physical phenomenon acting in a transformer is electromagnetic induction. Figure 8 shows a simple transformer: two coils of conducting wire are wrapped around a toroidal ferromagnetic core. The alternating current driven in the primary coil by the input EMF $V_1(t)$ creates a strong alternating magnetic field $\mathbf{B}(t)$, according to Ampère's law. Because of the ferromagnetism, $\mathbf{B}$ extends around the toroid. Then the changing magnetic flux through the secondary coil induces an EMF by Faraday's law, $\varepsilon = -d\Phi/dt$; this EMF is the output voltage V_2 . If N_1 is the number of turns of wire in the primary coil and N_2 is the number of turns in the secondary coil, then the ratio of the voltages is $V_2/V_1 = N_2/N_1$.

Transformers are a crucial part of the electric power grid. An electric power plant creates a three-phase alternating EMF supplied to the grid. But the electric appliances that will use the power are located far from the power plant. Therefore a current must occur over very long transmission lines. The amount of lost power P_{lost} , dissipated in resistance of the transmission wires, is given by the formula $P_{\text{lost}} = I^2 R$ where I = current and R = resistance of the line. The power supplied to users is $P_{\text{used}} = IV$, where V is the transmitted EMF. The ratio of power lost to power used is

$$\frac{P_{\text{lost}}}{P_{\text{used}}} = \frac{(P_{\text{used}}/V)^2 R}{P_{\text{used}}} = \frac{P_{\text{used}} R}{V^2},$$

that is, it is inversely proportional to the square of the line voltage. Therefore the transmission of electric power is most efficient if the voltage V is high.

A typical power-plant generator might produce an alternating EMF with amplitude 168 kV (i.e., root mean square EMF 120 kV). But that voltage would be changed in a step-up transformer at the power plant to a very high voltage (e.g., 345 kV). High-voltage transmission lines carry the electric power across large distances. Then step-down transformers are required at the end of a transmission line to reduce the voltage for applications. A transformer

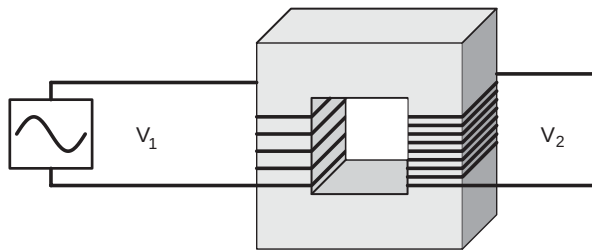


FIGURE 8 The design principle of a transformer. Two coils of conducting wire are wrapped around an iron core. The number of turns is different in the two coils, so the output EMF V_2 is different from the input EMF V_1 .

substation might reduce the voltage to 4.8 kV for overhead lines into residential areas. A final transformer mounted on the utility pole reduces the voltage to 110 V for the wires to a single house.

6. THE MAXWELL EQUATIONS OF THE ELECTROMAGNETIC FIELD

The mathematical theory of electromagnetism was developed and published in 1864 by James Clerk Maxwell. He described the known electric and magnetic effects in terms of four equations relating the electric and magnetic fields and their sources—charged particles and electric currents. The development of this theory was a supreme achievement in the history of science. Maxwell's theory is still used today by physicists and electrical engineers. The theory was further developed in the 20th century to account for the quantum theory of light. But even in quantum electrodynamics, Maxwell's equations remain valid although their interpretation is somewhat different from the classical theory. In any case, Maxwell's theory continues today to be an essential part of theoretical physics.

A knowledge of calculus and vectors is necessary for a full understanding of the Maxwell equations. However, the essential structure of the theory can be understood without going into the mathematical details. Each equation is expressed most powerfully as a partial differential equation relating variations of the electric and magnetic fields, with respect to variations of position or time, and the charge and current densities in space.

Gauss's Law. Gauss's law, written as a field equation, is $\nabla \cdot \mathbf{E} = \rho/\epsilon_0$. (The symbol $\nabla \cdot$ denotes the divergence operator.) Here $\rho(\mathbf{x}, t)$ is the charge per unit volume in the neighborhood of $\mathbf{x}$ at time t ; ϵ_0

is a constant of nature equal to $8.85 \times 10^{-12} \text{ C}^2/\text{Nm}^2$. Gauss's law relates the electric field $\mathbf{E}(\mathbf{x}, t)$ and the charge density. The solution for a charged particle q at rest is $\mathbf{E}(\mathbf{x}) = kqe/r^2$, where r is the distance from the charge to $\mathbf{x}$, $\mathbf{e}$ is the direction vector, and $k = 1/(4\pi\epsilon_0)$; this is the familiar inverse square law of electrostatics. Electric field lines diverge at a point charge.

Gauss's Law for Magnetism. The analogous equation for the magnetic field is $\nabla \cdot \mathbf{B} = 0$. There are no magnetic monopoles, particles that act as a point source of $\mathbf{B}(\mathbf{x}, t)$. Unlike the electric field lines, which may terminate on charges, the magnetic field lines always form closed curves because magnetic charges do not exist. There is no divergence of magnetic field lines.

Faraday's Law. The field equation that describes Faraday's law of electromagnetic induction is $\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t$. The quantity $\nabla \times \mathbf{E}$, called the curl of $\mathbf{E}(\mathbf{x}, t)$, determines the way that the vector field $\mathbf{E}$ curls around each direction in space. Also, $\partial \mathbf{B}/\partial t$ is the rate of change of the magnetic field. This field equation expresses the fact that a magnetic field that varies in time implies an electric field that curls around the change of the magnetic field. It is equivalent to Faraday's statement that the rate of change of magnetic field flux through a surface S is equal to an electromotive force (EMF) around the boundary curve of S .

The Ampère-Maxwell Law. In a system of steady electric currents, the magnetic field is constant in time and curls around the current in directions defined by the right-hand rule. The field equation that expresses this field $\mathbf{B}$ (Ampère's law) is $\nabla \times \mathbf{B} = \mu_0 \mathbf{J}$, where $\mathbf{J}(\mathbf{x})$ is the current per unit area at $\mathbf{x}$ and μ_0 is a constant equal to $4\pi \times 10^{-7} \text{ Tm/A}$. (The units are T = tesla for magnetic field and A = ampere for electric current). But Ampère's law is incomplete, because it does not apply to systems in which the currents and fields vary in time. Maxwell deduced from mathematical considerations a generalization of Ampère's law,

$$\nabla \times \mathbf{B} = \mu_0 \mathbf{J} + \mu_0 \epsilon_0 \frac{\partial \mathbf{E}}{\partial t},$$

in which the second term on the right side is called the displacement current. The displacement current is a necessary term in order for the system of four partial differential equations to be self-consistent. The Ampère-Maxwell law implies that $\mathbf{B}$ curls around either electric current ($\mathbf{J}$) or changing electric field ($\partial \mathbf{E}/\partial t$). The latter case is analogous to electromagnetic

induction but with $\mathbf{E}$ and $\mathbf{B}$ reversed; a rate of change in one field induces circulation in the other field.

Maxwell's introduction of the displacement current was a daring theoretical prediction. At that time there was no experimental evidence for the existence of displacement current. Laboratory effects predicted by the displacement current are very small, and their observation was not possible with the apparatus available at that time. However, the Maxwell equations, including the displacement current, make a striking prediction—that light consists of electromagnetic waves. The fact that Maxwell's theory explains the properties of light, and other forms of electromagnetic radiation, provides the evidence for the existence of the displacement current.

7. ELECTROMAGNETIC WAVES

Electromagnetism includes phenomena involving charges, currents, magnets, and the fields $\mathbf{E}$ and $\mathbf{B}$. Light is also an electromagnetic phenomenon. Light and other forms of radiation are described by field theory as electromagnetic waves. Therefore optics—the science of light—is a part of electromagnetism.

The Maxwell equations describe the behavior of the electric field $\mathbf{E}(\mathbf{x}, t)$ and magnetic field $\mathbf{B}(\mathbf{x}, t)$. The time-dependent fields influence each other, even in a vacuum where no charge or current is present, through Faraday's law ($\nabla \times \mathbf{E} = -\partial \mathbf{B}/\partial t$) and the displacement current ($\nabla \times \mathbf{B} = \mu_0 \epsilon_0 \partial \mathbf{E}/\partial t$). The four field equations are all satisfied in vacuum if both $\mathbf{E}(\mathbf{x}, t)$ and $\mathbf{B}(\mathbf{x}, t)$ are waves traveling in space. However, the electric and magnetic waves are codependent (i.e., intricately related in several ways).

Figure 9 shows a snapshot in time of the fields for a section of an electromagnetic wave in vacuum. The field vectors are shown at points along a line parallel to the direction of propagation of the wave (z axis). The wavelength is constant (at least for the section of wave shown). The field vectors $\mathbf{E}$ and $\mathbf{B}$ vary in magnitude as a function of position along the line of propagation. The magnitudes are related by $B = E/c$, where c is the wave speed. The directions of $\mathbf{E}$ and $\mathbf{B}$ are orthogonal, in the x and y directions, respectively. The field components oscillate sinusoidally with z . An electromagnetic wave is a transverse wave because the fields oscillate in directions orthogonal to the direction of propagation.

Figure 9 shows the field vectors along only a single line parallel to the direction of propagation $\hat{z}$. The full wave fills a three-dimensional volume. In a

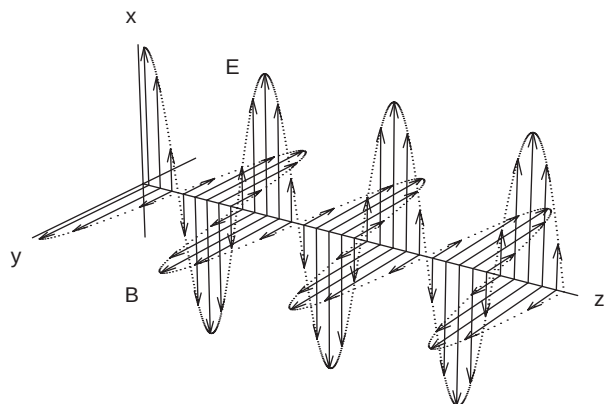


FIGURE 9 Electric and magnetic field vectors in a section of a linearly polarized electromagnetic wave.

linearly polarized plane wave the field vectors are constant on planar wave fronts orthogonal to $\hat{z}$ —that is, with $\mathbf{E}$ and $\mathbf{B}$ pointing in the $\hat{x}$ and $\hat{y}$ directions, respectively, uniformly over the wave front. All properties of polarized light are then described by the wave theory. As another 3D example, a short radio antenna radiates a spherical wave; far from the antenna $\mathbf{E}$ and $\mathbf{B}$ are orthogonal to one another on spherical wave fronts.

A wave is a structure that is extended in space. The parts of the wave also vary in time. Figure 9 is a snapshot showing the wave as a function of position at an instant of time. As time passes, the fields at any fixed position $\mathbf{x}$ will change. For a linearly polarized plane wave, $\mathbf{E}$ at $\mathbf{x}$ oscillates in time between orientations parallel and antiparallel to a fixed polarization direction ($\hat{x}$), which is orthogonal to the line of propagation ($\hat{z}$); $\mathbf{B}$ varies similarly in a direction ($\hat{y}$) orthogonal to both $\mathbf{E}$ and the line of propagation. The combined variations in space and time imply that the wave structure moves as a whole in the direction of propagation. So a snapshot at a later time would look just the same as Fig. 9 except translated by some distance in the direction of propagation. In other words, the shape is constant but the positions of the nodes (points where $\mathbf{E} = 0$ and $\mathbf{B} = 0$) move in space.

7.1 The Electromagnetic Spectrum

Electromagnetic waves have an infinite range of wavelength λ . Visible light has λ from 400 nm (violet)

to 700 nm (red). Ultraviolet and infrared light have shorter and longer wavelengths, respectively. Micro-waves have λ from 1 mm to 1 m, with important technological uses (e.g., communications and radar). X-rays have very short wavelengths, $\lambda < 1$ nm. All these forms of electromagnetic radiation have the same wave speed in vacuum, which can be evaluated from the field theory, $c = 1/\sqrt{\mu_0\epsilon_0} = 2.998 \times 10^8$ m/s.

8. SUMMARY

Our scientific understanding of electricity and magnetism has developed over a period of centuries. Phenomena that we understand today as examples of electricity, magnetism, or electromagnetism have been observed since ancient times: For example, lightning is an electric discharge in the atmosphere; a lodestone attracts iron; light consists of electromagnetic waves. Such common phenomena, while easily observed, were unexplained until the scientific revolution. Today we are able to accurately describe and explain all variety of electromagnetic phenomena, based on the mathematical theory of fields. With this knowledge has come the ability to create technologies, such as the distribution of electric power, that affect our daily lives.

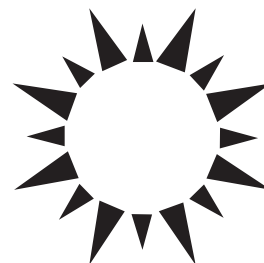
SEE ALSO THE FOLLOWING ARTICLES

Electrical Energy and Power • Electricity Use, History of • Electric Motors • Magnetic Levitation • Magnetohydrodynamics

Further Reading

- Griffiths, D. J. (2000). "Introduction to Electrodynamics." 3rd ed. Prentice-Hall, Engelwood Cliffs, NJ.
- Jackson, J. D. (1998). "Classical Electrodynamics." 3rd ed. John Wiley, New York.
- Lorrain, P., Corson, D. R., and Lorrain, F. (2000). "Fundamentals of Electromagnetic Phenomena." W. H. Freeman, New York.
- Meyer, H. W. (1972). "A History of Electricity and Magnetism." MIT Press, Norwalk, CT.
- Pollack, G. L., and Stump, D. R. (2002). "Electromagnetism." Addison-Wesley, San Francisco.
- Whittaker, E. T. (1960). "A History of the Theories of Aether and Electricity." Harper, New York.

Emergy Analysis and Environmental Accounting



MARK T. BROWN

University of Florida
Gainesville, Florida, United States

SERGIO ULGIATI

University of Siena
Siena, Italy

-
1. Introduction
 2. Emergy Evaluation of the Biosphere and its Processes
 3. Unit Emergy Values for Fuels and Some Common Products
 4. Case Studies: Evaluation of Energy Conversion Systems
 5. Summary

Glossary

emdollar (or EM\$) A measure of the money that circulates in an economy as the result of some process. In practice, to obtain the emdollar value of an emergy flow or storage, the emergy is multiplied by the ratio of total emergy to gross domestic product for the national economy.

emergy An expression of all the energy used in the work processes that generate a product or service in units of one type of energy. Solar emergy of a product is the emergy of the product expressed in the equivalent solar energy required to generate it. Sometimes it is referred to as energy memory.

emjoule The unit of measure of emergy, “emergy joule.” It is expressed in the units of energy previously used to generate the product; for instance, the solar emergy of wood is expressed as joules of solar energy that were required to produce the wood.

empower Emergy per unit time (sej/time)

energy Sometimes referred to as the ability to do work. Energy is a property of all things that can be turned into heat and is measured in heat units (BTUs, calories, or joules)

nonrenewable emergy The emergy of energy and material storages like fossil fuels, mineral ores, and soils that are consumed at rates that far exceed the rates at which they are produced by geologic processes.

production Production measured in emergy is the sum of all emergy inputs to a process.

renewable emergy The emergy of energy flows of the biosphere that are more or less constant and reoccurring, and that ultimately drive the biological and chemical processes of the earth and contribute to geologic processes.

transformity The ratio obtained by dividing the total emergy that was used in a process by the energy yielded by the process. Transformities have the dimensions of emergy/energy (sej/J). A transformity for a product is calculated by summing all of the emergy inflows to the process and dividing by the energy of the product. Transformities are used to convert energies of different forms to emergy of the same form.

This article presents a brief review of the concepts of energy hierarchy and definitions of emergy and related quantities. Tables of data are given on global emergy flows from which emergy and transformities of most products and processes of the biosphere are calculated. Then tables of transformities for many products are provided. Finally, several case studies of evaluations of energy technologies are given. In a summary discussion emergy analysis (EMA) and embodied energy analysis (EEA) are compared describing the strengths and weaknesses of each approach related to the evaluation of energy systems.

1. INTRODUCTION

1.1 Energy and the Emergy Hierarchy

Emergy is a universal measure of the work of nature and society made on a common basis. The work of nature and society results in energy transformations

that when viewed in totality are interconnected in webs of energy flow. All energy transformations of the geobiosphere can be arranged in an ordered series to form an energy hierarchy. For example, many joules of sunlight are required to make a joule of organic matter, many joules of organic matter to make a joule of fuel, several joules of fuel are required to make a joule of electric power, and so on. Since different kinds of energy are not equal in their contribution to processes, work is made comparable by expressing each form of energy in units of one form. To accomplish this, the available energy of different forms of energy is evaluated by means of the energy of one type previously required to produce it. This quantity of one type of energy previously used is called *emergy* (spelled with an “m”).

1.2 Definitions

The following paragraphs contain definitions and a brief explanation of *emergy* concepts. A more complete introduction can be found in H. T. Odum's 1996 text, *Environmental Accounting: Emergy and Environmental Decision Making*.

Emergy is the availability of energy (*exergy*) of one kind that is used up in transformations directly and indirectly to make a product or service. The unit of *emergy* is the *emjoule*, a unit referring to the available energy of one kind consumed in transformations. For example, sunlight, fuel, electricity, and human service can be put on a common basis by expressing them all in the *emjoules* of solar energy that is required to produce each. In this case, the value is a unit of solar *emergy* expressed in solar *emjoules* (abbreviated *sej*). Although other units have been used, such as coal *emjoules* or electrical *emjoules*, in most cases, all *emergy* data are given in solar *emjoules*.

Unit *emergy* values are calculated based on the *emergy* required to produce them. There are three main types of unit *emergy* values as follows:

- *Transformity* is one example of a unit *emergy* value and is defined as the *emergy* per unit of available energy (*exergy*). For example, if 4000 solar *emjoules* are required to generate a joule of wood, then the solar *transformity* of that wood is 4000 solar *emjoules* per joule (abbreviated *sej/J*). Solar energy is the largest but most dispersed energy input to the earth. The solar *transformity* of the sunlight absorbed by the earth is 1.0 by definition.

- *Specific energy* is the unit *emergy* value of matter defined as the *emergy* per mass, usually

expressed as solar *emergy* per gram (*sej/g*). Solids may be evaluated best with data on *emergy* per unit mass for its concentration. Because energy is required to concentrate materials, the unit *emergy* value of any substance increases with concentration. Elements and compounds not abundant in nature therefore have higher *emergy/mass* ratios when found in concentrated form since more work was required to concentrate them, both spatially and chemically.

- *Emergy per unit money* is a unit *emergy* value used to convert money payments into *emergy* units. Since money is paid to people for their services and not to the environment, the contribution to a process represented by monetary payments is the *emergy* that people purchase with the money. The amount of resources that money buys depends on the amount of *emergy* supporting the economy and the amount of money circulating. An average *emergy/money* ratio in solar *emjoules/\$* can be calculated by dividing the total *emergy* use of a state or nation by its gross economic product. It varies by country and has been shown to decrease each year. This *emergy/money* ratio is useful for evaluating service inputs given in money units where an average wage rate is appropriate.

Emergy accompanying a flow of something (energy, matter, information, etc.) is easily calculated if the unit *emergy* value is known. The flow expressed in its usual units is multiplied by the *emergy* per unit of that flow. For example, the flow of fuels in joules per time can be multiplied by the *transformity* of that fuel (*emergy* per unit energy in solar *emjoules/joule*), or the mass of a material input can be multiplied by its specific *emergy* (*emergy* per unit mass in solar *emjoules/gram*). The *emergy* of a storage is readily calculated by multiplying the storage quantity in its usual units by the *emergy* per unit.

Unit *emergy* values are a kind of efficiency measure, since they relate all the inputs to an output. The lower the *transformity* or specific *emergy*, the more efficient the conversion. It follows from the second law that there are some minimum unit *emergy* values for processes, which are consistent with maximum power operations. While there is no way to calculate them directly, the lowest *transformity* found in long-operating systems is used as an approximation. When estimating a theoretical potential of some system, it is appropriate to use the best (lowest) *transformity* known.

Empower is a flow of *emergy* (i.e., *emergy* per time). *Emergy* flows are usually expressed in units of solar *empower* (solar *emjoules* per time).

1.3 Emergy Evaluation Procedure

Emergy accounting uses the thermodynamic basis of all forms of energy, materials, and human services but converts them into equivalents of one form of energy. Emergy accounting is organized as a top-down approach where first a system diagram of the process is drawn to organize the evaluation and account for all inputs and outflows. Tables of the actual flows of materials, labor, and energy are constructed from the diagram and all flows are evaluated. The final step of an emergy evaluation involves interpreting the quantitative results. In some cases, the evaluation is done to determine fitness of a development proposal. In others, it may be a question of comparing different alternatives. The evaluation may be seeking the best use of resources to maximize economic vitality. So the final step in the evaluation is to calculate several emergy indices that relate emergy flows of the system being evaluated with those of the environment and larger economy within which it is embedded and that allow the prediction of economic viability, carrying capacity, or fitness.

This evaluation process has been termed emergy synthesis. Synthesis is the act of combining elements into coherent wholes. Rather than dissect and break apart systems and build understanding from the pieces upward, emergy synthesis strives for understanding by grasping the wholeness of systems. Emergy is a systems concept, context driven, and cannot be fully understood or utilized outside of systems. By evaluating complex systems using emergy methods, the major inputs from the human economy and those coming “free” from the environment can be integrated to analyze questions of public policy and environmental management holistically.

1.3.1 Left-Right Emergy Systems Diagram

Systems diagrams are used to show the inputs that are evaluated and summed to obtain the emergy of a resulting flow or storage. The purpose of the system diagram is to conduct a critical inventory of processes, storages, and flows that are important to the system under consideration and are therefore necessary to evaluate. Components and flows within diagrams are arranged from left to right reflecting more available emergy flow on the left, decreasing to the right with each successive emergy transformation. For example, abundant solar emergy is utilized in successive transformations in ecological, agricultural, and technoeconomic subsystems to support a small amount of high-quality emergy of humans, their

government, and information. A simple diagram of the global system including humans is shown in Fig. 1. The left-to-right organization also corresponds to increasing scale of territory and turnover time. As illustrated in Fig. 1, every emergy transformation box has more than one input, including larger emergy flows from the left, lesser amounts from units in parallel, and small but important controlling emergies feeding back from the right.

1.3.2 Preparation of an Emergy Evaluation Table

Tables of the actual flows of materials, labor, and energy are constructed from the diagram. Raw data on flows and storage reserves are converted into emergy units and then summed for a total emergy flow to the system. Inputs that come from the same source are not added, to avoid double counting. Only the larger input is accounted for. If the table is for the evaluation of a process, it represents flows per unit time (usually per year). If the table is for the evaluation of reserve storages, it includes those storages with a turnover time longer than a year.

Separate tables are constructed for evaluations of flows and storage. Tables are usually constructed in the same format, as given by the column headings and format shown in Table I.

- Column 1 is the line item number, which is also the number of the footnote found below the table where raw data sources are cited and calculations are shown.
- Column 2 is the name of the item, which is also shown on the aggregated diagram.
- Column 3 is the raw data in joules, grams, dollars or other units. The units for each raw data item are shown in column 4.
- Column 5 is the emergy per unit used for calculations, expressed in solar emergy joules per unit. Sometimes, inputs are expressed in grams, hours, or dollars, therefore an appropriate conversion ratio is used (sej/hr; sej/g; sej/\$).
- Column 6 is the solar emergy of a given flow, calculated as raw input times the transformity (column 3 times column 5).
- Column 7 is the emdollar value of a given item for a given year. This is obtained by dividing the emergy in Column 6 by the emergy-to-money ratio (EMR) for the country and selected year of the evaluation (units are sej/\$). The EMR is calculated independently. The resulting values in this column express the amount of economic activity that can be supported by a given emergy flow or storage.

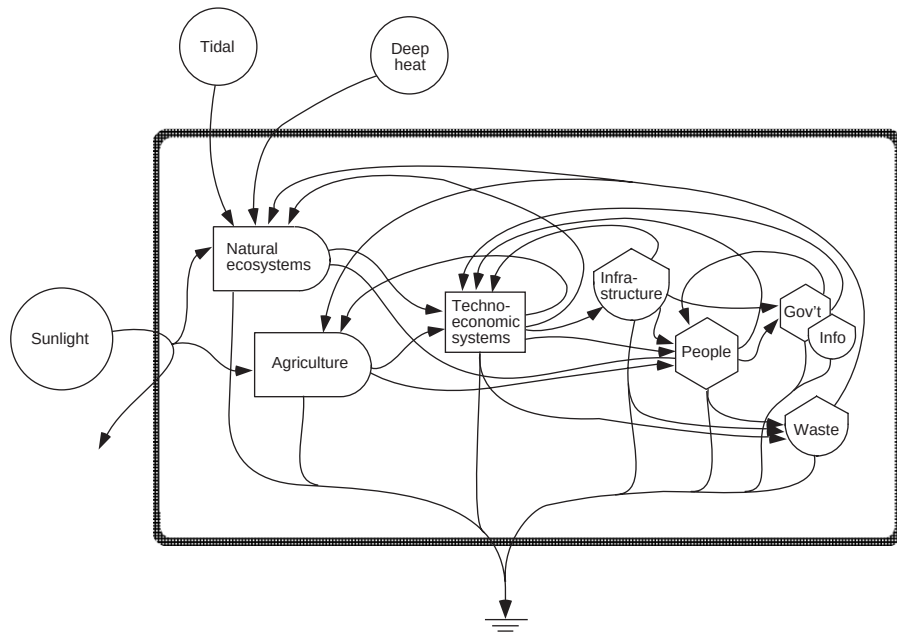


FIGURE 1 Successive energy transformations supporting human civilization. Large energy flows from the right support smaller and smaller amounts of high-quality energy. With each succeeding energy transformation, some energy is degraded some is passed on to the next level (to the left) and some is fed back in control actions. Explanations of symbols may be found in Odum (1996).

TABLE I

Example Emergy Evaluation Table

1	2	3	4	5	6	7
Note	Item	Data	Units	Emergy/unit (sej/unit)	Solar emergy (E + 15 sej/year)	em\$ Value (1998 em\$/year)
1.	First item	xx.x	sej/year	xxx.x	xxx.x	xxx.x
2.	Second item	xx.x	g/year	xxx.x	xxx.x	xxx.x
..						
..						
n.	nth item	xx.x	sej/year	xxx.x	xxx.x	xxx.x
O.	Output	xx.xx	sej or g/year	xxx.x	$\sum_n^1 Em$	xxx.x

1.3.3 Emergy of Storages

When calculating the emergy of stored quantities (storages)—for instance, wood biomass in a forest ecosystem or buildings in a city—it is necessary to sum the emergy of each of the inputs for the time of its contribution. Input emergy inflows are multiplied by the time it takes to accumulate the storage and exported yield, if any.

1.3.4 Evaluations Based on Averaged Inputs

Apparently all systems pulse with time intervals and pulse strength that increase with scale. To evaluate a process on one scale of time and space usually means

using averages for each of the inputs from smaller scales where pulses are of high frequency. For example, for an evaluation of phenomena on the scale of human economy, yearly averages are often appropriate. On this scale average solar energy and average tidal energy are used. For calculations of global processes over a longer scale of time, an average of inputs from the deep earth may be used. Thus, for many purposes emergy evaluations are made with averaged inputs. The result is emergy evaluation as if the system was in a steady state. If the table is an evaluation of a storage, the result is in solar emjoules. If the table is an evaluation of a

product flow, the result is in solar emjoules per time (or empower).

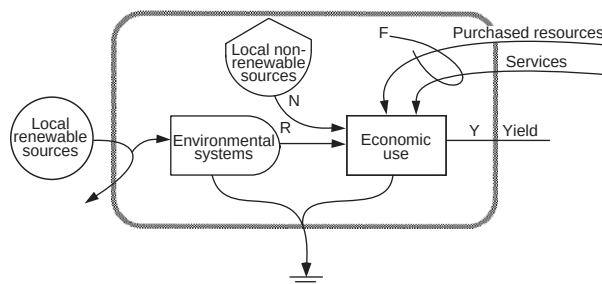
Dynamic energy evaluations of storages and flows are also used. When a storage is growing, decreasing, or oscillating, its stored energy contents and transformity are changing as well. Where a system is adequately represented by a dynamic model, an algorithm can be added to the model which calculates a time sequence of energy and transformity. A simulation model of this type produces a running calculation according to the energy definition, summing the inputs to storages, multiplying these by their unit energy, and subtracting energy removals.

1.3.5 Calculating Unit Energy Values

After a table is prepared that evaluates all the inputs, unit energy values of products can be calculated. The output or product (row "O" in Table I) is evaluated first in units of energy, exergy, or mass. Then the input energy is summed and unit energy value for the product calculated by dividing the energy by the units of the output. The unit values that result are useful for other energy evaluations. Thus, any energy evaluation generates new energy unit values.

1.3.6 Performance Indicators

The systems diagram in Fig. 2 shows nonrenewable environmental contributions (N) as an emergy storage of materials, renewable environmental inputs (R), and inputs from the economy as purchased (F) goods and services. Purchased inputs are needed for the process to take place and include human service and purchased nonrenewable energy and material brought in from elsewhere (fuels, minerals, electricity, machinery, fertilizer, etc.). Several ratios, or indices are given in Fig. 1 that are used to evaluate



Yield $(Y) = R + N + F$
 Emery yield ratio $= Y/F$
 Emery investment ratio $= F/(R + N)$
 Environmental loading ratio $= (F + N)/R$
 Empower density $= (R + N + F)/\text{area}$

FIGURE 2 Emery indices used as performance indicators.

the global performance of a process as follows:

Emery yield ratio. The ratio of the total emery (local and imported) driving a process or system to the emery imported. The ratio is a measure of the potential contribution of the process to the main economy, due to the exploitation of local resources.

Environmental loading ratio. The ratio of non-renewable and imported emery use to renewable emery use. It is an indicator of the pressure of a transformation process on the environment and can be considered a measure of ecosystem stress due to a production (transformation activity).

Emery sustainability index. The ratio of the emery yield ratio to the environmental loading ratio. It measures the potential contribution of a resource or process to the economy per unit of environmental loading.

Emery investment ratio. The ratio of emery fed back from outside a system to the indigenous emery inputs (both renewable and non-renewable). It evaluates if a process is a good user of the emery that is invested, in comparison with alternatives.

Empower density. The ratio of total emery use in the economy of a region or nation to the total area of the region or nation. Renewable and non-renewable emery density are also calculated separately by dividing the total renewable emery by area and the total nonrenewable emery by area, respectively.

Several other ratios are sometimes calculated depending on the type and scale of the systems being evaluated.

Percent renewable emery (%Ren). The ratio of renewable emery to total emery use. In the long run, only processes with high %Ren are sustainable.

Emprice. The emprice of a commodity is the emery one receives for the money spent. Its units are sej/\$.

Emery exchange ratio. The ratio of emery exchanged in a trade or purchase (what is received to what is given). The ratio is always expressed relative to one or the other trading partners and is a measure of the relative trade advantage of one partner over the other.

Emery per capita. The ratio of total emery use in the economy of a region or nation to the total population. Emery per capita can be used as a measure of potential, average standard of living of the population.

1.3.7 Understanding Emergy Indicators: What They Indicate

Lengthy discussion of the performance indicators were given in a previous publication by the authors titled “Emergy Based Indices and Ratios to Evaluate Sustainability—Monitoring Economies and Technology towards Environmentally Sound Innovation.” Additional background and further discussion follows.

Transformity only measures how much emergy it takes to generate one unit of output, regardless of whether or not the input is renewable. It is a measure of efficiency on the global spatial and timescale of the biosphere. It indicates the hierarchical position of an item in the thermodynamic scale of the biosphere and can be regarded as a quality factor from the point of view of biosphere dynamics.

The ELR = $(N + F)/R$ is designed to compare the amount of nonrenewable and purchased emergy ($N + F$) to the amount of locally renewable emergy (R). In the absence of investments from outside, the renewable emergy that is locally available would have driven the growth of a mature ecosystem consistent with the constraints imposed by the environment and would be characterized by an $ELR = 0$. Instead, the nonrenewable imported emergy drives a different site development, whose distance from the natural ecosystem can be indicated by the ratio $(N + F)/R$. The higher this ratio, the bigger the distance of the development from the natural process that could have developed locally. In some ways, the ELR is a measure of the disturbance to the local environmental dynamics, generated by the development driven from outside. The ELR is clearly able to make a difference between nonrenewable and renewable resources, thus complementing the information that is provided by the transformity (Brown and Ulgiati, 1997). From past experience gained in previous case studies investigated, it appears that low ELRs (around two or less) are indicative of relatively low environmental impacts (or processes that can use large areas of a local environment to “dilute impacts”). ELRs between three and ten are indicative of moderate environmental impacts, while ELRs ranging from ten up to extremely high values indicate much higher environmental impacts due to large flows of concentrated nonrenewable emergy in a relatively small local environment.

The emergy yield ratio, $EYR = (R + N + F)/F$, is a measure of the ability of a process to exploit and make available local resources by investing outside resources. It provides a look at the process from a different perspective, its openness. It provides a measure of the appropriation of local resources by

a process, which can be read as a potential additional contribution to the economy, gained by investing resources already available. The lowest possible value of the EYR is 1, which indicates that a process delivers the same amount of emergy that was provided to drive it, and that it is unable to usefully exploit any local resource. Therefore, processes whose EYR is 1 or only slightly higher do not provide significant net emergy to the economy and only transform resources that are already available from previous processes. In so doing, they act as consumer processes more than creating new opportunities for system’s growth. Primary energy sources (crude oil, coal, natural gas, uranium) usually show EYRs greater than 5, since they are exploited by means of a small input from the economy and return much greater emergy flows, which have been generated by previous geologic and ecological activities that accumulated these resources over past millennia. Secondary energy sources and primary materials like cement and steel show EYRs in the range from 2 to 5, indicating moderate contribution to the economy.

If the ELR (nonrenewable versus renewable emergy flows) and EYR (outside versus local emergy flows) are combined, a sustainability index is created—that is, an aggregated measure of the potential contribution to the larger system (EYR) per unit of loading imposed on the local system (ELR). This indicator, called emergy index of sustainability (EIS) is usefully applicable to measure openness and loading changes occurring over time in both technological processes and economies. In principle, the lowest possible value of the EIS is zero (when $EYR = 0$ and $ELR < \infty$ or when $EYR \neq 0$ and $ELR \rightarrow \infty$), while the theoretical upper limit ($\rightarrow \infty$) is only possible for untouched, mature ecosystems. According to the results of several case studies investigated, EISs lower than 1 appear to be indicative of consumer products or processes and those greater than 1 indicative of products that have net contributions to society without heavily affecting its environmental equilibrium. As it relates to economies, an EIS lower than 1 is indicative of highly developed consumer-oriented systems, EISs between 1 and 10 have been calculated for what have been termed “developing economies,” while EISs greater than 10 indicate economies that have not yet significantly started any industrial development.

Finally, the emergy density, ED, measures the amount of emergy invested on one unit of land and suggest land be a limiting factor to any kind of development or process. Higher EDs characterize

city centers, information centers such as governmental buildings, universities, and research institutions, and industrial clusters, while lower EDs are calculated for rural areas and natural environments.

2. EMERGY EVALUATION OF THE BIOSPHERE AND ITS PROCESSES

Figure 3 shows an aggregated system diagram of the energy transformation network of the biosphere arranged with decreasing energy from left to right. In Fig. 3, all the components interact and are required by the others. As a result, the total emergy driving the biosphere (the sum of solar, tidal, and deep heat) is required by all processes within, and thus the emergy assigned to each of the internal pathways is the same. After millions of years of self-organization, the transformations of the driving energies by the atmosphere, ocean, and land are organized simultaneously to interact and contribute mutual reinforcements. Therefore, the energy flow of each jointly necessary process is the sum of the emergy from the three sources.

2.1 Annual Budget of Emergy Flow (Empower) Supporting the Geobiosphere

An emergy evaluation table of the main inputs to the geobiosphere of the earth (omitting, for the moment, the emergy use from nonrenewable resources) is given in Table II. The annual budget of emergy flow (empower) supporting the geobiosphere (atmosphere, ocean, and earth crust) includes solar energy, tidal

TABLE II

Annual Emergy Contributions to Global Processes* (after Odum *et al.*, 2000)

Input	Units	Inflow units/year	Emergy/unit sej/unit	Empower E24 sej/year
Solar insolation,	J ^a	3.93 E24	1.0	3.93
Deep earth heat,	J ^b	6.72 E20	1.20 E4	8.06
Tidal energy,	J ^c	0.52 E20	7.39 E4	3.84
Total	—	—	—	15.83

Abbreviations: sej=solar emjoules; E24 means multiplied by 10^{24} .

*Not including nonrenewable resources.

^aSunlight: solar constant $2 \text{ gcal/cm}^2/\text{min} = 2 \text{ Langley per minute}$; 70% absorbed; earth cross section facing sun 1.27 E14 m^2 .

^bHeat release by crustal radioactivity 1.98 E20 J/year plus 4.74 E20 J/year heat flowing up from the mantle (Sclater *et al.*, 1980). Solar transformity 1.2 E4 sej/J based on an emergy equation for crustal heat as the sum of emergy from earth heat, solar input to earth cycles, and tide (Odum, 2000a).

^cTidal contribution to oceanic geopotential flux is 0.52 E20 J/year (Miller, 1966). Solar transformity of 7.4 E4 sej/J is based on an emergy equation for oceanic geopotential as the sum of emergy from earth heat, solar input to the ocean, and tide following Campbell (1998) (Odum, 2000a).

energy, and heat energy from the deep earth. Other inputs from space, such as the high-energy radiation of solar flares, cosmic rays, meteorites, and stellar dust, are not evaluated. All of these vary with oscillations and pulses, and their emergy values vary with their intensities.

Total emergy contributions to the geobiosphere are about $15.83 \text{ E24 sej/year}$ based on a reevaluation

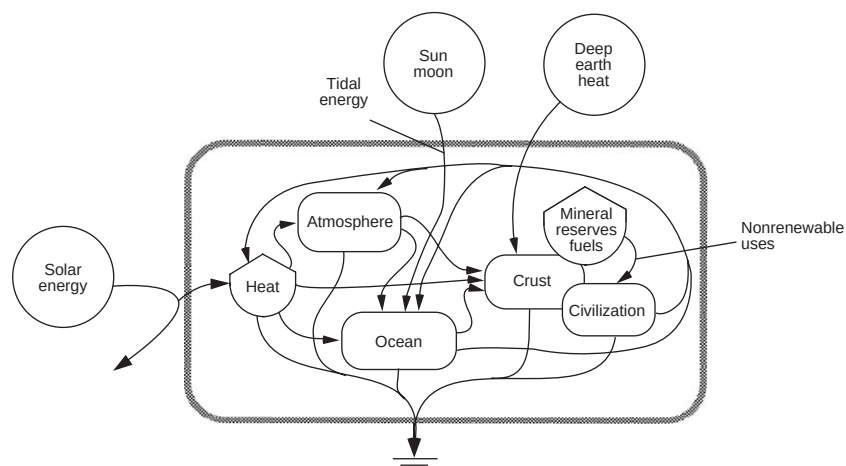


FIGURE 3 Energy transformation network of the biosphere. The interconnected pathways of energy flow, inputs, and all main components are necessary to all the other components and processes.

and subsequent recalculation of energy contributions done in the year 2000. Prior to that date, the total emergy contribution to the geobiosphere that was used in calculating unit emergy values was $9.44 \text{ E24 sej/year}$. The increase in global emergy reference base to $15.83 \text{ E24 sej/year}$ changes all the unit emergy values that directly and indirectly were derived from the value of global annual empower. Thus, unit emergy values calculated prior to that year are multiplied by 1.68 (the ratio of $15.83/9.44$).

2.2 Average Emergy Unit Values for Main Global Processes

Table III calculates unit emergy values for some main flows of the earth. The total emergy input to the geobiosphere in solar emergy (15.83 E24 sej/yr from Table II) is divided by each of the global product's ordinary measure (number of joules or grams). The unit values that result are useful for other emergy evaluations where global averages can be used.

2.3 Temporary Emergy Inputs to the Geobiosphere

In the 19th and 20th centuries, the production and consumption processes of human civilization using the large emergy in the geologic stores of fuels and minerals reached a scale with global impact. Because these storages are being used much faster than they are being generated in geologic cycles, they are often called nonrenewable resources. They are actually very slowly renewed resources. Table IV summarizes these additional components of the global emergy budget.

At present, the emergy contribution through the human civilization to the geobiosphere from the slowly renewed resources is greater than the inputs from renewable sources. A symptom of this surge of temporary emergy inflow is the carbon-dioxide accumulation in the atmosphere that causes greenhouse effects that change the pattern and intensity of weather.

It is important to note that the unit emergy values given in Table III are average values for global processes. It is well understood that there is no single

TABLE III

Emergy of Products of the Global Energy System (after Odum et al., 2000)

Product and units	Emergy* E24 sej/year	Production units/year	Emergy/unit sej/unit
Global latent heat, J^a	15.83	1.26 E24	12.6 sej/J
Global wind circulation, J^b	15.83	6.45 E21	2.5 E3 sej/J
Global precipitation on land, g^c	15.83	1.09 E20	1.5 E5 sej/g
Global precipitation on land, J^d	15.83	5.19 E20	3.1 E4 sej/J
Average river flow, g^e	15.83	3.96 E19	4.0 E5 sej/g
Average river geopotential, J^f	15.83	3.4 E20	4.7 E4 sej/J
Average river chem. energy, J^g	15.83	1.96 E20	8.1 E4 sej/J
Average waves at the shore, J^b	15.83	3.1 E20	5.1 E4 sej/J
Average ocean current, J^i	15.83	8.6 E17	1.8 E7 sej/J

*Main empower of inputs to the geobiospheric system from Table II not including nonrenewable consumption (fossil fuel and mineral use).

^aGlobal latent heat = latent heat of evapotranspiration 1020 mm/year (1020 mm/year) ($1000 \text{ g/m}^2/\text{mm}$) (0.58 kcal/g) (4186 J/kcal) (5.1 E14 m^2) = 1.26 E24 J/year .

^bGlobal wind circulation, 0.4 watts/m^2 (Wiin-Nielsen and Chen, 1993) ($0.4 \text{ J/m}^2/\text{sec}$) (3.15 E7 sec/year) ($5.12 \text{ E14 m}^2/\text{earth}$) = 6.45 E21 J/year .

^cGlobal precipitation on land = $1.09 \text{ E11 m}^3/\text{year}$ (Ryabchikov, 1975) (1.09 E14 m^3) (1 E6 kg/m^3) = 1.09 E20 g/year .

^dChemical potential energy of rainwater relative to seawater salinity (1.09 E20 g/year) ($4.94 \text{ J Gibbs free energy/g}$) = 5.19 E20 J/year .

^eGlobal runoff, $39.6 \text{ E3 km}^3/\text{year}$ (Todd, 1970) ($39.6 \text{ E12 m}^3/\text{year}$) (1 E6 g/m^3) = 3.96 E19 g/year .

^fAverage river geopotential work; average elevation of land = 875 m ($39.6 \text{ E12 m}^3/\text{year}$) (1000 kg/m^3) (9.8 m/sec^2) (875 m) = 3.4 E20 J/year .

^gChemical potential energy of river water relative to seawater salinity (3.96 E19 g/year) ($4.94 \text{ J Gibbs free energy/g}$) = 1.96 E20 J/year .

^hAverage wave energy reaching shores, (Kinsman, 1965) ($1.68 \text{ E8 kcal/m/year}$) ($4.39 \text{ E8 m shore front}$)(4186 J/kcal) = 3.1 E20 J/year .

ⁱAverage ocean current: 5 cm/sec (Oort et al., 1989); 2-year turnover time (0.5) ($1.37 \text{ E21 kg water}$) (0.050 m/sec) (0.050 m/sec/ (2 year) = 8.56 E17 J/year .

TABLE IV

Annual Emergy Contributions to Global Processes Including Use of Resource Reserves (after Brown and Ulgiati, 1999)

Inputs and Units	Inflow (J/year)	Emergy/unit* (sej/unit)	Empower E24 sej/year
Renewable inputs ^a	—	—	15.8
Nonrenewable energies released by society			
Oil, J ^b	1.38 E20	9.06 E4	12.5
Natural gas (oil eq.), J ^c	7.89 E19	8.05 E4	6.4
Coal (oil eq.), J ^d	1.09 E20	6.71 E4	7.3
Nuclear power, J ^e	8.60 E18	3.35 E5	2.9
Wood, J ^f	5.86 E19	1.84 E4	1.1
Soils, J ^g	1.38 E19	1.24 E5	1.7
Phosphate, J ^h	4.77 E16	1.29 E7	0.6
Limestone, J ⁱ	7.33 E16	2.72 E6	0.2
Metal ores, g ^j	9.93 E14	1.68 E9	1.7
Total nonrenewable empower			34.3
Total global empower			50.1

Abbreviations: sej = solar emjoules; E3 means multiplied by 10³; t = metric ton; oil eq. = oil equivalents.

*Values of solar emergy/unit from Odum (1996) and modified to reflect a global resource base of 15.83 E24 sej/year.

^aRenewable Inputs: Total of solar, tidal, and deep heat empower inputs from Odum (1996).

^bTotal oil production = 3.3 E9 Mt oil equivalent (British Petroleum, 1997) Energy flux = (3.3 E9 t oil eq.) (4.186 E10 J/t oil eq.) = 1.38 E20 J/year oil equivalent.

^cTotal natural gas production = 2.093 E9 m³ (British Petroleum, 1997) Energy flux = (2.093 E12 m³) (3.77 E7 J m³) = 7.89 E19 J/year.

^dTotal soft coal production = 1.224 E9 t/year (British Petroleum, 1997) Total hard coal production = 3.297 E9 t/year (British Petroleum, 1997) Energy flux = (1.224 E9 t/year) (13.9 E9 J/t) + (3.297 E9 t/year)(27.9 E9 J/t) = 1.09 E20 J/year.

^eTotal nuclear power production = 2.39 E12 kwh/year (British Petroleum, 1997); Energy flux = (2.39 E12 kwh/year)(3.6 E6 J/kwh) = 8.6 E18 J/year electrical equivalent.

^fAnnual net loss of forest area = 11.27 E6 ha/year (Brown et al., 1997) Biomass = 40 kg m²; 30% moisture (Lieth and Whitaker, 1975) Energy flux = (11.27 E6 ha/year) (1 E4 m²/ha) (40 kg m²)(1.3 E7 J/kg)(0.7) = 5.86 E19 J/year.

^gTotal soil erosion = 6.1 E10 t/year (Oldeman, 1994; Mannion, 1995) Assume soil loss 10 t/ha/year and 6.1 E9 ha agricultural land = 6.1 E16 g/year (assume 1.0% organic matter), 5.4 kcal/g Energy flux = (6.1 E16 g)(.01) (5.4 kcal/g)(4186 J/kcal) = 1.38 E19 J/year.

^hTotal global phosphate production = 137 E6 t/year (USDI, 1996) Gibbs free energy of phosphate rock = 3.48 E2 J/g Energy flux = (137 E12 g)(3.48 E2 J/g) = 4.77 E16 J/year.

ⁱTotal limestone production = 120 E6 t/year (USDI, 1996) Gibbs free energy of phosphate rock = 611 J/g Energy flux = (120 E12 g)(6.11 E2 J/g) = 7.33 E16 J/year.

^jTotal global production of metals 1994: Al, Cu, Pb, Fe, Zn (World Resources Institute, 1996): 992.9 E6 t/year = 992.9 E12 g/year.

unit emergy value for any given product, since no two processes are alike. This also holds for the processes of the biosphere. For instance, there are many transformities for rain depending on location and even time of year. Precipitation varies with altitude, is affected by mountains, and depends on the weather systems in complex ways. The evaluations in Table III are for the whole earth with 70% ocean. If the land is regarded as a higher level in the hierarchical organization of the geobiosphere, then rain over the land represents a convergence of oceanic resources as well as those of the continents, and calculation of continental rain transformity includes all geobiosphere driving energy (see rows 3 and 4 in Table III). As a result, continental rainfall has a higher transformity compared to the global average. To carry this idea even farther, the

rainfall in any particular location may have a higher or lower transformity depending on the source area and intensity of the solar energy driving the cycles that produce it.

3. UNIT EMERGY VALUES FOR FUELS AND SOME COMMON PRODUCTS

Unit emergy values result from emergy evaluations. Several tables of unit emergy values for some common materials and energy sources follow. In Table V, unit emergy values are given for primary nonrenewable energy sources along with the emergy yield ratio (EYR). In some cases, the EYR is based on only one

TABLE V

Unit Emergy Values for Primary Nonrenewable Energy Sources (After Odum, 1996, Updated)

Item	Transformity (Sej/J)	Sej/g	EYR
Plantation pine (<i>in situ</i>)	1.1 E4	9.4 E7	2.8
Peat	3.2 E4	6.7 E8	3.0
Lignite	6.2 E4		6.8
Coal	6.7 E4		8–10.5
Rainforest wood (chipped, trans.)	6.9 E4	4.1 E8	2.2
Natural gas	8.1 E4		6.8–10.3
Crude oil	9.1 E4		3.2–11.1
Liquid motor fuel	1.1 E5		2.6–9.0
Electricity	3.4 E5		—

evaluation for instance plantation pine. In other cases, several evaluations have been done of the same primary energy but from different sources and presumably different technology, so a range of values are given. For instance, the range of values for natural gas (6.8 to 10.3) represent the difference between offshore natural gas (6.8) and on natural gas produce in the Texas field on shore (10.3). Obviously each primary energy source has a range of values depending on source and technology. By using data from typical production facilities (and actually operating facilities), the unit emergy values represent average conditions and can be used for evaluations when actual unit values are not known. If it is known that conditions where an evaluation is being conducted are quite different than the averages suggested here, then detailed evaluations of sources should be conducted.

Table VI lists the unit emergy values for some common products in the order of their transformity. Only a few products are given here, while many more evaluations leading to unit emergy values have been conducted and are presented in a set of emergy folios, published by the Center for Environmental Policy at the University of Florida (www.ees.ufl.edu/cep).

Calculations of emergy production and storage provide a quantitative basis for making choices about environment and economy. The unit emergy values done previously aid new evaluations, since they may save time and energy of authors attempting other evaluations because they do not need to reevaluate all inputs from scratch. In light of this, a series of emergy folios have been published, and others planned, that provide data on emergy contents, the computations on which they were based, and comparisons (www.ees.ufl.edu/cep/emerg

TABLE VI

Unit Emergy Values for Some Common Products (After Odum, 1996, Updated)

Item	Transformity (Sej/J)	Specific emergy (Sej/g)
Corn stalks	6.6 E4	
Rice, high energy ^a	7.4 E4	1.4 E9
Cotton	1.4 E5	
Sugar (sugarcane) ^b	1.5 E5	
Corn	1.6 E5	2.4 E9
Butter	2.2 E6	
Ammonia fertilizer	3.1 E6	
Mutton	5.7 E6	
Silk	6.7 E6	
Wool	7.4 E6	
Phosphate fertilizer	1.7 E7	
Shrimp (aquaculture)	2.2 E7	
Steel ^b	8.7 E7	7.8 E9

^a After Brown and McKlanahan (1996).

^b After Odum and Odum (1983).

gydownloads.asp). To date there are five folios by as many authors, who take the initiative to make new calculations or assemble results from the extensive but dispersed literature in published papers, books, reports, theses, dissertations, and unpublished manuscripts. The tabulating of unit emergy values and their basis is the main purpose of the handbooks.

4. CASE STUDIES: EVALUATION OF ENERGY CONVERSION SYSTEMS

Several case studies of conversion systems are evaluated using embodied energy analysis and emergy analyses (hereafter EEA and EMA, respectively). The results from case studies are discussed jointly in the following sections of this article. Comparisons are made to highlight synergisms that support each other and differences that require EMA be used for a deeper understanding of energy systems performance. Evaluations of the case studies resulted from a joint research project between the authors with funding from the Italian Energy Agency (ENEA).

4.1 Description of the Power Systems

4.1.1 Oil-Powered Thermal Plant

A conventional oil-powered thermal plant (Fig. 4) mainly consists of a fuel storage area, boilers for

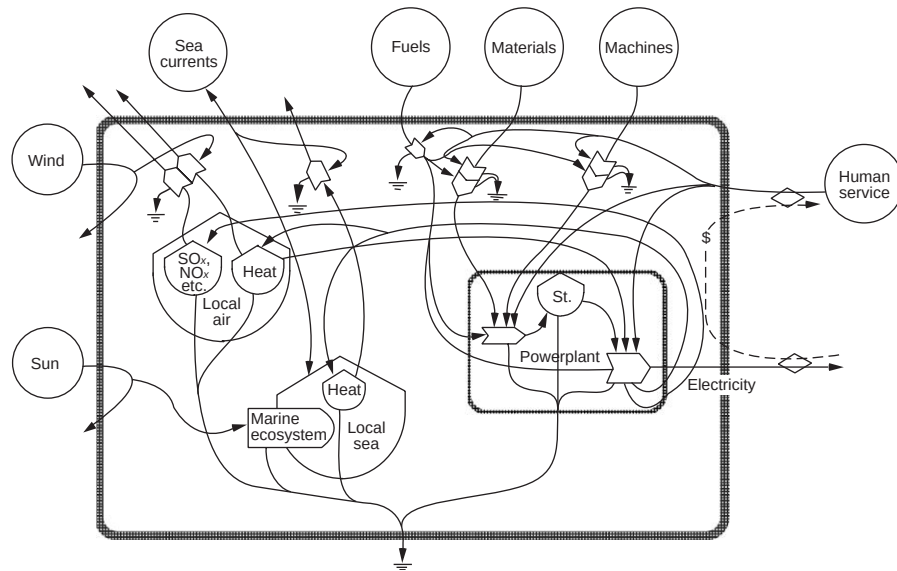


FIGURE 4 System diagram of a thermal power plant and its environment showing the use of nonrenewable energy as well as the use of the environment for cooling.

steam generation, steam turbines, electric generators, a set of electricity transformers from low to high voltage for easier transport to final users, one or more chimneys for the release of combustion gases to the atmosphere, and finally a set of heat exchangers and pumps to cool the plant off by means of river or sea water (44 m³/s) and condense the steam. Filters and scrubbers for partial uptake of pollutants at the chimney mouth as well as fuel desulfurization devices are also used. The data for the evaluation were provided by the Italian National Electric Company (ENEL) and relate to the actual average performance of the Piombino 1280 MWe power plant (Tuscany, Italy). The plant consists of four 320 MWe boiler-turbine-generator groups, cooled by the seawater from the nearby Mediterranean Sea. All inputs were listed and grouped according to their material components (steel, copper, etc.), then suitable energy and emergy intensity coefficients were used, according to previous studies available in the scientific literature. Labor and services are also accounted for in the emergy analysis only. Waste heat and chemicals released to the environment (seawater and atmosphere) were quantified and the amount of environmental services (cooling water, wind, etc.) needed for dispersal, dilution and abatement is calculated by means of the emergy accounting procedure. Fig. 4 shows the direct and indirect annual environmental flows supporting the construction of the plant structure (assuming a lifetime of 25 years) as well as its daily operation. The plant started to be fully

operative in the year 1979, which means that it is not far from the end of its life cycle and may require a significant upgrade investment or final decommissioning and replacement by a more modern technology.

4.1.2 Wind Turbine

The investigated case study (Fig. 5) consists of a 2.5 MW wind powered field, composed with 10 single-blade, 250 kW, wind turbines M30-A, sited in southern Italy (Casone Romano). The wind field, installed in the year 1995 by the Riva Calzoni Company, is presently operated by ENEL. The distance between two generators is 150 m. Each turbine is located on the top of a 33 m support tower. Operating wind speeds are between 4 and 25 m/s. For wind speeds outside of these extremes, the turbine automatically stops. The wind field does not require any local labor for its daily activity and is directly managed from the operating company by means of remote control. The energy systems diagram of a wind power plant is shown in Fig. 5.

4.1.3 Geothermal Plant

To optimize the exploitation of the rich geothermal fields available in central Italy (Tuscany), the National Electric Company designed a 20 MW power module to be efficiently and cost-effectively installed in several sites of the geothermal area (Fig. 6). The main reason supporting the development of such a module is the need for the reduction of the time between the discovery of a new geothermal site and

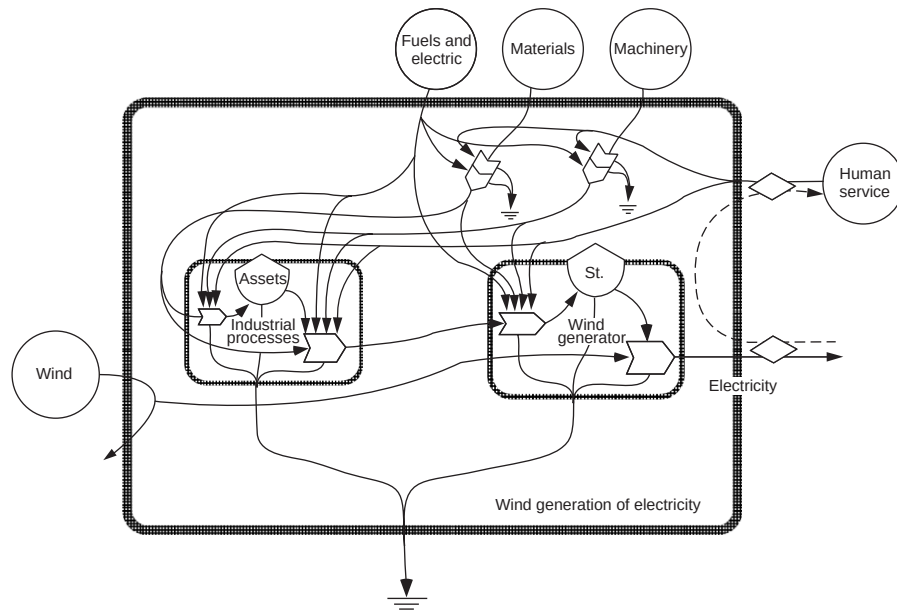


FIGURE 5 System diagram of wind power plant showing the industrial processes required to produce the generator and associated structure and the use of wind energy in the production of electricity.

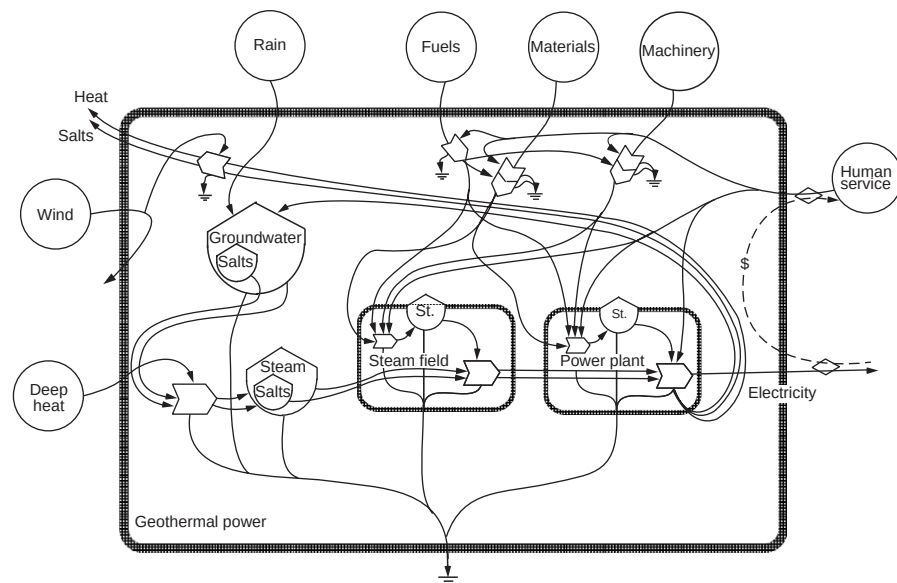


FIGURE 6 System diagram of a geothermal power plant showing the use of nonrenewable energies in developing the steam field and power plant, the use of geothermal steam, and release of heat and salts to the larger environment.

its actual exploitation. The module essentially consists of a turbine-generator group, receiving high-enthalpy fluids from nearby wells. A fraction of the extracted steam is then released to the atmosphere, through three cooling towers. The remaining fraction is, instead, condensed and reinjected into the underground reservoir to avoid their water depletion. The

steam pipelines are made with steel and are insulated with rockwool and aluminum. Instead, the reinjection lines are made with glass fiber. The data used for the case study were provided by ENEL and relate the Cornia 2 power plant, sited close to Pisa and installed in the year 1994. No local labor is required. The plant is 24 hours per day distance-operated by the

company's operative station sited in Larderello, the site where the first electric plant running on geothermal heat was installed in 1904.

Geothermal fluids carry and release significant amounts of hydrogen sulphide, carbon dioxide, radioactive radon, arsenic, mercury, nitrogen, and boron. This involves plant corrosion problems as well as pollution problems in the area. For this reason, several devices have been designed for the uptake and abatement of the most dangerous chemicals. The plant as well as the environmental and technological flows supporting its activity are described by the systems diagram drawn in Fig. 6.

4.1.4 Hydroelectric Plant

The case study (systems diagram shown in Fig. 7) concerns the reservoir-dam electric plant Pollino Nord, sited in southern Italy. Its activity started in the year 1973. The reservoir has a capacity of 1000 m³ and is located at about 600 m above sea level. Water flows out of the dam through an iron pipeline reaching the turbines located downstream, at about 50 m above sea level. Two 51 MW turbines with vertical axis convert the water energy into electricity, generating a total actual electric power of about 85 MW. The plant started its activity in 1973. No labor is required for plant operation. It is distance-operated from the ENEL headquarters in Catanzaro (southern Italy).

4.1.5 Bioethanol Production

Figure 8 describes the main steps of the process: corn production, harvesting and transport to plant, and industrial processing. Corn is processed to ethanol and distillers dried grains with solubles (DDGS), to be used as animal feed. Agricultural residues were considered as a partial source of process heat (to substitute for some of the coal), so that a lower fossil energy input can be charged to the industrial phase. However, harvesting agricultural residues depletes the amount of nutrients in soil and requires significant energy expenditure for their replacement. Soil erosion is also accelerated, with significant consequences on soil fertility. Therefore, the fraction of agricultural residues that can be actually harvested is small and depends on a variety of parameters that are not easy to determine. The evaluation was conducted assuming that approximately 70% of agricultural residues were harvested, taking some of the above mentioned consequences (loss and replacement of nutrients in soil) into account. Corn production data refer to Italian agricultural standards, while corn-to-ethanol industrial conversion refers to average available conversion technologies.

4.1.6 Hydrogen from Steam Reforming of Natural Gas

Steam reforming of hydrocarbons (mainly natural gas) has been the most efficient, economical, and

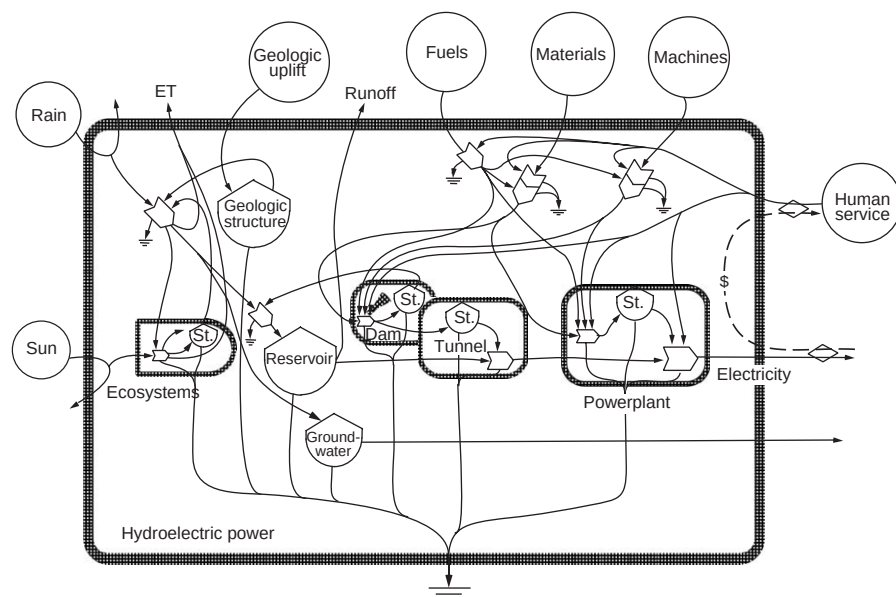


FIGURE 7 Systems diagram of a hydroelectric power plant showing the use of nonrenewable energy materials machines and human service in the construction operation and maintenance of the plant as well as the reliance of the system on geologic formation of the watershed and input of rain.

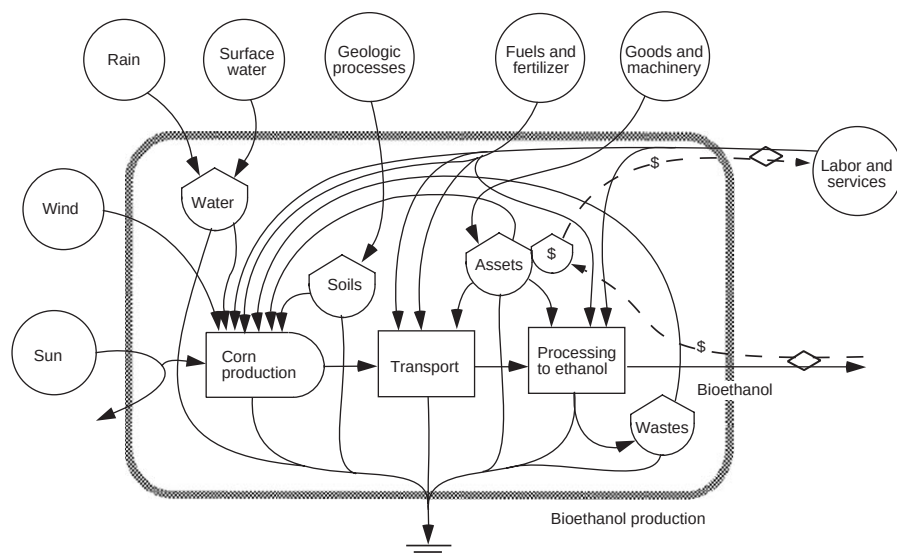
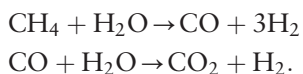


FIGURE 8 Systems diagram of bioethanol production showing the use of nonrenewable resources in the construction of the processing plant, transportation, and corn production as well as the environmental, renewable energies of sun, wind, and water.

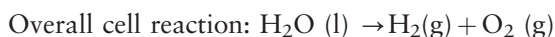
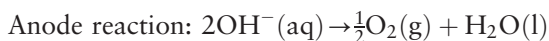
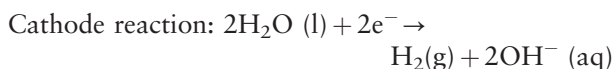
widely used process for hydrogen production. An aggregated energy systems diagram of a conventional steam reforming process is shown in Fig. 9A. The feedstock (i.e. natural gas) is mixed with process steam and reacts over a nickel-based catalyst contained inside a system of alloyed steel tubes. To protect the catalyst, natural gas has to be desulfurized before being fed to the reformer. The following reactions take place in the reformer:



The output of the reformer is fed to a series of shift reactors and a pressure swing adsorption unit, where hydrogen is separated and purified to over 99% purity. The reforming reaction is strongly endothermic and energy is supplied by combustion of natural gas or fuel oil. The reaction temperature is usually in the range 700 to 900°C.

4.1.7 Hydrogen from Water Electrolysis

Figure 9B shows an aggregated view of the hydrogen/water electrolysis process. Alkaline water electrolysis is a well-established industrial process in which electricity is used to split water into its component elements, generating hydrogen in a 1:1 molar ratio, with cell efficiencies in the order of 80 to 90%. The following reactions take place at the electrodes of an electrolysis cell (called an electrolyzer) filled with a suitable electrolyte (aqueous solution of KOH, NaOH, or NaCl) upon the application of a potential:



An electrolysis plant can operate over a wide range of capacity factors and is convenient for a wide range of operating capacities, which makes this process interesting for coupling with both renewable and nonrenewable energy sources. The needed electricity can be purchased from the grid (an interesting option in the case of excess electricity production from hydroelectric plants, for example) or produced on site. In this article, the electrolysis is assumed driven by the electricity generated by the oil-powered and wind-powered plants described earlier.

4.1.8 Ocean Thermal Energy Conversion (OTEC)

OTEC is an energy technology that converts indirect solar radiation to electric power (Fig. 10). OTEC systems use the ocean's natural thermal gradient—the fact that the ocean's layers of water have different temperatures—to drive a power-producing cycle. Commercial ocean thermal energy conversion plants must be located in an environment that is stable enough for efficient system operation. The temperature of the warm surface seawater must differ about 20°C (36°F) from that of the cold deep water that is no more than about 1000 meters (3280 feet) below the surface. The natural ocean thermal gradient necessary for OTEC operation is generally

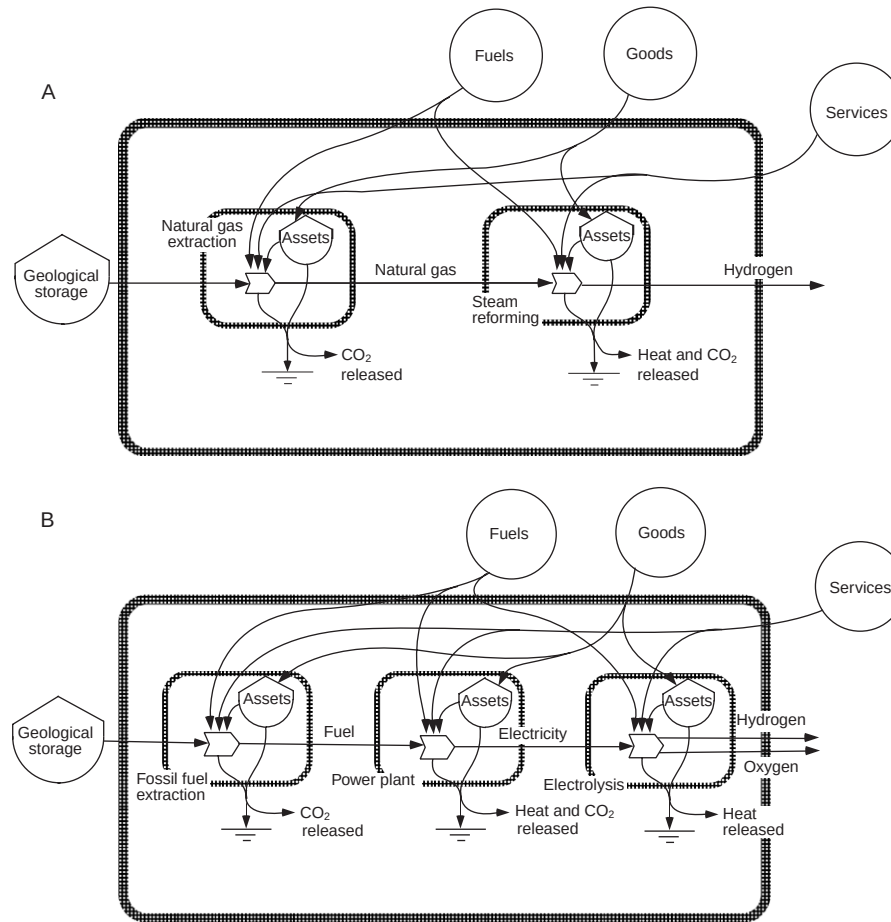


FIGURE 9 (A) Aggregated systems diagram of hydrogen production from steam reforming and (B) electrolysis.

found between latitudes 20 °N and 20° S. Within this tropical zone are portions of two industrial nations—the United States and Australia—as well as 29 territories and 66 developing nations. Of all these possible sites, tropical islands with growing power requirements and a dependence on expensive imported oil are the most likely areas for OTEC development. Electricity generated by plants fixed in one place can be delivered directly to a utility grid. A submersed cable would be required to transmit electricity from an anchored floating platform to land. Moving ships could manufacture transportable products such as methanol, hydrogen, or ammonia on board. The OTEC data for the evaluation are from Odum. He investigated the OTEC technology, performing both an energy and emergy assessment. His conclusions, based on emergy accounting, were that although OTEC shows a small net emnergy yield it is unlikely to be competitive with fossil powered technologies for electricity, due to the excessive

dilution of solar energy resulting into a very small heat gradient of marine water.

4.2 Methods

Apart from the OTEC plant, the case studies represent real plants, evaluated on the basis of the electricity actually generated and the fuel actually used. Therefore, calculated performance indicators may appear worse than those theoretically obtainable on the basis of nameplate power and efficiency, although the order of magnitude does not change significantly.

4.2.1 Data and Calculations

Unless specified differently, a lifetime of 25 years is assumed for each of the investigated plants or assets. Data used include construction materials, source energy, and labor in construction, as well as operation and maintenance, environmental inputs,

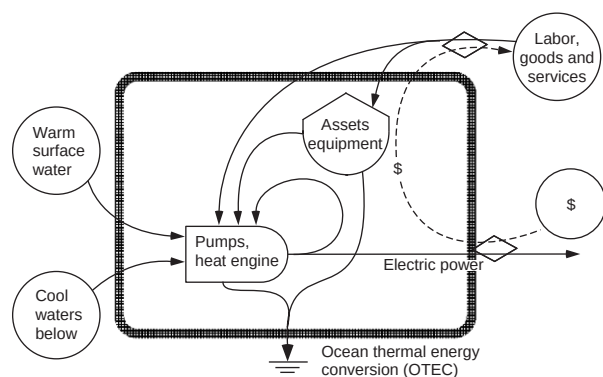


FIGURE 10 Aggregated systems diagram of ocean thermal energy conversion showing the use of labor, goods, and services (nonrenewable) and differential in surface and deep ocean water temperature.

and process outputs including electricity and pollutants. Input data for main plant components were kindly supplied by the companies producing and operating them.

All plants are analyzed using the same procedures as outlined earlier. Input data for each item (row) in the evaluation tables are the actual operational data of each plant and are converted into appropriate units using specific calculation procedures, most of which are described by Odum in his text *Environmental Accounting*. Input data to each investigated process are multiplied by suitable conversion coefficients (unit oil equivalents, $g_{oil}/unit$ of item, for energy evaluation; transformities, $sej/unit$ of item, for emergy evaluation), to yield energy and emergy values associated with each input item. Energy values (expressed as grams of oil equivalent) are then converted into grams of released carbon dioxide, by means of an average oil-to-carbon dioxide conversion coefficient derived from typical reaction stoichiometry. Energy, emergy, and CO_2 flows are then summed, yielding a total that is divided by the amount of the product of each plant to yield its energy cost, its transformity, and its associated carbon emissions. Finally, data are used to calculate several other energy-, emergy-, and carbon-flow indicators to obtain an overall picture of plant performance. All inputs to the electric production systems are expressed on a yearly basis. Fixed capital equipment, machinery, buildings, and so on are divided by their estimated useful life (10 to 30 years, depending on the case). Outputs are also expressed on an annual basis. CO_2 production is calculated for direct fuel combustion in thermal plants and estimated for goods and human services using

multipliers of indirect energy consumption for their production. These multipliers are derived from a standardized comparison of literature in this field.

4.2.2 Performance Indicators

EEA typically evaluates energy sources and the outputs from process in their heat equivalents. Two indicators commonly used in EEA are calculated for all the investigated case studies. The first one is related to CO_2 production and global warming potential and is the amount of CO_2 released per megawatt of energy delivered. The second indicator is the output/input energy ratio (also called energy return on investment, EROI, according to Hall *et al.*). It is calculated as the ratio of heat equivalent energies of outputs and input flows and is considered as a measure of the first law energy efficiency of a process. The total input energy in this study is derived by multiplying all goods and materials consumed in the construction (divided by the average life span of the structure and equipment) and the annual maintenance of plants by energy equivalent multipliers, and adding the annual energy consumed directly and indirectly.

The output as electricity is converted to joules using the standard conversion of 3.6×10^6 J/kWh, while the masses of bioethanol and hydrogen are multiplied by their higher heating value (J/g). CO_2 releases are evaluated by multiplying the total input energy (sum of energy used directly and indirectly) by a standard stoichiometric conversion, 3.2 g CO_2/g oil equivalent. Renewable energy plants (wind, hydro, and geothermal) obviously do not burn fossil fuels directly, so their direct CO_2 emissions are lower (zero for the hydro and wind plants, but not negligible for the geothermal power plant due to underground CO_2 contained in the geothermal fluid extracted).

EMA indices are calculated in a similar way. Each item in Table VII (bioethanol) and Table VIII (wind turbine) as well as in all the other tables related to the investigated case studies is multiplied by an appropriate transformity to get the emergy input from the item to the process. The total emergy (a measure of the total environmental support or ecological footprint to the process) is then calculated by summing all inputs according to the previously described emergy algebra. Finally, a new transformity for the process output is calculated by dividing the total emergy by the exergy (sometimes the mass) of the product. Further emergy-based indicators (emergy yield ratio, environmental loading ratio, emergy index of sustainability, emergy density) are also calculated.

4.3 Results and Discussion

Comparing EEA- and EMA-based performance indicators for each kind of energy conversion device makes it possible to ascertain what pieces of information can be obtained from each one of the two approaches and to better understand the space-time scale difference between them.

A careful reading of Tables VII to X provides a rich set of information. Tables VII and VIII are given as examples of how emergy calculations are performed. Table VII refers to bioethanol production from corn (Fig. 8), while Table VIII refers to a 2.5 MW wind electricity production field (Fig. 5). Tables IX and X provide a comparison among energy-, carbon-, and emergy- based indicators in all the conversion patterns investigated. Table IX refers to electricity generation, while Table X refers to selected energy carriers other than electricity. Conversion factors used and references are clearly listed in each table.

4.3.1 Energy and Emergy Evaluations

Emergy evaluations of bioethanol production from corn (Table VII) and electricity generation from wind (Table VIII) are given as examples for a clear description of the emergy approach. These two case studies do not focus on the best available technologies and it is possible that other wind turbines or biomass-to-ethanol processes exist showing better performances. These two cases are used as more or less average-technology case studies. The same rationale applies to the other case studies summarized in Tables IX and X but not shown in detail here.

Renewable flows to bioethanol production from corn (Table VII) account for about 15.3% of the total emergy driving the process. They can be identified as the rain driving the agricultural phase and the wind dispersing the pollutants that come out of the industrial step. Electricity and fuels (including coal) account for about 17.4%, goods and machinery for 42.6%, while labor and services represent 24.7% of the total emergy input. Goods and machinery are therefore the main emergy input to the process, due to the environmental support embodied in their manufacture from raw materials.

Environmental flows as well as labor and services are not included in the conventional energy accounting, therefore a comparison between EEA and EMA requires the percentages be calculated taking these flows out of the total emergy. In so doing, fuels and electricity come out to be respectively the 29% of total emergy and the 45% of total energy,

while goods and machinery account for 71% of total emergy and 55% of total energy. This means that EEA overestimates the role of direct energy inputs and underestimates the role of energy flows embodied in goods and machinery. A similar calculation can be performed for the wind turbine by using data presented in Table VIII. As far as wind plant is concerned, it can be noted that wind emergy accounts for about 82% of the total emergy input, while labor and services (4%) and goods and machinery (14%) play a much smaller role. A comparison with energy data is less significant here, since no direct energy inputs are required. It is crystal clear that EMA offers a more comprehensive picture of the whole process by taking into account the hidden environmental flows that support a given process and by evaluating their role on the scale of the biosphere, where they are compared by means of a common measure of environmental quality.

4.3.2 Embodied Energy Performance Indicators

The EEA results are synoptically shown in Tables IX and X. The EROIs of electricity generation range from a low 0.30 of the oil-powered plant to a high 23.7 of the hydroelectric plant. The geothermal and wind conversions also show high energy returns (respectively 20.7 and 7.7), while the OTEC plant only shows an energy return equal to 3.8, due to the low thermal gradient between deep and surface water temperatures. Similarly, Table X shows EROIs in the range 0.24 to 6.1 for energy carriers different than electricity (bioethanol and hydrogen). As expected, the lower value refers to hydrogen production via electrolysis with electricity generated by an oil-powered, thermal plant similar to the one described in Table IX. Instead, the highest EROI is shown by hydrogen generated via electrolysis by means of wind-powered electricity.

The net CO₂ release shows similar patterns, with higher values when direct fossil fuel combustion is involved and lower values when renewable sources drive the process, as expected. In the latter case, only indirect emissions can be calculated, linked to fuel use for the production of components and for machinery construction and operation. The surprising high emission from geothermal electricity is due to the CO₂ content of geothermal fluids.

The better performance of processes driven by renewables is clearly due to the fact that the EROI compares only commercial energies (i.e., calculates the ratio of the thermal energy delivered to the fossil or fossil equivalent energy that is provided to the

TABLE VII

Emergy Analysis of Bioethanol Production

Item	Unit	Amount (unit/ha * year)	Unit emergy values (seJ/unit)	Ref. for UEV	Solar emergy (seJ/ha * year)
Agricultural phase (corn production)					
<i>Renewable inputs</i>					
Sunlight	J	5.50E + 13	1.00E + 00	[1]	5.50E + 13
Rain water (chemical potential)	J	3.07E + 10	3.06E + 04	[2]	9.39E + 14
Earth cycle	J	3.00E + 10	1.02E + 04	[2]	3.05E + 14
<i>Nonrenewable inputs</i>					
Organic matter in topsoil used up	J	3.24E + 09	1.24E + 05	[2]	4.02E + 14
Nitrogen fertilizer	g	1.69E + 05	6.38E + 09	[2]	1.08E + 15
Phosphate fertilizer	g	8.20E + 04	6.55E + 09	[2]	5.37E + 14
Insecticides, pesticides, and herbicides	g	5.38E + 03	2.49E + 10	[3]	1.34E + 14
Diesel	J	6.67E + 09	1.11E + 05	[2]	7.40E + 14
Lubricants	J	1.64E + 08	1.11E + 05	[2]	1.82E + 13
Gasoline	J	3.00E + 00	1.11E + 05	[2]	3.33E + 05
Water for irrigation	J	1.36E + 09	6.89E + 04	[3]	9.37E + 13
Electricity for irrigation	J	2.00E + 09	2.52E + 05	[4]	5.04E + 14
Agricultural machinery (mainly steel)	g	1.36E + 04	1.13E + 10	[3]	1.53E + 14
Seeds	g	1.62E + 04	5.88E + 04	[7]	9.53E + 08
Human labor	years	1.30E-02	6.32E + 16	[6]	8.21E + 14
Annual services	US \$	8.90E + 02	2.00E + 12	[6]	1.78E + 15
<i>Additional inputs for harvest of 70% residues to be used as process energy</i>					
Nitrogen in residues harvested	g	7.88E + 04	6.38E + 09	[2]	5.03E + 14
Phosphate in residues harvested	g	1.82E + 04	6.55E + 09	[2]	1.19E + 14
Diesel for residues	J	9.03E + 03	1.11E + 05	[2]	1.00E + 09
Machinery for residues	g	2.59E + 03	1.13E + 10	[3]	2.92E + 13
Labor for residues	years	1.39E-03	6.32E + 16	[6]	8.78E + 13
Additional services	\$	8.90E + 01	2.00E + 12	[6]	1.78E + 14
Industrial phase (conversion of corn to bioethanol)					
<i>Renewable inputs for the dilution of airborne and waterborne pollutants</i>					
Wind	J	2.20E + 11	2.52E + 03	[2]	5.54E + 14
<i>Nonrenewable inputs</i>					
Diesel for transport	J	6.76E + 08	1.11E + 05	[2]	7.50E + 13
Transport machinery (mainly steel)	g	1.22E + 04	1.13E + 10	[3]	1.37E + 14
Plant machinery (mainly steel)	g	4.41E + 04	1.13E + 10	[3]	4.96E + 14

continues

Table VII continued

Item	Unit	Amount (unit/ha* year)	Unit emergy values (seJ/unit)	Ref. for UEV	Solar emergy (seJ/ha* year)
Cement in plant construction	g	7.84E + 04	3.48E + 09	[5]	2.73E + 14
Additional coal for hot water/steam generation	J	8.10E + 07	6.72E + 04	[2]	5.44E + 12
Process electricity	J	1.41E + 09	2.52E + 05	[4]	3.55E + 14
Process and cooling water	J	7.99E + 07	6.89E + 04	[3]	5.50E + 12
Gasoline denaturant	J	4.89E + 08	1.11E + 05	[2]	5.42E + 13
Ammonia	g	3.56E + 01	6.38E + 09	[2]	2.27E + 11
Lime	g	9.27E + 00	1.68E + 09	[2]	1.56E + 10
Electricity running the waste water treatment plant	J	4.35E + 07	2.52E + 05	[4]	1.10E + 13
Labor	years	1.68E-03	6.32E + 16	[6]	1.06E + 14
Annual capital cost and services	US \$	2.22E + 02	2.00E + 12	[6]	4.44E + 14
Main product of industrial phase					
Ethanol produced, without services	J	5.65E + 10	1.28E + 05	[7]	7.22E + 15
Ethanol produced, with services	J	5.65E + 10	1.73E + 05	[7]	9.78E + 15

process). Energies that are not purchased, but instead are provided for free by nature (e.g., the solar energy driving the wind or the solar energy that drives the photosynthesis for corn production) are not accounted for by EEA and therefore are not reflected in the EROI or in the CO₂ release.

4.3.3 Emergy-Based Performance Indicators

A series of emergy performance indicators are given in Tables IX and X for selected energy conversion devices. Some of these processes, driven by intensive use of primary nonrenewable resources, are secondary energy conversions, while others (wind, hydro and geothermal) should be considered primary energy exploitation patterns. Their values can be usefully compared to those of the primary fuels (oil, coal, natural gas) shown in Table X. Electricity generation in Table IX shows transformities that are, as expected, process specific and range from the low value of about 1.1×10^5 seJ/J for hydro- and wind-electricity up to much higher values around 3.5×10^5 seJ/J for geothermal and oil-powered electricity. The role of labor and services in these processes can be quantified as about 5 to 10% of the transformity. This means that the main contribu-

tions to any electricity generation process are the driving forces of fuels or renewable sources, followed by the emergy associated with plant components. As a consequence, a significant lowering of the total emergy demand cannot be expected by simply moving the plant to countries with lower cost of labor, as has become very common with goods production. The OTEC plant combines very small inputs in the form of solar energy via the temperature gradient of the oceans and technological equipment and, instead, a significant input in the form of financial investment, labor, and services, to which a huge emergy is associated. This makes this plant too dependent on nonrenewable, purchased emergy inputs, very high compared to the locally available renewable flow. The transformity of OTEC electricity is around 1.7×10^5 seJ/J, but declines by more than one order of magnitude when labor and services are not accounted for.

Transformities of energy delivered in the form of energy carriers other than electricity (Table X) show similar orders of magnitude. Hydrogen from water electrolysis shows higher or lower transformities depending respectively on oil- or wind-generated electricity. Hydrogen from steam reforming of

TABLE VIII

*Emergy Accounting of Wind Electricity Production in Italy**

Item	Unit	Amount	Unit emergy value (seJ/unit)	Ref. for UEV	Solar emergy (seJ)
Direct renewable inputs					
Wind	J	4.85E + 13	2.52E + 03	[2]	1.22E + 17
Nonrenewable and purchased inputs					
Concrete (basement)	g	5.62E + 06	2.59E + 09	[1]	1.45E + 16
Iron (machinery)	g	2.65E + 05	2.50E + 09	[2]	6.63E + 14
Steel (machinery)	g	8.61E + 05	5.31E + 09	[2]	4.57E + 15
Pig iron (machinery)	g	1.56E + 05	5.43E + 09	[2]	8.47E + 14
Copper	g	8.76E + 04	3.36E + 09	[3]	2.94E + 14
Insulating and miscellaneous plastic material	g	1.01E + 04	2.52E + 09	[4]	2.55E + 13
Lube oil	J	3.10E + 09	1.11E + 05	[5]	3.44E + 14
Labor	years	8.33E-02	6.32E + 16	[6]	5.26E + 15
Services	US \$	2.94E + 01	2.00E + 12	[6]	5.88E + 13
Electricity generated					
Electricity, with labor and services	J	1.35E + 12	1.11E + 05	[7]	1.49E + 17
Electricity, without labor and services	J	1.35E + 12	1.06E + 05	[7]	1.44E + 17

*Data on a yearly basis; 2.5 MW plant sited at Casone Romano, Foggia, Italy.

References for transformities (Tables VII and VIII):

[1] Brown and Buranakarn (2003).

[2] Bargigli and Ulgiati (2003).

[3] Lapp (1991).

[4] Average estimate based on selected case studies.

[5] Odum (1996).

[6] Ulgiati (2003).

[7] From calculation performed in this work.

natural gas shows a relatively low transformity. Unfortunately, only the value without services was calculated in the steam-reforming case study. However, since steam-reforming is a well-known technology, not particularly labor intensive, a dramatic increase of its transformity should not be expected when adding the emergy value of labor and services. The comparison with the transformities of primary sources in Table IV clearly shows that transformities depend on the length of the manufacturing chain. More process steps require more emergy inputs and may decrease the actual energy delivered at the end of the chain. Oil, coal, natural gas, hydro- and wind-electricity, and hydrogen from steam reforming all show transformities on the order of 10^4 seJ/J. However, when technologically heavy steps are added to further process the primary energy source, transformities grow by an order of magnitude to 10^5 seJ/J. It can be clearly understood that if bioethanol were used to generate electricity or if electricity from oil were used to refine metals, transformities would become even greater.

The other emergy indicators complement this picture, by showing low EYRs for oil-powered and OTEC electricity as well as for bioethanol and hydrogen from thermal electricity. Not surprisingly, all the other energy conversions have EYRs around 5. Similarly, high ELRs characterize oil-based patterns, as well as bioethanol and OTEC electricity, while very low ELRs (below 1) are shown by the processes relying to a larger degree on renewables. As a consequence, the aggregated sustainability measure expressed by the EIS ranges from a low 0.06 for oil-based processes up to a significantly high 25 for wind-based ones. Again, the OTEC plant shows a very low performance, quantified by an EIS lower than 0.01. However, let's recall that the evaluation of OTEC is based on a feasibility study and not on an actually existing plant.

4.3.4 Comparison of EEA and EMA Results

Results presented in Tables IX and X allow further considerations in addition to the already discussed features of the two approaches. Both EEA and EMA

TABLE IX

Energy and Emery-Based Indicators for Selected Electricity Power Plants

Item		Unit	Oil plant (1280 MWe) ^a	Wind (2.5 MWe) ^a	Geothermal (20 MWe) ^a	OTEC 4.57 MWe ^b	Hydro (85 MWe) ^a
Energy indicators							
E _{out}	Total electric energy delivered per year	J/year	2.35E + 16	1.35E + 13	3.28E + 14	1.19E + 14	3.94E + 14
E _{In}	Total energy invested per year	J/year	7.84E + 16	1.76E + 12	1.58E + 13	3.12E + 13	1.66E + 13
GWP	CO ₂ released/unit of energy delivered ^c	g/MJ	260.22	10.17	182.32	20.43	3.29
ER	Energy return on investment (energy output/input)		0.30	7.67	20.76	3.82	23.73
Emery indicators							
R ₁	Renewable input from the local scale	seJ/year	3.43E + 20	1.22E + 18	5.63E + 19	1.13E + 17	2.84E + 19
R ₂	Renewable input from the global scale ^d	seJ/year	1.72E + 21	0.00E + 00	2.64E + 19	0.00E + 00	0.00E + 00
N	Locally nonrenewable input ^e	seJ/year	0.00E + 00	0.00E + 00	7.74E + 18	0.00E + 00	7.47E + 18
F	Purchased plant inputs, including fuel	seJ/year	5.71E + 21	2.13E + 17	1.55E + 19	7.73E + 17	4.42E + 18
L	Labor and services ^f	seJ/year	5.36E + 20	5.33E + 16	4.17E + 18	1.95E + 19	3.73E + 18
Y	Yield (R ₁ + R ₂ + N + F), without labor and services	seJ/year	7.78E + 21	1.44E + 18	1.06E + 20	8.85E + 17	4.03E + 19
Y _L	Yield (R ₁ + R ₂ + N + F + L), with labor and services	seJ/year	8.32E + 21	1.49E + 18	1.10E + 20	2.04E + 19	4.40E + 19
Indices (including services for fuel supply and plant manufacturing							
Tr ₁	Solar transformity of electricity, with labor and services	sej/J	3.54E + 05	1.10E + 05	3.35E + 05	1.71E + 05	1.12E + 05
Tr ₂	Solar transformity of electricity, w/out labor and services	sej/J	3.31E + 05	1.06E + 05	3.23E + 05	7.44E + 03	1.02E + 05
EYR	Emery yield ratio, EYR = (Y _s)/(F + L)		1.33	5.59	5.60	1.01	5.41
ELR	Environmental loading ratio, ELR = (R ₂ + N + F + L)/R ₁		23.26	0.22	0.96	180.00	0.55
ED	Emery density, ED = (Y _L)/area of plant	sej/m ²	4.33E + 15	2.12E + 12	3.54E + 14	n.a.	2.84E + 13
EIS	EYR/ELR		0.057	25.64	5.87	0.006	9.84

n.a., not available.

^aBrown and Ulgiati, 2002, modified.^bData from H. T. Odum (2000b), partially modified by the authors.^cGlobal warming potential (GWP), includes direct CO₂ release from fuel combustion (if any), indirect CO₂ release embodied in inputs, and (for geothermal plant only) CO₂ from underground water, released through the cooling towers.^dIndirect environmental inputs from the larger scale, needed to dilute heat and chemicals that are released by the plant.^eIncludes only locally extracted fuel delivered to plant, as well as underground water and topsoil used up (if any).^fLabor accounts for hours actually worked in the investigated process. Services include human labor in all steps prior than plant use, quantified as the emery associated to the monetary cost (cost of item, \$, x country emery/GNP ratio, sej/\$).

TABLE X

Energy and Emery-Based Indicators for Selected Nonelectric Energy Carriers

Item		Unit	Bioethanol from corn ^a	Hydrogen from water electrolysis (oil powered) ^b	Hydrogen from water electrolysis (wind powered) ^b	Hydrogen from steam reforming of natural gas ^c
Energy indicators						
E _{out}	Total energy delivered per year	J/year	5.65E + 10	1.89E + 16	1.08E + 13	6.32E + 15
E _{in}	Total energy invested per year	J/year	3.77E + 10	7.84E + 16	1.77E + 12	8.18E + 15
GWP	CO ₂ released/unit of energy delivered ^d	g/MJ	52.05	323.56	12.78	100.96
ER	Energy return on investment (energy output/input)		1.50	0.24	6.10	0.77
Emery indicators						
R ₁	Renewable input from the local scale	seJ/year	9.37E + 14	3.43E + 20	1.22E + 18	
R ₂	Renewable input from the global scale ^e	seJ/year	1.08E + 15	1.73E + 21	0.00E + 00	
N	Locally nonrenewable input ^f	seJ/year	4.02E + 14	0.00E + 00	0.00E + 00	
F	Purchased plant inputs, including fuel	seJ/year	5.33E + 15	5.70E + 21	2.12E + 17	6.69E + 20
L	Labor and services ^g	seJ/year	3.65E + 15	5.48E + 20	6.85E + 16	
Y	Yield (R ₁ + R ₂ + N + F), without labor and services	seJ/year	7.75E + 15	7.77E + 21	1.43E + 18	6.69E + 20
Y _L	Yield (R ₁ + R ₂ + N + F + L), with labor and services	seJ/year	1.14E + 16	8.32E + 21	1.50E + 18	6.69E + 20
Indices (including services for fuel supply and plant manufacturing)						
Tr ₁	Solar transformity of product, with labor and services	sej/J	2.02E + 05	4.40E + 05	1.39E + 05	
Tr ₂	Solar transformity of product, w/out labor and services	sej/J	1.37E + 05	4.11E + 05	1.33E + 05	1.06E + 05
EYR	Emery yield ratio, EYE = (Y _s)/(F + L)		1.27	1.33	5.36	
ELR	Environmental loading ratio, ELR = (R ₂ + N + F + L)/R ₁		11.16	23.26	0.23	
ED	Emery density, ED = (Y _L)/area of plant	sej/m ²	1.12E + 12	4.33E + 15	2.13E + 12	
EIS	EYR/ELR		0.11	0.057	23.41	

^aUlgiati (2001), modified. Use of agricultural residues as source heat and an energy credit for coproducts (DDGS) accounted for.^bCalculated in this work. Performance coefficients for oil and wind powered electricity production from Table II. Water electrolysis performance characteristics from Bargigli *et al.* (2002).^cBargigli *et al.* (2002). Only the step of hydrogen generation has been investigated, therefore emery costs may be underestimated. Since this step has not direct environmental inputs, the environmental loading ratio and other indicators cannot be calculated.^dGlobal warming potential (GWP), includes direct CO₂ release from fuel combustion, indirect CO₂ release embodied in inputs, and (for geothermal plant only) CO₂ from underground water, released through the cooling towers.^eIndirect environmental inputs from the larger scale, needed to dilute heat and chemicals that are released by the plant.^fIncludes only locally extracted fuel delivered to plant, as well as underground water and topsoil used up (if any).^gLabor accounts for hours actually worked in the investigated process. Services include human labor in all steps prior than plant use, quantified as the emery associated to the monetary cost (cost of item, \$, x country emery/GNP ratio, seJ/\$).

calculate bad performance indicators for oil-based energy systems. EEA simply indicates that the ability of the investigated devices to operate the conversion chemical energy to work is low, as expected for any engine subject to the second law and the Carnot factor. EMA adds an indication of poor sustainability in both the global dynamics of the economy and the biosphere. The same consideration applies to the investigated bioethanol from corn. In this case, however, EEA indicates a small thermal energy return, which supports several claims of bioenergy supporters in favor of increased use of photosynthesis for energy. On the other hand, EMA jointly evaluates the environmental work diverted from its natural pattern in order to provide free renewable inputs, to replace fertile soil eroded by the intensive tilling, to supply fuels, goods and machinery, to support labor and services, and concludes that the global sustainability of this product is so low that it cannot be usefully pursued as an alternative energy source.

Much lower environmental work is required to deliver energy via hydro- and wind-powered turbines, which simply means that less environmental activity is diverted from its original patterns to provide electricity to the economy. This ensures that these technologies are not withdrawing too many resources from the global biosphere dynamics. Renewable energies (sun, wind, rain and deep heat) already have a role in nature's self-organization processes, so that diverting to much of them to support an economic process reduces their input to other natural processes. EEA assigns a more favorable energy return to hydro- than to wind-electricity, while EMA-based sustainability is much greater for wind conversion, due to equal EYRs and much lower ELR of wind compared to hydro.

Geothermal energy return is surprisingly high according to EEA, since the huge amount of deep heat (high-quality flow) is not accounted for as an input. EMA indicators still suggest a good performance for geothermal electricity, but the fact that a large fraction of high-quality deep heat is dispersed increases the transformity of the output. In addition, the load on the environment due to the release of chemicals extracted with the underground fluid increases the ELR and decreases the global EIS of the process. None of these considerations could be obtained by the traditional energy analysis.

A final remark about OTEC is also illuminating. The process was evaluated as a net energy supplier according to EEA, due to an energy return equal to 3.8. However, this favorable result cannot be

obtained without supplying a significant amount of resources in the form of capital, labor, and services. These resources, generated by previous investment of primary energy sources, decrease the global sustainability of the process to values that make its viability very unlikely, more than its actual feasibility.

4.3.5 EEA and EMA: Scale and Energy Quality

Energy analysis and embodied energy analysis treat the conceptual issues of scale and quality of energy very differently. From these differences, significant discussion has arisen over the years (see, for instance, the article in *Ecological Economics* titled "Embodied Energy Analysis and Energy Analysis: A Comparative View" by Brown and Herendeen). In the following discussion, two important issues are elucidated: temporal and spatial scale, and energy quality.

4.3.5.1 Issues of Scale It should be kept in mind that EEA and EMA have different windows of interest (although they also share several common features) and therefore are used to investigate different questions about a given process.

In particular, EEA provides a measure of the overall commercial (oil equivalent) energy invested to support a production process. Its outcome is the amount of oil equivalent energy required to drive a process or to generate a product. Its spatial scale is the actual scale of the process, although it may in principle include the spatial scale of the previous commercial processes from which the input flows came from (for instance, if a mineral from a faraway country is fed to the process, transport must also be accounted for). Its timescale is the time required to extract the raw materials and to actually make the product. In principle, time is only accounted for in order to calculate the fraction of embodied energy of assets and machinery that are used up in the process and that therefore should be assigned to the product. Since machinery and assets have little embodied energy relevance, this time and the related amount of assets are often disregarded by EEA analysts.

Instead, the outcome of EMA is a quantification of the environmental support provided by nature to a process, which may be or may be not under human control. The spatial scale of EMA is larger than the actual scale of the process, since the evaluation also includes free environmental flows from the larger scale of the biosphere that are not accounted for in energy analysis. In a similar way, the EMA timescale is the total time it took to make a resource via all natural processes involved (generation of minerals, fossil fuels, topsoil, water storages, etc.):

embodied in the transformity of a resource is the global environmental input required over space and time to drive the geological and biosphere processes, which converge to build a resource storage or to support a resource flow. Emergy-based indicators therefore have a built-in memory of the process history—that is, the real “trial and error” pathway that was followed by environment and humans to generate the product, including those pathways discarded by natural selection for maximum power output. The latter sentence may not completely apply to short timescale of technological processes and human economies, but results can be managed in such a way as to take into account the further uncertainty related to patterns that are not yet optimized.

4.3.5.2 Issues of Quality Energy has been defined as the ability to do work, based on the physical principle that work requires energy input. Energy is measured in units of heat, or molecular motion, the degree of motion resulting in expansion and quantified in degrees of temperature. All energies can be converted to heat at 100% efficiency; thus, it is relatively easy and accurate to express energies in their heat equivalents. The basic units of energy are the amount of heat required to raise a given amount of water a given number of degrees of temperature. Thus, the calorie is the amount of heat required to raise 1 cm³ of water 1°Celsius. A joule is equal to 4.187 calories.

Heat-equivalent energy is a good measure of the ability to raise water temperature. However, it is not a good measure of more complex work processes. Processes outside of the window defined by heat engine technology do not use energies that lend themselves to thermodynamic heat transfers. As a result, converting all energies of the biosphere to their heat equivalents reduces all work process of the biosphere to heat engines. Human beings, then, become heat engines and the value of their services and information is nothing more than a few thousand calories per day. Obviously, not all energies are the same and methods of analysis need reflect this fact.

Different forms of energy have different abilities to do work, not only in terms of amounts of work but also in terms of kind of work. It is therefore necessary to account for these different abilities if energies are to be evaluated correctly. A joule of sunlight is not the same as a joule of fossil fuel or a joule of food, unless it is being used to power a steam engine. Sunlight drives photosynthesis. A system

organized to use concentrated energies like fossil fuels cannot process a more dilute energy form like sunlight, joule for joule. Evaluation of energy sources is system dependent. The processes of the biosphere are infinitely varied and are more than just thermodynamic heat engines. As a result, the use of heat measures of energy that can only recognize one aspect of energy, its ability to raise the temperature of things, cannot adequately quantify the work potential of energies used in more complex processes of the biosphere. As in thermodynamic systems where energies are converted to heat to express their relative values, in the larger biosphere system as a whole energies should be converted to units that span this greater realm, accounting for multiple levels of system processes, ranging from the smallest scale to the largest scales of the biosphere, and accounting for processes other than heat engine technology. The ability of driving processes other than engine-like ones is a new quality aspect of resource flows, which is accounted for by emergy analysis.

Net energy evaluations of energy sources and transformation processes are designed to provide information concerning efficiency and potential yield under the engine-like point of view cited previously. This is something that energy analysis does appropriately. It is able to offer at least two different results:

1. As embodied energy analysis, it provides a measure of the commercial energy cost of a product (MJ or grams of oil equivalent per unit of product). When the product is energy, this measure is better expressed as EROI (joules of energy delivered per unit of energy provided to the process).

2. As exergy analysis at the process scale, it provides a measure of thermodynamic efficiency, indicates possible optimization patterns, and finally ranks the quality of the product from the user-side point of view.

Neither of these two energy analysis patterns is able to provide any significant insight into the quality of each input and output flow in the larger scale of the biosphere, as pointed out earlier. Neither of them takes into account the role of the environmental work supporting both ecosystems and human societies, in particular the role of unmonied inputs provided by nature to human economies. This is something that requires a scale expansion and a change of perspective, from local to global, i.e., the emergy approach.

5. SUMMARY

This article is a brief synopsis of the emergy accounting methodology. Emergy-based case studies of energy conversion systems are presented to allow both an increased understanding of how the method is applied as well as a critical comparisons with embodied energy analysis. Much like a short intensive course given in a limited time frame, this article has touched on what is most important. There is much literature explaining the theory and concepts of the emergy approach as well as applications of the method to a large number of case studies, ranging from ecosystems to socioeconomic systems, to evaluations of the biosphere, including energy processing, agriculture, industry, tourism, wastewater management, recycling patterns, generation and storage of culture and information, development of cities as emergy attractors, and scenarios for future societal development. References to emergy literature are given throughout. Starting on 1999, an international emergy conference takes place every other year. The books of proceedings of the first two emergy conferences provide a significant set of theoretical and applied papers for further reading. In all cases, the common thread is the ability to evaluate all forms of energy, materials, and human services on a common basis by converting them into equivalents of one form of energy, solar emergy, a measure of the past and present environmental support to any process occurring in the biosphere.

SEE ALSO THE FOLLOWING ARTICLES

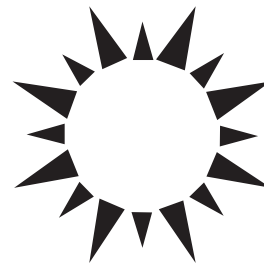
Earth's Energy Balance • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Exergy • Exergy Analysis of Energy Systems • Global Energy Use: Status and Trends • Human Energetics • Thermodynamics, Laws of

Further Reading

- Bargigli, S., Rauegi, M., and Ulgiati, S. (2002). Thermodynamic and environmental profile of selected gaseous energy carriers. Comparing natural gas, syngas and hydrogen. In "ECOS 2002," Book of Proceedings of the 15th International Conference on Efficiency, Costs, Optimization, Simulation and Environmental Impact of Energy Systems—Berlin (Germany), July 2–5 (G. Tsatsaronis, M. J. Moran, F. Czesla, and T. Bruckner, Eds.), pp. 604–612.
- Bargigli, S., Rauegi, M., and Ulgiati, S. (2003). A multi-scale and multi-criteria analysis of aluminium "greenness." *J. Cleaner Production*, submitted.
- Bargigli, S., and Ulgiati, S. (2003). Emergy and life-cycle assessment of steel production. In Proceedings of the Second Biennial Emergy Evaluation and Research Conference, Gainesville, FL.
- British Petroleum (1997). "BP Statistical Review of World Energy, 1997." The British Petroleum Company, London.
- Brown, L. R., Renner, M., and Flavin, C. (1997). "Vital Signs, 1997: The Environmental Trends That Are Shaping Our Future." W. W. Morton & Co., New York.
- Brown, M. T., and Arding, J. (1991). Tranformities Working Paper, Center for Wetlands. University of Florida, Gainesville, FL.
- Brown, M. T., Brandt-Williams, S., Tilley, D., and Ulgiati, S. (eds.). (2001). "Emergy Synthesis. Theory and Applications of the Emergy Methodology." Book of Proceedings of the First International Emergy Research Conference, Gainesville, FL, September 2–4, 1999. The Center for Environmental Policy, University of Florida, Gainesville, FL.
- Brown, M. T., and Buranakarn, V. (2003). Emergy indices and ratios for sustainable material cycles and recycle options. *Resources, Conservation and Recycling* 38, 1–22.
- Brown, M. T., and Herendeen, R. A. (1996). Embodied energy analysis and emergy analysis: A comparative view. *Ecological Economics* 19, 219–235.
- Brown, M. T., and McClanahan, T. (1996). Emergy Analysis Perspectives for Thailand and Mekong River Dam Proposals. *Ecological Modeling* 91, 105–130.
- Brown, M. T., Odum, H. T., Tilley, D., and Ulgiati, S. (eds.). (2003). "Emergy Synthesis. Theory and Applications of the Emergy Methodology–2." Book of Proceedings of the Second International Emergy Research Conference, Gainesville, FL, September 20–22, 2001. The Center for Environmental Policy, University of Florida, Gainesville, FL.
- Brown, M. T., and Ulgiati, S. (1997). Emergy based indices and ratios to evaluate sustainability—Monitoring economies and technology towards environmentally sound innovation. *Ecological Engineering* 9, 51–69.
- Brown, M. T., and Ulgiati, S. (1999). Emergy evaluation of the biosphere and natural capital. *Ambio* 28(6), 486–492.
- Brown, M. T., and Ulgiati, S. (2002). Emergy evaluation and environmental loading of electricity production systems. *J. Cleaner Production* 10, 321–334.
- Campbell, D. (1998). Emergy analysis of human carrying capacity and regional sustainability: An example using the state of Maine (Appendix). *Environmental Monitoring and Assessment* 51, 531–569.
- Hall, C. A. S., Cleveland, C. J., and Kaufman, R. (1986). "Energy and Resource Quality: The Ecology of the Economic Process." Wiley/Interscience, New York.
- Huang, S. -L., Lai, H. -Y., and Lee, C. -L. (2001). Emergy hierarchy and urban landscape system. *Landscape and Urban Planning* 53, 145–161.
- Kinsman, B. (1965). "Wind, Waves, Their Generation and Propagation on the Ocean Surface." Prentice-Hall, Englewood Cliffs, NJ.
- Lapp, C. W. (1991). Emergy analysis of the nuclear power system in the United States. M. S. thesis, University of Florida, Department of Environmental Engineering Sciences, Gainesville, FL.
- Mannion, A. M. (1995). "Agriculture and Environmental Change: Temporal and Spatial Dimensions." John Wiley & Sons, New York.
- Miller, G. A. (1966). The flux of tidal energy out of the deep oceans. *J. Geophysical Res.* 71, 2485–2489.

- Odum, H. T. (1996). "Environmental Accounting. Emergy and Environmental Decision Making." John Wiley & Sons, New York.
- Odum, H. T. (2000a). Handbook of Emergy Evaluation: A Compendium of Data for Emergy Computation Issued in a Series of Folios. Folio #2—Emergy of Global Processes. Center for Environmental Policy, Environmental Engineering Sciences, University of Florida, Gainesville, FL.
- Odum, H. T. (2000b). Emergy Evaluation of an OTEC Electrical Power System. *Emergy* 25, 389–393.
- Odum, H. T., Brown, M. T., and Ulgiati, S. (1999). Ecosystems as energetic systems. In "Handbook of Ecosystem Theories" (S. E. Jorgensen and F. Muller, Eds.), pp. 281–302. CRC Press, New York.
- Odum, H. T., Brown, M. T., Whitefield, L. S., Woihte, R., and Doherty, S. (1995). Zonal Organization of Cities and Environment: A Study of Energy System Basis for Urban Society. A Report to the Chiang Ching-Kuo Foundation for Intergovernmental Scholarly Exchange, Center for Environmental Policy, University of Florida, Gainesville, FL.
- Odum, H. T., Brown, M. T., and Williams, S. B. (2000). Handbook of Emergy Evaluation: A Compendium of Data for Emergy Computation Issued in a Series of Folios. Folio #1—Introduction and Global Budget. Center for Environmental Policy, Environmental Engineering Sciences. University of Florida, Gainesville, FL.
- Odum, H. T., and Odum, E. C. (2001). "A Prosperous Way Down: Principles and Policies." University Press of Colorado, Boulder, CO.
- Odum, H. T., and Odum, E. C. (eds.). (1983). Energy Analysis Overview of Nations. Working Paper WP-83–82. International Institute for Applied Systems Analysis, Laxenburg, Austria.
- Oldeman, L. R. (1994). The global extent of soil degradation. In "Soil Resilience and Sustainable Land Use" (D. J. Greenland and I. Szabolcs, Eds.). CAB International, Wallington, United Kingdom.
- Oort, A. H., Ascher, S. C., Levitus, S., and Peixoto, J. P. (1989). New estimates of the available potential energy in the World Ocean. *J. Geophysical Research* 94, 3187–3200.
- Ryabchikov, A. (1975). "The Changing Face of the Earth." Translated by J. Williams. Progress, Moscow.
- Scienceman, D. M. (1987). Energy and emergy. In "Environmental Economics" (G. Pillet and T. Murota, Eds.), pp. 257–276. Roland Leimgruber, Geneva, Switzerland.
- Slater, J. F., Taupart, G., and Galson, I. D. (1980). The heat flow through the oceanic and continental crust and the heat loss of the earth. *Rev. Geophys. Space Phys.* 18, 269–311.
- Tabacco, A. M., Bargigli, S., Raugei, M., and Ulgiati, S. (2003). Life cycle, exergy and emergy assessment of metallic chromium production from ore. In "Advances in Energy Studies: Reconsidering the Importance of Energy" (S. Ulgiati, M. T. Brown, M. Giampietro, R. A. Herendeen, and K. Mayumi, Eds.). SGE, Padova, Italy.
- Todd, D. K. (ed.). (1970). "The Water Encyclopedia," 2nd ed. Lewis, Chelsea, MI.
- Ulgiati, S. (2001). A comprehensive energy and economic assessment of biofuels: When "green" is not enough. *Critical Reviews in Plant Sciences* 20, 71–106.
- Ulgiati, S. (2003). "Emergy Analysis of Italy in the years 1984, 1989, 1991, 1995, and 2000." Unpublished manuscript, Energy and Environment Research Unit, Department of Chemistry, University of Siena, Italy (ulgiati@unisi.it).
- USDI (1996). "Mineral Commodity Summaries, January 1997." U. S. Department of Interior, Washington, DC.
- Wiin-Nielsen, A., and Chen, T. (1993). "Fundamentals of atmospheric energetics." Oxford Press, New York.
- World Resources Institute (1996). "World resources 1996–97." Oxford University Press, Oxford.

Energy Development on Public Land in the United States



BRIAN BLACK

Pennsylvania State University, Altoona College
Altoona, Pennsylvania, United States

-
1. Introduction
 2. Energy for All?
 3. General Phases of Public Land Administration
 4. Establishing the Legal Framework
 5. Making Legal Exception: Native American Communities
 6. History of Energy Development on Public Land
 7. Debating Ownership: The Sagebrush Rebellion
 8. A Symbolic Contest: Alaska
 9. Battle Lines in the Current Debate

Glossary

- application for permit to drill (APD)** An application to drill a well; submitted by a lessee or operator to the Bureau of Land Management.
- Bureau of Land Management (BLM)** The Department of the Interior agency responsible for managing most federal onshore subsurface minerals.
- environmental impact statement (EIS)** A written analysis of the impacts on the natural, social, and economic environment of a proposed project or resource management plan.
- lease** An authorization to possess and use public land for a period of time sufficient to amortize capital investments in the land.
- public lands** The lands owned by the federal government, including parks, monuments, wilderness areas, refuges, underground mineral reserves, marine sanctuaries, historic parks, forests, and seashores.
- withdrawal** An action that restricts the disposition of public lands and that holds them for specific public purposes.

Public lands in the United States have provided supplies of raw energy resources for over a century. The story of the harvest and administration of energy

from natural resources reveals the public debate and contest that often results when resources are owned in common. Recent developments suggest that future use of these resources will continue to be hotly debated.

1. INTRODUCTION

When Woody Guthrie wrote the song *This Land Is Your Land*, his rough-toned voice instructed Americans to recall that “this land was made for you and me.” Nowhere is this American ideal more obvious than in the nation’s publicly owned lands—or so it would seem. Yet the spirit of Guthrie’s idealism has rarely penetrated the administration of the more than 600 million acres of land that belong to each citizen. One-third of all national land is administered by the federal government and is owned collectively by the people of the nation. For a variety of reasons, including climate and the late entry of Euro-American settlers, most of the federal land can be found in the American West.

Public lands include parks, monuments, wilderness areas, refuges, underground mineral reserves, marine sanctuaries, historic parks, forests, and seashores. Throughout American history, the administration of this great national resource has been tied to politics and the powerful elements of economic development. The use of public land sites and particularly the harvest of natural resources existing on them have been consistently debated in recent years. Should these lands be viewed as vast storehouses of resources on which our nation depends? Or are federal lands intended more for preservation, regardless of valuable resources that might be found within their borders? And, of course, when development takes place, who should gain the financial benefit?

As a matter of existing law, mineral resources on public lands exist as national property, primarily due to specific strategies to foster development in the western United States. Outside variables, including the needs of war or domestic crisis or even presidential preference, influence the perceptions of mineral resources, particularly those resources related to energy production. Although the use and availability of such resources vary with the philosophy of each presidential administration, each American leader has viewed these energy sources as an important tool for economic development of communities in the western United States. The harvest of resources on public lands carries economic benefits locally and nationally (see Table I).

Although wood must be included as an energy resource that has been harvested from federal lands, the major sources in the 20th century have included: coal, oil, natural gas, shale, uranium, and geothermal. In recent years, there has additionally been a concerted effort to develop alternative energy resources (such as solar and wind) on public lands. By the end of the 20th century, the energy industry estimated that Americans received 30% of their fuels for energy production from public lands.

2. ENERGY FOR ALL?

The remaining oil, natural gas, and mineral deposits in the continental United States are concentrated in the American West, where they were created between 40 million and a few hundred million years ago. The abundance of such resources on public lands is a historical confluence of variables: lands that were unsettled and therefore taken by the federal government happen to have been the site, thousands of years ago, of ecosystems that today result in abundant energy resources. Western sites such as Fossil Butte National Monument (Wyoming) and Dinosaur National Monument (Utah) contain the fossilized remains that demonstrate abundant life many years ago, both in and out of the sea. Of course, as these ancient plants and animals decomposed over millions of years, extreme heat and pressure transformed them into resources that can be mined and burned to provide energy.

Such nonrenewable sources are just a fraction of the available energy resources. Today, the open space available on much of the federal lands has also become a resource for the development of another type of energy: renewable, including wind, solar, and water power. The U.S. Department of the Interior

TABLE I
Disposition of Public Lands, 1781–1997^a

Type of disposition	Acres
Unclassified ^b	303,500,000
Granted or sold to homesteaders ^c	287,500,000
Total unclassified and homestead	591,000,000
Support of common schools	77,630,000
Reclamation of swampland	64,920,000
Construction of railroads	37,130,000
Support of miscellaneous institutions ^d	21,700,000
Purposes not elsewhere classified ^e	117,600,000
Canals and rivers	6,100,000
Construction of wagon roads	3,400,000
Total granted to states	328,480,000
Granted to railroad corporations	94,400,000
Granted to veterans as military bounties	61,000,000
Confirmed as private land claims ^f	34,000,000
Sold under timber and stone law ^g	13,900,000
Granted or sold under timber culture law ^h	10,900,000
Sold under desert land ⁱ	10,700,000
Total miscellaneous dispositions	224,900,000
Grand total	1,144,380,000

^aData estimated from available records.

^bChiefly public, private, and preemption sales, but includes mineral entries, scrip locations, and sales of town sites and town lots.

^cThe homestead laws generally provided for the granting of lands to homesteaders who settled on and improved vacant agricultural public lands. Payment for the lands was sometimes permitted, or required, under certain conditions.

^dUniversities, hospitals, asylums, etc.

^eFor construction of various public improvements (individual items not specified in the granting acts), reclamation of desert lands, construction of water reservoirs, etc.

^fThe government has confirmed title to lands claimed under valid grants made by foreign governments prior to the acquisition of the public domain by the United States.

^gThe timber and stone laws provided for the sale of lands valuable for timber or stone and unfit for cultivation.

^hThe timber culture laws provided for the granting of public lands to settlers on condition that they planted and cultivated trees on the lands granted. Payments for the lands were permitted under certain conditions.

ⁱThe desert land laws provided for the sale of arid agricultural public lands to settlers who irrigated them and brought them under cultivation.

(USDI) oversees the leasing of lands for this development, including mining. Within USDI, the Bureau of Land Management (BLM) manages the development of all fossil fuels and minerals on nearly all of the lands, excluding areas of special jurisdiction such as the Naval Petroleum Reserves

TABLE II

Public Lands under Exclusive Jurisdiction of the Bureau of Land Management, Fiscal Year 1997^a

State	Vacant public lands (acres) ^b			Reserved lands (acres)		
	Outside grazing districts	Within grazing districts	Total	Lu ^c	Other	Grand total
Alabama	3117	—	3117	—	107,846	110,963
Alaska	86,908,060	—	86,908,060	—	—	86,908,060
Arizona	1,516,765	10,093,234	11,609,999	32,321 ^d	2,610,458	14,252,778
Arkansas	2059	—	2059	—	289,107	291,166
California	7,363,543	1,725,343	9,088,886	—	5,467,188	14,556,074
Colorado	480,331	6,781,321	7,262,065	—	998,241	8,296,512
Florida	1512	—	1512	—	23,765	25,277
Idaho	422,341	10,733,321	11,155,662	72,276	619,390	11,847,328
Illinois	3	—	3	—	224	227
Iowa	—	—	—	—	378	378
Louisiana	4351	—	4351	—	305,260	309,611
Michigan	47	—	47	—	74,807	74,854
Minnesota	6044	—	6044	—	145,211	151,255
Mississippi	1240	—	1240	—	55,971	57,211
Missouri	67	—	67	—	2094	2161
Montana	1,148,321	4,940,802	6,089,123	1,801,171	170,088	8,060,382
Nebraska	6580	—	6580	—	—	6,580
Nevada	3,140,726	44,493,239	47,633,965	3127	207,299	47,840,569
New Mexico	1,355,577	11,047,165	12,402,742	229,500	138,327	12,770,569
North Dakota	59,536	—	59,536	—	181	59,717
Oklahoma	2142	—	2142	—	—	2142
Oregon	585,675	12,455,100	13,040,775	78,124	3,104,840	16,233,739
South Dakota	272,277	—	272,277	—	7592	279,869
Utah	—	21,155,026	21,155,026	45,083	1,677,604	22,877,663
Washington	366,921	—	366,921	—	3189	370,110
Wisconsin	2521	—	2521	—	157,616	160,137
Wyoming	3,910,677	11,273,811	15,184,488	10,434	3,189,004	18,383,926
Total	107,560,433	134,698,775	242,259,208	2,308,242	19,355,680	263,929,258

^aLands under the exclusive jurisdiction of the Bureau of Land Management (BLM) (and resources) managed solely by the BLM. The table includes all Uperfected entries except for those involving reclamation and forest homesteads. Differences in fiscal year 1997 acreage from the previous year are due to land acreage acquired by the BLM; exchange land acreage received by the BLM; exchange BLM acreage (surface estate only) patented out by the BLM; withdrawn land acreage restored to BLM's jurisdiction; withdrawals, rights-of-way, acreage adjustments in carryover data.

^bThe following types of surveyed and unsurveyed public and ceded Native American lands are included: areas withdrawn under the executive orders of November 26, 1934, and February 5, 1935 (43 CFR 2400.0-3 *et seq.*); areas embraced in mineral withdrawals and classifications; area withdrawn for survey; and areas restored to entry within national forests (Act of June 11, 1906, 34 Stat. 233, 16 U.S.C., §§ 506–509), within reclamation projects (Act of June 17, 1902, 32 Stat. 388) and within power site reserves (Act of June 10, 1920, 41 Stat. 1063; 16 U.S.C., §§ 791). These lands are not covered by any nonfederal right or claim other than permits, leases, rights-of-way, and unreported mining claims.

^cLand Utilization project lands, purchased by the federal government under Title III of the Bankhead–Jones Farm Tenant Act and subsequently transferred by various executive orders between 1941 and 1958 from jurisdiction of the U.S. Department of Agriculture to the U.S. Department of the Interior, now administered by the Bureau of Land Management. Acreages are within grazing districts unless otherwise noted.

^dIncludes acreage on the Barry M. Goldwater Air Force Range.

and hydroelectric watershed development areas that are administered by the Army Corps of Engineers (see Table II).

The sheer amount of energy resources on federal lands is staggering. Of the 75.6 million acres of coal-bearing public lands, only 1% is currently being

mined. As for oil and gas, energy companies are actively working between 18,000 and 25,000 wells on public lands, with tens of thousands more sites held in reserve. A rarer mineral, uranium, is processed to create nuclear energy, which provides electricity to millions of homes, schools, and businesses. About one-third of all oil and gas comes from public lands. In addition, public lands are estimated to contain about 68% of all undiscovered oil reserves and 74% of all undiscovered gas reserves in our country. Nearly all reserves of uranium are found on federal land. Public lands are also used for harvesting hydroelectric power, although these areas are administered by the Army Corps of Engineers.

Regardless of the resource, the story of the harvest and administration of this energy supply reveals a theme that recurs in many publicly owned areas: public debate and contest. If studied over time, the use and management of the energy resources on public lands demonstrate important changes in the ethics guiding the management of federal lands. In addition, though, a survey of public land use demonstrates a public controversy that promises to rage deep into the 21st century.

3. GENERAL PHASES OF PUBLIC LAND ADMINISTRATION

Broader philosophical ideas, called ethics, toward land-use inform the administration of public lands. For the purposes of this discussion, the administration of public lands can be separated into four distinct eras or shifts in these larger, governing ethics. Such ethics, of course, are formed by a complex set of cultural, social, and political variables. As a publicly administered entity, public lands are subject to the influence of larger patterns within American society. In the case of energy development on public lands, the predominant cultural variable has been the ideological shifts caused by the evolution of the modern environmental movement in the 20th century. The ethic of the environmental movement has provided a nonutilitarian impulse toward the administration of the public domain.

Although the overall progression of public land use has moved toward incorporating an environmental ethic into the general management of the land, there have been a few crucial junctures in which the overall trajectory of this policy shift has been interrupted by a backlash of regressive efforts at development. After introducing the overall phases of public land administration, the focus here will be to

trace the general chronology of legal restrictions on public land policy, to explore how the harvest of specific resources has been affected by these laws, and, finally, to shed light on a few of these interruptions in the overall evolution of public lands policy. The data in Table III detail how public lands were allocated in 1996.

3.1 Settlement Phase

The settlement phase of land use grew out of the nation's emphasis on settling or reclaiming nearly 1.8 billion acres of original public domain. Hundreds of laws were passed between 1785 and 1878 that were designed to encourage settlement and development of the timber, mineral, and forage supplies. Typically, such policies said little about the native peoples who were being displaced by this disposal of land or about the fashion in which the lands were to be developed. This phase of land use had little to do with energy development, because it predominantly concerned an era in which there was little use of fossil fuels. Growing out of this approach to land use, however, Congress created the General Mining Law of 1872, which remains in effect to govern mining on public lands. The General Mining Law grants free access to individuals and corporations to prospect for minerals in public domain lands. Once anyone discovers a location, they can stake a claim on the mineral deposit. This claim provides the holder the right to develop the minerals and may be "patented" to convey full title to the claimant. In this stage of land use, the federal government clearly prioritized the stimulation of new use over any other consideration.

3.2 Resource Conservation Phase

Growing out of the American conservation movement of the late 1800s, the resource conservation era in public land use emphasized progressive ideals, particularly those expressed during the administration of President Theodore Roosevelt (1901–1909) by the president and his chief advisor on land management, Gifford Pinchot. This era marked the beginning of the nation's first period of public land stewardship. President William Henry Harrison previously began the forest reserve system with land appropriations before solidifying it the Organic Act of 1897.

Roosevelt and Pinchot emphasized forest management, expanding the reserves to 148 million acres. Energy resources, other than wood, attracted the attention of progressives with the Coal Lands

TABLE III

Applications, Entries, Selections, and Patents, Fiscal Year 1996

Type by state	Applications, entries, and selections				Patents issued		Purchase money (U.S. dollars) ^d
	Allowed		Approved				
	Number	Acres	Number	Acres	Number	Acres	
Homestead							
Montana					1	160	
Wyoming					2	304	
Total homestead					3	464	
Desert land							
Idaho					2	520	520
Nevada					8	2530	3162
Utah			1	39	1	39	40
Total desert land			1	39	11	3089	3722
Land sales (FLPMA) ^b							
Alabama					2	80	1050
Arizona					8 ^c	280	213,681
California					2	10	550
Colorado					2	38	64,906
Idaho					2	100	15,456
Nebraska					7	349	1600
Minnesota					2	232	
Nevada					9	444	2,717,375
New Mexico					9	18	60,892
Oregon					9	1744	203,459
South Dakota					2	1	1000
Utah					10	875	71,050
Wyoming					5	246	10,775
Total land sales					69	4417	3,361,794
Mineral							
Arizona					6	487	
Total mineral					6	487	
Native American new trust							
California					1	13	
Minnesota					6	275	
Wyoming					4	22	
Total Native American new trust					11	310	
Native American fee and reissue trust							
Colorado			1	52	1	52	
Idaho					7	640	
Montana			145	27,161	145	27,161	
Minnesota					2	48	
Nebraska					3	38	
North Dakota			35	6903	35	6903	
Oklahoma					25	1519	
Oregon			24	2091	24	2091	

continues

Table III continued

Type by state	Applications, entries, and selections						Purchase money (U.S. dollars) ^d
	Allowed		Approved		Patents issued		
	Number	Acres	Number	Acres	Number	Acres	
South Dakota			76	16,696	76	16,676	
Washington			49	2779	49	2779	
Total Native American fee and reissue trust			330	55,682	367	57,907	
Miscellaneous cash sales							
Color of title							
Alabama					2	80	1050
Arkansas					2	194	12,172
Colorado			1	1	1	1	3276
New Mexico					5	2	6
North Dakota			1	40	1	40	430
Oregon			1	40	1	40	3760
Total color of title			3	81	12	357	20,694
Recreation and public purposes							
Arizona			6	1465	8 ^d	1465	10,138
California					7	2548	523
Colorado			3	333	3	333	3558
Florida					1	1	
Minnesota					1	16	
Missouri					1	40	
Montana			1	160	1	160	
Nevada					8	460	14,812
New Mexico					2	84	50
Oklahoma					1	11	14
Oregon			1	3	1	3	50
Utah			6	314	6	314	67,800
Wyoming					2	243	6096
Total recreation and public purposes			17	2275	42	5678	103,041
Other							
Nevada					1 ^e	2	69,502
Washington					1 ^f	414	294,276
Total other					2	416	363,778
Total miscellaneous cash sales			20	2356	56	6451	487,513
Conveyance of federally owned mineral estates							
Arizona					7	5998	
California					1	561	
Colorado							50
Florida					2	7	225
Georgia					1	106	2000

continues

Table III continued

Type by state	Applications, entries, and selections						Purchase money (U.S. dollars) ^a
	Allowed		Approved		Patents issued		
	Number	Acres	Number	Acres	Number	Acres	
Minnesota					1	331	547
Montana					1	2	2600
New Mexico					1	1	1650
Utah					21	1443	584,798
Wyoming							50
Total conveyance of federally owned mineral estates					35	8449	591,920
Exchanges ^g							
Arizona					3	150	184,000 ^b
Arkansas					1	40	
California					42	17,981	23,500
Colorado					48	6208 ⁱ	— ^j
Idaho					32	7831	
Michigan					3	128	
Montana					11	12,058	
Nevada					20	45,039	10,360
New Mexico					10	80,476	
Oregon					17	11,541	74,000
South Dakota					3	707	
Utah					5	3963	
Washington					11	1925	150,000
Wyoming					3	5244	8311
Total exchanges					209	193,291	450,171
State grants							
Florida					1	175	
Total state grants					1	175	
Corrections							
Alabama					3	508	
Arkansas					1	320	
California					2	6548	
Colorado					3	74	
Indiana					1	64	
Louisiana					1	685	
Michigan					1	640	
Minnesota					3	380	
Mississippi					4	376	
Missouri					2	572	
Montana					1	15	
New Mexico					5 ^k		
North Dakota					2	640	
Ohio					4	560	
Oregon					2	91	
Utah					5	471	
Wisconsin					2	280	

continues

Table III continued

Type by state	Applications, entries, and selections						Purchase money (U.S. dollars) ^a
	Allowed		Approved				
	Number	Acres	Number	Acres	Number	Acres	
Wyoming					1	160	
Total corrections					43	12,384	
Other							
Alaska					364 ^l	600,204	400
Colorado					1 ^m	1	
Wyoming					2 ⁿ	256	
Total other					367	600,461	400
Grand total			351	58,077	1178	817,594 ^o	4,895,520

^aPurchase money includes all revenues collected during the reporting fiscal year at allowed and approved stages from all ongoing cases leading to patent; however, this money is not always received in the same year that the application is allowed or approved, or in the same year that the patent is issued. Purchase money enters into the land exchange process because exchanges must result in equal value given and received by both parties; this means that cash payments are sometimes made or received by the Bureau of Land Management to ensure an equitable exchange.

^bFLPMA, Federal Land Policy and Management Act of 1976.

^cPatents issued under Sections 203 and 209 of the Federal Land Policy and Management Act of 1976 (43 U.S.C. §§ 1713, 1719).

^dIncludes two Certificate of Transfer/Change of Use patents.

^ePublic land sale (2.5 acres), Santini-Burton Act of December 23, 1980 (94 Stat. 3881; Public Law 96-586).

^fPublic land sale patent under the Act of September 19, 1964 (78 Stat. 988; Public Law 88-608).

^gIncludes National Forest Service forest exchanges, Act of March 20, 1922 (42 Stat. 465; 16 U.S.C. § 485) as amended; Federal Land Policy and Management Act exchanges, Act of October 21, 1976 (90 Stat. 2756; 43 U.S.C. § 1716; and other exchanges.

^hThe U.S. Forest Service received a cash equalization payment under the Sikes Act of 1974 (16 U.S.C. § 484a); no land received by the United States.

ⁱDoes not include 960 acres of mineral estate only that was patented.

^jThe U.S. Forest Service paid \$666,010 in equalization payments to the exchange proponents.

^kRecordable Disclaimers of Interest in Lands, Section 315 of the Federal Land Policy and Management Act of 1976 (43 U.S.C. § 1745); no patents issued.

^lEight homesites (37 acres), Act of March 3, 1927 (44 Stat. 1364; 43 U.S.C. § 687a), as amended; headquarters Sites (15 acres), Act of March 3, 1927 (44 Stat. 1364; 43 U.S.C. § 687a), as amended; one trade and manufacturing Site (66 acres), Act of May 14, 1898 (30 Stat. 413), as amended August 23, 1958 (72 Stat. 730; 43 U.S.C. § 687a); 39 tentative approvals (355, 355 acres), Alaska Statehood Act of July 7, 1958 (72 Stat. 339), as amended; 31 interim conveyances (212, 333 acres), Alaska Native Claims Settlement Act of December 18, 1971 (43 U.S.C. § 1601); 281 native allotment certificates (32, 396 acres), Act of May 17, 1906 (34 Stat. 197), as amended August 2, 1956 (70 Stat. 954; 43 U.S.C. §§ 270-1 to 270-3); 1 highway lot (0.35 acres), Act of June 14, 1926 (44 Stat. 741; 43 U.S.C. § 971a).

^mClear Creek County, Colorado Public Lands Transfer Act of 1993 (108 Stat. 674), approved May 19, 1994.

ⁿAirport patents under Section 516 of the Airport and Airway Improvement Act of September 3, 1982 (49 U.S.C. § 2215).

^oThis total does not include 57,907 acres in Native American fee and reissue trust patents, and 12,384 acres in correction of patents, for a total of 70,291 acres.

withdrawal in 1906 that allowed the government to reserve mineral rights on 66 million acres. Initially, this reservation included oil and gas, though it would be further expanded to include many other minerals. Roosevelt also withdrew other public lands from settlement, using them to create national parks, monuments, wildlife refuges, and military bases. Within such sites, this era saw an increase in control over the use of the resources on public lands, including legislation such as the Mineral Leasing Act of 1920 and the Taylor Grazing Act of 1934. Pinchot's conservation ethic was reflected in laws that promoted sustainable use of natural resources.

These ideas were also of interest to President Franklin Delano Roosevelt, who used the Civilian Conservation Corps (CCC) during the 1930s to carry out many of these ideas on public lands.

3.3 Post-World War II Phase

During World War II and for most of two decades following it, the federal management of public lands went along with nationwide initiatives to stimulate national growth and international security. In 1946, the Grazing Service merged with General Land Office to form the BLM, the entity that oversees

the majority of the public lands in the western United States. The pressure to harvest resources from these lands intensified with the need for materials for war and the expansion of American consumption with the massive growth of its middle class.

With the beginnings of environmentalism in the late 1950s, many Americans grew more concerned that public lands were not subject to environmental assessment prior to use. This shift in overall ethics is generally accepted to have grown from the effort to prevent the damming of the Green River in Dinosaur National Park in 1954. Although the environmental movement influenced land use in national parks to a much greater degree, the public lands shifted more toward a compromise philosophy of “multiple use.”

3.4 Environmental Protection Phase

Although the era of environmental protection has met with significant interruptions, the overall pattern from 1960 to the present has been toward a more environmentally sustainable pattern of use on public lands. The Multiple Use Sustained Yield Act in 1960 was followed by the public outcry related to Rachel Carson’s publication of *Silent Spring* in 1962. The ensuing flurry of environmental legislation culminated in 1969 with the National Environmental Policy Act (NEPA), which created the Environmental Protection Agency (EPA) and required that any development on federally owned lands needed to be preceded by an environmental impact statement. Ecological concepts, such as ecosystems and watersheds, provided administrators with a new way of viewing public lands. No longer simply a storehouse of potential resources, each of these sites could now be seen as part of larger, complex natural systems. This awareness crystallized during a serious period of contestation over the use of public lands in the late 20th century.

4. ESTABLISHING THE LEGAL FRAMEWORK

The basic tool for defining each of the eras in public land use has been laws or regulations. Throughout the history of public lands, the primary rationale for their administration was their usefulness. If settlers had found reasons to build communities in certain areas, it is likely that those areas would not have become public lands in the first place. Vast tracts of the territories that became states in the late 1800s were simply unwanted by settlers and developers.

Similar to the Homestead Act, the General Mining Law of 1872 provided incentives to settlement and development (see Table IV). Typical of legislation of this era, the General Mining Law emphasizes individual rights with minimal government management. The law prioritizes free and open access to the land and relatively unlimited opportunity to explore and develop resources. Developers were offered access to permanent, exclusive resource rights and were exempted from paying royalties to the federal government. Clearly, the law is based on the simple motive to spur development.

The success of the effort of the General Mining Law to stimulate western development relied on allowing open exploration. During the late 19th century, little was done to stop exploration from growing rapidly into mine development and full-blown extraction. In these early years, miners were not required to share production, to create joint ventures, or to pay taxes or royalties. Many mining expeditions proceeded largely unnoticed on public lands. Today, the Mining Law continues to provide the structure for much of the western United States mineral development on public domain lands. The unfettered basis of use and ownership established by the 1872 law continues to fuel the debate over how to harvest energy and mineral resources from public lands. The basis for such law was the assumption that there were no alternative criteria for land valuation than the resources that it held. The basis for much of the debate at the end of the 20th century revolved around the simple question of what would happen when, for a variety of reasons, this land became desirable.

These debates in the 20th century slowed the unlimited development of the earlier years. Many of the public lands had acquired additional importance, independent of energy and mineral resources. Additionally, such development now rarely proceeded unnoticed. Some Americans began calling for energy companies to share revenues with the land’s owner through royalties. Other observers reflected a growing call for more stringent oversight of the extraction process. This call resulted in the 1920 Minerals Leasing Act. For the first time, this act closed some federal lands to open access. Most often, such lands then became available for leasing. The 1920 act established a broad framework for leasing and for the payment of royalties. Lands were to be designated as “known mineral areas” (KMAs) and then administered by one of two methods: (1) preference right leasing, which provided exploration permits that would expire if the search was unsuccessful, or

TABLE IV

Patents Issued with Minerals Reserved to the United States through Fiscal Year 1996^a

	Type of mineral reservation (acres)						
State	All minerals	Coal	Oil and gas	Phosphate	Oil and gas plus other minerals	Miscellaneous minerals and combinations	Total
1909–1948 ^b							
Alaska	6501	10,823	1095		773		19,192
Alabama	4412	63,586	9563		2889		80,450
Arizona	2,547,517	4403	27,497			101,880	2,681,297
Arkansas	1107	1520	15,043	85	40		17,795
California	2,352,070	3005	156,783		23	1864	2,513,745
Colorado	4,271,042	1,348,288	215,423		38,494		5,873,247
Florida	1154		2304	71,259	520		75,237
Idaho	1,291,163	11,749	4940	70,036	216,060		1,793,948
Illinois	634	120					754
Iowa	359						359
Kansas	54,384		1421				55,805
Louisiana	1223		17,105		3844		22,172
Michigan	1935		3261				5196
Minnesota	235			8			243
Mississippi	974		10,231				11,205
Missouri	166						166
Montana	3,993,640	6,658,554	987,472	11,290	17,788	150	11,668,894
Nebraska	72,964		3253				76,217
Nevada	242,717	960	1119		80	40	244,916
New Mexico	6,378,118	614,779	112,995	680	70,673	2,092,091	9,269,336
North Dakota	134,578	4,636,851	11,915		1164	40	4,784,548
Ohio	38					744	782
Oklahoma	48,781		10,917				59,698
Oregon	1,639,742	5598	14,369		480		1,660,189
South Dakota	1,565,802	187,722	6328				1,759,852
Utah	856,083	215,528	98,922	1576	8157	1680	1,201,946
Washington	262,444	14,535	2518		384	400	280,281
Wisconsin	1546						1546
Wyoming	9,541,179	2,297,363	376,906	32,037	17,341	257	12,265,083
Total through 1948	35,272,508	16,075,384	2,091,380	406,971	378,710	2,199,146	56,424,099
Fiscal period ^c							
1949–1950	363,466	24,868	18,287	1138	7662	4235	419,656
1951–1960	1,258,347	64,204	650,841	5075	109,323	36,586	2,124,376
1961–1970	2,103,335	44,578	893,114	559	498,973	30,663	3,571,222
1971–1980	2,914,664	9344	186,689	340	185,764	49,784	3,346,585
1981–1990	2,694,395	16,609	294,416	439	159,006	166,558	3,331,423
1991	38,353		1603		3990	5534	49,480

Continues

Table IV continued

State	Type of mineral reservation (acres)						Total
	All minerals	Coal	Oil and gas	Phosphate	Oil and gas plus other minerals	Miscellaneous minerals and combinations	
1992	49,341	698	71,773		2486	2261	126,559
1993	41,431	1788	3034		8963	4096	59,312
1994	30,322		11,781		11,061	1354	54,518
1995	37,612	1680	38,630		21,265	27,033	126,215
1996	36,311	2271	57,483		3532	1988	101,585
Total, 1949–1996	9,567,577	166,040	2,227,651	7551	1,012,025	330,092	13,310,931
Grand total	44,840,085	16,241,424	4,319,031	414,522	1,390,735	2,529,238 ^d	69,735,030

^aData prior to 1949 were tabulated from incomplete records. Data include patents issued on both BLM and Forest Service lands.

^bMinerals began to be reserved with the Agricultural Coal Lands Act of March 3, 1909; acts expanding on this and covering other minerals were subsequently enacted.

^cAs of June 30 through 1976; thereafter, as of September 30 each year.

^dDoes not include 165,712 stockraising homesteads comprising 70,362,925 acres that reserved all minerals to the United States. Some of these homesteads were entered but not patented.

(2) competitive bidding for designated tracts, which used lotteries, oral auctions, sealed bids, or other formats to disperse the rights to KMAs and transfer them into known leasing areas (KLAs).

Eventually, the Competitive Bidding Act would be made even more specific. For coal, the law differentiated between new or maintenance tracts, but still required bids for existing operators to continue mining an existing site. Oil and gas development used reservoir studies to determine the pattern of leasing and lottery systems. In 1987, lottery systems were discontinued for oil and gas and development and were replaced by a familiar pattern in which industry representatives nominated tracts, which were then assessed for their environmental acceptability.

Neither the 1872 Mining Law nor the 1920 Mineral Leasing Act contained any direct environmental controls, but mining claims are subject to all general environmental laws as a precondition for development. These restrictions and requirements grew out of additional legislation, particularly the seminal environmental policies of the 1960s and 1970s, including the Multiple Use Sustained Yield Act, Wilderness Act, National Forest Management Act, National Environmental Policy Act (NEPA), and Federal Land Policy and Management Act. To varying degrees, each of these policies addressed environmental protection, multiple use, and management of federal land generally. By imposing new requirements on agency actions, and by withdrawing some federal lands from development, these acts

have significantly influenced mineral development on public lands.

By far, the most critical policies relating to the administration related to the formation and structure of the BLM. When it was formed by merging two portions of the U.S. Department of the Interior (the General Land Office and the Grazing Service) in 1946, the BLM became entrusted primarily with the management of the remaining public domain and railroad/wagon road grant lands. The BLM also has the primary responsibility for the management oversight of the mineral domain that underlies the public lands. As with the sister agencies, the BLM is heavily involved in forest and rangeland planning, as well as keeping the official land status records of all public federal lands. Over time, the BLM followed the lead of the U.S. Department of Agriculture (USDA) Forest Service and adopted policies of multiple use and, eventually, of ecosystem management. In many cases, BLM lands are adjacent to or intermingled with other public lands, particularly National Forest land. The BLM essentially operated under these previously existing acts until congressional discussion over forest management in the Forest Service came to a head in the mid-1970s. This debate resulted in the passage of the Federal Land Policy and Management Act of 1976 (FLPMA), which provided the BLM with its “organic act”—albeit three decades after its founding.

FLPMA brought the greatest change to the administration of public lands in American history. Through FLPMA, BLM acquired an unequivocal

statutory basis for its public land management policy; however, debate continued over what exactly this meant. The 1934 Taylor Act had provided for management of natural resources on public land “pending its final disposal.” To some observers, the phrase spoke to the old policy of transfer of public lands into private ownership; to others, it provided an argument against federal investment in land administration or rehabilitation. FLPMA, however, clearly prescribed retention of public lands in federal ownership as the rule and made transfer a carefully circumscribed exception. FLPMA established or amended many land and resource management authorities, including provisions on federal land withdrawals, land acquisitions and exchanges, rights-of-way, advisory groups, range management, and the general organization and administration of BLM and the public lands. FLPMA also called for public lands and their resources to be periodically and systematically inventoried. The BLM was empowered to use the data to create a plan for each site’s present and future. Additionally, after many years of debate, FLPMA also specified that the United States must receive fair market value for the use of public lands and their resources.

Finally, BLM was required to administer public lands in a manner that protected the quality of scientific, scenic, historical, ecological, environmental, air and atmospheric, water resource, and archeological values. Where appropriate, BLM was also empowered to preserve and protect certain public lands in their natural condition, particularly when such preservation would provide food and habitat for fish and wildlife. Of course, BLM also remained responsible for providing sites for outdoor recreation and human occupancy and use of such sites. In short, FLPMA proclaimed multiple use, sustained yield, and environmental protection as the guiding principles for public land management. Thanks to FLPMA, BLM took on the responsibility of administering public lands so that they are utilized in the combination that will best meet the present and future needs of the American people for renewable and nonrenewable natural resources. Although this remains the role of BLM today, there have been many exceptions to this jurisdiction.

5. MAKING LEGAL EXCEPTION: NATIVE AMERICAN COMMUNITIES

There remain areas of federal lands that do not cohere to these legal regulations. For instance, it is

reported that 80% of the nation’s uranium reserves, 33% of the nation’s low-sulfur coal, and 3–10% of the recoverable oil and gas are located on Native American reservation land. Problems of jurisdiction and ownership have highlighted the history of resource extraction on native lands. More important, though, are the health, social, and environmental impacts that have been noted where such development has been allowed to take place.

Outside sources approached Native American reservations for energy resources as early as 1920. For instance, the Navajo’s prolonged involvement with oil development began with the influence of New Mexico’s own Secretary of the Interior Albert B. Fall (who will be discussed later for his pivotal role in the management of energy resources on federal land). With one of Interior’s charges being to administer Native American reservations, Fall soon looked toward reservation land as a vast, untapped (and unprotected) resource. Native peoples, Fall wrote, were “not qualified to make the most of their natural resources.” Using the General Leasing Act of 1920, Fall attempted to clarify the status of treaty lands for leasing so that the government could do the leasing for the Native Americans. Although Fall’s perspective grew from a desire to battle waste, public sentiment after World War I held that the United States would run out of oil within 10 years. Fall saw to it that oil companies won leases on reservation land, particularly that of the Navajos in New Mexico. In addition, he conspired with oil companies to provide native occupants with ridiculously low royalty rates. Historian Kathleen Chamberlain notes that oil intensified factions within the native community. In sum, she writes, “oil also expedited a shift from subsistence and barter to a wage economy.” Although oil generated some jobs for the community, it never neared the promised levels.

Throughout the history of energy development on Native American lands, the development has been found to undermine tribal and family values and identity. For instance, in Hopi and Navajo experiences, wealth from royalties was spread unevenly and the large influx of wealth disturbed traditional hierarchies within the culture. Additionally, the non-Native American workers, who were necessary to energy companies for reservation development of energy resources, represented a persistent threat to tribal sovereignty. Possibly the most invasive change that energy development has wrought on Native American groups is on the traditional political process. In nearly every case, the debate over leasing lands for extraction has bitterly divided tribal leadership,

ultimately resulting in long-term impacts on the political structure of the group. With such breakdown, the reservations become what scholars have called a “colonial reservation economy.” Some of the most troubling instances involved negotiations with the U.S. federal government. During World War II and through the Cold War, uranium was mined from Hopi and Navajo reservations in Arizona. The mine remnants as well as the remaining waste have contributed to significant rises in cancer and other health problems. Additionally, many Hopi and Navajo men worked in the mines and have therefore contracted a variety of ailments, particularly lung cancer.

Since the 1970s, Native American groups have attempted to use the courts to acquire fair royalty rates and recompense for health hazards. This has been an uphill battle, with various successes starting in the late 1980s. The Native American community’s lessons about the community impact of energy development blossomed into an organization, the Council of Energy Resource Tribes (CERT), which helps each tribe to complete independent management of its energy resources. This has helped many tribes to acquire fair royalty rates as well as to initiate extensive development of alternative sources on western reservation land.

6. HISTORY OF ENERGY DEVELOPMENT ON PUBLIC LAND

The laws and ideas relating to energy development on public lands have changed due to shifts in larger ethics within American society, but also due to the type of resource being extracted. Primarily, the pressure to mine energy resources grew as resources such as petroleum became more and more critical to American life and national security. The most significant change followed the increased use of petroleum for transportation after 1900. For instance, by World War I, the energy reserves on public land had been transformed from an attraction for potential settlement to a matter of national security. In World War I, fossil fuels became directly linked to the nation’s ability to protect itself and maintain global security. Petroleum, however, is only one example. In this section, three specific energy resources and the time period that they most influenced are considered.

6.1 The Beginnings with Coal

The first energy resource to be granted unique status on public lands was coal. The Coal Act of 1864 gave

the U.S. President authority to sell coal beds or fields for no less than \$20/acre. Lands not sold in this fashion could be available to the general public at the minimum price under the general land statutes. In 1873, Congress authorized citizens to take vacant coal lands of not more than 160 acres for a minimum of \$10/acre. This law governed the disposal of coal lands until 1920. Precious metals were governed much more freely. In all probability, the rationale behind this distinction lay in the relative ease of locating coal compared to that of most precious metals.

The importance of the coal supply to national development would make it a major political issue in the early 20th century. In rapid succession, the Roosevelt administration proposed controlling the federal grazing lands, increasing the national forests, and withdrawing the coal, oil, and potash lands from sale. In this same period, the administration also reserved potential waterpower sites and set out to reclaim the arid lands of the West. In 1906, Roosevelt faced up to the rumors of random exploitation of coal supplies on federal lands by withdrawing coal-bearing lands from public availability. He responded to industry’s complaint that the acreage limitations in the 1873 statute impeded the development of certain types of coal. As a result, many coal and railroad companies evaded these limits by using dummy entries or by entries processed under the agricultural land statutes. To keep certain companies from acquiring a monopoly and to conserve mineral supplies, the withdrawals began in November 1906. Located in Colorado, North Dakota, Montana, Oregon, Washington, Utah, Wyoming, and the Territories of New Mexico, 66 million acres of land with supplies of “workable coal” were withdrawn. In his autobiography, Roosevelt explained his logic as follows:

The present coal law limiting the individual to 160 acres puts a premium on fraud by making it impossible to develop certain types of fields....It is a scandal to maintain laws which sound well, but which make fraud the key without which great natural resources must remain closed. The law should give individuals and corporations under proper government regulation and control...the right to work bodies of coal large enough for profitable development. My own belief is that there should be provision for leasing coal, oil and gas rights under proper restrictions.

Many western congressmen reacted angrily to Roosevelt’s withdrawal. They claimed that Congress and not the Chief Executive possessed the right to dispose of the public domain in such a manner. Some

referred to Roosevelt's use of conservation as "state socialism." Roosevelt revised his argument slightly in 1907 to include mineral fuels with forests and navigable streams as "public utilities." Between 1906 and 1909, public coal lands sold for between \$75 and \$100 per acre. During these same years, over 4 million acres were segregated, though many would be returned to the public domain.

6.2 Contested Development: Petroleum and Teapot Dome

Following the 1859 strike of the first oil well in Pennsylvania, development in the United States moved to the Midwest. As oil exploration moved westward, it became evident that the public lands held great stores of "black gold." By the 1870s, oil fever had swept specifically into the public domain of California. In the early 1890s, oil was discovered near Coalinga, California and in 1892, E. L. Doheny drilled the first successful well in Los Angeles. The Placer Act of 1870 included "all forms of deposit" except veins of quartz. Though the Department of Interior debated whether petroleum qualified, Congress passed an 1897 Placer Act that upheld that oil lands were administered by placer location. In 1909, President Taft temporarily withdrew 3 million acres of potential oil lands in the public domain.

Similar to Roosevelt's coal withdrawals, Taft's act garnered a great amount of debate—particularly in the West. Congress then passed the Pickett Act in 1910 to allow presidents to withdraw public lands for examination and reclassification for various "public purposes." During the 1910s, this act enabled the President to withdraw nearly all of the known oil lands. In order to conserve an adequate supply of oil for the Navy, two naval reserves were established by executive order in 1912; one was in Elk Hills, California and the other was in Buena Vista Hills, California. Similar to a forest reserve, the oil supply, administrators believed, was best preserved if left in the ground. These withdrawals were then followed in 1915 by reserves at Teapot Dome, Wyoming and in 1916 by the naval oil shale reserves in Colorado and Utah. In 1923, a similar reserve was created in Alaska.

Although most of the public debate centered on the right of the executive branch to make such withdrawals, other issues emerged as well, including whether the land should be sold or leased to oil companies, whether land previously patented to railroad companies could be withdrawn by the federal government as well, and whether oil compa-

nies had any standing if the federal government wished to withdraw land that they had already leased. These issues, of course, took on greater importance as the United States observed and then participated in World War I at the end of the decade. Petroleum had many military uses, including powering many of the ships of the new Navy.

Historian John Ise was one of the loudest critics of such government policy. His argument, however, differed from those of most of the critics. He wrote in 1926 that "during this time there was always overproduction of oil from privately owned lands and there was never any need or justification for opening any public lands." Although this argument shifted to the fore by the end of the 20th century, there was little public support for the idea of conservation. Western lawmakers focused on ensuring federal supplies of petroleum on public lands when it passed the Mineral Leasing Act of 1920, otherwise known as the Smoot Bill. Congress intended this act to bring relief to some of these issues and to offer more equal treatment to interests of the western United States. In the naval reserves, leases were given only on producing wells, unless the President chose to lease the remainder.

Within 6 months after the passage of the Mineral Leasing Act, the Secretary of the Navy was granted extensive jurisdiction over the naval reserves. In 1921, Albert B. Fall, a senator from New Mexico, was appointed Secretary of the Interior by Warren G. Harding. In short order, Fall convinced Secretary of the Navy Edwin Denby to relinquish control of the reserves and had Harding sign a secret presidential order to this effect. Within a few months, word leaked to the public that Fall had leased reserves in California and Wyoming to his friends in the oil business. The episode that transpired in full view of the public became the primary test of jurisdiction and law regarding energy development and the West.

In the scandal known as Teapot Dome, a congressional committee found that Fall had acted secretly and without competitive bidding to lease the Teapot Dome to Harry Sinclair and the California tract to E. L. Doheny. Allowing a private company to develop naval oil reserves, Fall went outside the boundaries of existing law. As a result, he was one of the first members of a presidential cabinet to serve a jail term; he was convicted of bribery and ordered to serve a year in prison as well as to pay a \$100,000 fine.

Fall represented the western interests who felt betrayed by the overly zealous control of the federal government over the development of public lands. Before Warren Harding appointed him as Secretary of the Interior, Fall had earned his spurs in farming and

ranching. In New Mexico, he owned a ranch that spanned more than 55 miles long and 35 miles wide. He had been a judge and a sheriff. Throughout his public life in New Mexico, Fall referred to the territory as “corporation country.” He hated government interference in business, and he did not have to struggle with torn loyalties when he became a federal employee. “I have long believed,” Fall said in 1918, “that the laws should protect...capital and should be so enforced as to offer an inducement for its investment.” In giving business the advantage of governmental favors, Fall had, however, acted as a man of a certain faith. The core of his faith was simple: because nature’s resources were unending, there was no need for humanity to restrain its ambitions and desires. “All natural resources should be made as easy of access as possible to the present generation,” Fall declared. “Man cannot exhaust the resources of nature and never will.” This, of course, included those resources falling within the public domain.

When Americans began to hear about Teapot Dome in news reports in the 1920s, much of the public participated in a growing appreciation of nature. Fall’s feelings were well known to many observers. On his appointment in 1921, one newspaper wrote that “To us Fall smells of petroleum and we know what that means among...politicians.” Another paper, also prior to the Teapot Dome scandal, described the new administrator of federal lands as a man who found “wise conservation of natural resources...hateful to his every instinct.” Although Fall would be held accountable for his abuse of power on the federal lands, his basic idea of unfettered development and unending abundance remains on the table. This has been a major portion of the debate over petroleum development on federal lands. Although Congress would act in 1928 to give the Secretary of the Navy exclusive jurisdiction of the petroleum reserves, the matter would continue to be debated throughout the 20th century.

6.3 Energy and National Security: Uranium

Similar debate followed the post-World-War II search for uranium to be used in nuclear weapons and eventually in nuclear power generation. The demand for fissionable source materials after the war and the urgency of a government monopoly in them prompted the enactment of the Atomic Energy Act of 1946, which reserved all such ores in public domain land to the United States and required that they also be reserved in any mineral patent.

The uranium boom reached its height around 1954, when prospectors flooded western lands in fashion similar to the California gold rush. Private prospectors made millions of dollars leasing and locating uranium supplies through the end of the 1950s. Applicable to both uranium and oil and gas deposits, the Multiple Mineral Development Act of 1954 clarified how lessees were to deal with dormant mining claims that may cloud their own title. The new act provided a procedure in which title examiners may be used. After the lessee filed in the proper county office, he or she then waited for 90 days before filing with the BLM.

7. *DEBATING OWNERSHIP: THE SAGEBRUSH REBELLION*

Picking up on the ideas of Albert Fall, a group of 11 western states rallied in the 1970s to gain control of the development of the resources on their public lands. In what has become known as the “Sagebrush Rebellion” and the “Wise Use Movement,” these groups have gained national attention and have often been able to bypass FLPMA and other government regulations.

In the late 1970’s, the movement flared up with the support of Adolph Coors and other westerners. Prominent in this legal effort, the Rocky Mountain Defense Fund employed talented lawyers James Watt and Ann Gorsuch. Prior to his election, Ronald Reagan espoused support for the Sagebrush Rebellion. When he was elected in 1980, Reagan wasted no time in appointing Watt the Secretary of the Interior and Gorsuch to head the Environmental Protection Agency. Watt pressed his ideas forward and was able to open thousands of acres of public land to coal mining. However, with his forced resignation in the 1983, the movement lost a great deal of its support from the Reagan White House.

The same ideas, however, have found new life in the administration of George W. Bush. Guided by leaders such as the Assistant Secretary of the Interior J. Steven Griles, who served under Watt and then spent years as a consultant to the energy industry, the Bush administration pushed to open up vast new tracts of federal land to oil, gas, and coal development.

8. *A SYMBOLIC CONTEST: ALASKA*

With the growing strength of the environmental movement after 1970, the utilitarian view of public

lands met public debate at nearly every juncture. This has been most evident in an ongoing symbolic fight over federal lands in Alaska.

The energy crisis of the 1970s, which was exacerbated by the 1973 Oil Embargo carried out by members of the Organization of Oil Exporting Countries (OPEC), caused a significant shift in feelings about oil development in federally owned areas of Alaska. The stalled debate over drilling in northern Alaska and constructing a trans-Alaska pipeline (TAP) to Valdez gathered congressional support and was ultimately carried out in the late 1970s. American leaders assured the public that the TAP would be a step toward energy independence. Specifically, the Nixon administration assured Americans that a TAP would cause oil imports from the Middle East to fall from 27% of the U.S. supply to 20% by 1990.

President Jimmy Carter's willingness to allow these developments was predicated on his effort to formalize restrictions on further such activities in undeveloped areas of Alaska. From 1978 to 1980, President Jimmy Carter established the Alaska National Wildlife Refuge (ANWR) and set aside 28% of Alaska as wilderness. Carter announced this watershed legislation as follows: "We have the imagination and the will as a people to both develop our last great natural frontier and also preserve its priceless beauty for our children." With the passage of the Alaska National Interest Lands Conservation Act (ANILCA) in 1980, the BLM was ordered to oversee 11 million acres of Alaska as wilderness. This included the Alaska National Wildlife Refuge.

Debate on further development in these areas of Alaska awaited the administration of George W. Bush. As part of the administration's energy plan, new priority was placed on developing energy resources on federal lands. Secretary of the Interior Gale Norton specifically zeroed in on ANWR in 2000. Congress rebuffed the first effort to open ANWR to oil development, but other efforts are sure to follow. With the Republican victory in the 2002 congressional elections, most environmentalists foresee little to stand in the way of energy developments in ANWR.

9. BATTLE LINES IN THE CURRENT DEBATE

Thanks to FLPMA, most new energy extraction requires an environmental impact statement (EIS), which is a detailed study of the proposed plan and

how it will affect wildlife, plant life, and the quality of the water, air, and soil. The EIS also examines how the resource extraction will affect other uses of the public lands, such as recreation, cultural resource conservation, or grazing. If the impacts of the proposed well, dam, or mine seem too great, then alternatives are examined. These range from scrapping the plan entirely to using advanced technology to extract the resource, such as directional drilling, whereby the well is sited in a less vulnerable area but the drill still reaches the underground reserves.

The debate over ANWR crystallizes many of the issues that now make up the larger debate about how energy development should proceed on federal lands. During the administration of President William Clinton, oil and gas leasing increased considerably over previous administrations, but with new environmental restrictions on industries that sought to develop energy resources on public lands (see Table V). President Clinton also designated 21 western national monument areas and 56 million acres of roadless area. Such changes, however, were susceptible to any changes in national politics. That change came in the election of 2000.

"Environmental issues today," instructed Bush's Interior Secretary Gale Norton, "are more complex and subtle than the ones we faced in the 1960s and 70s. With the growth of our population and expansion of our economy, pressures have increased correspondingly on our undeveloped land, water resources, and wildlife." Norton intends to put "collaboration ahead of polarization, markets before mandates...and to transcend political boundaries." In order to do so, she and George Bush coined the phrase "new environmentalism" to meet the U.S. need "for a vibrant economy and energy security—while at the same time protecting the environment." Thus far, environmental organizations have been disappointed by Bush and Norton's "new environmentalism." "We're dismayed that Pres. Bush wants to turn the oil and gas industry loose on virtually all the lands in our national forests, national wildlife refuges, and other public systems," said Wilderness Society representative Bill Meadows. "We're beginning to wonder if James Watt is whispering in his ear," he said.

In addition to ANWR, federal lands along the Rocky Mountain front in northwestern Montana to southern Wyoming's Red Desert and Colorado's Vermillion Basin are being explored for energy development. The BLM collected data on oil and gas deposits and their availability throughout the

TABLE V

Continuing Oil and Gas Activities on Federal Lands as of September 30, 1996

State	Number of units in effect ^a	Number of producible and service holes ^b	Number of producible and service completions ^b	Number of producible leases ^a	Number of acres in producing status ^a
Alabama	4	40	43	18	6222
Alaska	9	150	161	38	67,350
Arkansas	2	311	334	158	73,353
California	26	5693	6122	350	79,102
Colorado	136	3098	3331	2122	1,397,557
Florida	—	4	4	2	3468
Illinois	1	13	14	5	964
Kansas	8	476	512	436	110,464
Kentucky	—	135	145	43	24,132
Louisiana	12	386	415	95	38,289
Maryland	—	9	10	—	—
Michigan	20	70	75	57	29,900
Mississippi	7	195	210	85	45,268
Montana	50	3318	3568	1307	679,084
Nebraska	6	36	39	20	6507
Nevada	16	91	98	57	44,300
New Mexico	231	24,367	26,201	6122	3,512,145
New York	—	2	2	4	1010
North Dakota	32	691	743	490	307,774
Ohio	—	308	331	125	19,162
Oklahoma	19	269	289	775	121,376
Pennsylvania	—	54	58	67	6967
South Dakota	7	181	195	76	48,921
Tennessee	—	10	11	3	2366
Texas	14	207	223	163	68,931
Utah	112	1919	2063	1236	914,203
Virginia	—	30	32	9	5126
West Virginia	—	280	301	174	104,065
Wyoming	518	13,984	15,037	6049	2,514,920
Total	1230	56,327	60,567	20,086	10,514,920

^aData from the Bureau of Land Management Case Recordation System.^bService holes and completions are not necessarily located on producible leases. Data from the Bureau of Land Management Automated Inspection Record System.

public lands in order to support the prodevelopment stance of the Bush administration. Many local landowners and environmentalists have fought these efforts. Norton has requested new financing to streamline permits and study sites to drill on federal lands. The Bush administration has also shown some interest in increasing the development of alternative energy sources on federal lands as well. The future of the public lands is certain to contain open debate over the domain's role in America's energy future.

Acknowledgments

The author thanks the Center of the American West, University of Colorado, Boulder, and the William and Flora Hewlett Foundation for assistance in preparing this article.

SEE ALSO THE FOLLOWING ARTICLES

Arid Environments, Impacts of Energy Development in • Development and Energy, Overview • Ecological

Risk Assessment Applied to Energy Development • Environmental Injustices of Energy Facilities • Oil-Led Development: Social, Political, and Economic Consequences • Polar Regions, Impacts of Energy Development • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Wetlands: Impacts of Energy Development in the Mississippi Delta

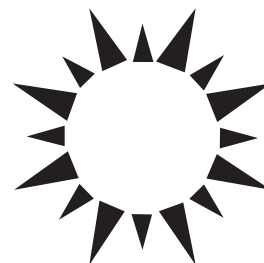
Further Reading

- Carstensen, V. (1963). "The Public Lands; Studies in the History of the Public Domain." University of Wisconsin Press, Madison.
- Center of the American West (2003). "What Every Westerner Should Know about Energy." <http://www.centerwest.org/>
- Chamberlain, K. (2000). "Under Sacred Ground: A History of Navajo Oil, 1922–1982." University of New Mexico Press, Albuquerque.
- Clawson, M. (1987). The Bureau of Land Management, 1947–1953. *Cruiser* [Newsletter of the Forest History Society]. 10(3), 3–6.
- Gates, P. W. (1968). "History of Public Land Law Development." U.S. Government Printing Office for the Public Land Law Review Commission, Washington, D.C.
- Gerard, D. (1997). "1872 Mining Law: Digging a Little Deeper." Political Policy Research Center (PERC) Policy Series PS-11. PERC, Bozeman, Montana.
- Gordon, R., and VanDorn, P. (1998). "Two Cheers for The 1872 Mining Law." CATO Institute, Washington, D.C.
- Hage, W. (1989). "Storm Over Rangelands: Private Rights in Federal Lands." Free Enterprise Press, Bellevue, Washington.
- Ise, J. (1926). "The United States Oil Policy." Yale University Press, New Haven.
- Leshy, J. D. (1987). "The Mining Law: A Study in Perpetual Motion." Resources for the Future, Washington, D.C.
- Muhn, J., and Stuart, H. R. (eds.). (1988). Opportunity and Challenge: The Story of Blm. U.S. Government Printing Office, Washington, D.C.
- National Research Council, Committee on Hardrock Mining on Federal Lands. (1999). "Hardrock Mining on Federal Lands." National Academy Press, Washington, D.C.
- Robbins, R. M. (1976). "Our Landed Heritage: The Public Domain, 1776–1970." University of Nebraska Press, Lincoln.
- Rohrbough, M. J. (1968). "The Land Office Business the Settlement and Administration of America's Public Lands, 1789–1837." Oxford University Press, New York.
- Voight, W., Jr. (1976). "Public Grazing Lands: Use and Misuse by Industry and Government." Rutgers University Press, New Brunswick, New Jersey.
- Williams, T. T. (2002). "Red: Passion and Patience in the Desert." Vintage Press, New York.

Energy Efficiency and Climate Change

ARTHUR ROSENFELD, PAT McAULIFFE, and
JOHN WILSON

California Energy Commission
Sacramento, California, United States



-
1. Introduction
 2. Improving Trends in Energy Efficiency
 3. The Combined Impact of Complimentary Efficiency Improvements
 4. Efficiency Improvements Reduce Generation Requirements
 5. Historic Primary Energy Trends in the United States
 6. Energy Intensity Trends in the United States and Other Regions
 7. Future Primary Energy Trends for the World
 8. The Effect of Efficiency Improvements on Primary Energy Demand
 9. Pathways to Improving Energy Intensity

Glossary

energy efficiency Useful output divided by energy input.
energy intensity Energy consumption per unit of gross domestic or gross world product.
final energy Energy available for use in homes, factories, and for transportation after accounting for losses in conversion (for example, waste heat from an electric power plant).
global warming Increases in the earth's surface temperature brought about by the release of greenhouse gases.
greenhouse gases Gases in the atmosphere that tend to trap heat and contribute to global warming (for example, carbon dioxide).
gross domestic product (GDP) Total value of goods and services in a country.
gross world product (GWP) Total of gross domestic product for all countries.
kWh Unit of electrical energy equal to 3,413 Btus.
million tons of oil equivalent (Mtoe) 40 million Btus.
primary energy Energy available from sources such as petroleum and coal.
quads 10^{15} British thermal units.

Large quantities of fossil fuels are burned to meet the world's energy needs. The combustion of these fuels produces carbon dioxide, contributing to increasing concentrations of carbon dioxide in the earth's atmosphere. According to the Intergovernmental Panel on Climate Change, the increasing concentrations of carbon dioxide and other greenhouse gases in the atmosphere due to human activity are causing the average surface temperature of the earth to increase. Such human influences are expected to continue to contribute to rising greenhouse gas emissions and additional global warming through this century.

1. INTRODUCTION

Humans combust these fuels to provide energy for needs such as heating, transportation, refrigeration, and industrial processes. The efficiency with which fossil fuels can be converted to meet these needs determines how much primary energy will be combusted. The greater the end-use efficiency, the less fuel is needed and the less the impact that such human activities will have on global warming, all other factors remaining the same.

Over the past century, the efficiency of energy use has improved. For example, lighting has changed from the burning of candles that produced 1 lumen for every 6 watts of candle or wax burned, to incandescent electric lights that produced 17 lumens with each watt of electric input, to fluorescent lighting that produces 100 lumens per watt. So the energy intensity index, defined as the amount of primary energy needed to produce a unit of gross domestic product, continues to improve. If this index can continue to improve at a rate of 2% per year, the

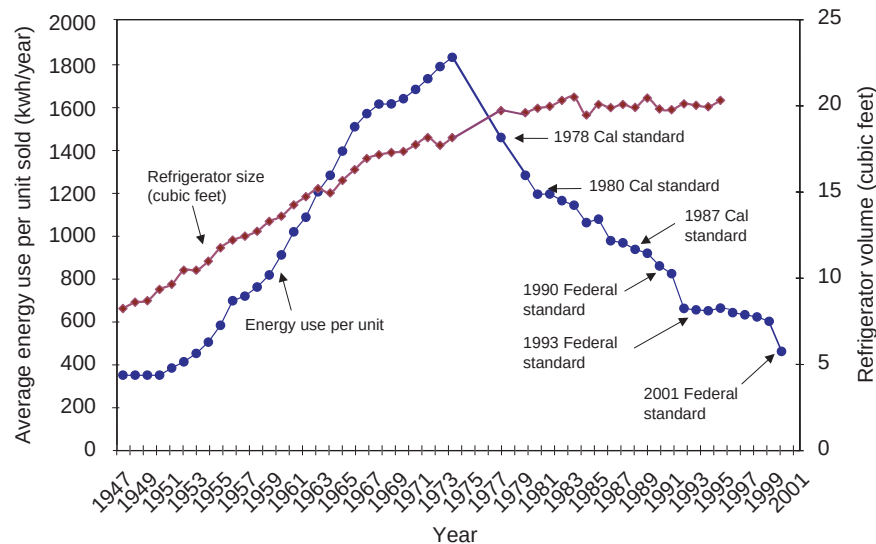


FIGURE 1 United States refrigerator use versus time. Annual drop from 1974 to 2001 is 5% per year. From David Goldstein of the Natural Resources Defense Council.

primary energy needs of the world can be limited even while increasing access to goods and services. Such an improvement in energy intensity also will lead to a reduction in carbon dioxide emissions and limit the extent of global warming.

2. IMPROVING TRENDS IN ENERGY EFFICIENCY

The efficiency of energy use is an important determinant of how much energy will be required to warm a home to a particular temperature or drive a car a certain distance. In our daily lives we often encounter the idea of energy efficiency when figuring how many miles per gallon our automobiles use while driving in town or on a freeway. Clearly, the more miles a car can travel on a gallon of gasoline, the less gasoline will be needed. The same concept applies to all uses of energy.

Since 1973, when the United States first experienced rapid increases in fossil fuel prices, considerable attention has been paid to improving the efficiency of energy use of various appliances and other products. The most effective path toward energy efficiency has been standards for autos, buildings, appliances, and equipment. Figure 1 shows the remarkable gains in refrigerators. Even while increasing in size, refrigeration has improved efficiency 5% every year since the late 1970s.

Figure 2 shows the results of these improvements. With an estimated 150 million refrigerators and

freezers in the United States, the difference in consumption at 1974 efficiency levels compared to 2001 levels is the equivalent to avoiding 40 GW of power plants. A typical unit at a nuclear power plant is 1 GW, while the typical natural gas-fired combined cycle plant is 300 MW. This is based on saving 1400 kWh per year with the more efficient appliance and assuming power plant operation of 5000 hours per year.

Significant gains in efficiency also have been achieved with improvements in other appliances. Figure 3 illustrates improvements in three difference appliances (gas furnaces, central air conditioning, and refrigerators) compared to their respective consumption in 1972. This figure is taken from

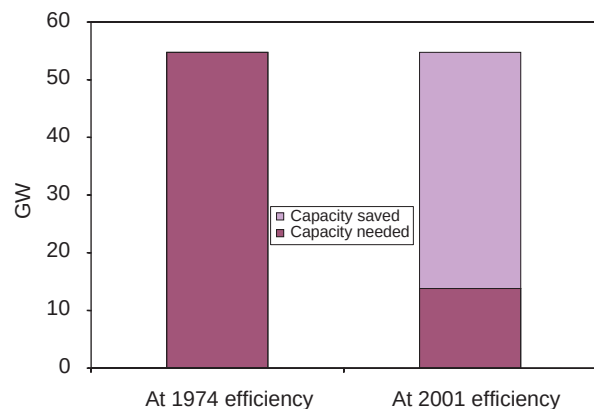


FIGURE 2 Electricity generating capacity for 150 million refrigerators and freezers in the United States.

Steven Nadel in the ECEEE 2003 Summer Study. Compared to consumption in 1972, new gas furnaces use 75%, central air conditioning units use 60%, and refrigerators only 25%. The figure also indicates that early standards were adopted by states. United States' national standards were not adopted until the late 1980s. Although these trends are encouraging, some other uses of energy have shown little improvement in efficiency since the 1980s. For example, average fuel efficiency for cars and truck remains at the same level as in 1985, even though there are no engineering obstacles blocking significant improvements in fleet efficiency. This was after significant improvement in CAFE standards from 18 miles per gallon for model year 1978 cars to 27.5 mpg for 1985 cars.

However, even consideration of increasing CAFE standards remains a controversial topic.

3. THE COMBINED IMPACT OF COMPLIMENTARY EFFICIENCY IMPROVEMENTS

Efficiency improvements in appliance and building design and construction can often be combined to result in even larger energy savings. Figure 4 provides an example of this combined impact. Air conditioning operates to cool a home or a portion of a home based on the temperature setting of the

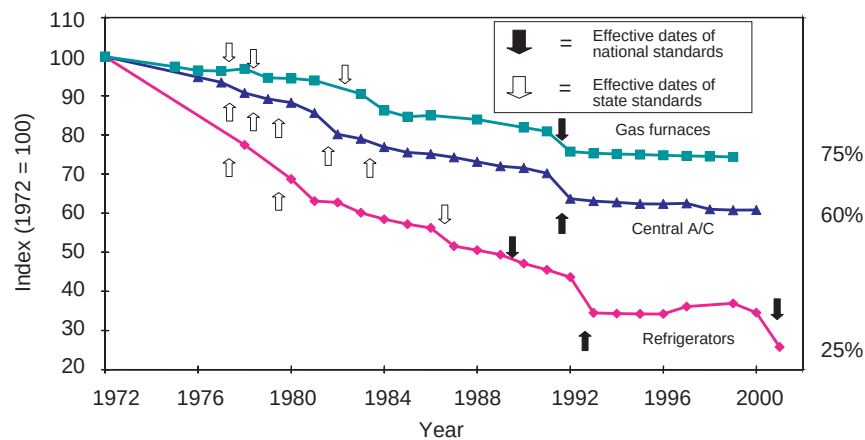


FIGURE 3 Impacts of standards on efficiency of three appliances. From S. Nadel, ACEEE, in ECEEE 2003 Summer Study, www.eceee.org.

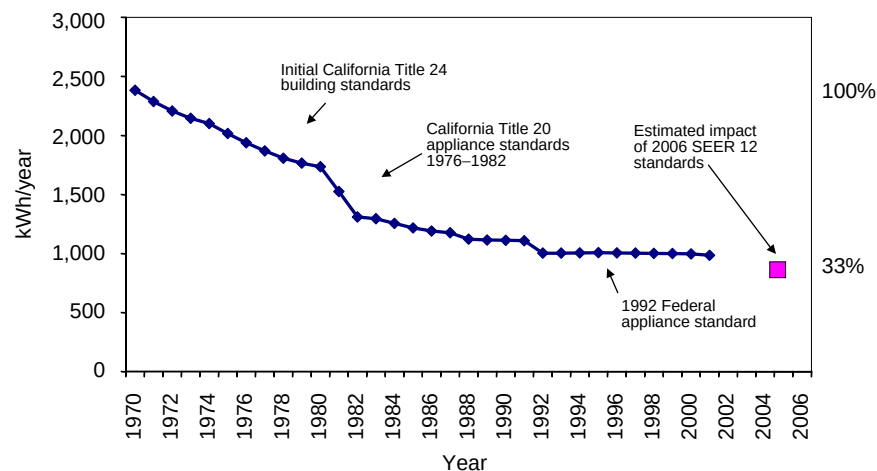


FIGURE 4 Annual usage of air conditioning in new homes in California. Annual drop averages 4% per year. From CEC Demand Analysis Office.

thermostat. The efficiency of the air conditioning equipment (for example, the compressor and air handlers) itself is known at the time of purchase and is required to meet a specific federal standard. However, how much it actually runs is also dependent on the characteristics of the building or home where the equipment is located. Prior to the 1970s many homes, especially in milder climates, had little or no insulation and single pane windows. As building standards were introduced, new homes were built to more stringent standards, resulting in lower requirements for heating and cooling to maintain the same temperature within the structure. As indicated in Fig. 4, the effect of combining both air conditioning standards and building standards has been to reduce energy use in a new California home to 40% of the usage in the early 1970s. With the implementation of newer standards in 2006, use is expected to decline to 33% of the 1970s levels.

4. EFFICIENCY IMPROVEMENTS REDUCE GENERATION REQUIREMENTS

Since the early 1970s, efficiency improvements have lowered energy demands and the need for peaking electrical power. To highlight the magnitude of such savings, we now consider how the U.S. demand for peak electrical power has been reduced, compared to what it might have been in the absence of efficiency improvements. The requirement for peaking power to meet air conditioner loads in the United States is about 250 GW. Standards only apply to the 200 GW of residential central systems and commercial rooftops air conditioning systems. By moving to current efficiency standards, equivalent to an energy efficiency ratio (EER) of 11, the United States has avoided about 135 GW of peak electricity load. This is about twice the peak load of the entire state of California and roughly equal to the total nuclear capacity in the United States. In addition, white roofs could save an additional 20 GW.

Enhancements in refrigerator and freezer efficiency save 200 billion kWh per year compared to consumption if refrigerators had remained at 1974 efficiency levels. That amount of electricity is nearly equal to annual electricity consumption in the state of California, about twice the amount of annual generation from existing renewables (not including hydroelectric) in the United States, and about one-third the quantity of electricity generated at nuclear stations in the United States in a year. Since the value

of electricity at the point of consumption is worth nearly three times its value at the point of production, the value of improvements in refrigerator and freezer efficiency is worth the same as all the nuclear generation in the United States in a typical year.

Clearly, efficiency improvements have reduced the requirements for all forms of energy, even though some sectors, such as automobile transport, lag in this regard. Emissions of greenhouse gases have been proportionally reduced due to these efficiency gains. Yet it is generally accepted that the earth's surface temperature is on the rise, in great part due to human activity involving the combustion of fossil fuels. We now turn to past and estimated future trends in primary energy requirements.

5. HISTORIC PRIMARY ENERGY TRENDS IN THE UNITED STATES

Even with efficiency gains, energy requirements continue to increase. Figure 5 illustrates this trend. Primary energy use (not just electricity) increased by a factor of three over the time period of 62 years. Carbon dioxide emissions increased at nearly the same rate, as the vast majority of primary energy is provided from fossil fuels with only limited contributions from nuclear, hydroelectric, and other renewable energy sources.

Figure 6 shows a steady increase in per capita energy consumption from 1949 until the early 1970s. The flattening in 1974 is the impact of the OPEC oil embargo and high oil prices until the collapse of OPEC in 1985. Comparing Figures 5 and 6, we can see that much of the increase in primary energy is due to an expanding population, especially during the 1990s when per capita consumption was relatively flat.

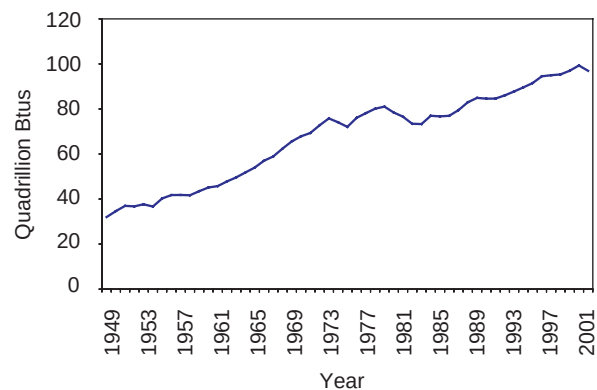


FIGURE 5 United States energy consumption, 1949 to 2001. From Table 1.5, *Annual Energy Review*; data for 2001 is preliminary.

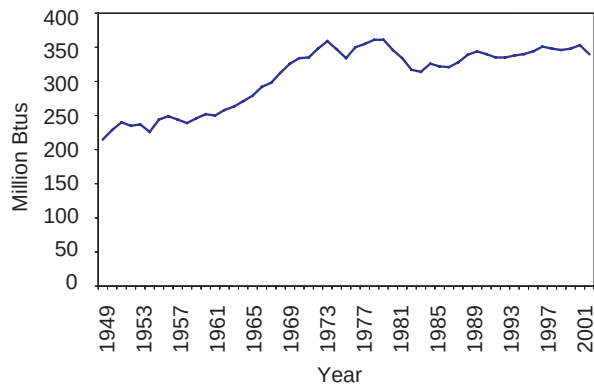


FIGURE 6 Energy consumption per person, 1949 to 2001. From Table 1.5, *Annual Energy Review*; data for 2001 is preliminary.

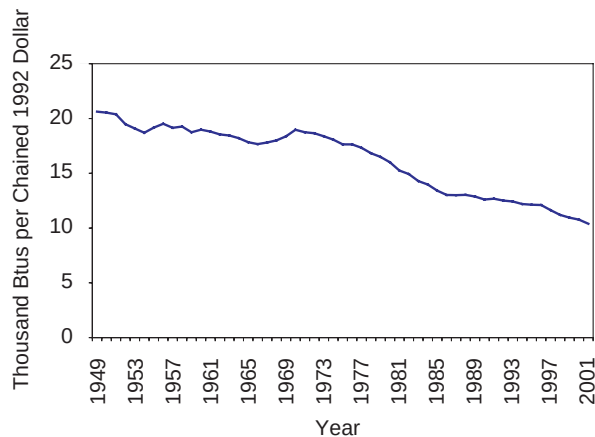


FIGURE 7 Energy consumption per dollar of gross domestic product, 1949 to 2001. From Table 1.5, *Annual Energy Review*; data for 2001 is preliminary.

6. ENERGY INTENSITY TRENDS IN THE UNITED STATES AND OTHER REGIONS

Another measure of interest to economists and energy analysts is how much energy is used per dollar of gross domestic product (GDP). This measure is referred to as energy intensity. The lower the number, the more efficient in terms of energy per dollar of GDP. Figure 7 shows that the United States improved its energy intensity at a rate of about 1.3% per year through this entire period. However, the rate of improvement is by no means uniform, as can be seen in the figure. Beginning in the early 1970s, the slope of the line changes as the rate of improvement in energy intensity increases.

Figure 8 provides similar data on energy intensities by geographic region. Five regions plus the world are included. Changes in energy intensity are not uniform throughout the world. Western Europe and North America have shown continual improvement over the time period, while other regions have remained flat or increased. China has shown dramatic and sustained improvement.

The next four figures illustrate annual rates of change in energy intensity data available from the United States Energy Information Agency (EIA). The EIA uses market exchange rate to compare gross domestic products across different countries. Market exchange rate is based on trading currency in international monetary markets.

Figures are provided for the United States, Western Europe, China, and the entire World. The

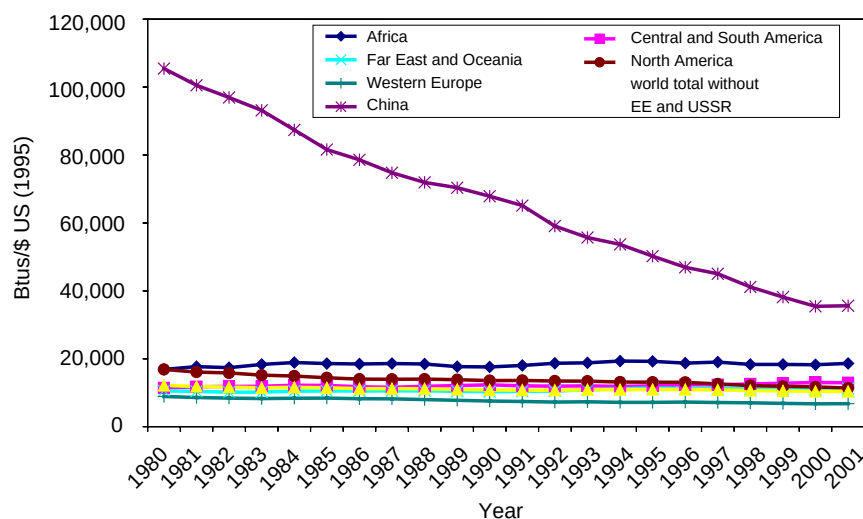


FIGURE 8 Energy intensity by geographic region including China, 1980 to 2001 (Btus/\$US 1995). From Energy Information Agency (EIA) data.

vertical lines divide the time period (1981 to 2001) into three eras. Era 1 covers the latter years of high OPEC oil prices. Era 2 starts soon after the collapse of OPEC and the consequent stagnation of corporate automobile fuel economy (CAFE) standards. Era 3 begins with the significant impact of information technologies on business and the economy. The numbers written above each era reflect the average rate of change in E/GDP for that era. We introduce each of these figures and discuss trends of interest as illustrated in the figure.

Figure 9 tracks year to year changes in energy intensity and shows different rates of improvement in E/GDP in each of the three eras discussed earlier. In the early 1980s, the United States experienced improvements averaging -3.4% per year (improvements are indicated as negative numbers on these figures) and only -0.7% between 1986 and 1996. Improvements from 1997 through 2001 averaged -2.7% . Although many factors are at work, a number of analysts argue that the era improvements are the result of structural changes in the economy. The United States has moved away from energy intensive industry, such as steel production, and moved toward service and information-based firms, especially computer and internet firms. Such changes may portend not only less energy use for processes but improvements in scheduling, shipping, and other aspects of business practice due to the additional information availability.

Figure 10 shows a nearly constant improvement in the range of -1.3% per year and much less

variability than in the United States. Europe already is considerably more efficient than the United States (see Fig. 8). In Fig. 11, we find nearly constant improvements in E/GDP in the range of -5.0% per year. Such improvement is dramatic. According to IEA's World Energy Outlook 2002, China's GDP has grown at about 8% per year since the 1970s to become the world's second largest consumer of primary energy. Finally, in Fig. 12 we find improvements in energy intensity averaging about -1.0% since the 1970s. Many factors contribute to the energy intensity of a county or a region within a country. As indicated by Figs. 8 to 12, the rate of change of energy intensity varies significantly throughout the world.

For energy intensity (E/GDP) to improve—that is, get smaller—GDP needs to increase more rapidly than primary energy, presuming that both are increasing. For example, primary energy use in the United States grew by 2.4% between 1999 and 2000. GDP grew by 3.7% , resulting in an improvement in energy intensity of 1.3% from 1999 to 2000. This article uses the convention of showing this as -1.3% . From the perspective of using energy better, the United States improved. Yet carbon emission also increased since the majority of primary energy comes in the form of fossil fuel—primarily coal, natural gas, and petroleum. If one is concerned about global warming, it seems that only three options are available to limit carbon emissions: (1) reduce the amount of primary energy requirements or limit the rate of growth in primary energy demand; (2) find

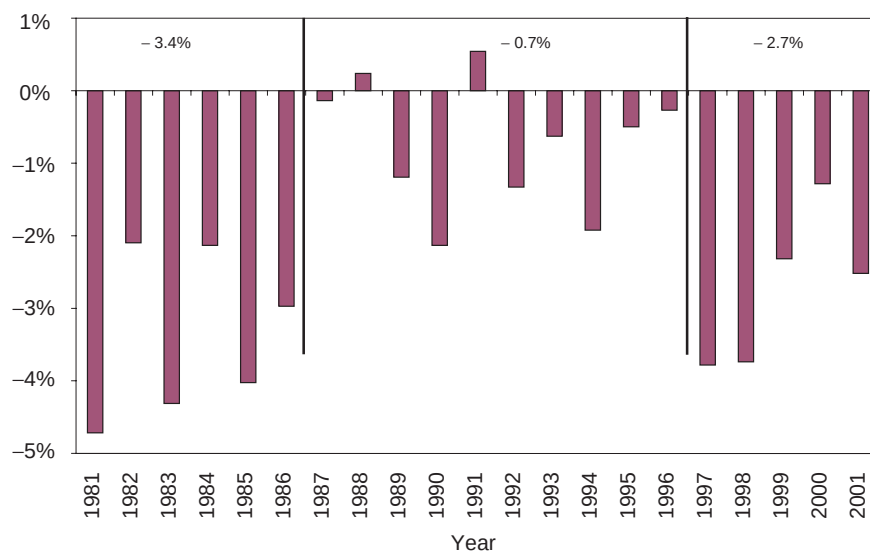


FIGURE 9 Annual rate of change in energy/GDP for the United States.

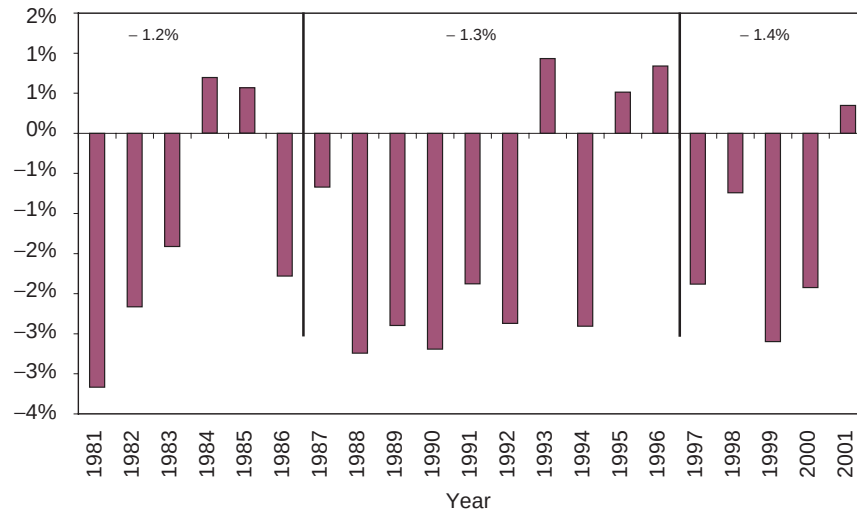


FIGURE 10 Annual rate of change in energy/GDP for Europe.

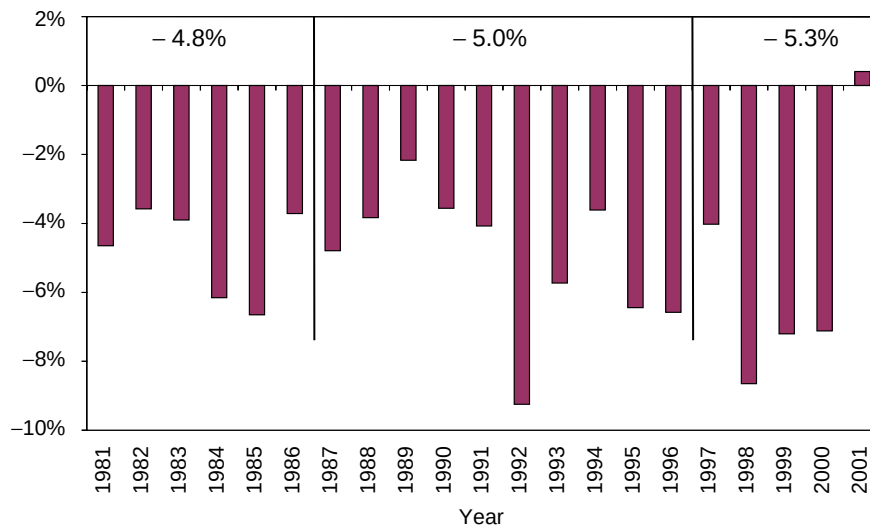


FIGURE 11 Annual rate of change in energy/GDP for China.

primary fuel sources that are not carbon based; and (3) sequester the carbon that is emitted. Not only are these options not mutually exclusive, but we expect all might well be necessary if we are to stem the tide of rising carbon concentrations in the atmosphere and global warming.

Even with improving E/GDP, concerns remain regarding primary energy demands for the world through the coming century and the amount of carbon that may be released into the atmosphere as fossil fuels are burned to meet these needs.

7. FUTURE PRIMARY ENERGY TRENDS FOR THE WORLD

Many different organizations have addressed the issue of how much energy might be needed in the future. Some have tried to forecast this amount; others have developed a host of scenarios to attempt to bound the magnitude of the problem. Estimates of future primary energy requirements vary considerably. Figure 13 illustrates the range of such scenarios. The scenarios all start at current levels on world energy demand, about 400 quads. By 2100, the range of

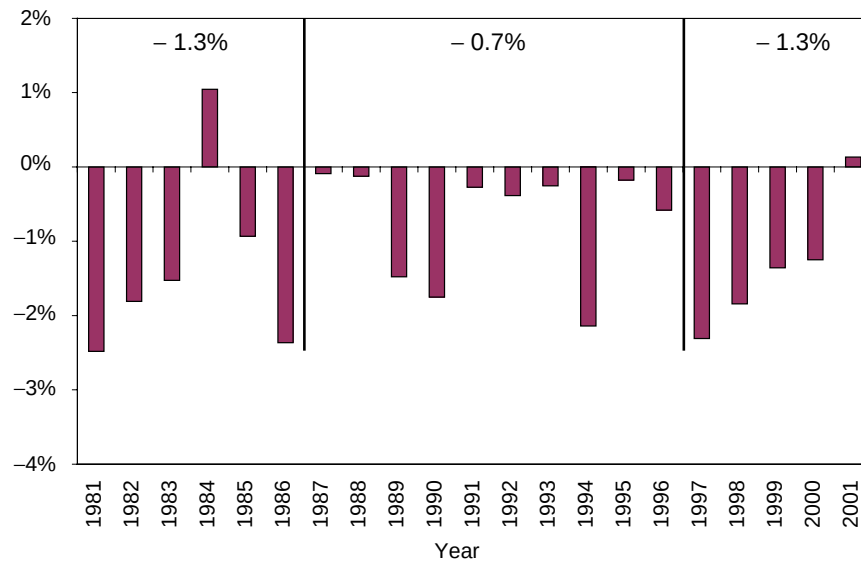


FIGURE 12 Annual rate of change in energy/GDP for the world. Note that Russia not included.

world primary energy demand is from 500 quads all the way to 2600 quads per year. These scenarios are predicated on different estimates of population growth, economic development, and energy intensity.

8. THE EFFECT OF EFFICIENCY IMPROVEMENTS ON PRIMARY ENERGY DEMAND

Efficiency improvements can drastically alter primary energy demand. We now discuss other cases that differ primarily in assumptions regarding the rate of improvement of energy intensity. Using the

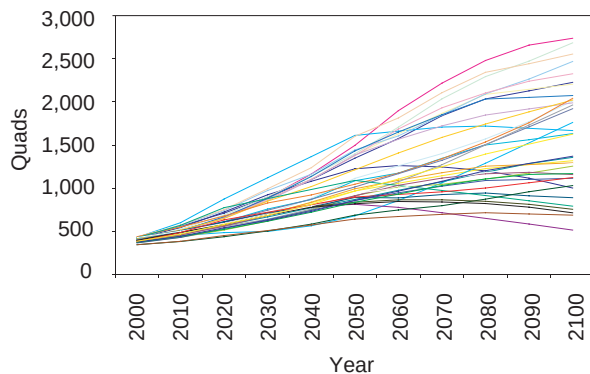


FIGURE 13 Primary energy 2000 to 2100. This figure is taken from data provided in IPCC Special Report on Emissions Scenarios 2001, Appendix VII.

IPCC's IS 92a as the base case, Fig. 14 compares this case with two alternative trajectories of the world's primary energy consumption. The IS 92a case assumes E/GDP improving at -0.8% per year through 2020 and -1.0% per year thereafter through 2100. The second trajectory is taken from EIA's Annual Energy Outlook 2003. In the third case, energy intensity is assumed to improve at -2.0% annually through 2100. In assuming this improvement, total gross domestic product (or gross world product) is held at the same level as in the IS 92a case, and primary energy is set so that E/GDP improves by -2.0% per year.

Since the early 1990s when the IS 92a case was developed, primary energy demand has grown more slowly than anticipated. This explains the differences between the base case assumptions and the EIA International Energy Outlook as shown in Fig. 14. These differences result in a reduction of forecasted primary energy demand in 2025 of about 7.5%. However, if we assume that world energy intensity improves at a rate of -2% per year, world energy demand rises to only 510 quads by 2025 and then stays nearly flat through the end of the century. World demand is 400 quads per year.

The impact of improvements in energy intensity is dramatic. Primary energy demand is cut by more than 60% compared to the IS 92a base case. Put more dramatically, IS 92a requires, by 2100, in units of today's energy infrastructure, growth from 1.0 to 3.5, whereas the 2% per year scenario requires growth only from 1.0 to 1.3, at great savings of

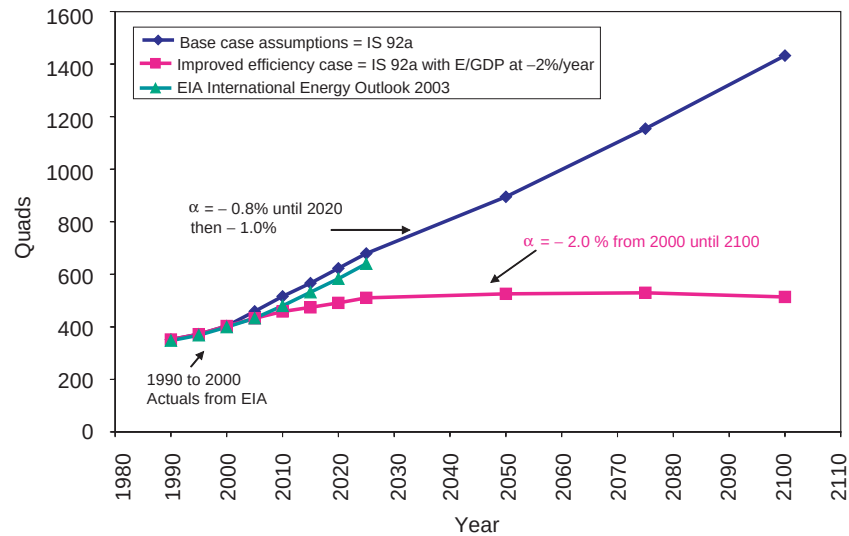


FIGURE 14 Primary energy demand in IPCC base case compared to other cases. α is the annual growth in energy intensity, E/GWP negative α indicates reduced energy intensity.

economic and environmental cost. While all this is occurring, per capita income rises to a worldwide average of nearly 25,000 (in 1995 US\$). This is a level consistent with standards in developed European countries and Japan.

9. PATHWAYS TO IMPROVING ENERGY INTENSITY

We have seen that world energy intensity (E/GWP) is spontaneously dropping 1.3% per year, and that gaining another 1% per year for the coming century will cut demand and supply by another factor of 2.7. This extra 1% per year can be achieved by adopting and enforcing building, appliance, and vehicle fuel economy standards, and by ploughing about 1% of utility revenues back into efficiency programs designed to beat the standards. Examples of such programs are rebates for the purchase of super-efficient (“energy star”) appliances, buildings, cars and trucks, and technical assistance to industries, architects, building owners and operators, and so on.

Standards are usually designed to reduce utility bills fast enough to repay a slightly higher first cost in 5 to 10 years, which is typically half the service life of the product. We call this a simple payback time (SPT) of 5 to 10 years. Efficiency measures to retrofit existing buildings and industry will also

have SPTs that are safely shorter than the service life of the products installed. In contrast, investment in energy supply—power plants, transmission lines, refineries, dams, mines, and wells—typically has an SPT (from sale of energy) of 20 to 40 years, so efficiency is a better investment (although, of course, we need both), and the savings from efficiency, ploughed back into the economy, can stimulate jobs and growth.

Developed countries, which already have low energy intensity, can improve further by tightening and enforcing standards and extending their scope and by expanding efficiency programs. Reductions of CO₂ will then be an attractive by-product.

But later developing countries (LDCs), all of them lacking adequate energy supply, tend to focus on expanding their supply, and indeed rich countries should help them finance new supply. But first, for their own economic development and to reduce CO₂, more developed countries should attach conditions. Before financing supply with a 30-year payback time, developed countries should require that the host country develop standards and programs to capture all measures with a payback time of 3 to 5 years. Also, rich countries can help LDCs with technical assistance and training. These conditions should be adopted worldwide by donor countries, international banks, and host countries and their utilities. This is the cheapest way for the rich countries to delay climate change.

SEE ALSO THE FOLLOWING ARTICLES

Carbon Taxes and Climate Change • Climate Change and Energy, Overview • Climate Change and Public Health: Emerging Infectious Diseases • Climate Change: Impact on the Demand for Energy • Climate Protection and Energy Policy • Consumption, Energy, and the Environment • Economics of Energy Efficiency • Energy Efficiency, Taxonomic Overview • Environmental Change and Energy • Greenhouse

Gas Emissions from Energy Systems, Comparison and Overview • Obstacles to Energy Efficiency

Further Reading

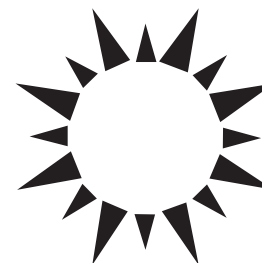
“Climate Change 2001: The Scientific Basis.” (2001). Contributions of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK.

Stewart, R. B., and Jonathan B. W. (2002). “Reconstructing Climate Policy Beyond Kyoto.” AEI Press, Washington, DC.

Energy Efficiency, Taxonomic Overview

AMORY B. LOVINS

Rocky Mountain Institute
Snowmass, Colorado, United States



1. Definition and Importance of Energy Efficiency
2. Benefits of Energy Efficiency
3. Engineering vs Economic Perspectives
4. Diminishing vs Expanding Returns to Investments in Energy Efficiency
5. Market Failures and Business Opportunities
6. Old and New Ways to Accelerate Energy Efficiency

Glossary

conversion efficiency The physical ratio of desired output to total input in an energy conversion process, such as converting fuel to electricity or fuel to heat. Undesired or nonuseful outputs are excluded, making the ratio less than unity for most devices (except in such cases such as heat pumps and heat-powered chillers, where it can exceed unity and is called a coefficient of performance). The definition is thus based partly on what each individual finds useful. Synonymous with the thermodynamic concept of First Law efficiency, but counts only the quantity of energy, not also the quality, and hence differs from Second Law efficiency, which counts both.

customer The ultimate recipient of energy services, regardless of intermediate transactions.

delivered energy Secondary energy provided at the place where it is used to provide the desired energy service (e.g., electricity or fuel entering the end-use device that performs that final conversion). Further losses between that device and the customer may or may not be included. Delivered energy is net of distribution efficiency (1) to the end-use device, but may or may not be net of distribution efficiency (2) between end-use device and customer.

distribution efficiency (1) The fraction of centrally supplied energy shipped out (such as electricity from a power station, petroleum products from a refinery, or natural gas from a treatment plant) that is delivered to the end-use device, net of energy “lost” or “consumed” in its delivery. For electricity, this conversion into

unwanted forms, chiefly low-temperature heat, comprises transmission as well as distribution losses, and is conventionally measured from the generator’s busbar to the customer’s meter. (2) The fraction of energy services produced by the end-use device that reaches the customer (e.g., the fraction of the heat produced by a furnace that provides warmth in a room, net of nonuseful heat escaping from pipes or ducts).

end use (1) The category of desired physical function provided by an energy service, such as heating, cooling, light, mechanical work, electrolysis, or electronic signal processing. (2) The physical quantity of such end use delivered to the customer, whether or not it is useful energy.

end-use device Equipment converting delivered energy into energy service.

end-use efficiency The physical ratio of end use (2) provided to delivered energy converted in the end-use device.

energy conservation An ambiguous term best avoided; considered by some as synonymous with increased energy efficiency but to many others connoting the opposite: privation, curtailment, and discomfort, i.e., getting fewer or lower quality energy services. The degree of confusion between these meanings varies widely by individual, culture, historic period, and language spoken. Some analysts, chiefly outside the United States, embrace energy conservation as an umbrella term for energy efficiency plus changes in personal habits plus changes in system design (such as spatial planning or design for product longevity and materials recovery/reuse).

energy efficiency Broadly, any ratio of function, service, or value provided to the energy converted to provide it. Herein, energy efficiency and its components all use (a) physical rather than economic metrics and (b) engineering, not economic, definitions (this physical convention can, however, become awkward with multiple inputs or outputs). Energy efficiency may or may not count thermodynamic quality of energy (ability to do work); see the distinction between First Law and Second Law efficiency.

energy intensity Energy (primary, delivered, or otherwise defined) “used” per unit of service or value provided.

Intensity can be expressed in economic or in physical terms (e.g., the very crude and aggregated metric of primary energy consumption per dollar of real gross domestic product).

energy service The desired function provided by converting energy in an end-use device (e.g., comfort, mobility, fresh air, visibility, electrochemical reaction, or entertainment). These functions, which together contribute to a material standard of living, are generally measured in physical units, not by energy used or money spent. Because diverse services are incommensurable, they cannot be readily added to express a meaningful “total end-use efficiency” for a person, firm, or society. Economic surrogates for such totals are seldom satisfactory.

extractive efficiency The fraction of a fuel deposit that is recovered and sent out for processing and use, net of energy employed to conduct those extractive processes.

hedonic (functional) efficiency How much human happiness or satisfaction (equivalent to economists’ welfare metrics) is achieved per unit of energy service delivered. Some analysts similarly distinguish the task (such as delivering heat into a house) from the goal it seeks to achieve (such as the human sensation of indoor comfort in winter).

primary energy (1) Fossil fuel extracted and then converted, typically at a central facility, into a secondary form (e.g., crude oil into refined products or coal into electricity) for delivery to end-use devices. (2) Nuclear, hydroelectric, and renewable energy captured for such delivery; if electric, conventionally expressed as the imputed amount of fossil fuel used to produce that much electricity in a typical thermal power station. (3) The quantity of energy described in (1) or (2). Most analysts exclude from primary energy consumption the metabolic energy in human food, but include the nonsolar energy needed to grow, process, deliver, and prepare it.

secondary energy Any processed, refined, or high-quality form of useful energy converted from primary energy, such as electricity, refined petroleum products, dry natural gas, or district heat. Excludes undesired and nonuseful conversion products.

Second Law efficiency The ratio of First Law thermodynamic efficiency to its maximum theoretically possible value; equivalently, the ratio of the least available work that could have done the job to the actual available work used to do the job. For a device that produces useful work or heat (not both), such as a motor, heat pump, or power plant, Second Law efficiency is the amount of output heat or work usefully transferred, divided by the maximum possible heat or work usefully transferable for the same function by any device or system using the same energy input. This maximum is defined by the task, not the device: to maximize how much heat is delivered from fuel into a building, an ideal fuel cell and an ideal heat pump would be used. Second Law efficiency thus measures the

effectiveness of a device in approaching the constraints of the First and Second Laws of thermodynamics. First and Second Law efficiencies are nearly equal for a power plant, but are very different when high-quality energy is converted into a low-energy useful form: a 60%-efficient (First Law) furnace delivering 43°C heat into a house in a 0°C environment has a Second Law efficiency of only 8.2%.

service substitution Providing a desired energy service by a different means (e.g., providing illumination by opening the curtain in the daytime rather than turning on an electric light).

useful energy The portion of an energy service that is actually, not just potentially, desired and used by customers (e.g., lighting an empty room, or overheating an occupied room to the point of discomfort, is seldom useful).

Efficient use of energy is in all countries the most important, economical, prompt, underused, overlooked, and misunderstood way to provide future energy services. It is rapidly becoming even larger, faster, and cheaper as technologies, delivery methods, and integrative design improve. Whole-system design can indeed make very large energy savings cost less than small ones. But capturing energy efficiency’s remarkable potential requires careful terminology, prioritization, attention to engineering details and to market failures, and willingness to accept measured physical realities even if they conflict with economic theories. If well done, such energy efficiency can displace costly and disagreeable energy supplies, enhance security and prosperity, speed global development, and protect Earth’s climate—not at cost but at a profit.

1. DEFINITION AND IMPORTANCE OF ENERGY EFFICIENCY

Energy efficiency is generally the largest, least expensive, most benign, most quickly deployable, least visible, least understood, and most neglected way to provide energy services. The 39% drop in U.S. energy intensity (primary energy consumption per dollar of real gross domestic product) from 1975 to 2000 represented, by 2000, an effective energy “source” 1.7 times as big as U.S. oil consumption, three times net oil imports, five times domestic oil output, six times net oil imports from Organization of Petroleum Exporting Countries (OPEC) members, and 13 times net imports from Persian Gulf countries. It has lately increased by 3% per year, outpacing the

growth of any source of supply (except minor renewables). Yet energy efficiency has gained little attention or respect for its achievements, let alone its far larger untapped potential. Physical scientists, unlike engineers or economists, find that despite energy efficiency's leading role in providing energy services today, its profitable potential has barely begun to be tapped. In contrast, many engineers tend to be limited by adherence to past practice, and most economists are constrained by their assumption that any profitable savings must already have occurred.

The potential of energy efficiency is increasing faster through innovative designs, technologies, policies, and marketing methods than it is being used up through gradual implementation. The uncaptured "efficiency resource" is becoming bigger and cheaper even faster than oil reserves have lately done through stunning advances in exploration and production. The expansion of the "efficiency resource" is also accelerating, as designers realize that whole-system design integration (see Section 4) can often make very large (one or two order-of-magnitude) energy savings cost less than small or no savings, and as energy-saving technologies evolve discontinuously rather than incrementally. Similarly rapid evolution and enormous potential apply to ways to market and deliver energy-saving technologies and designs; research and development can accelerate both.

1.1 Terminology

Efficiency unfortunately means completely different things to the two professions most engaged in achieving it. To engineers, efficiency means a physical output/input ratio. To economists, efficiency means a monetary output/input ratio, though for practical purposes many economists use a monetary output/physical input ratio. Only physical output/input ratios are used here, but the common use of monetary ratios causes vast confusion, especially to policymakers using economic jargon.

Wringing more work from energy via smarter technologies is often, and sometimes deliberately, confused with a pejorative usage of the ambiguous term energy conservation. Energy efficiency means doing more (and often better) with less—the opposite of simply doing less or worse or without. This confusion unfortunately makes the honorable and traditional concept of energy conservation no longer useful in certain societies, notably the United States, and underlies much of the decades-long neglect or suppression of energy efficiency. However, deliber-

ately reducing the amount or quality of energy services remains a legitimate, though completely separate, option for those who prefer it or are forced by emergency to accept it. For example, the 2000–2001 California electricity crisis ended abruptly when customers, exhorted to curtail their use of electricity, cut their peak load per dollar of weather-adjusted real gross domestic product (GDP) by 14% in the first 6 months of 2001. Most of that dramatic reduction, undoing the previous 5–10 years of demand growth, was temporary and behavioral, but by mid-2002, the permanent and technological fraction was heading for dominance. Even absent crises, some people do not consider an ever-growing volume of energy services to be a worthy end in itself, but seek to live more simply—with elegant frugality rather than involuntary penury—and to meet nonmaterial needs by nonmaterial means. Such choices can save even more energy than can technical improvements alone, though they are often considered beyond the scope of energy efficiency.

Several other terminological distinctions are also important. At least five different kinds of energy efficiency can be measured in at least five different stages of energy conversion chains; these are discussed in Section 1.2. Also, technical improvements in energy efficiency can be broadly classified into those applicable only to new buildings and equipment, those installable in existing ones (retrofitted), those addable during minor or routine maintenance (slipstreamed), and those that can be conveniently added when making major renovations or expansions for other reasons (piggybacked).

Efficiency saves energy whenever an energy service is being delivered, whereas "load management" (sometimes called "demand response" to emphasize reliance on customer choice) changes only the time when that energy is used, either by shifting the timing of the service delivery or by, for example, storing heat or coolth so energy consumption and service delivery can occur at different times. In the context chiefly of electricity, demand-side management, a term coined by the Electric Power Research Institute, comprises both of these options, plus others that may even increase the use of electricity. Most efficiency options yield comparable or greater savings in peak loads; both kinds of savings are valuable, and both kinds of value should be counted. They also have important but nonobvious linkages: for example, because most U.S. peak electric loads are met by extremely inefficient simple-cycle gas-fired combustion turbines, saving 5% of peak electric load in 2000 would have saved 9.5% of total natural gas

consumption, enough to reequilibrate high 2003 gas prices back to $\sim \$2/\text{GJ}$ lower historic levels.

Conflating three different things—technological improvements in energy efficiency (such as thermal insulation), behavioral changes (such as resetting thermostats), and the price or policy tools used to induce or reward those changes—causes endless misunderstanding. Also, consider that the theoretical potential for efficiency gains (up to the maximum permitted by the laws of physics) exceeds the technical potential, which exceeds the economic potential based on social value, which exceeds the economic potential based on private internal value, which exceeds the actual uptake not blocked by market failures, which exceeds what happens spontaneously if no effort is made to accelerate efficiency gains deliberately; yet these six quantities are often not clearly distinguished.

Finally, energy statistics are traditionally organized by the economic sector of apparent consumption, not by the physical end uses provided or services sought. End uses were first seriously analyzed in 1976, rarely appear in official statistics even a quarter-century later, and can be difficult to estimate accurately. But end-use analysis can be valuable because matching energy supplies in quality and scale, as well as in quantity, to end-use needs can save a great deal of energy and money. Supplying energy of superfluous quality, not just quantity, for the task is wasteful and expensive. For example, the United States now provides about twice as much electricity as the fraction of end uses that economically justify this special, costly, high-quality form of energy, yet from 1975 to 2000, 45% of the total growth in primary energy consumption came from increased conversion and grid losses in the expanding, very costly, and heavily subsidized electricity system. Much of the electric growth, in turn, provided low-temperature heat, a physically and economically wasteful use of electricity.

Many subtleties of defining and measuring energy efficiency merit but seldom get rigorous treatment, such as the following losses, services, or metrics:

- Distribution losses downstream of end-use devices (an efficient furnace feeding leaky ducts yields costlier delivered comfort).
- Undesired or useless services, such as leaving equipment on all the time (as many factories do) even when it serves no useful purpose.
- Misused services, such as space-conditioning rooms that are open to the outdoors.

- Conflicting services, such as heating and cooling the same space simultaneously (wasteful even if both services are provided efficiently).
- Parasitic loads, as when the inefficiencies of a central cooling system reappear as additional feedback cooling loads that make the whole system less efficient than the sum of its parts.
- Misplaced efficiency, such as applying energy-using equipment, however efficiently, to a task that does not need it—say, cooling with a mechanical chiller when groundwater or ambient conditions can more cheaply do the same thing.
- Incorrect metrics, such as measuring lighting by raw quantity (lux or footcandles) unadjusted for its visual effectiveness, which may actually decrease if greater illuminance is improperly delivered.

To forestall a few other semantic quibbles, physicists (including the author) know energy is not “consumed,” as the economists’ term “consumption” implies, nor “lost,” as engineers refer to unwanted conversions into less useful forms. Energy is only converted from one form to another; yet the normal metaphors are clear, common, and adopted here. Thus an 80%-efficient motor converts its electricity input into 80% torque and 20% heat, noise, vibration, and stray electromagnetic fields; the total equals 100% of the electricity input, or roughly 30% of the fuel input at a classical thermal power station. (Note that this definition of efficiency combines engineering metrics with human preference. The motor’s efficiency may change, with no change in the motor, if changing intention alters which of the outputs are desired and which are unwanted: the definition of “waste” is as much social or contextual as physical. A floodlamp used to keep plates of food warm in a restaurant may be rather effective for that purpose even though it is an inefficient source of visible light).

More productive use of energy is not, strictly speaking, a physical “source” of energy but rather a way to displace physical sources. Yet this distinction is rhetorical, because the displacement or substitution is real and makes supply fully fungible with efficiency. Also, energy/GDP ratios are a very rough, aggregated, and sometimes misleading metric, because they combine changes in technical efficiency, human behavior, and the composition of GDP (a metric that problematically conflates goods and services with “bads” and nuisances, counts only monetized activities, and is an increasingly perverse measure of well being). Yet the two-fifths drop in U.S. energy intensity and the one-half drop in oil

intensity during the period 1975–2001 reflect mainly better technical efficiency. Joseph Romm has also shown that an important compositional shift of U.S. GDP—the information economy emerging in the late 1990s—has significantly decreased energy and probably electrical energy intensity, as bytes substituted for (or increased the capacity utilization of) travel, freight transport, lit and conditioned floorspace, paper, and other energy-intensive goods and services.

The aim here is not to get mired in word games, but to offer a clear overview of what kinds of energy efficiency are available, what they can do, and how best to consider and adopt them.

1.2 Efficiency along Energy Conversion Chains

The technical efficiency of using energy is the product of five efficiencies successively applied along the chain of energy conversions: (1) the conversion efficiency of primary into secondary energy, times (2) the distribution efficiency of delivering that secondary energy from the point of conversion to the point of end use, times (3) the end-use efficiency of converting the delivered secondary energy into such desired energy services as hot showers and cold beer. Some analysts add another term at the upstream end, (4) the extractive efficiency of converting fuel in the ground or power from wind or from sun in the atmosphere, etc. into the primary energy fed into the initial conversion device, and another term at the downstream end, (5) the hedonic efficiency of converting delivered energy services into human welfare. (Delivering junk mail with high technical efficiency is futile if the recipients did not want it.) Counting all five efficiencies permits comparing

ultimate means, the primary energy tapped, with ultimate ends, the happiness or economic welfare created. Focusing only on intermediate means and ends loses sight of what human purposes an energy system is to serve. Most societies pay attention to only three kinds of energy efficiency: extraction (because of its cost, not because the extracted fuels are assigned any intrinsic or depletion value), conversion, and perhaps distribution. End-use and hedonic efficiency are left to customers, are least exploited, and hence hold the biggest potential gains.

They also offer the greatest potential leverage. Because successive efficiencies along the conversion chain all multiply, they are often assumed to be equally important. Yet downstream savings—those nearest the customer—are the most important. Figure 1 shows schematically the successive energy conversions and losses that require about 10 units of fuel to be fed into a conventional thermal power station in order to deliver one unit of flow in a pipe. But conversely, every unit of flow (or friction) saved in the pipe will save approximately 10 units of fuel, cost, and pollution at the power station. It will also make the pump's motor (for example) nearly two and a half units smaller, hence cheaper. To save the most primary energy and the most capital cost, therefore, efficiency efforts should start all the way downstream (see Section 4.2), by asking: How little flow can actually deliver the desired service? How small can the piping friction become? How small, well matched to the flow regime, and efficient can the pump be made? Its coupling? Its motor? Its controls and electrical supplies?

Analyses of energy use should start with the desired services or changes in well being, then work back upstream to primary supplies. This maximizes the extra value of downstream efficiency gains and

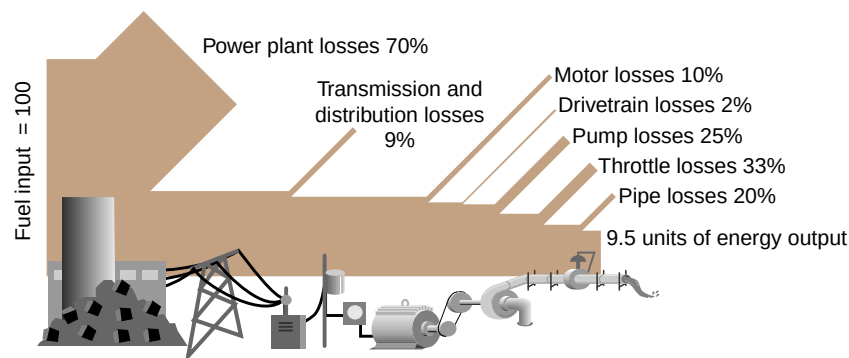


FIGURE 1 To deliver one unit of flow in the pipe requires about 10 units of fuel at the power plant, thus those 10-fold compounding losses can be turned around backward, yielding 10-fold compounding savings of fuel for each unit of reduced friction or flow in the pipe. From the E SOURCE “Drivepower Technology Atlas,” courtesy of Platts Research & Consulting.

the capital-cost savings from smaller, simpler, cheaper upstream equipment. Yet it is rarely done. Similarly, most energy policy analysts analyze how much energy could be supplied before asking how much is optimally needed and at what quality and scale it could be optimally provided. This wrong direction (upstream to downstream) and supply orientation lie at the root of many if not most energy policy problems.

Even modest improvements in efficiency at each step of the conversion chain can multiply to large collective values. For example, suppose that during the years 2000–2050, world population and economic growth increased economic activity by six- to eightfold, in line with conventional projections. But meanwhile, the carbon intensity of primary fuel, following a two-century trend, is likely to fall by at least two- to fourfold as coal gives way to gas, renewables, and carbon offsets or sequestration. Conversion efficiency is likely to increase by at least 1.5-fold with modernized, better-run, combined-cycle, and cogenerating power stations. Distribution efficiency should improve modestly. End-use efficiency could improve by four- to sixfold if the intensity reductions sustained by many industrial countries, when they were paying attention, were sustained for 50 years (e.g., the United States decreased its primary energy/GDP intensity at an average rate of 3.4%/year from 1979 to 1986 and 3.0%/year from 1996 to 2001). And the least understood term, hedonic efficiency, might remain constant or might perhaps double as better business models and customer choice systematically improve the quality of services delivered and their match to what customers want. On these plausible assumptions, global carbon emissions from burning fossil fuel could decrease by 1.5- to 12-fold despite the assumed six- to eightfold grosser world product. The most important assumption is sustained success with end-use efficiency, but the decarbonization and conversion-efficiency terms appear able to take up some slack if needed.

1.3 Service Redefinition

Some major opportunities to save energy redefine the service being provided. This is often a cultural variable. A Japanese person, asked why the house is not heated in winter, might reply, “Why should I? Is the house cold?” In Japanese culture, the traditional goal is to keep the person comfortable, not to heat empty space. Thus a modern Japanese room air conditioner may contain a sensor array and move-

able fans that detect and blow air toward people’s locations in the room, rather than wastefully cooling the entire space. Western office workers, too, can save energy (and can often see better, feel less tired, and improve esthetics) by properly adjusting venetian blinds, bouncing glare-free daylight up onto the ceiling, and turning off the lights. As Jørgen Nørgård remarks, “energy-efficient lamps save the most energy when they are turned off”; yet many people automatically turn on every light when entering a room. This example also illustrates that energy efficiency may be hard to distinguish from energy supply that comes from natural energy flows. All houses are already ~98% solar-heated, because if there were no Sun (which provides 99.8% of Earth’s heat), the temperature of Earth’s surface would average approximately -272.6°C rather than $+15^{\circ}\text{C}$. Thus, strictly speaking, engineered heating systems provide only the last 1–2% of the total heating required.

Service redefinition becomes complex in personal transport. Its efficiency is not just about vehicular fuel economy, people per car, or public transport alternatives. Rather, the underlying service should often be defined as access, not mobility. Typically, the best way to gain access to a place is to be there already; this is the realm of land-use management—no novelty in the United States, where spatial planning is officially shunned, yet zoning laws mandate dispersion of location and function, real-estate practices segregate housing by income, and other market distortions maximize unneeded and often unwanted travel. Another way to gain access is virtually, moving just the electrons while leaving the heavy nuclei behind, via telecommunications, soon including realistic “virtual presence.” This is sometimes an effective alternative to physically moving human bodies. And if such movement is really necessary, then it merits real competition, at honest prices, between all modes—personal or collective, motorized or human-powered, conventional or innovative. Creative policy tools can enhance that choice in ways that enhance real-estate value, saved time, quality of life, and public amenity and security. Efficient cars can be an important part of efficient personal mobility, but also reducing the need to drive can save even more energy and yield greater total benefit.

1.4 Historic Summaries of Potential

People have been saving energy for centuries, even millennia; this is the essence of engineering. Most savings were initially in conversion and end use:

preindustrial households often used more primary energy than modern ones do, because fuelwood-to-charcoal conversion, inefficient open fires, and crude stoves burned much fuel to deliver sparse cooking and warmth. Lighting, materials processing, and transport end uses were also very inefficient. Billions of human beings still suffer such conditions today. The primary energy/GDP intensities in developing countries average about three times those in industrialized countries. But even the most energy-efficient societies still have enormous, and steadily expanding, room for further efficiency gains. Less than one-fourth of the energy delivered to a typical European cookstove ends up in food, less than 1% of the fuel delivered to a standard car actually moves the driver, U.S. power plants discard waste heat equivalent to 1.2 times Japan's total energy use, and even Japan's economy does not approach one-tenth the efficiency that the laws of physics permit.

Detailed and exhaustively documented engineering analyses of the scope for improving energy efficiency, especially in end-use devices, have been published for many industrial and some developing countries. By the early 1980s, those analyses had compellingly shown that most of the energy currently used was being wasted—i.e., that the same or better services could be provided using severalfold less primary energy by fully installing, wherever practical and profitable, the most efficient conversion and end-use technologies then available. Such impressive efficiency gains cost considerably less than the long-run, and often even the short-run, marginal private internal cost of supplying more energy. Most policymakers ignore both these analyses, well known to specialists, and less well-known findings show even bigger and cheaper savings from whole-system design integration (see Section 4).

Many published engineering analyses show a smaller saving potential because of major conservatism, often deliberate (because the real figures seem too good to be true), or because they assume only partial adoption over a short period rather than examining the ultimate potential for complete practical adoption. For example, the American Council for an Energy-Efficient Economy estimates that just reasonable adoption of the top five conventional U.S. opportunities—industrial improvements, 40-mile per gallon (U.S. gallons; = 4.88 liters/100 km) light-vehicle standards, cogeneration, better building codes, and a 30% better central-air-conditioning standard—could save 530 million T/year of oil equivalent—respectively equivalent to the total 2000 primary energy use of

Australia, Mexico, Spain, Austria, and Ireland. But the full long-term efficiency potential is far larger, and much of it resides in innumerable small terms. Saving energy is like eating an Atlantic lobster: there are big, obvious chunks of meat in the tail and the front claws, but a similar total quantity of tasty morsels is hidden in crevices and requires some skill and persistence to extract.

The whole-lobster potential is best, though still not fully, seen in bottom-up technological analyses comparing the quantity of potential energy savings with their marginal cost. That cost is typically calculated using the Lawrence Berkeley National Laboratory methodology, which divides the marginal cost of buying, installing, and maintaining the more efficient device by its discounted stream of lifetime energy savings. The levelized cost in dollars of saving, say, 1 kWh, then equals $Ci/S[1-(1+i)^{-n}]$, where C is installed capital cost (dollars), i is annual real discount rate (assumed here to be 0.05), S is energy saved by the device (kilowatt-hours/year), and n is operating life (years). Thus a \$10 device that saved 100 kWh/year and lasted 20 years would have a levelized “cost of saved energy” (CSE) of 0.8¢/kWh. Against a 5¢/kWh electricity price, a 20-year device with a 1-year simple payback would have $CSE = 0.4¢/kWh$. It is conventional for engineering-oriented analysts to represent efficiency “resources” as a supply curve, rather than as shifts along a demand curve (the convention among economists). CSE is methodologically equivalent to the cost of supplied energy (e.g., from a power station and grid); the price of the energy saved is not part of the calculation. Whether the saving is cost-effective depends on comparing the cost of achieving it with the avoided cost of the energy saved. (As noted in Section 2, this conventional engineering-economic approach actually understates the benefits of energy efficiency.)

On this basis, the author's analyses in the late 1980s found, from measured cost and performance data for more than 1000 electricity-saving end-use technologies, that their full practical retrofit could save about three-fourths of U.S. electricity at an average CSE $\sim 0.6¢/kWh$ (1986 dollars)—roughly consistent with a 1990 Electric Power Research Institute analysis in which the differences were mainly methodological rather than substantive. Similarly, the author's analysis for Royal Dutch/Shell Group found that full use of the best 1987–1988 oil-saving end-use technologies, assuming turnover of vehicle stocks, could save about 80% of U.S. oil use at an average levelized CSE below \$2.5/barrel (1986 dollars). Both analyses have proven systematically

conservative: today's potential is even larger and cheaper. (The analyses explicitly excluded financing and transaction costs, but those would only slightly affect the results. There is also a huge literature accurately predicting and rigorously measuring the empirical size, speed, and cost of efficiency improvements delivered by actual utility and government programs.) Such findings are broadly consistent with equally or more detailed ones by European analysts: for example, late-1980s technologies could save three-fourths of Danish buildings' electricity or half of all Swedish electricity at \$0.016/kWh (1986 dollars), or four-fifths of German home electricity (including minor fuel switching) with a $\sim 40\%$ /year aftertax return on investment. Such findings with ever greater sophistication have been published worldwide since 1979, but have been rejected by nontechnological economic theorists, who argue that if such cost-effective opportunities existed, they would already have been captured in the marketplace, even in planned economies with no marketplace or mixed economies with a distorted one. This mental model—"don't bother to bend over and pick up that banknote lying on the ground, because if it were real, someone would have picked it up already"—often dominates government policy. It seems ever less defensible as more is learned about the reality of pervasive market failures (see Section 5) and the astonishing size and cheapness of the energy savings empirically achieved by diverse enterprises (discussed in Section 3). But by now, the debate is theological—about whether existing markets are essentially perfect, as most economic modelers assume for comfort and convenience, or whether market failures are at least as important as market function and lie at the heart of business and policy opportunity. To technologists and physical scientists, this seems a testable empirical question.

1.5 Discontinuous Technological Progress

This engineering/economics divergence about the potential to save energy also reflects a tacit assumption that technological evolution is smooth and incremental, as mathematical modelers prefer. In fact, although much progress is as incremental as technology diffusion, discontinuous technological leaps, more like "punctuated equilibrium" in evolutionary biology, can propel innovation and speed its adoption, as with $5\times$ -efficiency light vehicles (see Section 4.1).

Technological discontinuities can even burst the conventional boundaries of possibility by redefining

the design space. Generations of engineers learned that big supercritical-steam power plants were as efficient as possible ($\sim 40\%$ from fuel in to electricity out). But through sloppy learning or teaching, these engineers overlooked the possibility of stacking two Carnot cycles atop each other. Such combined-cycle (gas-then-steam) turbines, based on mass-produced jet engines, can exceed 60% efficiency and are cheaper and faster to build, so in the 1990s, they quickly displaced the big steam plants. Fuel cells, the next innovation, avoid Carnot limits altogether by being an electrochemical device, not a heat engine. Combining both may soon achieve 80–90% fuel-to-electric efficiency. Even inefficient distributed generators can already exceed 90% system efficiency by artfully using recaptured heat.

As another example, most authors today state that the theoretical efficiency limit for converting sunlight into electricity using single-layer photovoltaic (PV) cells is 31% ($\sim 50\%$ using multicolor stacked layers; the best practical values so far are around 25 and 30%). This is because semiconductor bandgaps have been believed too big to capture any but the high-energy wavelengths of sunlight. But those standard data are wrong. A Russian-based team suspected in 2001, and Lawrence Berkeley National Laboratory proved in 2002, that indium nitride's bandgap is only 0.7 eV, matching near-infrared ($1.77\mu\text{m}$) light and hence able to harvest almost the whole solar spectrum. This may raise the theoretical limit to 50% for two-layer and to $\sim 70\%$ for many-layer thin-film PVs.

Caution is likewise vital when interpreting Second Law efficiency. In the macroscopic world, the laws of thermodynamics are normally considered ineluctable, but the definition of the desired change of state can be finessed. Ernie Robertson notes that when turning limestone into a structural material, one is not confined to just the conventional possibilities of cutting it into blocks or calcining it at $\sim 1250^\circ\text{C}$ into Portland cement. It is possible instead grind it up and feed it to chickens, in which ambient-temperature "technology" turns it into eggshell stronger than Portland cement. Were we as smart as chickens, we would have mastered this life-friendly technology. Extraordinary new opportunities to harness 3.8 billion years of biological design experience, as described by Janine Benyus in *Biomimicry*, can often make heat-beat-and-treat industrial processes unnecessary. So, in principle, can the emerging techniques of nanotechnology using molecular-scale self-assembly, as pioneered by Eric Drexler.

More conventional innovations can also bypass energy-intensive industrial processes. Making artifacts that last longer, use materials more frugally, and are designed and deployed to be repaired, reused, remanufactured, and recycled can save much or most of the energy traditionally needed to produce and assemble their materials (and can increase welfare while reducing GDP, which swells when ephemeral goods are quickly discarded and replaced). Microfluidics can even reduce a large chemical plant to the size of a watermelon: millimeter-scale flow in channels etched into silicon wafers can control time, temperature, pressure, stoichiometry, and catalysis so exactly that a very narrow product spectrum is produced, without the side-reactions that normally require most of the chemical plant to separate undesired from desired products. Such “end-run” solutions (rather like the previous example of substituting sensible land-use for better vehicles, or better still, combining both) can greatly expand the range of possibilities beyond simply improving the narrowly defined efficiency of industrial equipment, processes, and controls. By combining many such options, it is now realistic to contemplate a long-run advanced industrial society that provides unprecedented levels of material prosperity with far less energy, cost, and impact than today’s best practice. The discussion in Section 4.1, drawn from Paul Hawken *et al.*’s synthesis in *Natural Capitalism* and Ernst von Weizsäcker *et al.*’s earlier *Factor Four*, further illustrates recent breakthroughs in integrative design that can make very large energy savings cost less than small ones; and similarly important discontinuities in policy innovation are summarized in Section 6.

In light of all these possibilities, why does energy efficiency, in most countries and at most times, command so little attention and such lackadaisical pursuit? Several explanations come to mind. Saved energy is invisible. Energy-saving technologies may look and outwardly act exactly like inefficient ones, so they are invisible too. They are also highly dispersed, unlike central supply technologies that are among the most impressive human creations, inspiring pride and attracting ribbon-cutters and rent- and bribe-seekers. Many users believe energy efficiency is binary—you either have it or lack it—and that they already did it in the 1970s. Energy efficiency has relatively weak and scattered constituencies, and major energy efficiency opportunities are disdained or disbelieved by policymakers indoctrinated in a theoretical economic paradigm that claims they cannot exist (see Section 3).

2. BENEFITS OF ENERGY EFFICIENCY

Energy efficiency avoids the direct economic costs and the direct environmental, security, and other costs of the energy supply and delivery that it displaces. Yet the literature often overlooks several key side-benefits (economists call them “joint products”) of saving energy.

2.1 Indirect Benefits from Qualitatively Superior Services

Improved energy efficiency, especially end-use efficiency, often delivers better services. Efficient houses are more comfortable; efficient lighting systems can look better and help you see better; efficient motors can be more quiet, reliable, and controllable; efficient refrigerators can keep food fresher for longer; efficient cleanrooms can improve the yield, flexibility, throughput, and setup time of microchip fabrication plants; efficient fume hoods can improve safety; efficient supermarkets can improve food safety and merchandising; retail sales pressure can rise 40% in well-daylit stores; and students’ test scores suggest 20–26% faster learning in well-daylit schools. Such side-benefits can be one or even two more orders of magnitude more valuable than the energy directly saved. For example, careful measurements show that in efficient buildings, where workers can see what they are doing, hear themselves think, breathe cleaner air, and feel more comfortable, labor productivity typically rises by about 6–16%. Because office workers in industrialized countries cost about 100 times more than office energy, a 1% increase in labor productivity has the same bottom-line effect as eliminating the energy bill, and the actual gain in labor productivity is about 6–16 times bigger than that. Practitioners can market these attributes without ever mentioning lower energy bills.

2.2 Leverage in Global Fuel Markets

Much has been written about the increasing pricing power of major oil-exporting countries, especially Saudi Arabia, with its important swing production capacity. Yet the market power of the United States—the Saudi Arabia of energy waste—is even greater on the demand side. The United States can raise its oil productivity more and faster than any oil exporter can adjust to reducing its oil output. This was illustrated

from 1977 to 1985, when U.S. GDP rose 27% while total U.S. oil imports fell by 42%, or 3.74 million barrels (bbl)/day. This was largely due to a 52% gain in oil productivity, causing U.S. oil consumption to fall by 17% and oil imports from the Persian Gulf to fall by 91%. That took away an eighth of OPEC's market. The entire world oil market shrank by a tenth; OPEC's share was slashed from 52 to 30%, and its output fell by 48%. The United States accounted for one-fourth of that reduction. More efficient cars, each driving 1% fewer miles on 20% less gasoline—a 7-mile per (U.S.) gallon gain in 6 years for new American-made cars—were the most important single cause; 96% of those savings came from smarter design, with only 4% deriving from smaller size.

Those 8 years around the second oil shock (1979) demonstrated an effective new source of energy security and a potent weapon against high oil prices and supply manipulations. The United States showed that a major consuming nation could respond effectively to supply disruptions by focusing on the demand side and boosting oil productivity at will. It could thereby exercise more market power than suppliers, beat down their prices (breaking OPEC's pricing power for a decade), and enhance the relative importance of less vulnerable, more diversified sources of supply. Had the United States simply continued its 1979–1985 rate of improvement of oil productivity, it would not have needed a drop of oil from the Persian Gulf after 1985. That is not what happened, but it could be if the United States chose to repeat and expand its previous oil savings.

2.3 Buying Time

Energy efficiency buys time. Time is the most precious asset in energy policy, because it permits the fullest and most graceful development and deployment of still better techniques for energy efficiency and supply. This pushes supply curves downward (larger quantities at lower prices), postpones economic depletion, and buys even more time. The more time is available, the more information will emerge to support wiser and more robust choices, and the more fruitfully new technologies and policy options can meld and breed new ones. Conversely, hasty choices driven by supply exigencies almost always turn out badly, waste resources, and foreclose important options. Of course, once bought, time should be used wisely. Instead, the decade of respite bought by the U.S. efficiency spurt of 1979–1985 was almost entirely wasted as attention waned, efficiency and alternative-supply efforts stalled,

research and development teams were disbanded, and underlying political problems festered. From the perspective of 2004, that decade of stagnation looked like a blunder of historic proportions.

2.4 Integrating Efficiency with Supply

To the first order, energy efficiency makes supply cheaper. But second-order effects reinforce this first-order benefit, most obviously when efficiency is combined with onsite renewable supplies, making them nonlinearly smaller, simpler, and cheaper. Consider the following examples:

- A hot-water-saving house can achieve a very high solar-water-heat fraction (e.g., 99% in the author's home high in the Rocky Mountains) with only a small collector, so it needs little or no backup, partly because collector efficiency increases as stratified-tank storage temperature decreases.
- An electricity-saving house needs only a few square meters of PVs and a simple balance-of-system setup (storage, inverter, etc.). This can cost less than connecting to the grid a few meters away.
- A passive-solar, daylit building needs little electricity, and can pay for even costly forms of on-site generation (such as PVs) using money saved by eliminating or downsizing mechanical systems.
- Such mutually reinforcing options can be bundled: e.g., 1.18 peak MW of photovoltaics retrofitted onto the Santa Rita Jail in Alameda County, California, was combined with efficiency and load management, so at peak periods, when the power was most valuable, less was used by the jail and more was sold back to the grid. This bundling yielded a internal rate of return of over 10% including state subsidies, and a customer present-valued benefit/cost ratio of 1.7 without or 3.8 with those subsidies.

2.5 Gaps in Engineering Economics

Both engineers and economists conventionally calculate the cost of supplying or saving energy using a rough-and-ready tool kit called “engineering economics.” Its methods are easy to use but flawed, ignoring such basic tenets of financial economics as risk-adjusted discount rates. Indeed, engineering economics omits 207 economic and engineering considerations that together increase the value of decentralized electrical resources by typically an order of magnitude. Many of these “distributed benefits,” compiled in the author's *Small Is Profitable*, apply as much to decentralized efficiency as to

generation. Most of the literature on the cost of energy alternatives is based solely on accounting costs and engineering economics that greatly understate efficiency's value. Properly counting its benefits will yield far sounder investments.

Efficient end use is also the most effective way to make energy supply systems more resilient against mishap or malice, because it increases the duration of buffer stocks, buying time to mend damage or arrange new supplies, and it increases the share of service that curtailed or improvised supplies can deliver. Efficiency's high "bounce per buck" makes it the cornerstone of any energy system designed for secure service provision in a dangerous world.

3. ENGINEERING VS. ECONOMIC PERSPECTIVES

Engineering practitioners and economic theorists view energy efficiency through profoundly different lenses, yet both disciplines are hard pressed to explain the following phenomena:

1. During the period 1996–2001, U.S. aggregate energy intensity fell at a near-record pace despite record low and falling energy prices. (It fell faster only once in modern history, during the record high and rising energy prices of 1979–1985.) Apparently, something other than price was getting Americans' attention.

2. During the period 1990–1996, when a kilowatt-hour of electricity cost only half as much in Seattle as in Chicago, people in Seattle, on average, reduced their peak electric load 12 times as fast, and their use of electricity about 3640 times as fast, as did people in Chicago—perhaps because Seattle City Light encouraged savings while Commonwealth Edison Co. discouraged them.

3. In the 1990s, DuPont found that its European chemical plants were no more energy efficient than its corresponding U.S. plants, despite long having paid twice the energy price—perhaps because all the plants were designed by the same people in the same ways with the same equipment; there is little room for behavioral change in a chemical plant.

4. In Dow Chemical Company's Louisiana Division during the period 1981–1993, nearly 1000 projects to save energy and reduce waste added \$110 million/year to the bottom line and yielded returns on investment averaging over 200%/year, yet in the latter years, both the returns and the savings were trending upward as the engineers

discovered new tricks faster than they used up the old ones. (Economic theory would deny the possibility of so much "low-hanging fruit" that has fallen down and is mashing up around the ankles: such enormous returns, if real, would long ago have been captured. This belief was the main obstacle to engineers' seeking such savings, then persisting after their discovery.)

5. Only about 25–35% of apartment dwellers, when told that their air conditioner and electricity are free, behave as economists would predict—turning on the air conditioner when they feel hot and setting the thermostat at a temperature at which they feel comfortable. The rest of the apartment dwellers show no correlation between air-conditioning usage and comfort; instead, their cooling behavior is determined by at least six other variables: household schedules, folk theories about air conditioners (such as that the thermostat is a valve that makes the cold come out faster), general strategies for dealing with machines, complex belief systems about health and physiology, noise aversion, and wanting white noise to mask outside sounds that might wake the baby. Energy anthropology reveals that both the economic and the engineering models of air-conditioning behavior are not just incomplete but seriously misleading.

6. The United States has misallocated \$1 trillion of investments to about 200 million refrigerative tons of air conditioners, and 200 peak GW (two-fifths of total peak load) of power supply to run them, that would not have been bought if the buildings had been optimally designed to produce best comfort at least cost. This seems explicable by the perfectly perverse incentives seen by each of the 20-odd actors in the commercial real-estate value chain, each systematically rewarded for inefficiency and penalized for efficiency.

7. Not just individuals but also most firms, even large and sophisticated ones, routinely fail to make essentially riskless efficiency investments yielding many times their normal business returns: most require energy efficiency investments to yield roughly six times their marginal cost of capital, which typically applies to far riskier investments.

Many economists would posit some unknown error or omission in these descriptions, not in their theories. Indeed, energy engineers and energy economists seem not to agree about what is a hypothesis and what is a fact. Engineers take their facts from tools of physical observation. Three decades of conversations with noted energy economists suggest

to the author that most think facts come only from observed flows of dollars, interpreted through indisputable theoretical constructs, and consider any contrary physical observations aberrant. This divergence makes most energy economists suppose that buying energy efficiency faster than the “spontaneous” rate of observed intensity reduction (for 1997–2001, 2.7%/year in the United States, 1.4%/year in the European Union, 1.3%/year worldwide, and 5.3%/year in China) would require considerably higher energy prices, because if greater savings were profitable at today’s prices, they would already have been bought; thus the engineers’ bottom-up analyses of potential energy savings must be unrealistically high. Economists’ estimates of potential savings at current prices are “top-down” and very small, based on historic price elasticities that confine potential interventions to changing prices and savings to modest size and diminishing returns (otherwise the economists’ simulation models would inconveniently explode). Engineers retort that high energy prices are not necessary for very large energy savings (because they are so lucrative even at present prices) but are not sufficient either (because higher prices do little without enlarged ability to respond to them).

Of course, engineering-based practitioners agree that human behavior is influenced by price, as well as by convenience, familiarity, fashion, transparency, competing claims on attention, and many other marketing and social-science factors—missing from any purely technological perspective but central to day-to-day fieldwork. The main difference is that they think these obstacles are “market failures” and dominate behavior in buying energy efficiency. Most economists deny this, and say the relatively slow adoption of efficiency must be due to gross errors in the engineers’ claims of how large, cheap, and available its potential really is. This theological deadlock underlies the debate about climate protection. Robert Repetto and Duncan Austin showed in 1997 that all mainstream climate-economics models’ outputs are hard-wired to the input assumptions, and that realistic inputs, conforming to the actual content of the Kyoto Protocol and its rules, show that climate protection increases GDP. Florentin Krause has shown that the main official U.S. government analyses, taken together, concur. Yet the official U.S. position at the end of 2003 was still that climate protection, even if desirable, cannot be mandated because it is too costly.

In fact, climate protection is not costly but profitable; its critics may have the amount about

right, but they got the sign wrong. The clearest proof is in the behavior and achievements of the smart companies that are behaving as if the United States had ratified the Kyoto Protocol, because energy efficiency is cheaper than the energy it saves. For example, large firms such as DuPont, IBM, and STMicroelectronics (ST; one of the world’s largest chipmakers) have lately been raising their energy productivity by 6%/year with simple paybacks of a few years. DuPont expects by 2010 to cut its greenhouse gas emissions by 65% below the 1990 level; ST expects to achieve zero emissions (despite making 40 times more chips). British Petroleum announced that its 10% reduction by 2010 had been achieved 8 years early at zero net cost; actually, the 10-year net-present-valued saving was \$650 million. Other examples abound; the Web sites www.cool-companies.org and www.pewclimate.org contain examples of the achievements of actively engaged businesses. The companies involved, many of them well known in the business world, are hardly naïve or deluded. Anyone ignoring this market reality is mistaking the econometric rearview mirror for a windshield. Econometrics measures how human populations behaved under past conditions that no longer exist and that it is often a goal of energy policy to change. Where price is the only important explanatory variable, econometrics can be a useful tool for extrapolating history into how price may influence near-term, small, incremental changes in behavior. But limiting our horizons to this cramped view of technical possibilities and human complexity rules out innovations in policies, institutions, preferences, and technologies—treating the future like fate, not choice, and thus making it so.

4. DIMINISHING VS. EXPANDING RETURNS TO INVESTMENTS IN ENERGY EFFICIENCY

Among the most basic, yet simply resolved, economic/engineering disagreements is whether investing in end-use efficiency yields expanding or diminishing returns. Economic theory says diminishing: the more efficiency we buy, the more steeply the marginal cost of the next increment of savings rises, until it becomes too expensive (Fig. 2). But engineering practice often says expanding: big savings can cost less than small or no savings (Fig. 3) if the engineering is done unconventionally but properly.

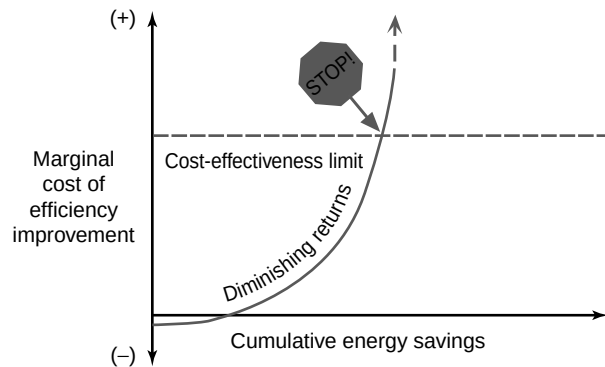


FIGURE 2 Diminishing returns (greater energy savings incur greater marginal cost) can be true for some (not all) components, but need not be true for most systems.

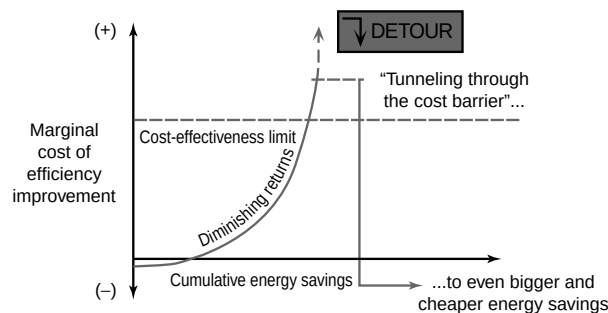


FIGURE 3 Optimizing whole systems for multiple benefits, rather than isolated components for single benefits, can often “tunnel through the cost barrier” directly to the lower-right-corner destination, making very large energy savings cost less than small or no savings. This has been empirically demonstrated in a wide range of technical systems.

4.1 Empirical Examples

Consider, for example, how much thermal insulation should surround a house in a cold climate. Conventional design specifies just the amount of insulation that will repay its marginal cost out of the present value of the saved marginal energy. But this is methodologically wrong, because the comparison omits the capital cost of the heating system: furnace, ducts, fans, pipes, pumps, wires, controls, and fuel source. The author’s house illustrates that in outdoor temperatures down to -44°C , it is feasible to grow bananas (28 crops at this writing) at 2200 m elevation in the Rocky Mountains with no heating system, yet with reduced construction cost, because the superwindows, superinsulation, air-to-air heat exchangers, and other investments needed to eliminate the heating system cost less to install than the heating system would have cost to install. The resulting $\sim 99\%$ reduction in heating energy cost is an extra benefit.

Similarly, Pacific Gas and Electric Company’s Advanced Customer Technology Test for Maximum Energy Efficiency (ACT²) demonstrated in seven new and old buildings that the “supply curve” of energy efficiency generally bent downward, as shown schematically in Fig. 3. For example, an ordinary-looking new tract house was designed to save 82% of the energy allowed by the strictest U.S. standard of the time (1992 California Title 24); if this design were widely implemented rather than restricted to a single experiment, it was estimated by PG&E that it would cost about \$1800 less than normal to build and \$1600 less (in present value) to maintain. It provided comfort with no cooling system in a climate that can reach 45°C ; a similar house later did the same in a 46°C -peak climate. Another example, the 350-m² Bangkok home of architecture professor Suntoorn Boonyatikarn, provided normal comfort, with 10% the normal air-conditioning capacity, at no extra construction cost. These examples illustrate how optimizing a house as a system rather than optimizing a component in isolation, and optimizing for life-cycle cost (capital plus operating cost, and preferably also maintenance cost), can make a superefficient house cheaper to build, not just to run, by eliminating costly heating and cooling systems. Similarly, a retrofit design for a 19,000-m² curtainwall office building near Chicago found 75% energy-saving potential at no more cost than the normal 20-year renovation that saves nothing, because the \$200,000 capital saving from making the cooling system four times smaller (yet four times more efficient), rather than renovating the big old system, would pay for the other improvements.

In a striking industrial example, a pumping loop for heat transfer originally designed to use 70.8 kW of pumping power was redesigned to use 5.3 kW, 92% less, with lower capital cost and better performance. No new technologies were used, but only two changes in the design mentality:

1. Use big pipes and small pumps rather than small pipes and big pumps. The friction in a pipe falls as nearly the fifth power (roughly -4.84) of its diameter. Engineers normally make the pipe just fat enough to repay its greater cost from the saved pumping energy. This calculation improperly omits the capital cost of the pumping equipment—the pump, motor, inverter, and electricals that must overcome the pipe friction. Yet the size and roughly the cost of that equipment will fall as nearly the fifth power of pipe diameter, while the cost of the fatter pipe will rise as only about the second power of

diameter. Thus conventionally optimizing the pipe as an isolated component actually pessimizes the system! Optimizing the whole system together will clearly yield fat pipes and tiny pumps, so total capital cost falls slightly and operating cost falls dramatically.

2. Lay out the pipes first, then the equipment. Normal practice is the opposite, so the connected equipment is typically far apart, obstructed by other objects, at the wrong height, and facing the wrong way. The resulting pipes have about three to six times as much friction as they would have with a straight shot, to the delight of the pipefitters, who are paid by the hour, who mark up a profit on the extra pipes and fittings, and who do not pay for the extra electricity or equipment sizing. But the owner would do better to have short, straight pipes than long, crooked pipes.

Together, these two design changes cut the measured pumping power by 12-fold, with lower capital cost and better performance. They also saved 70 kW of heat loss with a 2-month payback, because it was easier to insulate short, straight pipes. In hindsight, however, the design was still suboptimal, because it omitted seven additional benefits: less space, weight, and noise; better maintenance access; lower maintenance cost; higher reliability and uptime; and longer life (because the removed pipe elbows will not be eroded by fluid turning the corner). Properly counting these seven benefits would have saved not 92% but nearer 98% of the energy, and cost even less, so about a factor-four potential saving was left uncaptured.

Other recent design examples include a 97% reduction in air-conditioning energy for a California office retrofit, with attractive returns and better comfort; lighting retrofit savings upward of 90% with better visibility and a 1- to 2-year payback; an energy cost reduction >40% with a 3-year payback in retrofitting an already very efficient oil refinery; ~75% electrical savings in a new chemical plant, with ~10% lower construction time and cost; ~89% in a new data center at lower cost; and ~70–90% in a new supermarket at probably lower cost. The obvious lesson is that optimizing whole systems for multiple benefits, not just components for single benefits, typically boosts end-use efficiency by roughly an order of magnitude at negative marginal cost. These enormous savings have not been widely noticed or captured because of deficient engineering pedagogy and practice. Whole-system design integration is not rocket science; rather, it rediscovers the forgotten tradition of Victorian system engineering,

before designers became so specialized that they lost sight of how components fit together.

It is not even generally true, as economic theory supposes, that greater end-use efficiency costs more at the level of components. For example, the most common type of industrial motor on the 1996 U.S. market, the 1800-revolution per minute (rpm) totally enclosed fan-cooled (TEFC) motor (National Electrical Manufacturers' Association Design B), exhibited no empirical correlation whatever between efficiency and price up to at least 225 kW. (Premium-efficiency motors should cost more to build because they contain more and better copper and iron, but they are not priced that way.) The same is true for most industrial pumps and rooftop chillers, Swedish refrigerators, American televisions, and many other products. But even if it were true, artfully combining components into systems can definitely yield expanding returns.

Perhaps the most consequential example is in light vehicles. A small private company (see its Web site at www.hypercar.com) completed in 2000 the manufacturable, production-costed virtual design of a midsize sport utility vehicle (SUV) concept car that is uncompromised, cost-competitive, zero emission, and quintupled efficiency. It is so efficient (2.38 liters/100 km, 42 km/liter, 99 mpg, U.S. gallons) not just because its direct-hydrogen fuel cell is about twice as efficient as a gasoline-fueled Otto engine, but also because it is so lightweight (but crash-worthy) and so low in aerodynamic drag and rolling resistance that it can cruise at highway speed on no more power to the wheels than a conventional SUV uses on a hot day just for air conditioning. This design suggests that cars, too, can “tunnel through the cost barrier,” achieving astonishing fuel economy at no extra cost and with many other customer and manufacturing advantages. With aggressive licensing of existing intellectual property, such vehicles could start ramping up production as early as 2008. All major automakers have parallel development programs underway, totaling ~\$10 billion of commitments through 2000, since the basic concept was put into the public domain in 1993 to maximize competition.

A full U.S. fleet of such light vehicles of various shapes and sizes would save about as much oil as Saudi Arabia produces (~8 million bbl/day); a global fleet would save as much oil as OPEC sells. Moreover, such vehicles can be designed to plug into the grid when parked (which the average car is ~96% of the time), acting as a power station on wheels, selling back enough electricity and ancillary

services to repay most of its cost. A U.S. fleet of light vehicles doing this would have ~ 5 – 10 times as much electric generating capacity as all power companies now own. This is part of a wider strategy that combines hydrogen-ready vehicles with integrated deployment of fuel cells in stationary and mobile applications to make the transition to a climate-safe hydrogen economy profitable at each step, starting now (beginning with ultrareliable combined heat and power in buildings). The resulting displacement of power plants, oil-fueled vehicles, and fossil-fueled boilers and furnaces could decrease net consumption of natural gas, could save about \$1 trillion of global vehicle fueling investment over the next 40 years (compared with gasoline-system investment), and could profitably displace up to two-thirds of CO₂ emissions. It could also raise the value of hydrocarbon reserves, in which hydrogen is typically worth more without than with the carbon. It is favored by many leading energy and car companies today, and it is not too far off: over two-thirds of the fossil fuel atoms burned in the world today are already hydrogen, and global hydrogen production (~ 50 MT/year), if diverted from its present uses, could fuel an entire fleet of superefficient U.S. highway vehicles.

4.2 The Right Steps in the Right Order

Breakthrough energy efficiency results need not just the right technologies but also their application in the right sequence. For example, most practitioners designing lighting retrofits start with more efficient luminaires, improving optics, lamps, and ballasts. But for optimal energy and capital savings, that should be step six, not step one. First come improving the quality of the visual task, optimizing the geometry and cavity reflectance of the space, optimizing lighting quality and quantity, and harvesting daylight. Then, after the luminaire improvements, come better controls, maintenance, management, and training. Likewise, to deliver thermal comfort in a hot climate, most engineers retrofit a more efficient and perhaps variable-speed chiller, variable-speed supply fans, etc. But these should all be step five. The previous four steps are to expand the comfort range (by exploiting such variables as radiant temperature, turbulent air movement, and ventilative chairs); reduce unwanted heat gains within or into the space; exploit passive cooling (ventilative, radiative, ground coupling); and, if needed, harness nonrefrigerative alternative cooling (evaporative, desiccant, absorption, and

hybrids of these). These preliminary steps can generally eliminate refrigerative cooling. If refrigerative cooling is still nonetheless desired, it can be made superefficient (e.g., system coefficient of performance 8.6 measured in Singapore), then supplemented by better controls and by coolth storage. Yet most designers pursue these seven steps in reverse order, worst buys first, so they save less energy, pay higher capital costs, yet achieve worse comfort and greater complexity.

Whole-system engineering optimizes for many benefits. There are, for example, 10 benefits of superwindows and 18 benefits of premium-efficiency motors or dimmable electronic lighting ballasts, not just the one benefit normally counted. (The arch that holds up the middle of the author's house has 12 different functions but is paid for only once.) Superwindows cost more per window, but typically make the building cost less because they downsize or eliminate space-conditioning equipment. Similarly, 35 retrofits to a typical industrial motor system, properly counting multiple benefits, can typically save about half its energy (not counting the larger and cheaper savings that should first be captured further downstream, e.g., in pumps and pipes) with a 16-month simple payback against a 5¢/kWh tariff. The saving is so cheap because buying the correct seven improvements first yields 28 more as free by-products. Such motor-system savings alone, if fully implemented, could save about 30% of all electricity worldwide. Such design requires a diverse background, deep curiosity, often a transdisciplinary design team, and meticulous attention to detail. Whole-system design is not what any engineering school appears to be teaching, nor what most customers currently expect, request, reward, or receive. But it represents a key part of the "overhang" of practical, profitable, unbought energy efficiency absent from virtually all official studies so far.

5. MARKET FAILURES AND BUSINESS OPPORTUNITIES

In a typical U.S. office, using one-size-fatter wire to power overhead lights would pay for itself within 20 weeks. Why is that not done? There are several answers: (1) The wire size is specified by the low-bid electrician, who was told to "meet code," and the wire-size table in the National Electrical Code is meant to prevent fires, not to save money. Saving money by optimizing resistive losses requires wire

about twice as fat. (2) The office owner or occupant will buy the electricity, but the electrician bought the wire. An electrician altruistic enough to buy fatter wire is not the low bidder and does not win the job. Correcting this specific market failure requires attention both to the split incentive and to the misinterpretation of a life-safety regulation as an economic optimum. This microexample illustrates the range and depth of market failures that any skilled practitioner of energy efficiency encounters daily. A 1997 compendium, *Climate: Making Sense and Making Money*, organizes 60–80 such market failures into eight categories and illustrates the business opportunity each can be turned into. Some arise in public policy, some are at the level of the firm, and some are in individuals' heads. Most are glaringly perverse. For example, in all but two states in the United States, regulated distribution utilities are rewarded for selling more energy and penalized for cutting customers' bills, so naturally they are unenthusiastic about energy efficiency that would hurt their shareholders. Nearly all architects and engineers, too, are paid for what they spend, not for what they save; "performance-based fees" have been shown to yield superior design, but are rarely used. Most firms set discounted-cashflow targets for their financial returns, yet tell their operating engineers to use a simple-payback screen for energy-saving investments (typically less than 2 years), and the disconnect between these two metrics causes and conceals huge misallocations of capital. When markets and bidding systems are established to augment or replace traditional regulation of energy supply industries, negawatts (saved watts) are rarely allowed to compete against megawatts.

In short, scores of market failures—well understood but widely ignored—cause available and profitable energy efficiency to get only a small fraction of the investment it merits. Thus most of the capital invested in the energy system is being misallocated. The most effective remedy would be to put systematic "barrier-busting" atop the policy agenda, turning obstacles into opportunities and stumbling-blocks into stepping-stones, so market mechanisms could work properly, as economic theory correctly prescribes.

Using energy in a way that saves money is not only a perquisite of the rich, it is also arguably the most powerful key to global economic development for the poor. Using quintupled-efficiency compact fluorescent lamps in Bombay or installing superwindows in Bangkok takes about a thousand times less capital compared to expanding the supply of electricity to

produce the same light and comfort via inefficient lamps and office/retail air conditioners. The efficiency investment is also repaid about 10 times faster. The resulting ~10,000-fold decrease in capital requirements could turn the power sector, which now uses about one-fourth of global development capital, into a net exporter of capital to fund other development needs. This is also true at the microscale of a rural village: a package of photovoltaics and superefficient end-use devices (lamps, pumps, mobile phone, water purification, vaccine refrigerator, etc.), with normal utility financing and no subsidy, often incurs debt service lower than what the villagers were already paying for lighting kerosene, candles, and radio batteries, so they have a positive cash flow from day one. Conversely, when Chinese authorities imported many assembly lines to make refrigerators more accessible, the saturation of refrigerators in Beijing households rose from 2 to 62% in 6 years, but the refrigerators' inefficient design created unintended shortages of power and of capital to generate it (an extra half-billion dollars' worth). A Chinese Cabinet member said this error must not be repeated: energy and resource efficiency must be the cornerstone of the development process. Otherwise, resource waste will require supply-side investment of the capital meant to buy the devices that were supposed to use those resources. This realization contributed to China's emphasis on energy efficiency (halving primary energy/GDP elasticity in the 1980s and nearly re-halving it since), laying the groundwork for the dramatic 1996 initial shift from coal to gas, renewables, and efficiency. This greatest contribution of any nation so far to reducing carbon emissions was a by-product of two other domestic goals: eliminating the bottleneck in China's development and improving public health.

6. OLD AND NEW WAYS TO ACCELERATE ENERGY EFFICIENCY

6.1 Old but Good Methods

In the 1980s and 1990s, practitioners and policy-makers greatly expanded their tool kits for implementing energy efficiency. During the period 1973–1986, the United States doubled its new-car efficiency, and from 1977 to 1985, cut its oil use 4.8%/year. From 1983 to 1985, 10 million people served by Southern California Edison Company were cutting the decade-ahead forecast of peak load by

8½%/year, at ~1% of the long-run marginal cost of supply. In 1990, New England Electric System signed up 90% of a pilot market for small-business retrofits in 2 months. In the same year, Pacific Gas and Electric Company (PG&E) marketers captured a fourth of the new commercial construction market for design improvements in 3 months, so in 1991, PG&E raised the target, and got it all in the first 9 days of January.

Such impressive achievements resulted from nearly two decades of refinement of program structures and marketing methods. At first, utilities and governments wanting to help customers save energy offered general, then targeted, information, and sometimes loans or grants. Demonstration programs proved feasibility and streamlined delivery. Standards knocked the worst equipment off the market. (Congress did this for household appliances without a single dissenting vote, because so many appliances are bought by landlords, developers, or public housing authorities—a manifest split incentive with the householders who will later pay the energy bills.) Refrigerator standards alone cut electricity usage by new U.S. units by fourfold from 1975 to 2001 (5%/year), saving 40 GW of electric supply. In Canada, labeling initially did nearly as well. Utilities began to offer rebates—targeted, then generic—to customers, then to other value-chain participants, for adopting energy-saving equipment, or scrapping inefficient equipment, or both. Some rebate structures proved potent, such as paying a shop assistant a bonus for selling an energy-efficient refrigerator but nothing for selling an inefficient one. So did leasing (20¢ per compact fluorescent lamp per month, so that it is paid for over time), paying for better design, and rewarding buyers for beating minimum standards. Energy-saving companies, independent or utility owned, provided turnkey design and installation to reduce hassle. Sweden aggregated technology procurement to offer “golden carrot” rewards to manufacturers bringing innovations to market; once these new products were introduced, competition quickly eliminated their modest price premia. These engineered-service/delivery models worked well, often spectacularly well. Steve Nadel’s 1990 review of 237 programs at 38 U.S. utilities found many costing <1¢/kWh (1988 dollars). From 1991 to 1994, the entire demand-side management portfolio of California’s three major investor-owned utilities saved electricity at an average program cost that fell from about 2.8 to 1.9 current ¢/kWh (1.2¢ for the cheapest), saving society over \$2 billion more than the program cost.

6.2 New and Better Methods

Since the late 1980s, another model has been emerging that promises even better results: not just marketing negawatts (saved watts), maximizing how many customers save and how much, but making markets in negawatts, i.e., also maximizing competition in who saves and how, so as to drive quantity and quality up and cost down. Competitive bidding processes let saved and produced energy compete fairly. Savings can be made fungible in time and space; transferred between customers, utilities, and jurisdictions; and procured by “bounty hunters.” Spot, futures, and options markets can be expanded from just megawatts to embrace negawatts too, permitting arbitrage between them. Property owners can commit never to use more than x MW, then trade those commitments in a secondary market that values reduced demand and reduced uncertainty of demand. Efficiency can be cross-marketed between electric and gas distributors, each rewarded for saving either form of energy. Revenue-neutral “feebates” for connecting new buildings to public energy supplies (fees for inefficiency, rebates for efficiency) can reward continuous improvement. Standardized measurement and reporting of energy savings allow savings to be aggregated and securitized like home mortgages, sold promptly into liquid secondary markets, and hence financed easily and cheaply (see the Web site of the nonprofit International Performance Measurement and Verification Protocol, Inc., www.ipmvp.org). Efficiency techniques can be conveniently bundled and translated to “vernacular” forms, which are easily chosen, purchased, and installed. Novel real-estate value propositions emerge from integrating efficiency with on-site renewable supply (part of the revolutionary shift now underway to distributed resources) so as to eliminate all wires and pipes, the trenches carrying them, and the remote infrastructure to which they connect. Performance-based design fees, targeted mass retrofits, greater purchasing aggregation, and systematic barrier busting all show immense promise. Aggressively scrapping inefficient devices, paying bounties to destroy them instead of reselling them, could both solve many domestic problems (e.g., oil, air, and climate in the case of inefficient vehicles) and boost global development by reversing “negative technology transfer.”

Altogether, the conventional agendas for promoting energy efficiency—prices and taxes, plus regulation or deregulation—ignore nearly all the most effective, attractive, transideological, and quickly spreadable methods. And they ignore many

of the new marketing “hooks” just starting to be exploited: security (national, community, and individual), economic development and balance of trade, protection from disruptive price volatility, avoiding costly supply overshoot, profitable integration and bundling with renewables, and expressing individual values. Consider, for example, a good compact fluorescent lamp. It emits the same light as an incandescent lamp but uses four to five times less electricity and lasts 8–13 times longer, saving tens of dollars more than it costs. It avoids putting a ton of carbon dioxide and other pollutants into the air. But it does far more. In suitable numbers (half a billion are made each year), it can cut by a fifth the evening peak load that causes blackouts in overloaded Mumbai, it can boost profits of poor American chicken farmers by a fourth, or it can raise the household disposable cash income of destitute Haitians by up to a third. As mentioned previously, making the lamp requires 99.97% less capital than does expanding the supply of electricity, thus freeing investment for other tasks. The lamp cuts power needs to levels that make solar-generated power affordable, so girls in rural huts can learn to read at night, advancing the role of women. One light bulb does all that. It can be bought at the supermarket and self-installed. One light bulb at a time, the world can be made safer. “In short,” concludes Jørgen Nørgård, by pursuing the entire efficiency potential systematically and comprehensively, “it is possible in the course of half a century to offer everybody on Earth a joyful and materially decent life with a per capita energy consumption of only a small fraction of today’s consumption in the industrialized countries.”

6.3 Deemphasizing Traditionally Narrow Price-Centric Perspectives

The burgeoning opportunities for adoption of energy efficiency approaches suggest that price may well become the least important barrier. Price remains important and should be correct, but is only one of many ways to get attention and influence choice; ability to respond to price can be far more important. End-use efficiency may increasingly be marketed and bought mainly for its qualitatively improved services, just as distributed and renewable supply-side resources may be marketed and bought mainly for their distributed benefits. Outcomes would then become decreasingly predictable from economic experience or using economic tools. Meanwhile, disruptive technologies and integrative design methods are clearly inducing dramatic shifts of, not just along, demand

curves, and are even making them less relevant by driving customer choice through nonprice variables. Ultralight cars, for example, would do a complete end run around two decades of trench warfare in the U.S. Congress (raising efficiency standards vs. gasoline taxes). They would also defy the industry’s standard assumption that efficient cars must trade off other desirable attributes (size, performance, price, safety), requiring government intervention to induce customers to buy the compromised vehicles. If advanced cars can achieve not incremental but fivefold fuel savings as a by-product of breakthrough design integration, yet remain uncompromised and competitively priced, then the energy-price-driven paradigm becomes irrelevant. People will prefer such vehicles because they are better, not because they are clean and efficient, much as most people now buy digital media rather than vinyl phonograph records: they are simply a superior product that redefines market expectations. This implies a world in which fuel price and regulation become far less influential than today, displaced by imaginative, holistic, integrative engineering and marketing. In the world of consumer electronics—ever better, faster, smaller, cheaper—that world is already upon us. In the wider world of energy efficiency, the master key to so many of the world’s most vexing problems, it is coming rapidly over the horizon. We need only understand it and do it.

Acknowledgments

The author is grateful to many colleagues, especially Dr. David B. Brooks, Dr. Jørgen S. Nørgård, and Dr. Joel N. Swisher, for their insightful comments, and to the William and Flora Hewlett Foundation for supporting the preparation of this article.

SEE ALSO THE FOLLOWING ARTICLES

Conservation Measures for Energy, History of • Consumption, Energy, and the Environment • Economics of Energy Efficiency • Energy Efficiency and Climate Change • Industrial Energy Efficiency • Market Failures in Energy Markets • Obstacles to Energy Efficiency • Potential for Energy Efficiency: Developing Nations

Further Reading

American Institute of Physics (AIP). (1975). “Efficient Use of Energy.” AIP Conf. Procs. No. 25. AIP, New York.
Benyus, J. M. (1997). “Biomimicry: Innovation Inspired by Nature.” William Morrow, New York.

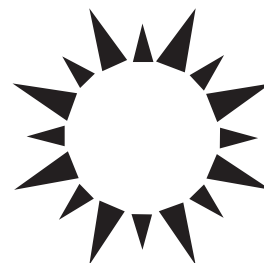
- Brohard, G. J., Brown, M. W., Cavanagh, R., Elberling, L. E., Hernandez, G. R., Lovins, A., and Rosenfeld, A. H. (1998). "Advanced Customer Technology Test for Maximum Energy Efficiency (ACT²) Project: The Final Report. Proceedings, Summer Study on Energy-Efficient Buildings." American Council for an Energy-Efficient Economy, Washington, D.C.
- Daly, H. E. (1996). "Beyond Growth—The Economics of Sustainable Development." Beacon Press, Boston.
- Drexler, K. E. (1992). "Nanosystems: molecular machinery, manufacturing, and computation." John Wiley & Sons, New York.
- E SOURCE (Platts Research & Consulting). (2002). "Drivepower Technology Atlas" and "Electronic Encyclopedia." E SOURCE, Boulder, Colorado. [Available through their Web site at www.esource.com.]
- Fickett, A. P., Gellings, C. W., and Lovins, A. B. (1990). Efficient use of electricity. *Sci. Am.* 263(3), 64–74.
- Hawken, P. G., Lovins, A. B., and Lovins, L. H. (1999). "Natural Capitalism: Creating the Next Industrial Revolution." Little Brown, New York. [See also the Web site at www.natcap.org.]
- International Project for Sustainable Energy Paths (IPSEP). (1989–1999). "Energy Policy in the Greenhouse." IPSEP, El Cerrito, California. [www.ipsep.org]
- International Project for Sustainable Energy Paths (IPSEP). (2001). "Cutting Carbon Emissions at a Profit: Opportunities for the U.S." IPSEP, El Cerrito, California. [www.ipsep.org]
- Johansson, T. B., Bodlund, B., and Williams, R. H. (eds.). (1989). "Electricity." Lund University Press, Lund, Sweden.
- Lovins, A. B. (1992). "Energy-Efficient Buildings: Institutional Barriers and Opportunities." Strategic Issues Paper III. E SOURCE (Platts Research & Consulting), Boulder, Colorado. [www.esource.com]
- Lovins, A. B. (1994). Apples, oranges, and horned toads. *Electricity J.* 7(4), 29–49.
- Lovins, A. B. (1995). The super-efficient passive building frontier. *ASHRAE J.* 37(6), 79–81.
- Lovins, A. B. (2003). "Twenty Hydrogen Myths." Available from the Rocky Mountain Institute Web site at www.rmi.org.
- Lovins, A. B., and Cramer, D. R. (2004). Hypercars[®], hydrogen, and the automotive transition. *Int. J. Vehicle Design* (in press).
- Lovins, A. B., and Gadgil, A. (1991). "The Negawatt Revolution: Electric Efficiency and Asian Development." RMI Publication #E91-23, Rocky Mountain Institute, Snowmass, Colorado.
- Lovins, A. B., and Lovins, L. H. (1991). Least-cost climatic stabilization. *Annu. Rev. Energy Environ.* 16, 433–531.
- Lovins, A. B., and Lovins, L. H. (1996). Negawatts: Twelve transitions, eight improvements, and one distraction. *Energy Policy* 24(4), 331–344.
- Lovins, A. B., and Lovins, L. H. (1997). "Climate: Making Sense and Making Money." Rocky Mountain Institute, Snowmass, Colorado.
- Lovins, A. B., and Lovins, L. H. (2001). Fool's gold in Alaska. *Foreign Affairs* 80(4), 72–85.
- Lovins, A. B., and Williams, B. D. (1999). "A Strategy for the Hydrogen Transition." Proceedings, 10th Annual Hydrogen Meeting., 8 April. National Hydrogen Assoc., Washington, D.C.
- Lovins, A. B., Datta, E. K., Feiler, T., Rábago, K. R., Swisher, J. N., Lehmann, A., and Wicker, K. (2002). "Small Is Profitable: The Hidden Economic Benefits of Making Electrical Resources the Right Size." Rocky Mountain Institute, Snowmass.
- Lovins, A. B., Koomey, J. G., Datta, E. K., Bustnes, O.-E., and Swisher, J. N. (2004). "Out of the Oil Box: A Roadmap for U.S. Mobilization." Rocky Mountain Institute, Snowmass, Colorado.
- Lovins, A. B., Lovins, L. H., Krause, F., and Bach, W. (1982). "Least-Cost Energy: Solving the CO₂ Problem." Brick House, Andover, Massachusetts. [Reprinted 1989 by Rocky Mountain Institute, Snowmass, Colorado.]
- Nadel, S. (1990). "Lessons Learned." Report #90-08. New York State Energy R&D Authority, Albany, with NY State Energy Office and Niagara Mohawk Power Corp. American Council for an Energy-Efficient Economy, Washington, D.C. [www.aceee.org]
- Nørgård, J. S. (2002). "Energy Efficiency and the Switch to Renewable Energy Sources. Natural Resource System Challenge II: Climate Change. Human Systems and Policy" (A. Yotova, ed.). UNESCO Encyclopedia of Life Support Systems. EOLSS Publ. Co., Oxford, U.K.
- Reddy, A. K. N., Williams, R. H., and Johansson, T. B. (1997). "Energy After Rio: Prospects and Challenges." United Nations Development Program, New York.
- Repetto, R., and Austin, D. (1997). "The Costs of Climate Protection: A Guide to the Perplexed." World Resources Institute, Washington, D.C. [www.wri.org]
- Romm, J. J., and Browning, W. D. (1994). "Greening the Building and the Bottom Line: Increasing Productivity through Energy-efficient Design." Rocky Mountain Institute, Snowmass, Colorado.
- Romm, J. J., Rosenfeld, A. H., and Herrmann, S. (1999). "The Internet Economy and Global Warming: A Scenario of the Impact of E-Commerce on Energy and the Environment." Center for Energy and Climate Solutions, Washington, D.C.
- Swisher, J. N. (2002). "The New Business Climate." Rocky Mountain Institute, Snowmass, Colorado.
- Swisher, J. N., Jannuzzi, G., and Redlinger, R. (1998). "Tools and Methods for Integrated Resource Planning: Improving Energy Efficiency and Protecting the Environment." UNEP Collaborating Centre on Energy and Environment, Denmark.
- von Weizsäcker, E. U., Lovins, A. B., and Lovins, L. H. (1995/1997). "Factor Four: Doubling Wealth, Halving Resource Use." Earthscan, London.
- Wilson, A., Uncapher, J., McManigal, L., Lovins, L. H., Cureton, M., and Browning, W. (1998). "Green Development: Integrating Ecology and Real Estate." John Wiley & Sons, New York. [See also "Green Developments," (2002) a CD-ROM published by the Rocky Mountain Institute, Snowmass, Colorado.]

Energy Futures and Options

JOHN ELTING TREAT

Pinnacle Partners

San Francisco, California, United States



-
1. Introduction
 2. Criteria for Successful Futures Markets
 3. Exchanges and Their Contracts
 4. Futures Present
 5. Trading Futures: A Primer
 6. Future Directions
 7. Summary

Glossary

forward contract A supply contract between a buyer and seller, whereby the buyer is obligated to take delivery, and the seller is obligated to provide delivery, of a fixed amount of a commodity at a predetermined price on a specified future date. Payment in full is due at the time of, or following, delivery.

futures contract A contract between a buyer and seller, whereby the buyer is obligated to take delivery, and the seller is obligated to provide future delivery, of a fixed amount of a commodity at a predetermined price at a specified location. Futures contracts are most often liquidated prior to the delivery date and are generally used as a financial risk management and investment tool rather than for supply purposes.

margin The amount of money or collateral deposited by a customer with a broker, or deposited by a broker with a clearing member, or by a clearing member with the clearinghouse, for the purpose of insuring the broker or clearinghouse against adverse price movement on open futures contracts.

spread (futures) The simultaneous purchase and sale of futures contracts for different months, different commodities, or different grades of the same commodity.

spread (options) The purchase and sale of options that vary in terms of type (call or put), strike prices, and/or expiration dates. May also refer to an options contract purchase (sale) and the simultaneous sale (purchase) of a futures contract for the same underlying commodity.

Energy futures trading in the 1980s focused on growth of the liquid petroleum markets for crude oil,

natural gas liquids, and the major refined products of gasoline, heating oil, and fuel oil. In the 1990s, the boundaries of the energy complex expanded to include, first, natural gas (in 1990), then electricity (in 1996), and to add a wide range of options contracts based on the underlying futures markets.

1. INTRODUCTION

In the second half of the 19th century, a “petroleum exchange” flourished in New York. Again, in the early 1930s, when market discipline was briefly disrupted by the explosive growth of oil production in Oklahoma and Texas and oil prices fell dramatically, an oil futures market (in West Texas Intermediate crude oil) was established in California. It soon collapsed as a formidable alliance of “big oil” and “big government” restored discipline to the marketplace. Nearly 40 years of relative price stability ensued, leaving little incentive for the emergence of an oil futures market.

Only with the traumatic price increases accompanying the Arab oil embargo of late 1973 was another attempt made, this time back in New York at the Cotton Exchange. The contract called for Rotterdam delivery (to avoid the constraints of U.S. price regulations). That attempt was stillborn, however, doomed by continuing U.S. government price controls and a skeptical oil industry. In the decade that followed, the commercial realities and—equally important—the perceptions of those realities by the international oil industry had gradually changed; the point at which oil futures could fulfill the industry’s need for risk management and could provide an outlet for the speculative impulses of investors had been reached. Investors’ interest in oil had been captured by the commodity’s new prominence in daily headlines and nightly newscasts.

The emergence of oil futures markets and their remarkable growth was a natural, indeed inevitable,

consequence of three concurrent, but only partly interrelated, trends in petroleum, financial, and commodity markets. By far the most important determinant was the structural change in oil markets. The nationalization of production by members of the Organization of Petroleum Exporting Countries (OPEC) and non-OPEC governments alike, and the subsequent pressures to eliminate large third-party crude resales, resulted in a disintegration of an oil market that had been highly integrated since the days of J. D. Rockefeller. In the 10 years following the 1973 Arab oil embargo, the crude oil available to the major companies fell by nearly 50%, from about 30 million barrels (bbl)/day to just over 15 million bbl/day. Equity oil available to the majors fell even more sharply, by some 75%. The net result was a drop in the major's share of internationally traded oil from 62% to just 37%.

In addition to the newly created national oil companies, a host of oil trading firms and independent refineries entered the picture. The links that had traditionally tied upstream and downstream (vertical) integration were weakened to the breaking point. The reduction of horizontal integration (as large third-party crude sales were curtailed and joint venture production was nationalized) further eroded the ability of the largest oil companies to exercise control over markets. Simply put, there were too many actors with divergent commercial and political interests to guarantee market stability. The consequences were not long in coming (see Table I).

After a decade of virtually universal confidence that oil prices would rise, prices began to weaken and fluctuate over an ever-wider range, climaxing in the dramatic events of 1986, when prices fell from nearly \$30/bbl to less than \$10/bbl in a period of only 9 months. Although many market observers feel stability returned in 1987, it is interesting to note that prices oscillated between \$15 and \$22/bbl between September and December of that year alone. The fact is that stability has yet to rear its hoary head in current-day oil markets.

Along with the structural change that was reshaping oil markets during the decade, a second important trend was emerging from the financial markets. High interest rates (along with high oil prices) at the beginning of the 1980s were making inventory maintenance very expensive, causing oil company management teams to rethink traditional approaches to inventory and risk management. Also, hedging of financial risks was increasingly becoming a fact of life in foreign currency and interest rate markets. These trends ensured that oil companies

TABLE I
Crude Oil Price Trends^a

Year	Price
1950	\$2.51
1960	2.88
1961	2.89
1962	2.90
1963	2.89
1964	2.88
1965	2.86
1966	2.88
1967	2.92
1968	2.94
1969	3.09
1970	3.18
1971	3.39
1972	3.39
1973	3.89
1974	6.74
1975	7.56
1976	12.17
1977	13.24
1978	13.30
1979	20.19
1980	32.27
1981	35.10
1982	32.11
1983	27.73
1984	27.44
1985	25.83
1986	12.52
1987	16.69
1988	13.27
1989	15.50

^a Average annual prices (U.S. dollars). Data from the U.S. Energy Information Administration (EIA), *Monthly Energy Review*, March 1989 (published June 1989). Prices shown for 1950–1975 are from the EIA's "U.S. Wellhead." After 1975, the EIA series is "FOB Cost of Imports" into the United States; 1989 prices are based on the first quarter.

were increasingly receptive to the hedging potential of the fledgling oil future markets. Finally, the third important factor that set the stage for the ultimate success of energy futures was the general growth and diversification of futures contracts in a wide variety of new markets, and the growing sophistication with which they were being used and modified to offer an ever wider range of hedging tools. For almost 100 years, futures markets (then commonly called commodity markets) were largely confined to the

traditional agricultural products (especially grains). In the past two decades, however, there has been an explosion in the variety of products served by these markets. The first waves of expansion brought in new agricultural contracts (especially meats) and precious metals. The second phase, starting in the 1970s, saw the introduction of financial instruments, including currency, interest rate, and stock index contracts. A third phase brought in oil and a number of other industrial products. The fourth stage saw the introduction and rapid acceptance of options on futures contracts. The fifth stage saw an explosion of trading in over-the-counter (OTC) derivatives, often in direct competition with exchange-traded instruments. The introduction and success of oil futures was a product of the first three trends. The growing volatility and loss of confidence in the future stability of oil prices demanded the emergence of new market structures and institutions. One obvious sign of the change was the rapid growth of spot markets and the trading companies that thrived on price volatility.

Prior to 1979, less than 5% of internationally traded crude moved at spot prices, that is, outside of official term supply contract arrangements. By the end of 1985, virtually all crude moved at some sort of market-related pricing, and experts estimated that oil companies were acquiring anywhere from 30 to 50% of their supplies on a spot, noncontract basis. Although the proportion of oil sold on a purely spot basis has subsequently shrunk, the price risks remain, because term contracts today almost universally call for market-related prices. The trading companies, independent refineries, and increasingly the larger companies developed trading techniques to cope with the growing price volatility of these markets. The first response was to create informal, forward markets, at first only 30 days out, then 60 days out, and more recently 90 days out. A second response was an explosive growth in the demand for rapid, often real-time, pricing and other market information. Against this backdrop, futures became inevitable: a time-proven and efficient technique for coping with broad market instability.

Although the breadth of energy markets has expanded, their fundamental purposes remain the same. Futures markets, basically spot markets for standardized forward contracts, serve three functions:

1. Price discovery, giving an instantaneous reading of marginal price movements.
2. Risk management, allowing companies to hedge their price risks for limited periods of time (however, the hedging opportunity rarely extends

more than 6 months forward as a result of market illiquidity in the more distant months).

3. Speculative opportunity, attracting additional risk capital to the market from outside the oil industry (low margin requirements—lower than in equity markets—enhance the attraction of futures as a vehicle for speculation).

These are the necessary but not sufficient conditions for a successful contract. In reality, however, new futures contracts often fail. The reason is that the criteria for a successful futures contract are simply too stringent, with too few physical markets that can actually meet those criteria.

2. CRITERIA FOR SUCCESSFUL FUTURES MARKETS

Assessing the suitability of any commodity/market for futures trading requires analysis of a number of conditions. These are considered in the following sections.

2.1 Price Volatility

Price volatility is perhaps the single most important criterion for assessing futures trading. It provides the basic economic justification for futures trading, which is to provide protection to the hedger against adverse price fluctuations. If a commodity is characterized by a relatively stable—or at least predictable—price, there would be little associated risk and there would be no need for a futures market. Price volatility is also necessary to attract risk capital from speculators and is essential to ensure sufficient liquidity to maintain the market.

Quantitative indicators: Variations of $\pm 20\%$ per annum are assumed to be the minimum necessary to sustain futures trading. In general, the greater the degree of volatility, the more likely a futures market will survive.

2.2 Uncertain Supply and Demand

Uncertainty in supply and demand is generally the cause of price volatility and therefore is generally present when price volatility is found.

Quantitative indicators: In energy markets, which typically display a rather high inelasticity of price demand, variations of $\pm 10\%$ during a 2-year period should be sufficient to sustain futures trading.

2.3 Sufficient Deliverable Supplies

Sufficiency of supply is the Catch 22 of futures trading. If sufficient deliverable supplies of the commodity meeting the quality specification of the contract are not available, futures trading will fail. However, there must be some uncertainty about the sufficiency of supplies if the previous conditions are to be met. In the United States, this dilemma is heightened by the regulatory requirements of the Commodity Futures Trading Commission (CFTC), whose fear of market squeezes at times forces exchanges to overstate deliverable supplies in order to gain government approval.

Quantitative indicators: Storage capacity equal to at least 30 days average demand is highly desirable.

2.4 Product Homogeneity

Product homogeneity is another prerequisite. Futures contracts are traded on the premise that product taken on a delivery will meet certain quality specifications. The commodity must therefore have certain key characteristics that are quantifiable, allowing the clear differentiation of product from other grades. Standardized tests and generally accepted procedures are essential. For example, the various American Petroleum Institute (API), German Institute for Standardization, and American Society for Testing and Materials (ASTM) standards generally provide the necessary references for oil. In addition, the existence of generally trusted independent inspection agencies or inspectors to administer these tests is an important aspect. A range of different products (for example, several types of crude oil) may be suitable for delivery, if the price differences between the various grades are relatively stable and if the technical characteristics of the various deliverable grades are sufficiently close to one another. This is often a difficult aspect of contract design, because the price variation between various grades of products fluctuates from time to time. For example, it may be desirable to allow several grades to be deliverable, perhaps with price adjustments for quality, in order to ensure sufficient deliverable supplies. However, if buyers are uncertain of what grades they will receive and if they place different values on the quality differences among the grades, they may be deterred from trading.

Quantitative indicators: The quality of the product must be capable of being described by objective, quantifiable standards.

2.5 Product Perishability

The perishability of a product can be a deterrent to trading. In general, a product should have a shelf life sufficiently long enough to permit storage and delivery as called for under the contract. In addition, the maintenance of inventories of the commodity will both facilitate deliveries and provide a ready pool of potential hedgers. While perishability is not usually a major concern in oil and natural gas markets, the stability of some oil product blends is an issue. Long storage of gasoline, for example, can result in separation of blended product.

Quantitative indicators: Products should have a minimum shelf or stock life of 6–12 months.

2.6 Market Concentration

Market concentration is a difficult factor to quantify. A successful futures market is a highly competitive market, marked by a large number of buyers and sellers. No one market participant, or plausible combination of market participants, should possess sufficient market power to exert unilateral control either on the supply or the demand for the commodity, either in the short or medium term. In oil markets, however, the existence of OPEC has not prevented the emergence of highly successful futures markets. The answer lies in the inability of OPEC to act decisively, and in the availability of alternative sources of supply and stocks that seriously limit OPEC's ability to achieve its stated objectives. However, the concentration of producers and/or consumers can be a serious obstacle in specific regional oil markets. Thus, a U.S. West Coast gasoline market would be risky, given the relative concentration of production in the hands of a small number of refiners. Similarly, an East Coast residual fuel market might be too heavily dominated by the demand from a small number of very large utilities to sustain liquid futures trading.

Quantitative indicators: In general, the market share of the top five firms should be less than 50%, and the top 10 firms should have less than 80%.

2.7 Availability of Price Information

Readily available price information is critical to market success. It should be noted that the opening of a futures market will probably, in turn, stimulate a rapid growth of price information services. However, at the outset, market participants must have a sufficiently broad base of price information to permit evaluation of spot prices and their relationship to

futures prices. Convergence between these two prices as the delivery period approaches is essential. A market in which all products are traded on the basis of long-term contracts for which prices remain undisclosed would be a very difficult market in which to establish futures trading.

Quantitative indicators: Daily cash market prices should be available from at least two independent sources.

2.8 Uniqueness

The uniqueness of a trading opportunity is another key factor. If an existing market for a commodity has reasonable liquidity and is serving its customers well, it is extremely difficult to launch a copycat contract. Inertia, habit, and personal relationships will tend to keep the traders loyal to the preexisting market. In addition, even if there is no active market at present, recent failures of similar contracts can be a substantial (but not fatal) deterrent.

Quantitative indicators: The ideal candidate would be a commodity that is not currently traded on any futures exchange in the world and has not been the subject of a failed attempt in the previous 5 years. However, special circumstances may override these concerns.

2.9 Market Timing

Market timing and blind luck are often critical to the success or failure of a contract. However, forecasting is often impossible. Ideally, contracts should be introduced to coincide with periods of high volatility and high levels of cash market activity. For example, a heating oil or natural gas contract would be best introduced in the fall months, when physical trading is at its yearly high. Conversely, a gasoline contract would be best introduced in the spring, prior to an anticipated surge of summer driving.

Quantitative indicators: Contracts should be introduced to coincide with high levels of cash market activity, to the extent that this is predictable. Alternatively, an astrologer might just as well be consulted.

3. EXCHANGES AND THEIR CONTRACTS

Two exchanges currently dominate trade of energy futures contracts: the New York Mercantile Ex-

change and the London International Petroleum Exchange. In addition several smaller exchanges also have offered energy futures contracts, albeit with limited and often transitory success, including the Singapore International Monetary Exchange (SIMEX; currently known as the Singapore Exchange Ltd.) and the Kansas City Board of Trade. The future of open-outcry trading on exchange floors is increasingly being challenged by the advent of electronic trading. Although the exchanges and their floor traders still have the upper hand, the growth of electronic trading and the success of all-electronic marketplaces (such as the Internet) pose serious challenges for the exchanges and their trading floors.

3.1 New York Mercantile Exchange

Founded more than 100 years ago, the New York Mercantile Exchange (NYMEX), or the “Merc,” as it is often called, has enjoyed a diverse and colorful history. Evolving from a lower Manhattan produce exchange that dealt in contracts for butter, eggs, and even apples, NYMEX began to diversify some two decades ago, adding precious metals and—briefly—even stocks to its portfolio. Nevertheless, by the end of the 1970s, it was one of the smallest exchanges in the United States, outpaced by the growth of the large Chicago markets and by most other New York exchanges as well. After an abortive attempt to start a residual fuel oil contract in 1974, the Merc launched its first successful energy futures contract—New York heating oil—in 1978. Trading grew slowly but steadily while a companion contract for residual fuel was stillborn. With the addition of a leaded gasoline contract for New York delivery in 1981 (later replaced by an unleaded version of the same contract) and West Texas Intermediate crude oil in 1983, NYMEX achieved international prominence in the 1980s; its energy contracts grew at spectacular rates. The subsequent addition of options on crude oil, heating oil, and gasoline as well as futures contracts on propane, natural gas, and electricity added another dimension and further impetus to the growth of the NYMEX energy complex.

Since its merger with the Commodities Exchange (COMEX) in the mid-1990s, NYMEX has been known as the NYMEX Division of the merged exchange. Today, the NYMEX Division is the leading energy futures exchange and the third largest futures exchange in the world, following only the Chicago Board of Trade and the Chicago Mercantile Exchange. Although contracts for platinum and

palladium still survive on the Merc's trading floor, energy contracts regularly account for more than 90% of its turnover. NYMEX, for most of its life, was a nonprofit organization. However, in 2000, its members voted to "demutualize" and the exchange became a for-profit corporation, the first New York commodity exchange to do so.

3.2 International Petroleum Exchange

An independent outgrowth of London's loosely linked futures markets, the International Petroleum Exchange (IPE) was created in 1981 by a diverse coalition of oil traders and commodity brokerage firm members who saw the emerging success of the NYMEX heating oil contract in New York and were determined to build an energy futures market on the other shore of the Atlantic. Early success with a gas oil contract was followed by a series of unsuccessful attempts to trade crude oil. Finally, in 1988, the right set of circumstances and contract terms allowed IPE to launch a Brent crude oil market that has established good liquidity and a substantial following. Although IPE occupies second place among the world's energy exchanges, the gap between IPE and NYMEX is very large, with London's turnover averaging well under half of New York's volume. Following the NYMEX demutualization, the IPE too converted to for-profit status in 2001 and subsequently merged with the U.S.-based Intercontinental Exchange.

3.3 Singapore Exchange Ltd

The Singapore International Monetary Exchange was created in 1983 as a restructuring of the former Gold Exchange of Singapore (founded in 1978) with the strong support of the Chicago Mercantile Exchange (CME). CME and SIMEX offered customers an innovative mutual offset agreement, whereby positions on one exchange can be offset against positions on the other. Initially, the exchange concentrated on financial futures, but in February 1989 SIMEX launched a high-sulfur residual fuel oil contract that enjoyed a brief success. In the mid-1990s, they traded Brent crude oil for several years.

In 1999, SIMEX merged with the Stock Exchange of Singapore to form the Singapore Exchange Ltd. (SGX). Their futures trading is organized under the SGX Derivatives Trading Division (SGX-DT), which has recently launched a contract on Dubai crude oil—the second attempt to trade Middle Eastern crude. Like the other major exchanges, the SGX has also converted to a for-profit corporation.

3.4 Other Exchanges

From time to time, other futures exchanges, most recently Kansas City, have launched energy futures contracts. However, none of these has succeeded in building the liquidity needed to attract significant trading.

4. FUTURES PRESENT

Since 1974, there have been over 60 attempts to launch energy futures and options markets. The success rate has averaged about 20%, typical of the experience in other commodity markets. In spite of thorough research by the exchanges, often excruciating governmental reviews (particularly in the United States), and extensive marketing campaigns, roughly 80% of all future markets opened fail to reach the critical mass needed for takeoff (commonly defined as reaching an average open interest of 5000 contracts). Table II lists the various attempts to launch new futures markets in energy from 1970 through 2002.

When first introduced in late 1978, heating oil futures attracted smaller, independent marketers and refiners, who turned to the New York Merc as an alternative source of supply. Physical deliveries were initially quite high, as these smaller firms sought alternatives in a marketplace dominated by the larger companies. These initial participants were quickly joined by the spot oil traders and by a growing number of pure speculators on and off the trading floor, drawn from other financial and commodity markets. This phase lasted until well into 1983. Then, with the introduction of crude oil futures and the increasing instability of prices, the larger refiners and integrated companies reluctantly entered the market.

By 1984, more than 80% of the 50 largest companies were using futures. Larger end-users, such as airlines and other major energy consumers, also appeared. In addition, a far wider range of speculators entered the scene, as trading volume and open interest rose high enough to meet the minimum liquidity requirement of the commodity funds. Finally, another phase, dating from 1986, brought in almost all the remaining holdouts among the large U.S. companies, more foreign participation, and a new group of traders, the Wall Street refiners. These were companies such as Morgan Stanley and Bear Stearns, which were attracted by the rising volatility of oil prices and the speculative opportunities presented by that price instability, particularly relative to other markets. As one Bear Stearns trader

TABLE II

Energy Futures Markets Authorized, Opened, or Proposed, 1970–2002^a

Year	Commodity ^b	Delivery point	Exchange ^c	Status
1974	Crude oil	Rotterdam	NYCE	Failed
	Residual fuel	Rotterdam	NYMEX	Failed
1978	Heating oil	New York	NYMEX	Active
	Residual fuel	New York	NYMEX	Failed
1981	Gas oil	Rotterdam	IPE	Active
	Heating oil	Gulf Coast	NYMEX	Failed
	Leaded gasoline	New York	NYMEX	Replaced
	Leaded gasoline	Gulf Coast	NYMEX	Failed
	Unleaded gasoline	Gulf Coast	CBOT	Failed
	Unleaded gasoline	Gulf Coast	NYMEX	Unopened
	Propane	Texas	NYCE	Failed
1983	WTI crude oil	Crushing, OK	NYMEX	Active
	LLS crude oil	St. James, LA	CBOT	Failed
	Brent crude oil	Rotterdam	IPE	Failed
	Heating oil	Gulf Coast	CBOT	Failed
1984	Unleaded gasoline	New York	NYMEX	Active
	Leaded gasoline	Gulf Coast	CME	Failed
	Unleaded gasoline	Gulf Coast	CME	Unopened
1985	Brent crude oil	North Sea	IPE	Failed
1986	WTI crude oil options	Cushing, OK	NYMEX	Active
1987	Heating oil options	New York	NYMEX	Active
	Brent crude oil options	Rotterdam	EOE	Failed
1988	Brent crude oil	Cash settlement	IPE	Active
	Propane	Texas	NYMEX	Active
1989	Residual fuel oil	Singapore	SIMEX	Active
	Unleaded gasoline options	New York	NYMEX	Active
	Brent crude oil options	Cash settlement	IPE	Active
	Gas oil options	Rotterdam	IPE	Active
	Brent crude oil	Rotterdam	ROEFEX	Failed
	Gas oil	Rotterdam	ROEFEX	Failed
	Residual fuel oil	Rotterdam	ROEFEX	Failed
	Residual fuel oil	Cash settlement	IPE	Lightly traded
	Residual fuel oil	Gulf coast	NYMEX	Failed
1990	Natural gas	Henry Hub, LA	NYMEX	Active
	Dubai crude oil	Cash settlement	SIMEX	Failed
	Dubai crude oil	Cash settlement	IPE	Failed
1992	Sour crude	Gulf Coast	NYMEX	Failed
	Gulf Coast gasoline	Gulf Coast	NYMEX	Failed
	Natural gas options	Henry Hub, LA	NYMEX	Active
1994	NY gasoline/crude crack spread	New York	NYMEX	Active
	NYH heating oil/crude crack spread	New York	NYMEX	Active
1996	COB electricity	COB	NYMEX	Failed
	Palo Verde electricity	Palo Verde	NYMEX	Failed
	COB electricity options	COB	NYMEX	Failed
	Palo Verde electricity options	Palo Verde, CA	NYMEX	Failed
	Permian basin natural gas	Permian Basin	NYMEX	Failed
	Alberta natural gas	Alberta, Canada	NYMEX	Failed

continues

Table II continued

Year	Commodity ^b	Delivery point	Exchange ^c	Status
1998	Cinergy electricity	Ohio	NYMEX	Failed
	Cinergy electricity options	Ohio	NYMEX	Failed
	Entergy electricity	Arkansas	NYMEX	Failed
	Entergy electricity options	Arkansas	NYMEX	Failed
	PJM electricity	New Jersey	NYMEX	Failed
1999	Middle East crude	United States	NYMEX	Failed
	Mid-Columbia electricity	United States	NYMEX	Failed
2001	Central Appalachian coal	United States	NYMEX	Failed
	Brent crude options	United Kingdom	NYMEX	Failed
2002	Middle East crude	Asia	SIMEX	Active

^a Year first traded or, if never opened, year first authorized. Does not include various contract modifications or innovations in trading, e.g., in 2001 NYMEX began to offer “strip” trading in natural gas contracts across several months and introduced Exchanges of Futures for Swaps (EFS).

^b Abbreviations: WTI, West Texas Intermediate; LLS, Light Louisiana Sweet; NYH, New York Harbor; COB, California/Oregon Border; PJM, Pennsylvania/New Jersey/Maryland.

^c Abbreviations: NYCE, New York Cotton Exchange; NYMEX, New York Mercantile Exchange; IPE, International Petroleum Exchange; CBOT, Chicago Board of Trade; CME, Chicago Mercantile Exchange; EOE, European Options Exchange; SIMEX, Singapore International Monetary Exchange; ROEFEX, Rotterdam Energy Futures Exchange.

put it, “Plywood was dead, so we looked around for some better action and found it in oil.” The low internal cost of capital for margin maintenance and a built-in trading infrastructure made these new entrants formidable competitors for the older oil trading and supply companies. However, even today, participation by independent producers and smaller end-users remains limited. Participation of the producers is limited by the lack of liquidity in the more distant months; participation of the end-users is limited by ignorance of how the markets operate, the high management cost of setting up a futures trading department, and, for a number of domestic as well as international companies, a very real basis-risk problem.

Futures trading has thus survived adolescence and entered a period of youthful maturity. But growth in the coming years will have to come from an expansion of futures trading opportunities in the form of new contracts rather than from bringing in new participants. In other words, to continue to grow, the exchanges will have to offer a bigger and more diverse menu, not just put more seats around the table. The recent success of options would seem to confirm this point of view. Today, after the smoke has settled from various attempts by exchanges in New York, Chicago, London, and Singapore, there are six well established energy markets in New York and London: (1) NYMEX West Texas Intermediate (WTI) crude oil, (2) NYMEX natural gas, (3) IPE Brent crude oil, (4) NYMEX heating oil, (5) NYMEX unleaded gasoline, and (6) IPE gas oil. In addition, options contracts have

been successful as extensions of those future markets. The latest innovations by the exchanges include the introduction of trading in smaller contracts (the e-miNY at NYMEX), the expansion of electronic and after-hours trading, and the introduction of OTC trading and clearing services. Trading volumes of the major contracts are shown on Table III.

5. TRADING FUTURES: A PRIMER

Many readers of this article will be thoroughly familiar with the basic mechanisms and concepts of futures trading. This section is not for them. However, for those who are new to any type of futures trading, it is important to understand a few fundamentals about futures markets.

Futures markets offer hedgers or commercials (i.e., those who use a particular commodity in their business) and speculators the opportunity to buy or sell standardized contracts for a given commodity. In many cases, the same exchanges also offer options contracts on those same commodities. Options contracts as presently traded are options on the futures contract for the same commodity, which is often called the underlying futures.

5.1 Contract Identification

Both futures and options contracts are identified not only by the particular type of commodity being

TABLE III

Growth in Energy Futures and Options Trading (Daily Average Volume of Contracts Traded (000))^a

Contract	Exchange	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Futures																						
Crude oil																						
WTI	NYMEX	—	1.7	7.4	15.9	33.3	57.9	74.8	81.8	94.4	83.0	83.8	99.5	106.8	94.4	93.6	98.3	121.1	151.4	148.1	149.0	182.7
Brent	IPE	—	0.0	0.0	0.0	0.0	0.0	2.2	6.6	16.2	20.8	24.5	35.1	40.0	38.8	42.4	40.9	53.6	66.5	—	—	—
Brent	SIMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	0.5	0.2	0.1	0.1	—	—	—	—
MidEast crude	SIMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.1
Products																						
Heating oil	NYMEX	6.9	7.4	8.4	8.8	13.1	17.0	19.6	22.9	25.4	26.4	31.8	34.5	35.8	33.1	33.2	33.2	35.3	36.8	38.6	41.7	42.8
Gasoil	IPE	2.5	2.4	2.1	2.0	3.7	4.4	6.2	7.8	10.3	11.3	13.7	14.3	15.0	17.8	17.3	16.0	19.5	25.5	—	—	—
Unleaded gas	NYMEX	—	—	—	0.5	1.8	8.2	13.1	17.9	20.7	21.8	26.5	29.6	29.8	28.3	25.1	29.7	31.9	34.8	34.7	38.0	43.9
Resid	SIMEX	—	—	—	—	—	—	—	4.6	0.8	0.7	1.1	1.2	0.7	0.0	0.0	—	—	—	—	—	—
Propane	NYMEX	—	—	—	—	—	0.2	0.1	0.1	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1
Natural gas																						
Natural gas	NYMEX	—	—	—	—	—	—	—	—	0.5	1.7	7.6	18.7	25.3	32.3	35.1	47.3	63.6	76.6	71.8	47.5	97.4
Natural gas	IPE	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.3	1.3	1.1	—	—	—
Electricity																						
Palo Verde	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.1	0.6	0.6	0.2	—	—	—
California Oregon border	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.3	0.5	0.6	0.3	—	—	—
Cinergy	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.4	0.1	—	—	—
Entergy	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.4	0.1	—	—	—
Coal																						
Central Appalachian coal	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Options/Spreads																						
Crude oil																						
WTI	NYMEX	—	—	—	—	4.5	12.4	21.7	22.7	20.9	19.6	26.0	28.6	22.6	15.9	21.0	23.0	28.9	32.6	30.0	27.8	80.7
Brent	IPE	—	—	—	—	—	—	—	0.2	0.6	0.9	3.1	4.2	2.1	2.3	1.5	1.0	11.8	9.3	—	—	—
Products																						
Heating oil	NYMEX	—	—	—	—	—	1.1	0.5	1.2	1.6	3.4	5.0	3.2	2.8	2.8	4.4	4.8	2.6	2.8	5.6	4.2	2.4
Gasoil	IPE	—	—	—	—	—	0.0	0.1	0.1	0.4	0.4	0.8	0.9	0.5	0.5	0.4	0.3	7.4	3.5	—	—	—
Gasoline	NYMEX	—	—	—	—	—	—	—	1.3	1.7	2.3	3.4	2.6	2.3	3.1	2.6	4.1	3	2.4	4.1	4.5	2.9
Natural gas																						
Natural gas	NYMEX	—	—	—	—	—	—	—	—	—	—	0.3	1.4	2.0	3.7	4.9	8.3	10.3	15.4	21.4	15.7	43.9
Spreads																						
Gasoline–Crude oil spread	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	0.2	0.3	0.1	0.2	0.1	0.2	0.1	0.1	0.1
Heating oil–Crude oil spread	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	0.2	0.3	0.2	0.1	0.1	0.2	0.2	0.2	0.1
Electricity																						
Palo Verde	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.1	0.1	0.1	—	—	—	—
California Oregon border	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.3	0.1	—	—	—	—	—
Cinergy	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Entergy	NYMEX	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—

^a Annual averages, except for years on contract introduction, when averages computed on actual number of days traded. Minimal trading volumes entered as 0.1.

traded (e.g., heating oil, unleaded gasoline, Brent crude oil), but also by the delivery month called for in the contract. In practice, traders often abbreviate the names of months, so that it is common to hear references to “Feb Brent” or “Jan gas.” Options contracts are further identified by their strike prices and whether they are options to buy (call) or sell (put). Thus an options trader will talk about “Jan gas 55 puts,” meaning options to sell January unleaded gasoline futures contracts at 55 ¢/gallon. A buyer of a commodity contract is said to be long while he holds that contract. A seller is said to be short. The contracts traded are highly standardized with respect to volume, quality, and delivery terms. Exchanges do, however, change the terms of these contracts from time to time, to keep pace with changes in the physical market. The appropriate exchange can provide a copy of the latest contract terms.

5.2 Placing Orders

Except in the case of exchange members operating on their own account, all transactions must be conducted through a member of the exchange, who must also be registered to accept customer orders by the Commodities Futures Trading Commission in the United States or its counterparts in other countries. Orders can be placed anytime a broker is willing to answer the telephone, but exchange trading hours tend to fall between 9:00 a.m. and 5:00 p.m., with the New York exchanges closing earlier and the London exchanges closing later.

A buyer normally places an order by telephone to a broker, who may be located anywhere in the world. The broker in turn executes this order by telephone through exchange members on the floor of the appropriate exchange. Buyers can place various conditions on their orders, including price limits or time limits, and they may also simultaneously request a broker to close out the position if losses exceed a certain amount. Although brokers will generally accept such conditions on orders, they usually offer no guarantees that they can execute the order as given. Only a market order, in which the buyer (or seller) agrees to accept the prevailing market price, is virtually guaranteed for execution. Brokers will most often execute orders through the employees of their own firm on the floor. But in order to camouflage larger orders, execution will sometimes be shared with independent floor brokers, who execute orders on behalf of others.

On the exchange floor, all trading must be by open outcry, giving all present—at least in theory—an equal

opportunity to take the other side of the trade. Assuming a willing seller is then found to meet the buyer's order, the trade is posted with the exchange. In practice, exchanges publish price quotations over the various electronic information services to provide an up-to-date record of pricing trends even before the official record of the transaction is entered. The actual trade is usually entered into the exchange's computer within a few minutes of the transaction. The buyer and seller, however, are not matched permanently. At the end of the day, each broker is assigned an appropriate long or short position with the exchange's clearing house. Thus, although there must be an equal number of buyers and sellers each day, their respective positions are maintained totally independently, and each buyer and seller is free to close his/her position at any time. To do so, the buyer will simply sell his/her contract back into the market, effectively clearing his position from the exchange's books.

5.3 Spreads

In addition to straightforward orders to buy or sell a single commodity for a single month, many traders take spread positions, which are positions in several different contracts to profit from the relative price movements between those contracts. For example, spreads can be placed between contracts for different delivery months for a single commodity. Or they can cover different commodities for delivery in the same month. Or they can cover different commodities and different months. One popular type of spread position in the energy contracts is the “crack spread,” in which a position in crude oil is balanced against positions in both gasoline and heating oil, to approximate the refining process (in which crude oil is transformed by catalytic cracking into refined products). NYMEX has offered option contracts of such spreads since 1994. A newer type of spread trading is the “spark spread,” which pairs positions in natural gas with those in electricity to approximate the gross margin of power generation using natural gas.

5.4 Margins and Clearing Houses

Exchanges collect margins (or deposits, as they are called in England) from each broker on behalf of the customer (and in most cases, the broker in turn collects similar funds from the customer). These margins, which are usually in the range of 5–10% of the contract's total face value, are normally designed to be equal to the average daily fluctuation in value

of the contract being traded. Exchanges will therefore tend to lower margins in times of low price volatility and to raise them in times of high price volatility. Every night, based on the final closing or settlement price, the exchange calculates the effect of that price on each position and either requests additional margin or pays excess margin to each broker. Thus, if prices go up from one day to the next, a buyer's margin is credited with a gain and a seller's margin is debited. The rules of margin maintenance by customers to brokers vary considerably from country to country.

The exchange and its clearinghouse are therefore always in a very strong position to guarantee all outstanding positions. Moreover, the clearinghouse holds its member brokers—not the ultimate customer—responsible for performing under the contracts. In the unlikely event that an individual broker is unable to perform as called for under the contract, all the members of the clearinghouse are called upon to guarantee performance. Futures markets therefore offer several levels of financial performance guarantees.

Prices on an exchange are freely determined by the interplay of buyers and sellers. However, exchanges do place certain limits on both the minimum and maximum amounts of fluctuation that can occur in a given time period. The minimum price change is referred to as a tick and, in the oil contracts, is typically equal to 1.00 ¢/bbl, 25.00 ¢/MT, or 0.01 ¢/gallon in New York. In all cases, there are no limits on the spot contract, which is the contract that is next scheduled to go to delivery, but all other contracts face limits. In New York, these limits are typically \$1/bbl, \$15/MT, or 2.00 ¢/gallon. In a given day, no trades may take place outside these ranges. However, if a limit is reached on one day, the limits are expanded by 50% for the next day's trading, and so on, up to a maximum of \$2/bbl or 4.00 ¢/gallon. In London, the limits do not apply for a full day, but rather trigger cooling-off periods before trading is resumed.

5.5 Delivery

These markets should always be thought of primarily as financial markets, being used in parallel with physical movement of oil and natural gas. Nevertheless, delivery does take place and serves to ensure that the prices on futures markets remain closely linked to the real world. The standardization of contracts and their delivery terms are often unnecessarily rigid for the commercial participants, who prefer greater flexibility in their day-to-day opera-

tions. As a consequence, delivery typically occurs in only about 2% or less of all futures contracts. In simplest form, all those holding positions in a given contract at the closing bell on the last day of trading for a given contract are automatically required to take or make delivery of the specified commodity. The timing and methods of delivery are clearly spelled out in each contract and in the exchange's rules. The exchanges' staffs match buyers and sellers and the matched companies then must meet their respective obligations. Exchanges have found it useful, however, to permit several variations of this simple process. Prior to the exchange-matching process (typically the day after the end of trading), any two market participants may agree to an exchange for physicals (EFP), transferring title to oil (or natural gas) by mutual agreement in lieu of closing out their position on the exchange. EFPs can also be used to establish future positions by mutual agreement. In fact, in the U.S. crude markets, this mechanism is widely used as a routine means of buying and selling crude, because it has the attraction of the exchange's financial performance guarantees. A recent innovation, introduced by NYMEX, is an Exchange of Futures for Swaps (EFS), which adds additional flexibility.

Once trading in a given contract has ended and participants are matched, the two matched companies may elect to use an alternative delivery procedure (ADP), which also allows the two to make alternative arrangements. In the case of both EFPs and ADPs, the exchanges are relieved of any responsibility for guaranteeing performance.

5.6 Regulation

Exchanges are self-regulating and (since 2000) for-profit corporations, owned by their members/shareholders. The degree of governmental oversight and regulation has traditionally been most extensive in the United States (by the Commodities Futures Trading Commission) and least intrusive in the United Kingdom.

The exchanges maintain active compliance and market surveillance programs to enforce trading rules and to detect any evidence of market manipulation. Traders caught violating rules are typically fined, and in relatively infrequent instances, barred from trading. If evidence of market manipulation is uncovered, exchanges possess a wide range of powers to remedy the situation. These powers include the right to order a given participant to reduce or even eliminate his or her position, to substitute alternative delivery points or additional

supplies (i.e., by broadening quality specifications), or even to impose a cash settlement in place of physical delivery (assuming the contract calls for such delivery). These powers are not often used, but their very existence serves as a powerful disincentive to would-be market manipulators.

Perhaps the most controversial aspect of futures trading (particularly in the United States) is the permitting of dual trading; that is, allowing the same individuals to trade for their own account while simultaneously executing orders for customers as a floor broker. Many critics have argued that this practice provides opportunities for floor brokers to jump ahead of large customer orders, profiting from the market movements that those large orders are likely to provoke. Although such actions are a clear violation of exchange rules, detection is not always easy. Exchanges counter with the argument that dual trading promotes liquidity and that exchange enforcement activities are sufficient to prevent serious abuses. As periodic prosecutions in both Chicago and New York have demonstrated, there will always be temptations. Clearly, exchanges can improve their rules and surveillance. At a minimum, exchanges that want to allow dual trading have an obligation to create clear audit trails so that violations are easier to eventually detect. It seems likely however, that dual trading will be prohibited in futures trading as it is in securities trading. But the exchanges and their floor communities can be expected to resist this development until the bitter end.

6. FUTURE DIRECTIONS

The world oil market is relatively mature, but will continue to evolve as the supply/demand balances shift. On the supply side, this implies an increasing dependence on the Middle East and the Caspian Basin, creating pressures for the evolution of new marker crudes to replace the gradually diminishing supplies of WTI and Brent. On the oil products side, the inevitable growth of demand in the developing world will also challenge the traditional role of Rotterdam and the U.S. Gulf of Mexico as the pricing reference points. This may in turn increase the importance of Singapore and eventually lead to a more favorable market for the evolution of futures and options in that city.

Natural gas is still an emerging market. The growth of liquid natural gas (LNG) and interregional pipeline trade in the world should provide new opportunities for futures markets. Electricity con-

tinues to be a challenge for the futures exchanges. As of 2003, the attempts to launch and sustain electricity futures markets had been hampered by a variety of obstacles, especially the Enron scandal. But it seems likely that the exchanges will continue to seek opportunities to serve the electricity trading market. In addition, price volatility seems unlikely to retreat, so the need for risk management instruments, both on and off exchanges, will remain a constant in these markets for decades to come.

7. SUMMARY

The oil markets led the way for the entire futures complex. This was no accident. As the largest and most global energy source, oil is the natural leader. In addition, government regulations were first removed from oil before deregulation hit the natural gas and power markets. Although the natural gas and power markets have distinct characteristics, there is no doubt that the experience of oil markets in deregulation provides an important indicator of future directions for all energy markets. However, the rapid growth of e-commerce in natural gas and power has propelled those markets into a period of rapid growth, temporarily leaving oil the laggard. The explosive growth of electronic commerce is rapidly closing the gap between physical trading and futures. Within the first years of the 21st century, the convergence of wet and paper markets will be completed as the old "analog" (open-outcry) physical markets evolve to "digital" (Web-based) exchanges. At some point in the not too distant future, risk management tools will simply be a mouse click or a pull-down menu away, as companies and traders make their deals and fulfill them on the net.

Among energy markets, natural gas and electricity have leapt ahead to embrace e-commerce, leaving the oil patch to catch up. Ironically, the early deregulation of most national oil markets and the far-flung nature of global oil markets brought innovation to those markets in the 1970s and 1980s. However, the relatively quick deregulation of natural gas and electricity markets, the high level of market inefficiency spawned by decades of government economic regulation, and the greater homogeneity of product quality have encouraged those markets to move more quickly into the Internet era. Electronic exchanges blossomed in the late 1990s, and energy was one of the largest categories of business-to-business e-commerce. Electronic commerce will have a growing impact on the energy industry, including reducing

costs and improving service all along the supply chain, enhancing existing businesses, especially in the area of customer service, transforming existing business models, especially through the creation of “infomediaries” (companies that provide and analyze market information for other participants), and creating new businesses.

By lowering entry costs and facilitating access, Web-based exchanges can expand the range of energy products traded. The process has already started in the chemical industry, and Web sites (e.g., Chemdex, ChemConnect, and e-Chemicals) have been launched to compete for business. There will, however, remain two challenges for these sites. First, markets will remain focused on the liquidity of new products. Traders have notoriously short attention spans and will “tune out” if sites do not offer sufficient volume to move reasonable quantities quickly. Second, as these sites grow in importance and their impact on pricing grows, they will inevitably come under regulatory scrutiny, at least by the United States Government and possibly by the European Commission and other jurisdictions as well. The regulation (and taxation) of Internet transactions will likely be a major political and economic battle. Third, as these markets grow, there will be significant pressure to develop new risk management instruments as integral offerings. The growth of electronic trading should lower the hurdle rates for new risk management instruments. However, the launching of new futures and options instruments will always depend on their ability to attract liquidity early—as mentioned, traders are notoriously impatient and will not support illiquid markets. It is probable that the future instruments will therefore include a modest expansion of formal futures and options markets and a continued rapid expansion of off-exchange derivatives, perhaps with more standardized settlement terms and, eventually, common clearing mechanisms. The exchanges have realized the opportunity and NYMEX has been especially innovative in trying to capture more of these traditionally off-exchange activities.

Ideally, the market would like to have a complete suite of energy markets serving both the Atlantic and

Pacific basins, which would imply the development of additional futures and options markets for crude oil, products, gas, and electricity markets, especially in Asia. Obviously, the growth in interconnections of gas and power grids would tend to accelerate that trend, as would the continued expansion of international LNG trade. There seems little risk (or hope?) that volatility will soon depart the energy scene. And as long as there is volatility, there will be risk and a desire to manage that risk. However demand and production will shift, trade patterns will change, and product specifications will evolve, all requiring changes in contract design and continuous innovation.

Acknowledgments

Portions of this article originally appeared in *Energy Futures, Trading Opportunities* (John Elting Treat, ed.; PennWell Books, 2000). Excerpted by permission of the author.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Economics of Energy Demand • Economics of Energy Supply • Energy Services Industry • Inflation and Energy Prices • Innovation and Energy Prices • Oil Price Volatility • Prices of Energy, History of • Stock Markets and Energy Prices

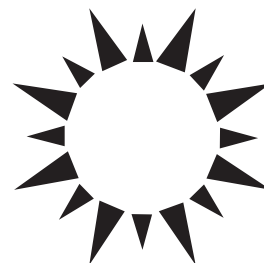
Further Reading

- Errera, S., and Brown, S. L. (2002). “Fundamentals of Trading Energy Futures and Options.” 2nd ed. PennWell Books, Tulsa, Oklahoma.
- Eydeland, A., and Wolyniec, K. (2002). “Energy and Power Risk Management: New Developments in Modeling, Pricing and Hedging.” John Wiley & Sons, New York.
- Fusaro, P. C. (ed.). (1998). “Energy Risk Management: Hedging Strategies and Instruments for the International Energy Markets.” McGraw-Hill, New York.
- Fusaro, P. C., and Wilcox, J. (eds.). (2000). “Energy Derivatives: Trading Emerging Markets.” Global Change Associates, Inc., New York.

Energy in the History and Philosophy of Science

ROBERT P. CREASE

Stony Brook University
Stony Brook, New York, United States



-
1. The Puzzle of Energy
 2. The Word Energy
 3. The Concept of Energy
 4. Energy Prior to 1800?

Glossary

argumentative analogy An analogy in which the secondary term is used to bring to bear an already-organized set of equations or network of concepts on the primary term, so that the primary term becomes a mode of the secondary.

constructivism The view that that nature is how we represent it and that the real depends on intersubjective agreement.

context of discovery The manner in which a theory or datum was historically discovered.

context of justification The manner in which a theory or datum is justified by reference to all the available evidence on hand.

corporeal template The use of the first-person experiences of a sentiently felt body to forge or model new concepts.

energeia A term in Aristotle's metaphysics constructed out of *ergon*, meaning act or deed, used to refer to the activity of tending toward or enacting a goal (literally, en-act-ment).

filtrative analogy An analogy in which the secondary term calls attention to certain of its conventionally understood features in the primary term.

paradigm A framework that provides the conditions, criteria, and examples for explaining problems and phenomena of a standard type.

realism The view that nature has a fixed structure prior to the way human beings represent it.

vis viva Living force, later formulated as mv^2 and thus an ancestor of the modern notion of kinetic energy, that some early modern philosophers argued existed in addition to quantity of motion or dead force (formulated as mv).

One important and controversial issue in the history and philosophy of science concerns whether the phenomena named by concepts precede the development of the concepts themselves. While realists claim that nature has a fixed structure and only concepts have histories, constructivists claim that, inasmuch as nature is how we represent it, natural phenomena are themselves histories. The history of energy sheds some light on this issue, exposing the ambiguities that lead to each position, by highlighting the significance of the fact that the environment in which scientists work is not simply a human-environment relation but one mediated by technology.

1. THE PUZZLE OF ENERGY

Few things in the history of science raise more hackles than the issue of whether the phenomena named by concepts precede the development of the concepts themselves. The natural and commonsensical realist assumption, which it is fair to say is shared by most people and nearly all scientists, is that nature has a fixed structure, and that while scientific concepts have histories—they arise, are transformed, and occasionally pass away—the phenomena to which these concepts refer do not. Conceptual evolution merely testifies to the weakness of the science of the past, and to the success of contemporary science at obtaining a better grip on the world.

The opposite constructivist extreme is the view that nature is how we represent it, and the real depends on intersubjective agreement. Sociologist Bruno Latour, for instance, has provocatively argued that things, not just words, have histories; has claimed that microbes (and not just the concept of a microbe) did not exist before Pasteur; and has remarked that Pharaoh Ramses II could not have

died of tuberculosis (as now thought) because the bacillus had no real existence before Robert Koch discovered it in 1882.

The history of energy is an intriguing episode in the history of science that sheds some light on this issue. It does so because the phenomenon of energy nicely illustrates the presence of factors such as the technological transformation of the world, and even the role of more metaphysical considerations such as the changing patterns of thought that sometimes have to occur for a concept to be taken as applying to nature.

2. THE WORD ENERGY

The word energy has a well-charted and uncontroversial history. Energy comes from the Greek *energeia*, or activity. For practical purposes one may say that the word was all but invented by Aristotle, who treats it as a noun form of a nonexistent verb. The word was constructed out of *ergon*, meaning act or deed; *en-erg-eia* might be literally translated as “en-act-ment.” In his substance ontology, Aristotle uses it to refer to the activity of tending toward or enacting a goal, a meaning much different from today’s. Aristotle saw bodies as composites of matter (*hyle*) and form (*eidos*). But for him, matter was only the possibility of a body, with no inherent structure and no perceptible or corporeal properties of its own—only a body can be soft or hard, hot or cold—and obtains a structure only via its form. Inextricably related to this distinction is another related to a body’s activity. Every existing thing, he said, has an *energeia*, which maintains it in being and is the activity of tending toward or enacting a goal, its end or function (*telos*) to which it is naturally fitted. Aristotle has various names for how that activity is related to its end. A body’s sheer potential or capacity for action he calls *dynamis* (potency), while its being at work en route to or at its *telos* he calls its *en-ergeia*. Every existing thing, then, has some activity or *energeia*, which maintains that thing in being. Aristotle seems only to have sensed, but not explicitly described, a difference between brute mechanical power and accumulated energy, or what we might call force and energy.

As philosophers such as Stephen Toulmin have emphasized, Aristotle’s view was a product of the everyday situations with which Aristotle was specifically concerned. When he did concern himself with situations involving what we would call force and energy, they were mechanical horse-and-cart like situations in which a body (such as a horse) moves

against a resistance (sources such as friction) to keep a body (cart) in motion. Aristotle compared other situations he was interested in to this type of example in formulating explanations of them. His analysis was therefore based on what Stephen Toulmin rather loosely (and prior to Thomas Kuhn’s more elaborate and technical account) called a particular paradigm. It would not have occurred to Aristotle, Toulmin notes, to explain a familiar horse-and-cart situation with reference to “idealized, imaginary states-of-affairs that never in practice occur, ideals to which even the motions of the planets can only approximate.” Unless we understand this context in which Aristotle was making and justifying explanations, we not only misunderstand him, but risk coming away thinking that he was guilty of elementary, unscientific blunders.

Aristotle’s conception of *energeia* is, of course, worlds away from the modern conception of energy as the sheer capacity to do work. *Energeia* is not found in a specific place, but suffused everywhere and in different forms. It cannot be created and stored, converted or dissipated; it has a directionality toward forms to be realized and cannot be detached from these ends. The notion that *energeia* could be set up independently over and against the objects on which it does work, or stored apart from them, or collected or harnessed, would not have made sense to Aristotle.

Energy, the eventual English anglicization of *energeia*, came to lose the technical meaning Aristotle had given it and, for a while, virtually any value as a scientific term. In the mid-18th century, philosopher David Hume, for instance, complained that the words power, force, and energy are virtually synonymous, and “obscure and uncertain.” As late as the seventh edition of the *Encyclopaedia Britannica* (1842), the word only merits a short entry: “ENERGY, a term of Greek origin, signifying the power, virtue, or efficacy of a thing. It is also used figuratively, to denote emphasis of speech.”

3. THE CONCEPT OF ENERGY

How, then, did the word come to refer to a precise technical concept? The word would become a technical term only by being linked, at the beginning of the 19th century, with a concept that itself had undergone a complex evolution, and this evolution was the product of a number of factors.

One factor was a metaphorical extension of the subjective experience individuals have of themselves as a bodily center of action, which was then projected into nonhumans as one of their properties. This

illustrates what philosopher Maxine Sheets-Johnstone calls “the living body serv[ing] as a semantic template,” a process which, she points out, is key to the emergence of many early scientific concepts. It is a classic case of the use of analogical thinking, or the use of the familiar to understand the unfamiliar. What is remarkable in this instance is that the familiar has its basis in the tactile-kinesthetic experiences of bodily life; thus, in a corporeal template.

Yale science historian Stanley Jackson, for instance, has written about how such a subjective experience was analogously extended into the concepts of early modern physics in the works of scientists such as Johannes Kepler and others. Kepler, like many scientists of his time, analogically projected a secular version of a soul-like animistic force into his mechanics. “If we substitute for the word ‘soul’ the word ‘force’ then we get just the principle which underlies my physics of the skies,” Kepler wrote, adding that though he now rejected such souls, he had come “to the conclusion that this force must be something substantial—‘substantial’ not in the literal sense but... in the same manner as we say that light is something substantial, meaning by this an unsubstantial entity emanating from a substantial body.”

What, then, was this unsubstantial entity? In the 17th century, this question sparked a furious metaphysical and scientific controversy on the existence, nature, and measure of force. René Descartes spoke of quantity of motion (mass times velocity, or mv), while Gottfried Leibniz argued that the force was not just a quantity but a quality of matter, called living force or *vis viva*, whose proper measure was mv^2 . The *vis viva* controversy, as it was known, continued through the 18th century and was the subject of young Immanuel Kant’s first essay. In 1807, Thomas Young, lecturing to the Royal Institution on “Collisions,” said that “the term energy may be applied, with great propriety, to the product of the mass or weight of a body, into the square of the number expressing its velocity.” He thereby tied the word, apparently for the first time, to its modern concept. But Young’s “energy” was not ours, only what we call kinetic energy, and not even our formulation of it ($\frac{1}{2} mv^2$).

The first half of the 18th century witnessed an intense discussion of the proper way to understand notions of activity, work, force, and *vis viva* among engineers and scientists including James Joule, Sadi Carnot, Hermann von Helmholtz, James Mayer, Emmanuel Clausius, William Thomson, and others. One of the most important motivations for this activity was to achieve a formulation adequate to the

phenomena that scientists and engineers of the day were occupying themselves with: most notably, steam and heat engines, which were revolutionizing the engineering and industry of the day. In particular, scientists and engineers were concerned with the question of how to evaluate and measure the efficiency of steam and heat engines.

A second, closely allied factor involved what might be called a metaphysical shift, which opened up the possibility of seeing phenomena that look and act very differently as but different modes of one force. Thomas Kuhn emphasizes this point in an article on energy conservation, saying that such a metaphysical shift helps to explain why “many of the discoverers of energy conservation were deeply predisposed to see a single indestructible force at the root of all natural phenomena.” The metaphysical shift in question itself may perhaps be specified more fully as the working out of another kind of analogical thinking, in this case, the use of an extended argumentative or scientific metaphor. Analogies and metaphors come in many different kinds, one common form being filtrative. In filtrative analogies or metaphors (“man is a wolf,” “love is a rose”) the secondary term (wolf, rose) calls attention to certain conventionally understood features of these things in the primary term (man, love). The focus is therefore on the primary term, and the secondary term only provides an assist. In argumentative or scientific analogies or metaphors, on the other hand, the priority of the two terms is reversed, and the secondary term is used to bring to bear an already-organized set of equations or network of concepts in order to understand the primary term—and the secondary term grows into the salient, technically correct term, often expanding its meaning in the process, with the primary term being merely one mode of the secondary. One classic illustration of an argumentative or scientific analogy occurred when Thomas Young and others began to call light (the primary term) a wave (the secondary term). Waves originally referred to a state of disturbance propagated from one set of particles in a medium to another set, as in sound and water waves. When Young and other early investigators called light a wave, they assumed that it, too, moved in some medium—but this proved erroneous, and scientists were soon saying that waves propagated in the absence of a medium. Despite this change, the same concepts and equations nevertheless governed light, which was now just another mode of an expanded conception of wave. Wave had grown into the technically correct term for characterizing light,

though in the process it became transformed and expanded (and the evolution would continue when waves turned up in quantum mechanics).

This kind of argumentative or scientific metaphor was at work in the beginning of the 19th century, in allowing engineers and scientists to see the presence of energy in quite different phenomena, from falling water and expanding air to heat and electricity. Only the kind of metaphorical sensitivity allowed the early investigators to realize the significance of the conversion processes that were so critical to the development of the concept of energy. It was doubtless key in allowing William Thomson, in a famous footnote (1849) to a work addressing a dispute between Joule and Carnot over conceptual and experimental problems in the heat formation of heat engines, to intimate that something involving both work and *vis viva* was conserved, though not yet visible “in the present state of science.”

The specific changes in state that then made the conserved something visible are discussed at length in other entries. If any date marks the emergence of this new conception of energy into visibility, it is 1868, with the publication of the mechanics textbook, *A Treatise on Natural Philosophy*, by William Thomson and Peter Tait. The full articulation of this insight, involving the recognition that heat was energy—and only one of many forms—revolutionized science, and by the time of the ninth edition of the *Britannica* (1899), the entry on energy was six pages long and full of technical terms and equations.

The factors just mentioned are extremely significant for the history and philosophy of science, for they illustrate an inversion of the traditional conception of the relationship between theory and practice, as well as the fact that the environment in which scientists work is not simply a static human-environment relation, but a dynamic, ever-changing one mediated by technology. The practical coping with technologies can precede and even help give birth to theoretical reflection. Technologies therefore can become a kind of philosophical springboard for scientists: familiarity with them reshapes the theoretical imagination, allowing scientists to perceive and project new kinds of concepts or extend them to larger terrains.

4. ENERGY PRIOR TO 1800?

These developments help to shed light on the puzzle, mentioned at the outset, concerning whether the phenomena named by concepts precede the develop-

ment of the concepts themselves and how the two vastly different positions mentioned can arise.

The puzzle is posed here by the question: Was there energy prior to 1800? This question would surely make most people, including and especially scientists, roll their eyes. Such people would interpret the question as asking whether the phenomenon named by the word existed. In this commonsense view, the answer is surely yes. Energy, after all, was not discovered or invented. It has powered the sun for billions of years, made organisms grow for millions, and driven industrial machines for hundreds. Our opinions, ideas, and representations of energy may have evolved, but energy apart from these—what we are trying to get at through them—does not. But if the question is interpreted as asking whether the abstract notion of a capacity to do work that comes in various forms convertible into each other existed prior to 1800, then the answer to the question is just as surely no.

Traditional philosophers of science would try to resolve the puzzle by appeal to the extremely influential distinction between the context of discovery, or the manner in which a theory or datum was historically discovered, and the context of justification, or the manner in which that theory or datum is justified by reference to all the available evidence on hand—and say that the convoluted, fortuitous, and often idiosyncratic process by which a discovery or concept enters science does not matter. There is some truth to this. But a term like energy cannot be justified in a bare environment, but only within a network or horizon of other historically given concepts (force, work, heat, etc.) and technologies (steam engines, instruments to measure temperature, etc.). These phenomena lead us to interpret other things in their light, changing our view of world and its findings, structures, and laws, leading us to project them as elements of the world—past, present, and future. This does not make scientific concepts arbitrary, an artifact of language or culture, but nor does it legitimate us thinking that these concepts refer to things apart from time and history.

The various interpretations of the question arise because the formulation of concepts such as energy is indebted both to a set of purely theoretical considerations and to a practical, technologically shaped world. This latter world is constantly changing, giving rise to new kinds of phenomena to be explained and related to phenomena that already appear. The network of theory and technologically shaped world forms a horizon in which scientific claims can be tested and be judged true or false.

If one emphasizes the values which permeate the horizon at the expense of the theoretical considerations, and the dependence on them that the invariance that phenomena such as energy show—if one emphasizes let us say the role of the development of steam and heat engines, the social and economic interest in creating more efficient types of such engines, and so forth; and if one emphasizes that a concept such as energy could not have emerged without such interests—it promotes a position such as Latour's.

On the other hand, if one emphasizes the theoretical considerations at the expense of the practical and technologically rich horizon thanks to which phenomena such as energy can show themselves as being invariant in myriad guises—if one emphasizes let us say the tightness and interdependence of the connection between concepts such as energy, power, work, and so forth—it promotes the view that science represents an ahistorical reality apart from a worldly context. But because the horizon is relatively stable, we tend to forget that it, too, can change, and that we justify things only in it.

So was there energy before 1800? If this means an abstractly defined concept of a quantity conserved in closed systems and related in precise quantitative ways to other concepts like power and work, no; if this means this is how we currently understand the world and its phenomena, yes. The history of the term energy is an episode in the history of science that helps to reveal the complexities of the history of science.

The scientific tradition as it has developed up to the present has shaped the contemporary notion of energy such that everything is a modality of energy. In cosmology, for instance, starting with radiant energy, some radiant energy becomes particles (thus

mass energy), some particles become atoms, which in turn become molecules and condensed matter, and so forth. Energy is, as it were, the dynamic existential *materia prima* of the universe. This has both a theory side that gives to the present an illusion of wild nature that exceeds all representations and is beyond human creation, and a practical, representable, and controllable side that gives to the present an illusion of being merely socially constructed.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • *Early Industrial World, Energy Flow in* • *Economic Thought, History of Energy in* • *Forms and Measurement of Energy* • *Geographic Thought, History of Energy in* • *Thermodynamic Sciences, History of* • *Work, Power, and Energy.*

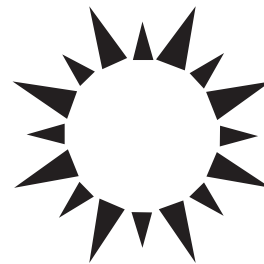
Further Reading

- Crease, R. (2002). What does energy really mean? *Physics World*, July.
- Jackson, S. (1967). Subjective experiences and the concept of energy. *Perspect. Biol. Med.* 10, 602–626.
- Kuhn, T. (1977). Energy conservation as an example of simultaneous discovery. In “The Essential Tension: Selected Studies in Scientific Tradition and Change” (Thomas, Kuhn, Ed.), pp. 66–104. University of Chicago Press, Chicago.
- Schönfeld, M. (2000). The *vis viva* debate. In “The Philosophy of the Young Kant” (M. Schönfeld, Ed.), pp. 17–35. Oxford University Press, New York.
- Sheets-Johnstone, M. (1990). “The Roots of Thinking.” Temple University Press, Philadelphia.
- Toulmin, S. (1961). “Foresight and Understanding.” Indiana University Press, Bloomington, IN.

Energy Ladder in Developing Nations

RICHARD H. HOSIER

United Nations Development Program (UNDP)
New York, New York, United States



-
1. Introduction
 2. Conceptual Foundation
 3. The Energy Transition: Drivers and Constraints
 4. Empirical Evidence for the Energy Ladder Transition
 5. Criticisms of the Concept
 6. Conclusions

Glossary

commercial fuels Fuels for which every step of the supply chain involves a purchase or sale involving a monetary transaction.

dung Animal fecal matter—typically from cattle, buffalos, or camels—that is frequently burnt as fuel.

end use The final purposes to which energy is put. In the household sector, the end uses can also be called energy services and are typically considered to be cooking, lighting, and heating.

fuelwood Firewood, or woody biomass burnt as fuel.

interfuel substitution The substitution of one fuel for another in carrying out the same end use or fulfilling the same need.

logit analysis A statistical technique intended to predict the choice that would be made between a variable with only two possible values. Multinomial logit analysis is a multiequation version of logit analysis designed to predict the choice between a number of different options.

modernization The process by which a society becomes modern, current, or up to date. It involves economic development, industrialization, and urbanization among other forces.

noncommercial fuels Typically, traditional fuels that are not bought or sold in the marketplace.

Tobit analysis A statistical technique named for the economist James Tobin who developed it to factor into regression analyses dependent variables where a large portion of the sample being used had a value of zero (0) for the dependent variable.

traditional fuels Fuels that have been used by households since time immemorial (such as firewood or crop

residues or dung). They are frequently gathered by individuals but are occasionally bought and sold in markets.

The energy ladder is a concept that has been used to describe the transition of households from utilizing traditional energy carriers for their energy service needs to utilizing more modern, technologically sophisticated energy carriers to meet those needs. It draws an analogy between household fuel choice and a ladder. Households using traditional fuels—such as firewood or dung—are assumed to be on the lower rungs of the ladder. Households using modern, commercial fuels—such as LPG, natural gas, or electricity—are assumed to be poised on the upper rungs of that ladder. Although this transition is of most concern to households in developing countries, a similar process is assumed to have taken place in developed countries as they went through the process of industrialization. The idea is that as a household unit's status changes—either through an increase in income, a shift in status, or a change in the society around it—it will begin to utilize a different mix of energy carriers to perform its energy needs. A household that chooses to begin utilizing kerosene or liquefied petroleum gas (LPG) for cooking instead of fuelwood is said to be moving “up” the energy ladder. In contrast, a household unit that makes the reverse decision—shifts from utilizing kerosene or LPG to using fuelwood—is said to make a move “down” the energy ladder.

1. INTRODUCTION

Since it became used commonly in the early 1980s, this analogue model has become a tool to studying

the energy-use patterns prevalent in developing countries. Researchers and policy analysts have used the concept to better understand household fuel choice and to better design energy policies by anticipating their impacts on energy-using households. Typically, poorer households and rural households have limited access to energy resources and fuels and occupy lower rungs on the energy ladder. Wealthier households and urban households generally have greater access to energy resources and occupy higher rungs on the energy ladder.

The same distinctions that have been made initially at the household level can also be applied at the level of a region or a nation. In fact, underlying much of the conceptual background on the energy ladder at the household level is the understanding that as societies as a whole develop, they not only begin to utilize a greater quantity of energy (as measured in calories, toe, or joules) but their energy mix consists of a greater percentage of modernized fuels, such as petroleum, natural gas, and electricity. While this direct analogue of the individual household unit to the national level may introduce certain fallacies of composition, as individual energy consumption patterns sum up to provide a national energy consumption pattern, the individual-national analogue is quite reasonable: national energy use patterns do reflect, *inter alia*, the sum total of the individual energy-use patterns of their citizens.

As most research on the energy ladder in developing countries has focused on either confirming the existence of the household-level energy transition or specifying its form in various contexts, the household level discussion will form the focus for most of this article. However, when the discussion reverts to the use of cross-national comparisons and analyses to generalize using national-level data, the jump from the household or agent level to the national level will be clearly delineated.

2. CONCEPTUAL FOUNDATION

The energy ladder concept is built on the idea that every fuel that can be used by households has its own set of advantages and disadvantages depending on the various end uses in which it is deployed. Modern fuels are considered to have more advantages than traditional fuels and are therefore considered to be higher rungs on the ladder. In developing countries, the primary end uses or energy services that households need to fulfil are cooking, space heating, water heating, and lighting. Household decision making

then must consider the qualities of all the available energy sources in carrying out these end uses in order to arrive at the final household decision regarding which fuels to use for which purposes. In many instances, the decisions are quite simple as the opportunity set facing the households is limited. In other instances—notably the peri-urban and urban areas of developing countries—households face a large range of options that make the household energy use decisions quite complex.

Leach argued that our understanding of this energy transition is based on energy surveys in developing countries that confirm “that most families have an ideal fuel preference ‘ladder’ running from biomass fuels up through kerosene to bottled gas (LPG), natural gas where it is available and electricity” (p. 118). Traditional fuels, such as firewood and dung, are relatively inconvenient to collect and store, require constant management while in use, and emit large quantities of smoke that has a detrimental impact on the health of the users. Modern fuels, such as natural gas and electricity are more convenient, cleaner to use, have power outputs that are easily controlled, and can be delivered directly to the kitchen (once the basic transmission and distribution infrastructure is built). The energy ladder or energy transition embodies the idea that as household income rises through the development process, they choose to utilize modern fuels and choose not to utilize the traditional fuels.

Each household makes energy-use decisions based on its unique circumstances, its preferences, and its environment. When aggregated across households, the resulting observed energy consumption patterns can become quite complex. In general, lower-income groups within a given area tend to rely more heavily on traditional biomass fuels for cooking. If they are able to afford lighting services at all, they will typically use kerosene or candles for lighting. Wealthier groups within that same rural area may also rely on traditional biomass fuels for cooking, but may supplement that biomass with kerosene or liquefied petroleum gas (LPG). For lighting needs, wealthier households may rely on electricity if their household is connected to the grid, but if not, they will also utilize kerosene, LPG, candles, or even battery-powered flashlights or torches. Thus, poorer and wealthier households in both urban and rural areas of developing countries demonstrate different energy use patterns resulting from their different decision made with respect to their energy consumption patterns.

Figure 1 illustrates diagrammatically the range of end uses and fuels faced by households at different

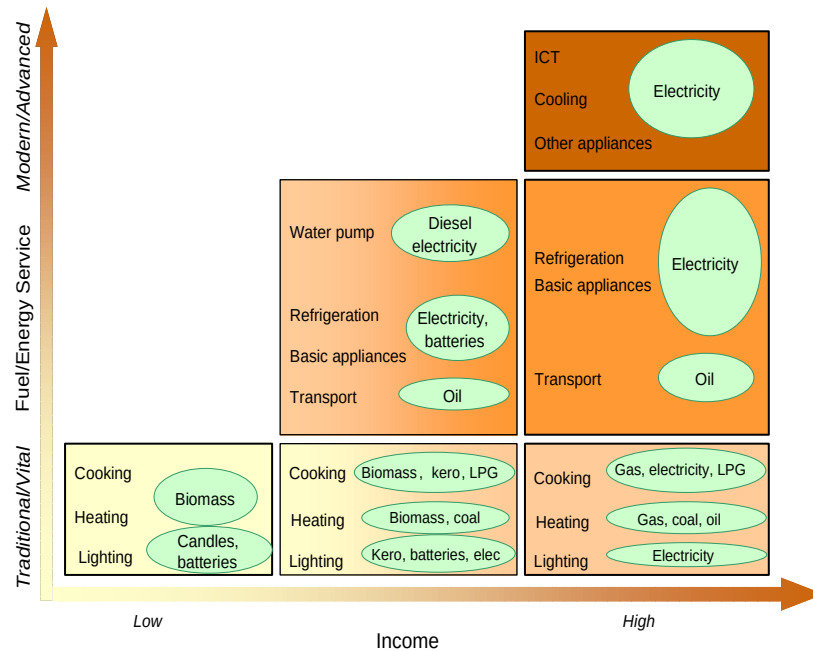


FIGURE 1 End uses and fuels faced by households at different income levels. From World Energy Outlook (2002), International Energy Agency (IEA). Used with permission.

income levels in developing countries. The horizontal axis represents income and the vertical axis represents the fuel used to fulfill an energy service need, moving from the most traditional to the most modern. As represented in this diagram, all households at all economic status levels must utilize energy for cooking, heating and lighting. In poorer households, the options normally used for these purposes are traditional biomass for cooking, and either candles, kerosene, or batteries for lighting. In middle-income households, the options to fulfill these services expand to include biomass, kerosene, and LPG for cooking; biomass, coal, or charcoal for heating; and kerosene, batteries, and candles for lighting. However, at middle-income levels, another set of end uses or energy services begin to be considered. These additional end uses include refrigeration, transport, basic appliances such as radio and television for entertainment, and even water pumping. The fuel options available for providing these services are petroleum-based fuels, batteries, and electricity. Finally, for upper-income households, the options for cooking extend to natural gas, LPG, and electricity, and the options for heating may include gas, coal, oil, or electricity. At this higher income level, electricity will provide in all likelihood provide all of the lighting needs. At the highest income level, electricity will likely provide all of the refrigeration and appliance needs while petroleum-fuels continue

to fill the transport needs. In addition, the higher-income households begin to pursue a range of additional energy needs, such as air conditioning, for which the only option is electricity.

Thus, as household incomes increase, the household tends to move “up” the energy ladder to more sophisticated fuels. Not only does the choice of fuels change, but the absolute quantity of energy effectively utilized by the household increases as it begins to pursue a broader range of energy end uses or energy services. For nations as a whole, as the economy develops and the population becomes wealthier, its citizens begin to use a larger share of modern fuels, especially electricity, to provide a larger number of energy services. This results generally in a larger consumption of energy at the national level—hence, the well-documented traditional link between energy consumption and gross national product.

3. THE ENERGY TRANSITION: DRIVERS AND CONSTRAINTS

The resource allocation decisions resulting in a household’s fuel-use patterns and, therefore its position on the energy ladder, are determined not just by income but also by a large number of other

factors. Because these factors influence the household's decision-making process, they also are considered to be important either as policy levers or as indicators of policy impact. In the following section, the factors driving or constraining the household's energy-use decisions are briefly discussed with respect to both their impact on the energy ladder decision and their importance as policy measures or as indicators of likely policy success. Some of these may be considered to be supply considerations; others may be characteristics of the fuels; and still others may have to do with the either the physical or economic environment faced by the household in a particular market.

3.1 Household Income

As has already been discussed, household disposable income is one of the most important factors determining household energy decisions in all countries, both developing and developed alike. In general, total energy use is positively correlated with income. In the context of the energy ladder discussion, wealthier households will show a greater predilection to using more sophisticated fuels than will poorer households. Expressed in microeconomic terms, the income elasticity of demand for most modern fuels (electricity, natural gas, LPG) is positive whereas for traditional fuels (over a wide range of incomes) it tends to be negative.

Income becomes an important policy discussion concerning household energy use. As incomes increase, so will the demand for modern fuels, other things being equal. Likewise, households with falling incomes are liable to move down the energy ladder to more traditional fuels. Perhaps a bigger concern for policymakers in developing countries revolves around the energy needs of the extremely poor household—that is, households living on less than \$2 per day. Whether or not the incomes and other resources available to these households will enable them to obtain adequate energy remains a concern for energy policymakers. Unfortunately, in such cases the policy options may be so limited that policymakers are unable to influence the energy mix used by these extremely poor households.

3.2 Urbanization

Urbanization is an important indirect influence on a household's energy-use decisions. Households living in urban areas face a different opportunity set than those living in rural areas in two major respects.

First, their ability to collect firewood or traditional fuels from their surrounding will tend to be more limited than that of their rural counterparts. Second, because their ability to engage in subsistence agriculture is more limited than that of their rural counterparts, they require a steady cash income from paid employment to be able to meet their budgetary needs. As a result, their monetary income tends to be higher than that of rural households and they will demonstrate a higher reliance on commercial or semicommercial fuels than their rural counterparts. In addition, urban infrastructure is normally more advanced than rural infrastructure meaning that households in urban areas tend to have better access to the electric or natural gas distribution grids as well as to markets for LPG bottles, kerosene, and even commercialized traditional fuels, such as charcoal. Therefore, an urban household will tend to occupy a higher rung on the energy ladder or use a larger share of modern, commercialized energy sources to meet its energy needs than will a rural household of comparable status.

Urbanization is an important consideration when discussing household energy policy and situation with respect to the energy transition. In most developing countries, more policy options have relevance to urban areas because all urban households are engaged in market activities for their livelihoods, supply-focused options are easier to design and implement, and price controls or subsidies are easier to enforce. For example, electrification of urban households or the extension of a natural gas distribution grid to urban households tends to be a far more cost-effective policy option to enhance household access to modern sources of energy sources than is applying those same policies to rural populations. In fact, given the low population densities of most rural areas in developed countries, grid extension as a policy intervention is often too expensive and cost-ineffective a policy option to consider pursuing at all.

3.3 Smoke Emissions and Exposure

One of the reasons that households in developing countries are presumed to prefer modern fuels over traditional ones is that the former are cleaner, in terms of both storage and handling and emissions. Traditional fuels, such as dung, crop residues, and fuelwood, emit noxious fumes when burned contributing respirable particulates to the household environment. Many households in developing countries burn these fuels for cooking indoors and often have no effective

means of venting the emissions from these fuels. This leaves the cooks—typically women—and other household members exposed to extremely high levels of particulate concentrations. Smith has documented that the rural indoor environment in many developing countries may represent the world's single most pervasive and intensive air pollution problem in terms of actual human exposure to particulates. In terms of total human exposure to pollutants, exposure to biomass (or coal) smoke from indoor cooking in developing countries poses a far more serious health threat than second-hand tobacco smoke or any source of ambient industrial emissions. Evidence from developing country households shows that concentrations of respirable particulates can run from 500 to more than 15,000 $\mu\text{g}/\text{m}^3$ in many indoor environments in developing countries. This level of pollution concentration from the indoor use poses very serious health threats to the exposed population. Although the medical records linking exposure to disease are rather weak in most developing countries, acute respiratory infections in children and chronic obstructive lung disease and *cor pulmonale* in adults exposed to smoke from biomass cooking have all been linked to exposure to cooking smoke in developing countries. Although other anticipated health effects from exposure to biomass cooking smoke—such as adverse pregnancy complications and increased cancer rates—have not yet been sufficiently documented through medical research, they must be assumed to be significant.

Smoke pollution clearly provides a significant incentive for households to move away from using these dirty fuels—such as firewood and coal—toward using more modern, cleaner fuels with fewer negative health impacts. Although many rural households and poorer urban households in developing countries may have no alternative to burning biomass fuels or coal in unventilated cooking devices, for those who do have some alternative, it provides a significant incentive to move “up” the energy ladder to cleaner, more modern fuels. For policymakers, the likely negative health impacts of continued reliance on these traditional fuels should create an incentive to find cooking alternatives that do not threaten the health of women and children in developing countries. One logical policy response might be to devise programs to encourage the use of LPG, natural gas, or electricity for cooking in areas where they are available and affordable to the local population. Government policies to support moves “up” the energy ladder in most developing countries could certainly be justified on the basis of the negative health impacts of continued traditional fuel use.

3.4 Appliance Costs

One of the obstacles to adopting modern fuels is the fact that a specialized piece of equipment—an appliance—is required to utilize them. The cost of these appliances to use more modern fuels may prove to be prohibitive for many poorer households in developing countries. Such costs serve as a transaction barrier to the adoption of more modern fuels by households in developing countries. The cost of LPG bottles, kerosene cooking stoves, LPG stoves, or electrical stoves may represent a major cost obstacle to households adopting and utilizing the more modernized fuels instead of the traditional biomass-based fuels found on the lower rungs of the energy ladder. For many households accustomed to cooking with wood over a traditional three-stone hearth, the additional cost of a fuel-efficient woodstove may prove prohibitive, even in cases where the fuel savings more than justify the additional costs.

Policies may be devised to assist some households in overcoming the financial barrier to purchasing appliances for modern fuels. While financing might help households amortize appliance cost over a longer period of time, such payments will still increase the costs of the more modern fuels, placing it at a disadvantage to the traditional fuel for which no improved appliance is required. Some governments have tried subsidizing or providing tax credits for the purchase of appliances designed to utilize modern fuels, thereby encouraging households to adopt and use the more modern fuels. Several African countries claim to have successfully encouraged a move away from fuelwood and charcoal by subsidizing LPG bottles and cooking devices while encouraging suppliers to provide better distribution networks for this relatively clean and modern fuel. These represent some of the few cases of desirable moves up the energy ladder that have been stimulated by deliberate government policy.

3.5 Relative Fuel Costs and Prices

Relative prices per delivered energy unit play a crucial role in determining what fuels a household will use as well as how much of those fuels to use to meet its own energy needs. In the case of traditional fuels, such as fuelwood, the costs may be measured in terms of the time spent gathering the fuel rather than in monetary units. How a household may choose to make this trade-off—between labor and monetary resources—involves a complex calculation made more difficult because of the different roles found

within the household unit: women may contribute the labor to gather the firewood, but men often control the money required to purchase the fuels. In general, the relative fuel costs and prices will be reflected in the household energy use patterns and will determine where on the energy ladder a household determines itself.

Policymakers have frequently utilized price to steer household energy use decisions. The most common case involves subsidized electricity prices: governments have frequently regulated electricity prices to be less than their actual costs in order to encourage its use and obtain favor with the electricity-consuming populace. By subsidizing connection fees and controlling energy charges at levels below their actual costs, governments encourage households to utilize more electricity than is economically sustainable. The net result is that more households move up the energy ladder than should, consuming more electricity than they can actually afford to purchase. In such cases, the electricity utility's balance sheet suffers.

A second example of governments using prices would be the kerosene subsidy. Many governments have sought to control by regulatory fiat the price at which kerosene can be sold in the market, either through price or controls a cross-subsidy provided from a surcharge on other petroleum fuels. While the goal of such a policy is admirable as it is intended to ensure that poorer households have access to a relatively modern and clean fuel—kerosene—for lighting and cooking, it also has proven to be problematic. Studies have shown that most of the benefits of these subsidies accrue to middle and upper-income families, not to the poor. In addition, the surcharge may be administratively cumbersome and the controlled price quickly becomes unsustainable. The costs of such policies frequently outweigh the benefits and governments should be aware of this prior to implementing such systems.

A strikingly different example of using pricing policy to shape household fuel-use decisions is the case of the stumpage fee imposed on trees being harvested for fuel use. Many governments in Africa have been convinced that there is inadequate control over the harvesting of firewood for sale to households and for the production of charcoal for sale to peri-urban and urban households. The proper response in such cases is to impose a tax or a stumpage fee to compensate for the fact that the harvester has not paid for the planting, protection, and caring for the tree over its lifetime. If a government imposes such a stumpage fee either at

the point of harvest, the point of production, or the point of importation into an urban area, it has two relevant effects. First, it raises funds to be able to replace the tree. Second, it serves to raise the price of the fuel creating an incentive for the household to either utilize less wood or charcoal or to find another alternative, one either lower or higher on the energy ladder. Thus, pricing policies can stimulate movements either up or down the energy ladder. Such movements may be intentional or unintentional, depending on the situation.

3.6 Fuel Availability

Perhaps the single most important supply-side determinant of household fuel choice for traditional fuels has to do with fuel availability. While the availability of fuel sounds like a relatively straightforward matter to determine, it actually becomes quite a complex matter to evaluate in any given location. It may be determined by the forest cover, the predominant land-use system, the nature of land tenure system, cropping patterns, and the distance traveled to collect fuel. For rural households accustomed to using firewood, even convenient surrogates such as the distance traveled to collect firewood or the amount of time required to collect a headload may be a misleading measure of fuel availability. While much of the literature on the household energy use in developing countries originally emerged in response to concerns regarding increasing deforestation and the “fuelwood crisis,” scientific work has delinked fuelwood use by rural households from deforestation. This does not mean, on the one hand, that in localized areas, households do not experience a severe shortage of fuelwood for domestic purposes or, on the other hand, that fuelwood use has never contributed to tree harvesting that is detrimental to the environment. Rather it means that energy and forest planners must maintain balance in their view and proposed plans for rural resource management and needs. Fuelwood availability, or rather fuelwood's limited availability in certain rural areas of developing countries, may create an incentive for households to choose to utilize fuels either higher (i.e., kerosene, LPG) or lower (crop residues, dung) on the energy ladder.

For those fuels that are lower on the energy ladder than fuelwood, such as crop residues or dung, they present their own challenges of availability. Crop residues are normally only available to rural households at certain post-harvest seasons during the year. As they are frequently used for fodder in rural areas,

their use for cooking may detract from the ability to raise livestock. Dung availability is dependent on livestock ownership in a given area and is easier to collect when livestock are penned in a small corral instead of being allowed to wander freely. As is the case with crop residues, dung has an alternative use as fertilizer, and using significant quantities of dung for cooking fuel will eventually contribute to soil exhaustion.

In many rural areas, modern fuels may simply be unavailable. While kerosene is fairly easy to distribute through formal and informal channels, LPG bottles are often viewed as cumbersome to manage meaning that LPG is not always available for household use. A decision to use electricity is entirely dependent on grid connections and the availability of power. While the availability of traditional fuels may be more a question of ecological status, the availability of modern fuels will be determined by the state of market development.

The question of fuel availability not only influences household fuel-use decisions in rural areas but also helps determine the markets for household fuels in urban and peri-urban areas. In many African countries, urban households rely on charcoal for the domestic cooking needs as firewood is relatively bulky and economically inefficient to transport and charcoal stoves are frequently more efficient than firewood stoves. An added advantage is that while charcoal combustion emits carbon monoxide, it does not emit the same type of noxious, respirable, particulate-laden smoke that firewood does. Hence, urban households using charcoal avoid much of the dangerous ground-level emissions from biomass smoke to which their rural counterparts using firewood are exposed.

In reviewing the question of availability of biomass fuels and its relevance to household fuel choice decisions, Barnes and Floor developed a transition model encompassing three stages that cities pass through in this transition. Stage 1 represents situations where the urban dwellers still rely heavily on wood use—encompassing situations where wood may still be free to the case where it is purchased and its price is fairly high. Stage 2 differs depending on local resource availability. In some parts of Africa (e.g., Senegal, Kenya, Tanzania) and the Caribbean (e.g., Haiti, Dominican Republic, Jamaica), the second stage encompasses heavy dependence on charcoal. For other cases (e.g., Indonesia and China), the second stage involves heavy reliance on coal and kerosene. The final stage, Stage 3, occurs when income in the city is fairly high

and virtually all households are dependent on LPG, natural gas, or electricity. Virtually all cities in developed countries fit into this third category, and increasingly so do many cities in middle-income developing countries. The progression of urban areas through this urban energy transition corresponds well to the movements of individual households in those cities up and down the energy ladder.

On the policy front, for rural households, much effort has focused on supporting the planting of trees either in woodlots, on farmland, or in forest reserves. Policy efforts have also focused on ensuring proper ownership of and access to trees on public and private lands to guarantee fuel availability. While these options are intended to ensure continued sustainability of traditional woody biomass for use as a cooking fuel, there is a need to maintain balance between the use of trees for fuel and their use for nonenergy purposes. Trees have many more functions in the rural ecology than just supplying cooking fuel needs.

Other policy efforts in both rural and urban areas have sought to guarantee that households have access to modern fuels. Governments have lent support to the development of marketing and distribution networks for LPG and kerosene to ensure that households have an expanded array of modern fuel options. Rural electrification and urban electrification (i.e., connecting previously unconnected urban households to the electric grid) are both policies designed to provide households with access to electricity, the fuel typically at the top of the energy ladder. Although results demonstrate that households newly connected to the grid own neither appliances nor the wherewithal to use electricity for all of their domestic purposes, gradual accumulation and consolidation of household status is presumed to expand future household use of electricity.

4. EMPIRICAL EVIDENCE FOR THE ENERGY LADDER TRANSITION

The concept of the energy ladder has been used as a foundation for many of the studies of household energy use in developing countries that have been undertaken since the early 1980s. While not all of them have provided confirmation for the linear transformation envisioned in the narrowest interpretation of the energy ladder concept, most of them do demonstrate that households undergoing the development process do shift their energy use

patterns in response to both internal and external conditions. The following sections focus on small fraction of the empirical literature on household energy use and its documentation of ladder-like household energy transitions. Section A discusses the results from various surveys that have been implemented, and Section B examines some of the statistical modeling that has been undertaken to model household energy behavior.

4.1 Survey Evidence

Household surveys undertaken in rural Africa have demonstrated that fuelwood remains the predominate fuel for rural household energy use throughout the continent. However, the determinants of fuel switching or movements up or down the energy ladder may be somewhat difficult to predict in the rural context. A rural household survey undertaken in Kenya in 1981 shows that fuelwood is by far the single most important fuel used by households for cooking. When it was unavailable, households would be forced to switch either to crop residues or kerosene. Fuelwood availability is an extremely localized phenomenon and may be linked to diverse factors, such as land tenure, population density, settlement density, and crop types. On the national level, income may prove less useful as a predictor of household energy use patterns than might be other factors such as level of engagement with the cash economy. Subsistence farmers showed a relatively heavier reliance on firewood and a lesser consumption of kerosene, LPG, and electricity than did cash-crop farmers or even farm workers or schoolteachers. As most of rural Kenya remains unelectrified, the final rung in the energy ladder was occupied by a very small fraction of households. Rural household energy use patterns remain somewhat problematic and difficult to predict given the wide range of other confounding variables found to influence household fuel choice.

A survey undertaken by Davis in several South African rural areas again confirms that fuelwood is the predominant fuel of choice for all households at all income levels. There is a tendency to shift from collecting firewood to purchasing it as incomes increase, but firewood remains relatively constant as the most important cooking fuel in rural areas. Kerosene again emerges as a transitional fuel for cooking. Candles, kerosene, and electricity are the predominant fuels used for lighting in rural areas. For households connected to the grid, there is a decline in expenditures on kerosene and candles for lighting. But there is also a tendency for the lowest

income households to continue purchasing and using candles even after they have grid electricity. This is attributable either to limited indoor wiring or to limited cash-flows prohibiting the purchase of cards for pre-payment meters. What does emerge from the South African evidence is the relatively large number of households that use multiple fuels for a single end use. For example, 25% of the electrified households and 56% of the unelectrified ones used both kerosene and fuelwood for cooking. In short, the authors conclude that a household energy transition from traditional to modern fuels is taking place in rural South Africa. However, they express reservations about whether it will result in a simple transition to electricity use or an increase in the use of multiple fuels for a single end use.

In urban areas across Africa, the situation appears to be more predictable. Hosier and Kipondya present data from urban Tanzania showing that charcoal is the predominant fuel used in urban households across all income groups. Although this predominance of charcoal may vary according to the city, its role varies little per income category. Kerosene consumption demonstrates the role of a transitional fuel whose use rises with small increases in income and then decreases at higher income levels. Electricity use increases across all income levels. The urban household energy transition appears to be proceeding slowly in Tanzania, despite the country's rather sluggish economic performance during the period under examination. Policy efforts—especially cross-product subsidies to kerosene and artificially low electricity prices—might have played an important role in ensuring that more urban households were able to move up the energy ladder than down during the period of examination.

A detailed ESMAP-sponsored survey of rural household energy use in six counties of China demonstrated that most households used fuelwood, agricultural residues, or coal for their domestic cooking needs. Households in the poorer counties tended to show greater reliance on firewood while those in the more industrialized counties used very little firewood and more coal. Coal was a major household fuel in four of the six counties, especially where fuelwood was scarce. In other locations where fuelwood was scarce, straw and crop waste were used. Electricity use was again positively related to income. On balance, income is seen to play an important role in driving households to use electricity, but fuelwood shortages—frequently very locally determined—tend to drive households either up or down the energy ladder to other fuels.

From an ESMAP-sponsored analysis of surveys of household energy use in six states of rural India, the evidence shows that there is a slow transition to modern fuels brought on both by increasing economic status of households and by an increasing scarcity of fuelwood. In a cross-sectoral analysis, higher-income households clearly used more electricity and more LPG than did lower-income households. Upper-income households also spent a larger absolute amount of money on their household energy needs, but due to the higher incomes, this represented a smaller fraction of their total income. While a higher fraction of upper-income rural households also utilized electricity for lighting and other purposes, the utilization of biomass energy appeared to decline only at the highest income levels in the rural areas. However, the time devoted to collecting fuelwood in rural areas declined with income as wealthier households would tend to purchase firewood instead of gathering it. From 1980 to 1996, over 8% of rural households acknowledged switching from using firewood for their domestic cooking chores to using some other fuels. Over 60% of these households moved to kerosene and LPG for cooking. The remaining households acknowledged fuel switches either to biogas or down the energy ladder to crop residues or dung. In both urban and rural areas, kerosene is seen to fulfill the role of “first step” up the energy ladder as many households begin using it to cook with prior to moving on to LPG or other modern energy carriers. In urban India, LPG has become the fuel of choice for cooking and has begun to be widely utilized. This has been attributable not just to an increased supply of the fuel, but also to the breaking up of the near-monopoly on LPG distribution maintained by the national petroleum companies. In general, these data from India show that there is a movement among both rural and urban households to more sophisticated modern fuels from traditional, biomass fuels.

In a review of energy use surveys in urban areas in Asia, Sathaye and Tyler presented household energy consumption patterns from cities in India, China, Thailand, the Philippines, and Hong Kong. While each country and indeed each city within a country demonstrated very different energy use patterns, their findings showed several tendencies. First, as incomes increased, households tended to rely less on biomass and more on modern fuels. Somewhat surprisingly, modern fuels may have been preferred over traditional biomass fuels not only because of their greater convenience but also because of they are cheaper when measured in terms of their cost per

unit of energy delivered. Second, as incomes increased, households not only made use of a larger quantity of electricity, but they also made use of electricity for a larger set of end uses. Urban households at all levels of income began purchasing and using electrical appliances, from light bulbs to refrigerators to color television sets. This relatively rapid dissemination of appliances throughout urban Asia would indicate that household electricity demand is likely to grow in the same rapid way as it has in the more developed, Organization for Economic Cooperation and Development (OECD) countries. It also poses an opportunity for electricity conservation: if these appliances can be made more efficient, then household standards of living can continue to rise with a less significant increase in electrical requirements. Finally, kerosene use—which is considered to be an inferior modern fuel—declined in importance through time and as incomes increased. While not confirming the existence of a rigid, linear, and universal energy ladder, the evidence presented from urban Asia confirmed that households do move to more modern fuels across time and as income allows.

4.2 Statistical and Econometric Evidence

Given the complexity of household fuel-use patterns and the large number of factors influencing household fuel choice, tables of descriptive statistics do not go very far in understanding household fuel choice. To push beyond the understanding available from using descriptive statistics, a number of authors have used regression-based and econometric analyses to analyze household fuel choice in developing countries. These analyses are important as they provide insights both into the factors that appear to determine household fuel-choice decisions and into the validity of the energy ladder transition itself.

From the analysis of a number of household energy surveys based in south Asia, Leach developed multiple regression equations designed to explain household energy utilization. Because household size is closely related to income, he chose to use per capita household energy consumption figures as the dependent variable. His analysis concluded that income, household size, and settlement size are the key factors determining household fuel use. Although the elasticities that he estimates demonstrate the expected sign in most cases, he argued that they are not much help to energy planners as the important independent variables are not under the planners' control. Other factors that are under more direct

control of energy policymakers—such as relative fuel prices, the cost of appliances, and the availability of commercial fuels—are still not shown to be crucial determinants of household fuel-use patterns, partly because of the inadequacy of the survey techniques deployed in gathering the data.

Using nationwide household energy data from Zimbabwe, Hosier and Dowd developed a model of household fuel choice. In the aggregate, their data showed that fuelwood and kerosene use decreased with income while electricity use increased with income. They then formulated a multinomial logistic model of fuel choice, where the dependent variable represented the probability of a household choosing one fuel over another to cook with. Their data made a distinction between gathered wood and purchased wood as being distinct alternatives. They also allowed for a separate “transitional” category for households that used both wood and kerosene or electricity. Their system of four equations was able to predict nearly 90% of the household fuel choices correctly. Their results demonstrated that an increase in income and localization in an urban area will both increase the probability that a household will make a move up the energy ladder. They conclude that while energy policies might have an influence on household fuel decisions in the urban areas, fuel decisions in the rural areas appear to be insensitive to readily manipulated policy instruments.

Using urban household energy survey data from Java, Indonesia, Fitzgerald, Barnes, and McGranahan undertook two distinct sets of econometric estimations for cooking and lighting. For the analysis of cooking fuels, they first estimated the demand for energy for cooking from the three fuels used (wood, kerosene, LPG) for cooking in Java using a Tobit analysis to account for the fact that many households did not use at least one of the fuels for cooking. The analysis included household size, the prices of all fuels, the availability of all fuels, urban size, and an index measure of cooking size to estimate the amount of each fuel used for cooking. Their results show that firewood tends to be treated as an inferior good: as household income rises, fuelwood use tends to decline. However, wherever fuelwood is plentiful, all households—even wealthier ones—tend to use it for cooking. The results also show that LPG tends to be the fuel of choice for the highest income households surveyed, but again, its availability limits its distribution. Kerosene again serves as a bridging fuel—it is used sparingly by lower and upper-income households, but fairly heavily among the middle-income groups. Their second estimation examined

the demand for energy (measured in MJ) in all forms attempting to isolate the substitution effects between the various fuels. These estimates show that actual substitution effects are slightly different than those derived from technical evaluations of fuels and efficiencies. For lighting, the authors repeated similar analyses of lighting demand and total residential fuel demand. The fuel demand equations demonstrate that electricity is the fuel of choice for all households in urban Java. The lighting substitution analysis demonstrates that a household using kerosene for lighting will have to utilize 10 times as much energy (heating value) for lighting as a similar household lighting with electricity. In addition, a household switching from kerosene to electricity can easily cut its overall lighting expenditures in half while obtaining nearly six times the amount of lighting. Overall, the analysis shows that urban residents of Java show household fuel preferences that resemble those that would be predicted by an energy ladder, with the exception being perhaps that firewood is used across all income groups if it is locally plentiful.

For the cities of Dakar, Senegal, and Abidjan, Cote d’Ivoire, Koumoin studied household energy demand using cross-sectional, time-series data for the years 1978 to 1992 taken from the Food and Agriculture Organization (FAO) production yearbook, the electric utilities in the countries, and a set of World Bank studies of the region. A system of eight equations representing the demand for energy in the urban households of the two cities was estimated in the form of a generalized logit model. His results show that the demand for charcoal is relatively price and income inelastic, meaning that it is used by households at all income levels. Rather than seeing immediate interfuel substitution as might be expected on the basis of the energy ladder concept, the evidence seems to support a gradual transition process with multifuel utilization at the household level being an initial step toward the anticipated substitution. Because of the relative price inelasticity of charcoal, electricity demand may be that sector of the overall household energy consumption which is most sensitive to pricing policies and variations. Local considerations, such as household size, migrant labor, and fuel availability, also seem to be important in shaping the profile of household energy demand. The importance of fuel availability can be seen in that LPG demonstrates a higher than expected penetration because of long-standing policy efforts to promote it as a cooking fuel. In such an economically marginal economic environment, the author concludes that policy efforts might well focus

on ensuring stability in the market for traditional fuels so as not to further threaten the economic base of the population. Interfuel substitution and movements up the energy ladder will likely take place over an extended period of multiple-fuel use given the nature of the prevalent economic and environmental conditions in the cities under examination.

Using a 4-year long study of a village in Mexico supplemented by a household energy survey in four Mexican states, Masera *et al.* discovered that rather than households moving in a straightforward, linear fashion along the energy ladder, in fact, their transition may be from a simple, traditional fuel use to a situation where multiple fuels are used simultaneously. Wealthier households continued to cook with fuelwood indoors even when they owned and also used LPG cooking devices. Multiple fuel use appears to be pursued for two reasons. First, it is a risk-minimization strategy so that if the LPG supply fails, the household will face only minor disruption as they are still accustomed to cooking with firewood. Second, cooking with both wood and LPG is an accepted practice because of preferences for a diet of tortillas—the traditional staple food—which cannot be cooked efficiently on LPG appliances. This cooking pattern has a number of implications. First, households that adopt the more modern fuel, LPG, rarely cease to cook with wood. As a result, their firewood consumption levels drop by only 30% once they begin cooking with LPG. Second, these multifuel using households typically are wealthier members of the community, but they willingly face exposure to nearly twice as many particulates as lower-income households. Fuel stacking, the process by which households continues to use firewood (a traditional fuel) while also adopting LPG (a modern fuel), is seen as a logical resultant of cultural and risk factors found in Mexico, and it appears to continue despite awareness of the potential negative health impacts. Identifying policy solutions to the type of nonlinear, straightforward interfuel substitution found in these areas of Mexico will require much greater thought, study, and a paradigm slightly more complex than that of simplified energy ladder.

5. CRITICISMS OF THE CONCEPT

While it is clear that households facing the development transition will also undergo some form of transformation of their household energy use, the energy ladder concept has been criticized from a number of different perspectives. As it developed

from a general understanding of household budgetary decisions, a knowledge of the development process, and a familiarity with the differences in energy use patterns between more developed and less developed countries, it serves as a loose conceptual framework that can neither be proved or disproved, but that rather informs or underlies a discussion of the transition in household energy use patterns in a particular context. Nevertheless, several criticisms of the concept deserve mentioning as they might provide a foundation for conceptual models that will prove even more useful in understanding the household energy transition.

The first such criticism is that the energy ladder fails to account for the complexities of actual household fuel usage, assuming that all households—or at least most households—will move progressively from traditional fuels to more modern fuels in a predetermined succession. It fails to account for households that might skip rungs by jumping from firewood to LPG without using what appears to be the ubiquitous transitional fuel: kerosene. It also fails to account for multiple fuel use as anything other than a transitional option between two rungs even though many households even in developed countries use more than one fuel for cooking, both as risk-minimization strategies and as cultural preferences. Multiple fuel-use strategies might even result in the most efficient overall household energy-use profiles, as fuels are used for those activities for which they are most efficient. Multiple fuel use may be a logical long-term household fuel choice given cultural preferences, local fuel availability or unavailability, system-wide energy efficiency, and other location-specific considerations. Pursuing multifuel use further might lead to the concept of bifurcated household energy ladder or a different stage model incorporating platforms being maintained at various stages rather than merely a simple ladder whose only options are for remaining stationary, ascending, or descending. A more complex model would do a better job of explaining reality, but would also run the risk of sacrificing the main strength of the energy ladder concept: its intuitive appeal.

A second criticism centers on the value-laden nature of the energy ladder. Value-laden judgments regarding the modernity of society and the fuels associated with a society at a particular “stage” of development underpin much of the thinking behind the use of the energy-ladder concept. Modernization theory—wherein all countries are engaged in the homogenizing conveyor belt called “modernization”

resulting in their becoming more like developed countries—has been discredited as a model for development thinking for quite some time. Yet the energy ladder still builds on this paradigm. Because households in developed or OECD countries have followed a certain evolution in their energy-use patterns, households in developing countries have been assumed to fall into the same—or at least—a very similar pattern. It ignores questions about the conspicuous consumption or lifestyle issues that must be raised in the context of developed countries. Clearly, if all of the people in developing countries used energy in the same profligate manner that those in developed countries do, the earth's energy resources would be rapidly depleted. The energy ladder assumes that there exists but one final end point for the household energy transition, and that end point is the same for developed and developing countries alike. It thereby ignores the possibility that developing countries might pursue paths that are more environmentally friendly; more sustainable over the longer term; and consistent with their own cultural traditions.

A third criticism focuses on the fact that the energy ladder has been built mostly on household energy surveys providing a snapshot of energy consumption at a particular point in time. Its basis is cross-sectional in nature, building more on differences between households at one point in time rather than on longitudinal evidence documenting changes in time among the same households. Generalizing from cross-sectional data to draw time-series conclusions may lead to a misrepresentation of trends and relationships, including the underestimation of the importance of price and cost to influence household decision making. More careful analysis of trends over time would clearly help to strengthen the case for the household energy transition as a whole.

Fourth, the energy ladder concept makes no allowance for the dynamics of power and control within each household unit. In rural households, for example, it may not be unusual for women to be charged with collecting firewood and cooking, while men control financial resources and therefore make decisions requiring financial allocations. For the sake of convenience and reducing exposure to toxic smoke, women might want to move up the energy ladder, but are prevented from doing so by an unwilling husband who controls the household's financial resources. Thus, the predictive capability of the model may be limited in cases where such situations occur. Some would argue that using the

energy ladder as a concept informing the analysis may preclude the consideration of gender roles in the household energy sector altogether. As earlier discussions have shown, household energy discussion clearly could benefit from greater analysis of the disaggregated nature of household units and the different decision-making frameworks of the individual agents involved.

Finally, the energy ladder yields more predictive insights into the household energy transition are more consistent in urban areas rather than rural areas. By and large, rural household energy-use patterns appear to be overwhelmingly determined by local fuel availability rather than other factors that are readily predicted on the basis of economic theory or factored into regression equations. For the most part, policymakers exert relatively little control over rural household fuel-use patterns. In urban areas, the problem may seem more comprehensible and tractable using the microeconomic theory of consumer behavior. That is not to deny the many instances of documented upward or downward shifts on the energy ladder by rural households. Rather, it merely means that the complexity of rural household energy use may be more difficult to reduce to a simple analogue model than are urban energy-use patterns.

6. CONCLUSIONS

The concept of the energy ladder has informed energy studies and policies in Africa, Asia, and Latin America. Household fuel choices are influenced by a number of factors that are both internal and external to the households making the decisions. Statistical models have attempted to identify these factors and define the strength of their influence in determining household fuel choices. Although it has been used as a conceptual basis to inform much of the work on household energy in developing countries, criticisms have brought to light several limitations. To date, no one has posited an alternative conceptual framework describing household fuel choice and decisions in the process of economic development. As a result, the energy ladder concept will likely continue to be used as an analogue model informing discussions of household energy transitions in developing countries. Its strength lies in its analytical clarity and simplicity. Given the importance of energy to economic development and the continued challenge of providing energy for both the subsistence and development needs of the urban and rural poor in developing

countries, the issues raised in the discussion of the energy ladder will likely continue to be important for energy and development professionals for the foreseeable future.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Development and Energy, Overview • Economic Growth and Energy • Equity and Distribution in Energy Policy • Global Energy Use: Status and Trends • International Comparisons of Energy End Use: Benefits and Risks • Population Growth and Energy • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use • Women and Energy: Issues in Developing Nations • Wood in Household Energy Use

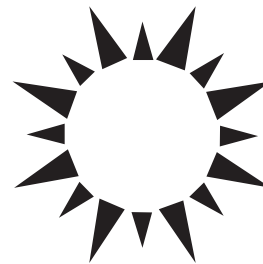
Further Reading

- Barnes, D. F., and Floor, W. (1999). Biomass energy and the poor in the developing world. *J. International Affairs* 53(1), 237–259.
- Davis, M. (1998). Rural household energy consumption: The effects of access to electricity—evidence from South Africa. *Energy Policy* 26(3), 207–217.
- DeWees, P. A. (1989). The woodfuel crisis reconsidered: Observations on the dynamics of abundance and scarcity. *World Development* 17(8), 1159–1172.
- Eberhard, A., and van Horen, C. (1995). “Poverty and power: Energy and the South African state.” Pluto Press, East Haven, CT.
- Energy Sector Management Assessment Program (ESMAP) (1996). “Energy for Rural Development in China: An Assessment Based on a Joint Chinese/ESMAP Study for Six Counties.” UNDP/World Bank ESMAP Program, Washington, DC.
- Energy Sector Management Assessment Program (ESMAP) (2002). “Energy Strategies for Rural India: Evidence from Six States.” United Nations Development Program/World Bank ESMAP Program, Washington, DC.
- Fitzgerald, K., Barnes, D., and McGranahan, G. (1990). “Interfuel Substitution and Changes in the Way Households Use Energy: The Case of Cooking and Lighting Behavior in Urban Java.” Industry and Energy Department Working Paper, Energy Series Paper Number 29. The World Bank Industry and Energy Department, PRS, Washington, DC.
- Hosier, R. H. (1985). “Energy Use in Rural Kenya: Household Demand and Rural Transformation.” *Energy, Environment and Development in Africa*, Volume 7. Beijer Institute and Scandinavian Institute of African Studies, Stockholm.
- Hosier, R. H., and Dowd, J. (1987). Household fuel choice in Zimbabwe: An empirical test of the energy ladder hypothesis. *Resources and Energy* 9, 337–361.
- Hosier, R. H., and Kipondya, W. (1993). Urban household energy use in Tanzania: Prices, substitutes and poverty. *Energy Policy* 21, 454–473.
- International Energy Agency (2002). “World Energy Outlook 2002,” especially Chapter 13, “Energy and Poverty.” Organization for Economic Cooperation and Development and IEA, Paris.
- Koumoin, M. C. (1998). Global carbon initiative and CO₂ mitigation in urban Sub-Saharan Africa: Misreading the empirical evidence? *World Resources Rev.* 10(4), 607–632.
- Leach, G. (1987). “Household Energy in South Asia.” Elsevier Applied Science, London and New York.
- Leach, G. (1992). The Energy Transition. *Energy Policy* 20(February), 116–123.
- Masera, O. R., Saatkamp, B. D., and Kammen, D. M. (2000). From linear fuel switching to multiple cooking strategies: A critique and alternative to the energy ladder model. *World Development* 28(12), 2083–2103.
- Sathaye, J., and Tyler, S. (1991). Transitions in household energy use in urban China, India, the Philippines, Thailand, and Hong Kong. *Annual Rev. Energy and the Environment* 16, 295–335.
- Smith, K. R. (1987). “Biofuels, Air Pollution, and Health: A Global Review.” Plenum, New York.
- Smith, K. R. (1993). Fuel combustion, air pollution exposure, and health: The situation in developing countries. *Annual Rev. Energy and the Environment* 18, 529–566.

Energy Services Industry

EDWARD L. VINE

Lawrence Berkeley National Laboratory
Berkeley, California, United States



-
1. Introduction to the Energy Services Industry and Energy Service Companies
 2. Origins and Growth of the U.S. ESCO Industry
 3. Future Prospects for the U.S. ESCO Industry
 4. International ESCO Activities
 5. Conclusions

Glossary

demand-side management (DSM) bidding An auction in which a utility generally solicits proposals from energy service companies interested in achieving specified amounts of DSM savings (e.g., 1000 kW of demand reduction).

energy service company (ESCO) A company that is engaged in developing, installing, and financing comprehensive, performance-based projects, typically 5–10 years in duration, centered around improving the energy efficiency or load reduction of facilities owned or operated by customers.

guaranteed savings The ESCO guarantees to the customer that the project's realized savings will equal (or exceed) the payments made (or financing costs absorbed) by the customer over the term of financing. Also known as *chauffage* in Canada and Europe.

investment-grade audit A detailed audit that is used as the basis for financing a project as well as for projecting energy and other savings on a performance contract.

non- and limited-recourse project financing Based on payments made exclusively from the cash flow (revenues) of the project and not by the other nonproject assets of the customer or ESCO.

off-balance sheet financing The building or facility owner can receive investments in his or her facility but with no financial liability recorded on the organization's balance sheet.

performance-based projects The ESCO's compensation, and often the project's financing, are meaningfully tied to the amount of energy actually saved, and the ESCO assumes the risk in linking its compensation directly to results.

shared savings projects The ESCO's payments are typically structured as a percentage of the energy savings actually realized by the end user over the length of the contract.

super ESCOs ESCOs that provide traditional energy services and supply gas and/or electricity (and/or other fuels) to customers.

third-party financing Funding of energy savings investments by an outside company, using energy savings to pay for the project.

The energy services industry represents an important group for promoting energy efficiency in developed and developing countries. This article provides an overview of energy service companies (ESCOs), starting with a description of their basic characteristics, and then tracing their origin and growth in the United States. After examining the future prospects for the U.S. ESCO industry, the article ends by briefly examining the ESCO industry in other countries.

1. INTRODUCTION TO THE ENERGY SERVICES INDUSTRY AND ENERGY SERVICE COMPANIES

1.1 ESCOs and Performance Contracting

An energy service company (ESCO) is a company that is engaged in developing, installing, and financing comprehensive, performance-based projects, typically 5–10 years in duration, centered around improving the energy efficiency or load reduction of facilities owned or operated by customers. Projects are performance-based when the ESCO's compensation, and often the project's financing, are meaningfully tied to the amount of energy actually saved, and the ESCO assumes the risk in linking its compensation directly to results. Performance contracting is another term for the performance-based approach.

Energy performance contracts can be distinguished from two contracting approaches commonly utilized by other types of energy-efficiency service providers: (i) contracts for installation of efficient equipment for a specified fee, which is typically backed by a manufacturer's equipment warranty contract, and (ii) energy consulting services, such as building audits and advice on potential energy-efficiency improvements. Contracting based on performance does not necessarily have to be undertaken by an ESCO, but in practice ESCOs have been the pioneers and major users of performance contracting for energy-efficiency projects. ESCOs used performance contracting primarily as a method for assuring their clients that they would confidently achieve energy and cost savings—there were no other companies in the marketplace taking this risk. For this reason, ESCOs are fundamentally different from consulting engineers and equipment contractors specializing in efficiency improvements; the former are typically paid a fee for their advice, whereas the latter are typically paid a fee for their equipment, and both accept no project risk. Such risk creates the highest possible motivation to properly specify, design, engineer, install, and maintain savings over the length of the contract.

1.2 ESCO Diversity

Each ESCO differs in terms of ownership, target markets, technology focus/expertise, in-house capabilities, etc. The following are some of the key areas in which ESCOs can be differentiated:

Ownership: ESCOs may be privately owned (independent) companies, utility subsidiaries or affiliates, nonprofits, joint ventures, manufacturers, or manufacturers' subsidiaries.

Target market: ESCOs typically focus on market niches, by subsector (e.g., hospitals, schools, and government) and by project size (e.g., large customers).

Service specialization: Some ESCOs perform project installation using in-house expertise, whereas others specialize in engineering design and analysis and subcontract other services (e.g., equipment installation).

Technology: Most ESCOs have a technological bias [e.g., lighting or heating, ventilation, and air-conditioning (HVAC) controls], whereas others are comprehensive by seeking to achieve energy savings from the widest possible array of cost-effective measures in a given facility. Also, some ESCOs focus more on electrical end uses than those using other

forms of energy (primarily because most projects have involved lighting), but comprehensive ESCO projects almost always involve thermal measures.

Geographic preference: Some ESCOs have established a preference to conduct business in specific geographic regions of one country, whereas others are national and a few have done some international work.

Project financing: ESCO financing capabilities vary with the financial strength of the ESCO. Those with "deep pocket" parent companies typically have a greater capacity to offer off-balance sheet financing. Some ESCOs have significant, well-established financing capabilities, some have limited in-house financing capabilities, and others have none at all and must arrange financing through lenders (the typical case). However, all ESCOs rely to some extent on third-party financing since most ESCOs do not have internal financing capabilities.

1.3 ESCO Capabilities and Skills

ESCOs are diverse: There is no real prototypical ESCO. However, most ESCOs typically have the following capabilities and skills: project development, engineering and design, feasibility analysis, energy analysis, general contracting, the ability to finance directly or arrange third-party financing, project and construction management, purchase and installation of equipment, risk management, monitoring and verification (M&V) of savings, training, operations and maintenance services for the installed equipment, and administrative services. Most of these capabilities reside among the staff of the company (in-house), but some of these skills are contracted out (e.g., installation of equipment). In addition, ESCOs may provide additional services, such as cogeneration and alternative power production.

Some additional comments on these activities are warranted. First, ESCOs use working capital for general corporate purposes such as marketing. They use project financing to construct a project; they also reimburse themselves through project financing for the amount of working capital expended to develop a particular project. All or most of the activities that ESCOs conduct for a project are bundled into that project's cost, which is usually financed entirely with debt and lease financing (i.e., no equity contribution is required to be made by the customer since payments come only from realized savings); project financing using ESCO or third-party equity contributions is rare. In sum, ESCOs attempt to remove the first-cost disincentive and technical uncertainty by

providing outside sources of capital and guaranteeing the performance of the equipment. The first-cost disincentive is a real barrier to investments in energy efficiency, whereas the perception of technical uncertainty is a more formidable barrier than the actual performance of energy-saving measures. The cost of the ESCO service is typically paid for from the stream of energy cost savings resulting from the project.

Second, because ESCO projects are performance based, ESCOs must measure and verify the energy savings, and these savings must be monitored over the project's life to make sure the performance-based savings are accurate and persist. M&V is typically conducted by periodically comparing a customer's energy bills with an established pre-project baseline and/or through energy metering.

Third, maintenance is usually the final component of most performance-based energy-efficiency projects. The maintenance is for all (or some portion) of the new equipment over the life of the contract. The cost of this ongoing maintenance may be folded into the overall cost of the project. Thus, throughout the life of the performance contract, the customer receives not only the benefit of reduced energy costs but also the benefit of reduced maintenance costs. The maintenance component is also important since it enables the ESCO to ensure that the equipment is maintained in a way that preserves optimal performance. As an additional service on most contracts, the ESCO also provides any specialized training needed so that the customer's maintenance staff is equipped to take over at the end of the performance contract period (the period may be 5, 10, or 15 years, and sometimes longer).

Fourth, each ESCO entering into an energy-savings performance contract tries to educate its customer on energy use and seeks to bring that customer into an energy-efficiency partnership whose purpose is to achieve the maximum energy savings available from the customer's facility. This is not only good for the customer but also beneficial for the ESCO as it seeks to maintain a long-term contract with the customer and obtain additional business.

In sum, although none of the discrete skills that an ESCO employs are particularly unique, the added value an ESCO brings is its ability to integrate a wide variety of skills and apply them efficiently to projects ranging from several hundred thousand dollars to tens of millions of dollars. Thus, ESCOs view themselves as project developers and integrators.

1.4 ESCO and Customer Responsibilities in Performance Contracting

In an energy performance contract, project responsibilities are divided between the ESCO and the customer. Once a customer has been contacted, an ESCO has the following responsibilities:

- Conducts walk-through energy audit and preliminary engineering study.
- Prepares engineering proposal and project development agreement based on audit results.
- Presents proposal and agreement to customer for comments and approval.
- Conducts an investment-grade audit.
- Finances purchase of energy-saving and measurement equipment (unless financed by customer or by third-party bank or leasing company).
- Installs equipment.
- Maintains equipment over life of contract (if maintenance provided).
- Delivers agreed level of energy savings.
- Monitors, measures, and verifies energy savings.

Customer responsibilities include the following:

- Reviews and approves engineering proposal (if proposal is not approved, customer pays ESCO a fee).
- Operates host facility as specified.
- Compensates ESCO as appropriate to type of financing used.
- Pays value of share of energy savings (if ESCO financing secured by shared savings).
- Permits ESCO to maintain and measure savings.
- Pays for services (if third-party financing).
- Maintains equipment over life of contract (if maintenance not provided).

1.5 Financing of ESCO Projects

Most ESCOs provide financing for projects from their own funds (working capital); however, they generally are only able to finance the initial stages of a project. Even though the ESCO eventually has no equity investment in the project, the ESCO does provide its equity to fund the initial risk equity by investing its working capital to develop and finance a paid-from-savings project for the customer. Thus, ESCOs seeking to use performance contracting must develop or arrange sources of debt and equity to finance their projects. More often, the ESCO's role is to arrange

financing for its customers with leasing companies, institutional investors, and commercial banks.

The customer usually enters into separate agreements for energy services and for financing. Typically, the ESCO may arrange financing but is not a party to the finance agreement. The end-user credit risk is assumed by the financier, not the ESCO. The ESCO and the end user enter into a turnkey contract or energy services agreement whereby the ESCO provides engineering, equipment installation, and other services such as operations, maintenance, and savings verification. If the savings do not materialize as projected, the end user has recourse against the ESCO.

Many ESCO projects are financed using nonrecourse or limited-recourse financing. Since no capital is required up front and financiers are repaid through energy savings, ESCOs and customers are allowed to develop projects that are much larger than their net worth. Although nonrecourse financing has important advantages (e.g., a positive cash flow immediately for the customer), it also introduces significant business risks into the financing and higher project costs.

The two types of project financing commonly used by ESCOs in the United States are guaranteed savings and shared savings. In a guaranteed savings agreement, the ESCO guarantees to the customer that the project's realized savings will equal (or exceed) the payments made (or financing costs absorbed) by the customer over the term of financing. In guaranteed savings contracts, ESCOs typically guarantee approximately 80–85% of the savings they expect to get out of a project. In these projects, the customer assumes the obligation to repay the project costs to a third-party financier (usually introduced by the ESCO), which is very often a commercial bank or leasing company. If the customer finances the project itself, the customer repays itself by saving energy costs.

If the savings realized fall short of the payments made by the customer due to a failure of the project to perform as guaranteed, the ESCO pays the customer the difference. If, however, actual savings exceed the payments, excess savings are normally returned to the customer. In some cases, the customer and ESCO negotiate the sharing of savings in excess of the amount guaranteed.

In shared savings projects, the ESCO's payments are typically structured as a percentage of the energy savings actually realized by the end user over the length of the contract. The percentage is predetermined, but the percentage of savings allocated to each party can vary over the period of the contract. The ESCO must ensure that the proceeds it receives

will cover the costs it has incurred in implementing the project.

In exchange for the ESCO providing project financing and for the ESCO's assumption of performance and credit risk, the customer agrees for a fixed term to pay the ESCO the value of a set percentage of those savings. The portion of savings paid to the ESCO that is attributed to debt service is always higher for shared savings than for guaranteed savings projects, reflecting the ESCO's significantly greater risk and expense. The ESCO uses its payments from the customer to repay debt associated with the project, conduct equipment maintenance, measure and monitor savings, and earn a return. As with guaranteed savings projects, the ESCO measures energy savings throughout the term of the project agreement.

2. ORIGINS AND GROWTH OF THE U.S. ESCO INDUSTRY

2.1 ESCO Driving Forces

The three key driving forces affecting the ESCO industry in the United States during the past 15 years have been energy prices, tax policies, and utility demand-side management (DSM) bidding programs. The U.S. ESCO industry began in the late 1970s and early 1980s, a period during which energy prices were dramatically rising in response to the 1973 oil embargo and the Iranian revolution in 1979, which presented the opportunity for fledgling ESCOs to make a business out of reducing customers' growing energy costs. For example, when energy prices were high (1980–1983), the annual growth rate in ESCO investments during this period was estimated to be 35%; the rising electricity prices from 1989 to 1994 led to another period of increased growth rate (25%) in ESCO investments.

During its formative years, changes in tax policy were also critical in shaping the ESCO industry. For example, the energy tax credit (available for energy-efficiency investments from 1980 to 1983) and the investment tax credit (for general capital investment, including efficiency improvements, from 1980 to 1986) affected the rate of ESCO project investment between 1980 and 1986. After the energy tax credits ended in 1983, ESCO investments decreased by 10% in 1984.

Finally, utility DSM bidding programs significantly impacted the ESCO industry in its early years. In a DSM bidding program, the utility enters into a

long-term conservation purchase agreement with the provider (usually an ESCO) to reduce demand or provide firm energy savings during specified periods. The utility pays for these savings on a unit basis, often simply per kilowatt-hour, much as it would pay for private power under a purchase agreement. The first utility DSM bidding program began in late 1987, initiating a period of substantial utility contributions to the pool of investment capital underwriting performance-based efficiency projects. Most observers believe, however, that the influence of these utility programs on the U.S. ESCO industry peaked in 1994: Some utilities froze or canceled DSM programs in the atmosphere of uncertainty that commenced with the California restructuring proposal in April 1994.

2.2 Types of ESCOs

Four main groups of ESCOs dominate the U.S. ESCO industry: original ESCOs, engineer-based ESCOs, manufacturer-based ESCOs, and utility-based ESCOs. Whatever their origins, all ESCOs are characterized by their project development capabilities and their assumption of performance risk for their projects.

The original ESCOs entering the marketplace were start-up ventures established specifically to pursue energy performance contracting. The engineer-based ESCOs grew out of businesses providing engineering services for a fee. Most of these consulting engineers specialized in performing detailed energy audits of clients' buildings, proposing designs for new energy systems, and suggesting improvements to operating practices. Some engineering firms fundamentally altered their business strategy and changed from fee-based energy services consulting to project development. These firms were willing and able to back their designs with performance guarantees and to secure financing for projects with the savings these projects were guaranteed to produce.

Manufacturer-based ESCOs are manufacturers of building control systems designed to regulate energy-using equipment in buildings and facilities. The earliest firms offered control systems specifically designed to reduce their customers' energy bills. Like the consulting engineers, these manufacturers extended their traditional business activity (e.g., controls) to assume the risks associated with project development. In the early 1980s, these companies established energy service divisions to develop performance-based projects centered around their control systems. These companies

further evolved by developing comprehensive projects that include the full range of available energy-efficiency measures.

Utility-based ESCOs include those ESCOs formed in response to the growth of utility DSM programs, either by the utilities or by private investors. Numerous utilities have launched DSM programs that complement ESCO projects: For example, utility customers have used rebate programs to improve the cost-effectiveness and comprehensiveness of their performance-based projects. Similarly, some utilities have offered DSM bidding programs in which ESCOs have been encouraged to participate. Nevertheless, the utility-created ESCO has historically been rare. Utility entry into the ESCO market or acquisition of existing ESCOs have accelerated significantly as utilities seek to position themselves to compete effectively in a retail energy services environment.

2.3 The Maturation and Growth of the ESCO Industry

The U.S. ESCO industry has matured in the past decade. The more positive image of the industry and its resultant growth stem from a combination of factors, including (i) the trend toward guaranteed energy savings, (ii) the development of standardized monitoring and verification protocols, (iii) the involvement of the National ESCO Industry Association (NAESCO) in accreditation and M&V protocols, (iv) DSM bidding programs, (v) the facilitation of energy performance contracting in the federal government, and (vi) improvements in performance and reductions in costs of energy-efficiency measures.

There are no firm data on the ESCO industry's size in the United States, and traditional financial measures are difficult to apply in an industry composed primarily of privately held companies and large public firms that do not report the financial results of their efficiency divisions separately. Nevertheless, there are more than 60 national and regional ESCOs actively operating in the United States. Approximately 12 companies account for 75% of the industry revenues. ESCOs typically develop projects in the \$600,000 to \$2 million range, and they install between \$800 million and \$1 billion worth of energy-efficient equipment (including labor and profit) each year. Since 1990, the installed cost of all projects implemented by ESCOs in the United States is \$6–7.2 billion. During the late 1990s, the ESCO industry grew by approximately 25% per year, mainly driven by rising electricity prices during

from 1984 to 1994. In 2002, ESCO industry revenues for energy-efficiency-related services in the United States ranged from \$1.8–2.1 billion.

2.4 ESCO Markets and Marketing

ESCOs have established themselves in well-defined niche markets using either geographic or industry-specific market segmentation strategies. For example, ESCOs that have been selected in DSM bidding programs naturally target certain types of customers (e.g., hospitals and small retail) in specific utility service territories. ESCOs have also tended to focus on regions in which retail electricity prices are high in order to maximize differences between project cost and benefits to customers from energy-efficiency investments. Other ESCOs, particularly those affiliated with the large controls and equipment companies, utilize a decentralized and extensive branch sales force that targets certain industries or sectors (e.g., schools and state and local governments).

The focus of most ESCOs has been on medium to large commercial and institutional customers: Local and state government, schools, and universities account for approximately 65–75% of overall ESCO activity. ESCOs have also achieved success among other groups of commercial-sector customers (e.g., owner-occupied office buildings, shopping malls, hotels, and a few national franchises). The federal government, for which military projects have huge energy bills, is also targeted by many ESCOs.

Industrial customers have hosted fewer performance-based efficiency projects, but their share of projects has increased slightly over time as energy services are becoming integrated with other services. In the industrial sector, ESCOs have attempted to differentiate themselves through specialized expertise in certain technologies (e.g., adjustable-speed drives and controls) or processes (e.g., pulp/paper and automotive).

Because the typical ESCO project costs more than \$350,000, small commercial and industrial companies and residential customers are generally not served by ESCOs. Some ESCOs do work in certain niche markets (e.g., large multifamily projects and water and wastewater facilities), but these are rare.

The so-called MUSH institutional markets (municipal governments, universities, schools, and hospitals) have been the focus of ESCO marketing because they tend to be very capital constrained; have a longer investment horizon; are faced with an aging

infrastructure and equipment in need of modernization; and often have limited technical expertise to develop, coordinate, and manage a comprehensive energy-efficiency project. Furthermore, ESCOs have been able to arrange relatively attractive financing terms with the MUSH markets because financial institutions have regarded these customers as relatively good credit risks (i.e., they are unlikely to go out of business or relocate).

2.5 Energy Measures in ESCO Projects

In 1980, approximately 66% of ESCO investment was in electrical efficiency measures, and the rest was in thermal energy measures. From 1980 to 1987, ESCOs gradually invested in more thermal energy measures and fewer electricity measures, until the investment ratio in 1987 was 55% of investments in electricity measures and 45% of investments in thermal energy measures. Beginning in 1988, when utility DSM bidding programs were predominantly sponsored by electric utilities, investments in electricity measures increased, resulting in an ESCO investment ratio of 75% electrical versus 25% thermal. Recently, thermal energy measure investments have increased so that the relative shares in 1997 were approximately 65% electrical versus 35% thermal.

A number of technologies have come to form the core of the typical performance-based energy-efficiency project, including high-efficiency lighting, high-efficiency HVAC equipment, efficient motors and variable-speed drives, and centralized energy management systems. Along with the core technologies are a full array of weatherization services where appropriate and, sometimes, more innovative applications, such as integrated cogeneration or renewable energy technologies.

2.6 Measurement and Verification and Protocols for Measurement and Verification

All ESCOs measure and verify the energy saved by their projects over the projects' lives—as demanded by the customer and as the essence of energy performance contracting. The purpose of monitoring energy savings over time is to ensure that those savings persist and to identify immediately any anomalies in consumption patterns that may occur. It also serves to ensure the quality and effectiveness of ongoing maintenance, entailed in most ESCO projects.

The development of standardized procedures for the measurement and verification of savings has helped both end users and the financial community to understand this key dimension of performance contracting. ESCOs have led the effort to verify, rather than estimate, energy savings, particularly when public and utility ratepayer funds are paying for them.

U.S. ESCOs have adopted the U.S. Department of Energy's (DOE) International Performance Measurement and Verification Protocol (IPMVP) as the industry standard approach to measurement and verification. States ranging from Florida to New York require the use of the IPMVP for state-level energy-efficiency retrofits. The U.S. federal government, through the Department of Energy's Federal Energy Management Program, uses the IPMVP approach for energy retrofits in federal buildings.

2.7 NAESCO

NAESCO was established in 1983 to bring credibility to the ESCO industry. Its members include independent performance contracting ESCOs; ESCOs affiliated with electric and natural gas utilities; energy-efficiency equipment suppliers, manufacturers, and other energy service providers; and a "miscellaneous" group (e.g., consulting firms, financial companies, and public-sector members). The primary activities of NAESCO include conferences, regulatory intervention (i.e., lobbying), accreditation, and publications.

NAESCO has a model energy performance contract that has been in use for 10–15 years and that forms the basis for most energy performance contracts today. NAESCO also serves an important policing service, ensuring that ESCOs are doing business in a fair and equitable fashion. The NAESCO board holds its members to a high standard of performance, as reflected in the accreditation process. To receive the accreditation, an ESCO goes through a rigorous review of credentials, references, technical capabilities, financial capabilities, project history, and interviews of former and existing clients. Certification must be renewed every year.

2.8 Utility Programs and ESCOs

Utility DSM bidding programs have played a critical role in the development of the ESCO industry. Although not all ESCOs have participated in utility

DSM bidding programs and some ESCO participants in these programs have negative views of their experience, most observers of the ESCO industry agree that these programs have helped to promote the ESCO industry.

2.9 Energy Performance Contracting in the Federal Government

In the past 25 years, both federal and state governments have adopted legislation for improving energy efficiency. Although most of this legislation has had indirect impacts on ESCOs (e.g., the promotion of energy efficiency), the real impact has been on prompting governments as customers to seek assistance from ESCOs. Governmental facilities are prime candidates for ESCO projects: These facilities tend to use energy inefficiently, the owners of the buildings have constrained budgets, and there are many buildings.

ESCOs can contract directly with the government, mainly through a vehicle called an energy savings performance contract (ESPC). An ESPC can be a type of shared savings contract, with no government capital required up front. ESCOs can finance ESPC with their own internal funds or with third-party financing. Under an ESPC, an ESCO typically provides the government with design, installation, and ongoing maintenance services as well as the capital needed to finance efficiency improvements. Payments are made to the ESCO based on a share of savings that resulted from the energy-savings measures implemented (i.e., a shared savings contract).

The Federal Energy Management Program (FEMP) of the U.S. DOE developed the broad-area FEMP Energy Saver Performance Contract (Super ESPC). The first broad-area contract, covering eight states in DOE's western region, was awarded in 1997. Under these contracts, federal agencies award delivery orders from existing contracts and reduce the lead time and resources to award individual contracts. Performance-based energy services are available to federal agencies with government-owned facilities within a region. These contracts involve the competitive selection of a small number of contractors (i.e., ESCOs) that are allowed to conduct negotiations with individual facilities for the awarding of delivery orders under the contracts. The selected contractors are able to add work to their contracts for a limited time without the need for additional extensive, time-consuming competition and evaluation.

3. FUTURE PROSPECTS FOR THE U.S. ESCO INDUSTRY

3.1 Opportunities and Challenges in a More Competitive Utility Environment

The future prospects for ESCOs and other energy service providers will be affected by both the length of the transition period to a more competitive electricity industry and the ultimate organization and institutional structure of bulk power and retail service markets. The issues surrounding electricity industry restructuring are enormously complex and multifaceted. The pace of industry restructuring is likely to vary state by state given regional differences in electricity prices and resource endowments and the influence of state regulation. Because the pace and ultimate outcome of industry restructuring cannot be determined, ESCOs and other energy service providers face a retail market environment with significant uncertainties, many of which are regulatory (rather than market) driven.

Many observers believe that the restructuring of the utility industry will lead to lower power prices, resulting in a reduction in energy-efficiency programs and projects (e.g., energy-efficiency projects will be less attractive and paybacks will be longer if energy is sold at 5 rather than 9¢/kWh). Short-term impacts have already been felt: (i) Energy is receiving the attention of top management in businesses as well as utilities, (ii) gas and electricity savings due to lower prices are seen as more important to end users than energy-efficiency savings (i.e., low-cost energy providers are expected to prevail), (iii) energy is primarily seen as a commodity (rather than as a service), (iv) more emphasis is being placed on short-term contracts, and (v) customers are requiring shorter paybacks and shorter returns on projects. The end result is that customers will be less motivated to implement energy-efficiency projects.

On the other hand, despite the expectation of lower prices, there is a great opportunity for energy efficiency and new types of energy-efficiency services. In a competitive marketplace, as price differences among power suppliers narrow, value-added products and services (such as energy efficiency) will be offered to customers by various ESCOs and other types of energy service providers. In addition, state adoption of a system benefits charge will encourage and fund energy-efficiency activities during the transition to a competitive electric industry.

In summary, in a restructured electricity industry, it is likely that support for energy efficiency will

continue, although most likely on a diminished basis. For ESCOs that have historically targeted larger institutional and C&I customers and that have depended on utility DSM funding, ratepayer funding for energy efficiency in the form of financial incentives (e.g., rebates) may be unavailable or available only for a short period (e.g., 2–5 years).

3.2 New Customer Needs

Customers have an array of needs that ESCOs will be asked to address. Customers are also quickly realizing that they will have more options and choice in a restructured electricity industry and will demand more products and services. Furthermore, customers will want changes in the way ESCOs do business with them, such as more flexible contracts, less risk, more warranties, simpler and clearer agreements, more customer control and involvement in projects, more convenience, and bundling or unbundling of products and services.

3.3 New ESCO Products and Services

ESCOs have the potential to interact with customers with more creativity, and it is likely that ESCOs will offer packages of comprehensive energy services, different fuels, varieties of pricing plans, and other creative ideas (including energy efficiency) as the market develops and grows. Occasionally, ESCOs will need to unbundle products and services for clarity and competitiveness, whereas at other times they will need to bundle products and services for convenience. Some ESCOs may manage a customer's fuel and energy needs through bill consolidation, analysis, and payment. The more innovative ESCOs may also be asked to provide nontraditional, performance-based customer service, such as providing "total comfort" or lighting based on \$/ft² or \$/load.

Hence, ESCOs may find more opportunities in providing assistance in other areas besides utility-related issues, including facilities management and operations, capital equipment budgets, environmental concerns, and compliance with government regulations. For example, some ESCOs are considering guaranteeing environmental compliance as well as energy savings when they change a process to improve the environment (e.g., indoor air quality). They will be responsible for any mistakes (e.g., if the facility does not comply with environmental regulations or if savings do not occur), and they will fix the mistakes.

In sum, today's ESCO may become tomorrow's super ESCO—a fully integrated (full-service) energy

services company that brokers power and offers energy efficiency and other energy services. The ESCOs that respond to these challenges may emerge stronger than before and ahead of their new competitors as they provide “energy services” in its broadest sense.

3.4 Changes in ESCO Practices

In addition to providing new services, ESCOs will need to change their business practices to remain competitive. Many of them will form partnerships, joint ventures, mergers, and strategic alliances (both regionally and nationally). All of them will take more risks in providing new services, developing new alliances, being entrepreneurial and agile developers, energy brokering and marketing, and integrating energy efficiency with power supply and communication services. More ESCOs will be asked by customers to act as problem solvers to address multiple and complex customer demands and to identify resource efficiency opportunities. Accordingly, ESCOs will be looked upon as technology integrators, resource managers, and networkers. Finally, ESCOs are expected to continue to target large (especially institutional) customers and wholesale markets (federal government) that need engineering and financial assistance. However, they will also be aggregating smaller customers (small C&I and residential) to strategically position themselves at the forefront of the ESCO industry.

3.5 New ESCO Competitors (or Partners?)

The restructuring of the electric and natural gas industries will change the nature of competition and competitors in the energy-efficiency performance contracting market. In a more competitive electricity industry, the competitors to ESCOs will likely include power marketers, utilities, utility affiliates, financial institutions, equipment suppliers, property managers, cable companies, telecommunication companies, building control companies, high-technology companies, independent power producers, and energy traders.

ESCOs will need to determine whether they want to compete with these entities or form partnerships or joint ventures. Most non-comprehensive ESCOs will continue to seek to partner with other firms to develop business and to implement large-scale projects. Full-service providers will also be expected to partner with other firms to compete for customers by offering a range of potential energy services,

making it more difficult for specialized or small energy providers to develop their own work independently. Partnering or consolidation of full-service providers with utilities or large, well-capitalized firms can be expected to increase the level of competition and the effectiveness of competition for customers.

As they prepare for competition, many utilities are using the regulated distribution business unit as an incubator to test various customer and energy services, whereas others have located these activities in unregulated retail service companies. Increasingly, the markets sought and the services offered by utility distribution companies and ESCOs are becoming indistinguishable. This presents existing ESCOs with both a competitive challenge and a partnering opportunity, from ad hoc marketing agreements to formal joint ventures and acquisitions. In addition to retention of electricity and/or gas commodity sales, these utility/ESCO partnerships can sell comfort, light, power, reliability, and end-use commodities (e.g., refrigeration, steam, chilled water, and compressed air). They can also offer a variety of building-related services that enhance the initial energy-efficiency services, such as facilities management, operations and maintenance services, and energy management/control.

The benefits to ESCOs of these alliances are primarily reductions in the transaction costs of lead generation, qualification, and closing. The benefits to utilities are defensive (e.g., customer retention), offensive (e.g., value added to wholesale and eventually retail offers), and potential revenue enhancement (through technology applications, fuel switching, customer growth, etc.).

4. INTERNATIONAL ESCO ACTIVITIES

The ESCO industry has been developing in countries besides the United States, particularly since the mid-to late 1990s. In most cases, the number of ESCOs are small (less than 10) and revenue from the projects is generally less than \$10 million. The impediments to the successful development of an ESCO industry in other countries are similar to those encountered in the United States. In particular, the following appear to be formidable:

- The concept of ESCOs and energy performance contracting is new to nearly all of the parties that need to be familiar with it (i.e., there is a lack of

awareness): financiers, attorneys, contractors, insurers, utilities, government agencies, and customers.

- Government policies regarding energy performance contracting are uncertain.
- Legal and contractual issues specific to performance contracting are unclear.
- Economic conditions are adverse (e.g., stagnant economic development) for investments in energy efficiency in many countries.
- There is a lack of capital (equity financing or equity investments) for establishing ESCOs in many countries.

Nevertheless, several countries have developed an ESCO industry and have started to promote the development of a vibrant ESCO industry. Several countries have established associations of energy service companies (e.g., China, Egypt, Japan, Korea, and the Ukraine). Revenue from the ESCO industry in a few countries exceeds \$50 million (e.g., Brazil, Canada, China, and Japan). In 2001, ESCO industry revenues for energy-efficiency-related services outside the United States ranged from \$560–620 million.

5. CONCLUSIONS

The ESCO industry is expected to grow significantly in the United States and in selected countries in the near future. There will be numerous new entrants in the ESCO market, including supply-side bidders (e.g., independent power producers) and utility subsidiaries. In the United States, ESCOs with technical strengths and deep pockets (or access to sizable funding) will survive; small ESCOs that specialize will disappear in a period of consolidation, some of which may be acquired by utilities. In the United States, the consolidation of ESCOs and the trend for utilities to acquire or develop full-service ESCOs to compete for other utilities' customers (particularly large customers) have become more prominent during the past few years. These are expected to also occur overseas, where the energy industry is being restructured.

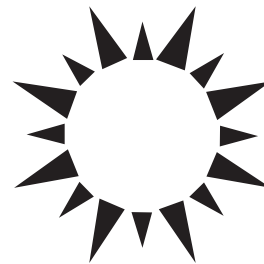
SEE ALSO THE FOLLOWING ARTICLES

Corporate Environmental Strategy • Derivatives, Energy • Economics of Energy Demand • Economics of Energy Supply • Energy Futures and Options • Investment in Fossil Fuels Industries • Trade in Energy and Energy Services

Further Reading

- Cudahy, M., and Dreessen, T. (1996). "A Review of the ESCO Industry in the United States." National Association of Energy Service Companies, Washington, DC.
- Dayton, D., Goldman, C., and Pickle, S. (1998). The energy services company (ESCO) industry: Analysis of industry and market trends. In "Proceedings of the 1998 ACEEE Summer Study," Vol. 6, pp. 29–45. American Council for an Energy-Efficient Economy, Washington, DC.
- Goldman, C., and Dayton, D. (1996). Future prospects for ESCOs in a restructured electricity industry. In "Proceedings of the 1996 ACEEE Summer Study," Vol. 10, pp. 59–69. American Council for an Energy-Efficient Economy, Washington, DC.
- Goldman, C., Jurgens, P., Fowle, M., Osborn, J., and Kawamoto, K. (2000). Historical performance of the U.S. ESCO industry: Results from the NAESCO project database. In "Proceedings of the 2000 ACEEE Summer Study," Vol. 5, pp. 89–102. American Council for an Energy-Efficient Economy, Washington, DC.
- Goldman, C., Osborn, J., Hopper, N., and Singer, T. (2002). "Market Trends in the U.S. ESCO Industry: Results from the NAESCO Database Project." LBNL-49601. Lawrence Berkeley National Laboratory, Berkeley, CA.
- Osborn, J., Goldman, C., Hopper, N., and Singer, T. (2002). Assessing U.S. ESCO industry performance and trends: Results from the NAESCO database project. In "Proceedings of the 2002 ACEEE Summer Study," Vol. 5, pp. 233–246. American Council for an Energy-Efficient Economy, Washington, DC.
- Shippee, G. (1996). Energy service companies: New trends in a changing industry. In "Proceedings of the 1996 ACEEE Summer Study," Vol. 10, pp. 153–156. American Council for an Energy-Efficient Economy, Washington, DC.
- U.S. Department of Energy (2000). "International Performance Measurement and Verification Protocol." U.S. Department of Energy, Washington, DC.
- Vine, E., Murakoshi, C., and Nakagami, H. (1998). International ESCO business opportunities and challenges: A Japanese case study. *Energy Int. J.* 23, 439–447.
- Vine, E., Nakagami, H., and Murakoshi, C. (1999). The evolution of the U.S. energy service company (ESCO) industry: From ESCO to super ESCO. *Energy Int. J.* 24, 479–492.

Entrainment and Impingement of Organisms in Power Plant Cooling



TODD P. CALLAGHAN

Massachusetts Office of Coastal Zone Management
Boston, Massachusetts, United States

-
1. Importance of Entrainment and Impingement
 2. Collecting Entrainment and Impingement Data
 3. Sources of Mortality to Entrained and Impinged Organisms
 4. Quantifying Impacts to Local Populations of Organisms
 5. Reducing Impacts

Glossary

air cooling The process by which radiator-type coils containing steam transfer heat to air passing over the coils; water is added to the cooling system only for periodic maintenance and cleaning (~200,000 gallons/day).

closed-cycle cooling The process by which cooling water cycles through a generating facility, absorbs heat from steam passing through a condenser, and then evaporates the heat to the atmosphere through cooling towers before it recycles through the facility; water is added to the cooling system to replace that lost by evaporation (~8 million gallons/day at a large facility).

condenser An apparatus used to convert steam coming from electricity-generating turbines back to water before the water is discharged from or recycled through a facility.

cooling water intake system (CWIS) The combination of intake cribs, screens, pressurized washes, debris and fish return sluices, and circulating water pumps used to withdraw cooling water.

delta T The temperature change of water passing through a cooling water system (delta T = discharge temperature – intake temperature).

diurnal cycle A response to daily changes in light intensity (e.g., the daily movement of aquatic organisms up and down in the water).

entrainment The uptake of small organisms or life stages during the process of water withdrawal.

fish return system An apparatus consisting of a pressurized wash that removes impinged fish from traveling screens

and a flume that conveys the fish back to the water body from which they were removed.

impingement The entrapment of organisms on intake screens used to prevent debris from entering cooling water intake structures.

once-through cooling The process of pumping water across a heat exchanger to condense steam from electricity-generating turbines; the heat-laden water is then discharged back to the surface water source, and the condensed steam is recycled to the facility's boilers (millions to billions of gallons of water/day are required).

traveling screen A mechanical device used to exclude organisms and debris from entering pump systems and condensers; screens are hung in front of a cooling water intake, partially above the water level, and are designed to rotate vertically to allow impinged organisms and debris to be removed when sprayed with pressurized wash water.

young-of-the-year A juvenile fish observed in the same calendar year as its hatching.

Many electricity-generating facilities withdraw cooling water from surface waters to absorb and transport waste heat. Incidental to the withdrawal of water is the concurrent uptake or entrapment of the organisms that live in those waters. Entrainment refers to the uptake of small organisms or life stages (e.g., eggs, larvae) during the process of water withdrawal, whereas impingement is the entrapment of larger organisms on screens used to prevent debris from entering cooling water intake structures.

1. IMPORTANCE OF ENTRAINMENT AND IMPINGEMENT

Entrainment and impingement susceptibility is contingent on an organism's size relative to the mesh

covering the intake (i.e., smaller organisms become entrained rather than impinged). The typical screen mesh size of a cooling water intake structures (CWIS) is 0.375 inch; therefore, entrained organisms are usually smaller than this, whereas impinged organisms are usually larger than 0.375 inch.

1.1 Factors Affecting Entrainment and Impingement Rates

Entrainment and impingement rates generally increase with increasing intake volumes and velocities unless the CWIS is equipped with a barrier to prevent organisms from approaching the CWIS. Entrainment rates are several orders of magnitude higher than impingement rates because smaller organisms and younger life stages (e.g., eggs, some larvae) are more abundant and have less capacity to overcome intake velocities compared with larger organisms. Ultimately, entrainment and impingement rates are influenced by a combination of the facility's intake flow, the location of the CWIS relative to the water body's physical characteristics, and the temporal and spatial distribution of organisms in the water body relative to intake rates.

Although entrainment and impingement can occur with any type of water withdrawal, the potential for the largest effects is at electricity-generating facilities that use once-through cooling because they typically withdraw water in significantly larger volumes than do other industries. Other cooling technologies (e.g., closed-cycle cooling, air cooling) require sig-

nificantly less water; therefore, electricity-generating facilities using these technologies have less potential for impacts to resident and migrating aquatic species.

In addition to intake flow rates, the siting of power plants and their CWIS affects the incidence of entrainment and impingement (Fig. 1). Most CWIS are located in nearshore areas of lakes, rivers, estuaries, and marine coasts. Whether in fresh or salt water, these areas have great ecological importance. Their shallow waters, submerged and emergent vegetation, and sheltered inlets provide feeding, resting, spawning, and nursery habitat for a wide variety of vertebrate and invertebrate species. In addition, when rock structures such as jetties are used to protect cooling water intakes, fish with habitat preferences for these structures (e.g., cunner, tautog) experience high rates of entrainment. CWIS have also been known to impinge waterfowl, seals, turtles, and crustaceans. Not surprisingly, when power plants that use once-through cooling withdraw their water directly from the shores of these productive waters, entrainment and impingement rates can be very high. On the other hand, offshore waters are typically cooler, provide less vegetation for shelter, and may be lower in dissolved oxygen, especially in deep, slow-moving freshwater systems. Although these offshore waters also harbor many species, most of them seek shallow nearshore areas to spawn. Consequently, offshore intakes may have lower entrainment and impingement rates.

Annual schedules of cooling water withdrawal also influence rates of entrainment. Egg and larvae

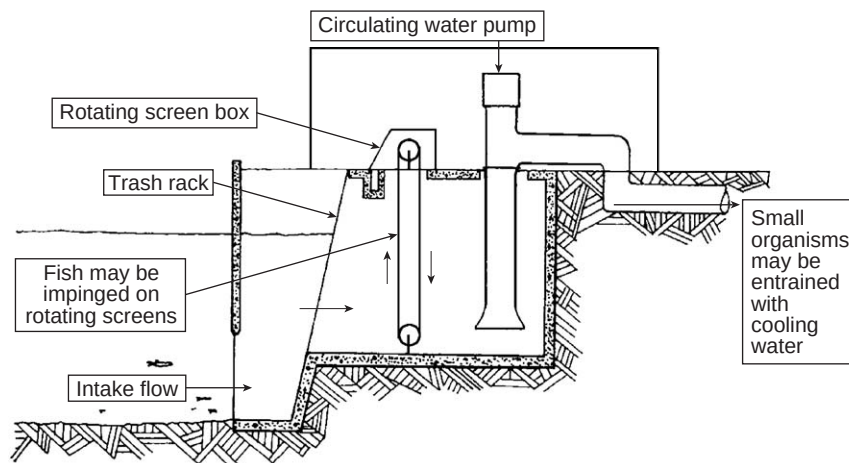


FIGURE 1 Schematic side view of a cooling water intake structure (CWIS). Circulating water pumps draw water into the CWIS. The trash rack and rotating screens keep large objects (including fish) from entering the cooling water system. Organisms that pass through the screens may be entrained into the plant's cooling system. Modified from Coutant, C., and Bevelhimer, M. (2003). Induced sweeping flows at CWIS for reducing fish impingement. In "Proceedings: Symposium on Cooling Water Intake Technologies to Protect Aquatic Organisms." U.S. Environmental Protection Agency, Washington, DC.

abundances are greatest in the water column during the spring when most organisms spawn. Consequently, facilities that withdraw large quantities of cooling water at this time are likely to have high rates of entrainment during the spring. Some facility managers have scheduled their annual maintenance outages (when cooling water withdrawal is minimal) in March or April to minimize entrainment during this biologically active period.

1.2 Importance of Measuring Entrainment and Impingement

Quantifying entrainment and impingement rates is an important step in understanding the losses to biological populations that can occur due to cooling water withdrawal. Large power plants can entrain and impinge a significant proportion of young organisms or life stages that would otherwise recruit to local biological populations. Although predators, meteorological events, fishing, and disease all contribute to mortality rates, the additional population reduction caused by power plants may contribute to the destabilization of natural oscillations in population abundances, potentially resulting in significant long-term reductions. Because entrainment and impingement rates are controllable variables, knowing the magnitude of losses due to power plants can lead to a variety of management decisions aimed at minimizing losses and maintaining biological populations.

In the United States, the Environmental Protection Agency (EPA) and state regulatory bodies are charged with protecting biological resources. The Federal Water Pollution Control Act of 1977 and its amendments (known collectively as the Clean Water Act) require that the location, design, construction, and capacity of a CWIS reflect the best technology available for minimizing adverse environmental impact. Because the law does not clearly define the term “adverse environmental impact,” vigorous and sometimes contentious debates have emerged as to the impact of cooling water withdrawal on local populations of aquatic organisms and how government agencies should regulate these impacts. The implementation of the Clean Water Act and other regulations may require facility owners to upgrade to withdrawal technologies with lower impacts (e.g., closed-cycle, air cooling). These technological upgrades or operational changes may cost tens or hundreds of millions of dollars. In some cases, implementation costs may prevent the construction of new intakes altogether. Under Section 316(b) of the Clean Water Act, a company faced with attaining

regulatory requirements can assert that the investment is “wholly disproportionate” to the environmental benefits to be gained from the modifications. Furthermore, some industry representatives contend that significant upgrades will affect the availability and price of energy across regional grid structures. It is the responsibility of the EPA to evaluate the balance of information and to determine whether technological upgrades are warranted. Clearly, managing entrainment and impingement impacts has important consequences for natural resources as well as for the energy industry.

2. COLLECTING ENTRAINMENT AND IMPINGEMENT DATA

With such important economic and resource protection decisions dependent on entrainment and impingement values, quantifying these parameters reliably is critical. However, there is no standard methodology or procedure for collecting entrainment and impingement data. Methods for estimating entrainment and impingement are determined on a site-by-site basis by the characteristics of the facility's CWIS, the flow rate through the facility, the life history of the species at risk, the characteristics of the water body, the priorities of the regulatory community, and the costs of the various methodologies.

2.1 Entrainment Sampling

Because of the patchy distribution of larvae and eggs, achieving the sampling effort necessary to calculate a reliable estimate of abundances can be challenging. To quantify entrainment impacts, estimates of the number of larvae and eggs going through the plant must be compared with estimates of the number of larvae and eggs of each species in the source water body adjacent to the plant. These ratios will provide species-specific estimates of the cropping of biological populations by the power plant.

Entrainment rates are highly dependent on the spatial and temporal distribution of larvae and eggs. Water currents, diurnal cycles, time from hatching, and species type all can affect entrainment estimates by influencing densities of larvae and eggs adjacent to intakes. In addition, flow rate and CWIS configuration influence the proportion of larvae and eggs adjacent to the intake that are drawn into a facility's cooling system. For example, increasing flow rate usually increases entrainment rate; however, larvae may be

able to avoid becoming entrained by avoiding the part of the water column from which a plant withdraws its cooling water. Because of the multitude of influences on entrainment rates, data collections need to occur frequently enough so that they are representative of the various combinations of conditions present in the source water body. Moreover, interannual variability is often great enough to warrant replication across years as well as within years.

Early efforts to quantify the mortality of larvae within cooling water systems used conical nets suspended in both the intake structures and discharge plumes of electricity-generating facilities. (It should be noted that larval sampling techniques also capture floating eggs; thus, entrainment of eggs and larvae are often assessed simultaneously.) However, one of the problems with this method is that the survival of larvae is influenced by the velocity of water flowing through the net, with higher velocities leading to increased mortality. Higher velocities also increase the risk of extrusion through the net and increase the limitations of the net to catch larvae, resulting in an incomplete sample. Furthermore, discharge plumes often have much greater velocities than do intakes. As a consequence, larvae collected in discharges may have higher mortality due to the force of the discharge water trapping them in collection nets. In short, traditional net sampling can add another source of mortality when only the mortality caused by the cooling system is of interest.

To reduce sampling-related mortality, some researchers have used a flume-like sampling station, floated in front of the intake or discharge, onto which water can be pumped. Unlike net sampling, a larval table equalizes sampling stresses for both intake and discharge samples by withdrawing water at a constant rate in both areas. However, the pump used to bring larvae onto the sampling table can cause a significant amount of mortality to the larvae.

To avoid pump-related mortality, samples may be taken using a pump that is downstream of the sampling screen. In this way, water and entrained organisms are pulled rather than pushed onto the sampler and its screens; therefore, no organisms have to go through the pump system. Other sampling efforts have eliminated the pump altogether, using the velocity of water in the discharge to bring water into the floating sampling station. However, this method does not work in intake structures where the velocity of water is not great enough to force water onto the floating sampling station.

Although innovations in sampling techniques have produced methods that increase the reliability of

larval mortality estimates, conical nets and pumps are still commonly used due to cost or design issues associated with sampling in intakes and discharges. For example, only one of the nine entrainment studies conducted for power plants along the coast of Massachusetts between 1994 and 2003 reported the use of a larval table, whereas all of the others used conical net sampling.

2.2 Impingement Sampling

Impingement is easier to measure than is entrainment because the presence or absence of larger organisms is more obvious and, therefore, easier to quantify. Impingement samples can be collected directly from traveling screens or from collection pens built into a fish return system. Often, impingement rates are assessed for only part of a 24-h period and long-term rates are extrapolated from the short-term assessment intervals. Impingement rates are usually reported as number of fish per year or number of fish per volume of flow through the power plant.

The most frequently impinged fish are usually young-of-the-year fish, although older juveniles and adults are also impinged. Schooling juvenile fish (e.g., menhaden) often follow shoreline margins and can become impinged in large numbers, especially during the late summer and fall. Impingement rates for juvenile schooling fish can exceed 1000 fish per hour and tens of thousands weekly.

3. SOURCES OF MORTALITY TO ENTRAINED AND IMPINGED ORGANISMS

Entrained and impinged organisms can incur several types of mortality associated with a CWIS. The risks of mortality arise due to physical, chemical, or thermal stresses within the cooling water system. In addition, there is the possibility of latent mortality after an organism has been discharged from the cooling water system.

3.1 Physical Stress

Physical stresses result in direct damage to the tissues of entrained organisms as they pass through the pumps and pipes of a cooling system. Within the system, they are subject to abrasion, compression, and shear stress. Physical impacts to impinged organisms may include suffocation while pressed

against the intake screens and tissue damage from pressurized screen washes and the velocity of the return sluice. Some sluices have been known to drop fish several vertical feet into the receiving water, whereas others have sent fish to rocky areas that are dry at low water. Such poorly designed sluices may stun or injure fish, leaving them more susceptible to predation. For example, birds have been known to frequent the outfalls of fish return systems, plucking the returned fish from the water. Furthermore, some sluices deliver fish directly to the discharge channel, where they experience turbulence as well as elevated temperatures that may contribute to mortality. In addition, sluices that return fish too close to intake structures may contribute to reimpingement, especially if the fish are stunned by the physical stress of impingement, high-pressure spray, and a turbulent return sluice.

3.2 Chemical Stress

Chemical mortality results from deleterious physiological effects induced by chemicals added to the cooling water as it flows through the facility. Biocides (typically chlorine) are used to keep the cooling water system free of fouling organisms such as algae, barnacles, and mussels. Biocides injected directly into cooling water will kill entrained larvae and eggs. Ultimately, these biocides are discharged to the receiving water body and, therefore, may also affect the organisms in the vicinity of the outfall. If biocides are used to keep traveling screens free of fouling organisms, impinged fish can also be dosed with toxic chemicals.

3.3 Thermal Stress

Thermal stress to entrained organisms occurs as the water that surrounds them absorbs heat from condenser tubes. Excessive temperatures cause the breakdown of proteins and can result in developmental problems or death. The delta T of a plant directly influences entrainment mortality. Minimizing delta T will increase the rate of survival because most organisms have the physiological ability to withstand small increases in temperature. However, for all organisms, there is a temperature threshold above which acclimation to higher temperatures will not occur and mortality will follow. The duration of exposure to elevated temperatures also affects survival. Therefore, the risk of mortality to organisms that have survived being entrained will increase as they spend

increasing amounts of time in a heated discharge or if they are reentrained into the power plant.

Increased temperatures may affect impinged organisms if a facility is designed to clean its screens by backflushing, which involves using heated water to dislodge debris and fouling organisms attached to the intake screens. This method can also physically damage impinged fish while subjecting them to excessive temperatures. In addition, impinged organisms can be thermally stressed if the return sluice deposits them in the plant's thermal effluent.

3.4 Latent Mortality

The combination of physical, chemical, and thermal stresses experienced during entrainment or impingement can make organisms more susceptible to pathogens, infections, or predation after they are returned to the water body. The mortality associated with these combined effects is called latent mortality. Latent mortality is difficult to quantify because long-term effects might not manifest themselves during the time frames of most studies. However, some researchers have attempted to quantify latent mortality by holding recently entrained larvae and impinged fish in tanks and monitoring their survivorship. Still, in practice, most assessments conservatively assume 100% through-plant mortality in the interest of expediency and cost savings, obviating the need for a latent mortality estimate.

4. QUANTIFYING IMPACTS TO LOCAL POPULATIONS OF ORGANISMS

The simplest method of evaluating cooling water withdrawal effects is to estimate the annual absolute losses of all life stages of species entrained or impinged. The magnitude of the impact can then be assessed relative to management criteria or historical trends in species abundances. Beyond calculating absolute losses, there are a number of methods for quantifying power plant impacts due to entrainment and impingement. These include equivalent adult analysis, calculating production forgone, calculating indirect losses to predators due to mortality to forage species, calculating conditional mortality, comparing equivalent adults or production forgone to commercial landings, and modeling populations.

4.1 Entrainment Impact Assessments

The equivalent adult method of assessing entrainment and impingement effects provides managers with the number of individuals (eggs, larvae, or juveniles) that would have survived to reproductive age if they had incurred only those mortality pressures independent of the generating facility. Equivalent adult losses can be calculated for those species whose life stage-specific mortality rates are known. Typically, these values are known for only a select few commercially or recreationally important fish species.

A second means of calculating impacts is called the production forgone method. Production forgone is similar to calculating equivalent adults except that it is used to determine the biomass (measured in pounds or kilograms) that would have been produced if the entrained and impinged organisms had experienced sources of mortality only in the absence of the plant. Equivalent adult analysis is usually applied to species that are part of a commercial or recreational fishery, whereas production forgone is usually used to measure impacts to nonfished species.

Another impact assessment method takes into account the indirect losses caused by the mortality of species that are prey to species that are part of a managed fishery. This assessment occurs in three steps. First, production forgone is calculated for the prey species. Next, for each predator species in the fishery, production forgone of their prey is multiplied by a food conversion efficiency factor that accounts for the biomass of prey consumed by the predators that results in an increase in predator biomass. Finally, the loss of predator biomass due to entrainment of young life stages of their prey is calculated by apportioning the biomass of prey to individual predator species based on the specific proportions of each prey species in a predator's diet and the relative biomass of each predator in the area of interest. In effect, this method converts biomass of prey species to biomass of predator species that have known commercial or recreational value.

Yet another assessment method, and perhaps the most controversial because it is the most subjective, is to determine conditional mortality rates. Conditional mortality is the fractional reduction of a population of organisms due only to entrainment and impingement. To estimate conditional mortality, one must (1) define the boundaries of the local populations of organisms entrained and impinged, (2) estimate the abundance of organisms in these populations, and (3) compare entrainment and impingement data for

individual species with population abundances adjacent to the intake. In the absence of genetic and behavioral information that could truly bound a population of interbreeding organisms, impact assessments rely on subjective semi-arbitrary boundaries defined by geographical features, hydrodynamics, and policy decisions by resource agencies. Increasing population boundaries, and thus a population's size, decreases the magnitude of impacts. Given that the magnitude of impacts can lead to expensive plant upgrades, spirited discussions can surround the definition of population boundaries.

A similar approach has been used to determine the percentage of larvae flowing past a plant that is entrained. In this assessment, the rate of larvae flowing through a plant (i.e., the density in the discharge multiplied by the plant's flow rate) is divided by the flux of larvae flowing past the intake structure (i.e., the density in the water body multiplied by the net current flow). This method provides an estimate of the plant's ability to affect the pool of larvae in its source water. The reliability of this estimate of impact is dependent on empirical data and accurate measurements of flow rate and larval density. In addition, because ocean currents and some rivers can flow in opposite directions depending on the tide, the opportunity for larvae to be entrained is dependent on how many times they flow back and forth in front of a CWIS.

Some impact assessments have compared entrainment and impingement losses with annual commercial landings of affected fish species. However, this estimate of impact can be misleading if the commercial landings data are from an area that is larger than the area of interest immediately adjacent to an intake structure. Another reason why commercial landings are not a good point of comparison is that they are influenced by fishing effort, management policies restricting effort, and the market price of various fish species. Because of the variability in these factors, commercial landings are not a reliable measure of species abundance.

After bounding the populations of interest and estimating current impacts, further analysis may involve predicting future impacts to populations using mathematical models. Models may be used to project how a population will respond to various operating or mitigation scenarios.

In models, the distribution of larvae and eggs is often represented as uniform, with their densities averaged across a section of water column. This simplification keeps model projections manageable while decreasing run time and financial costs.

However, in reality, the distribution of larvae and eggs is heterogeneous and the risk of higher density patches of organisms becoming entrained will vary over time. To take advantage of the simplicity of projection models and account for the real-world patchiness of eggs and larvae, cross-sectional averages of organism densities determined by models can be converted to more realistic densities by multiplying by a ratio called the withdrawal factor (W). W is the ratio of the density of entrained organisms in the intake structure to the density of entrainable organisms found in the adjacent water body.

Some models include a recirculation factor to account for the proportion of entrained organisms that reenter the CWIS because the discharge is close enough to the intake to reentrain discharge water. Recirculation can be determined through modeling or dye studies of the movement of the discharge plume. The amount of recirculation is site specific and may be as great as 25%. Incorporating recirculation of flow into a model tends to decrease estimates of impacts because a portion of the organisms in the cooling water is assumed to have already been killed by the plant's cooling system.

When the goal of an impact assessment is to quantify the specific sources of mortality to entrained organisms, also known as the entrainment mortality factor (f), the individual mortality rates due to the physical, chemical, or thermal stresses associated with passing through the cooling water system are estimated. Estimating f entails separating out mortality incurred through the sampling process (i.e., physical stresses due to pumps, nets, and handling time) from the mortality caused solely by the cooling water system. Because there are so many potential sources of mortality once eggs and larvae are in a cooling system, and because quantifying them reliably to the satisfaction of regulatory agencies can be an onerous and costly task, many impact assessments will conservatively estimate that 100% of all entrained eggs and larvae suffer mortality.

When the 100% entrainment mortality assumption is not used, through-plant mortality (f) is important to modeling operational procedures needed to reduce entrainment mortality because trade-offs exist between delta T and flow rate. When the flow of water passing through a cooling water system decreases, the relative amount of heat per volume of cooling water increases, thereby increasing delta T. Under certain operational scenarios, the potential increase in mortality to entrained larvae and eggs associated with an increased delta T can

offset reductions in entrainment achieved by flow reductions. However, when through-plant mortality is assumed to be 100%, flow rate becomes the most important factor because delta T has no effect on impact assessment. A further consideration is that there is also a trade-off between flow and thermal output to the receiving waters. Increasing flow through the plant can decrease delta T and reduce thermal impacts within the cooling system and in the receiving water.

4.2 Impingement Impact Assessments

Impacts due to impingement are relatively easier to quantify than are entrainment impacts. Representative samples of impinged organisms can be collected directly from traveling screens by special sampling trays or can be sampled from the return sluiceway. Although fish are usually the most frequently collected organisms, crustaceans, turtles, birds, and seals can also be impinged. However, in practice, most impact assessments focus on the number of fish impinged. In the overall assessment of a plant's impact, losses due to impingement are added to equivalent adult, production forgone, and conditional mortality estimates.

5. REDUCING IMPACTS

The most effective way of reducing impacts caused by cooling water withdrawal is to avoid the use of cooling water altogether by initially designing or upgrading a facility to utilize air cooling. However, in many cases, this is cost-prohibitive. Alternatively, closed-cycle cooling costs significantly less than does air-cooling and requires substantially less water than do once-through cooling systems; therefore, it has lower entrainment and impingement impacts. When air cooling or closed-cycle cooling systems are not an option due to design, siting, or cost constraints, power plant owners can reduce entrainment and impingement impacts through low-impact siting of a new CWIS, operational changes, technological upgrades, and/or biological mitigation. The efficacy of each of these methods has been debated within the CWIS permitting process and has been shown to be dependent on the site-specific characteristics of an individual facility and the distribution of organisms around its intake structure(s). Finally, the implementation of new and existing regulations can potentially reduce entrainment and impingement impacts on a regional scale.

5.1 CWIS Siting

When the use of once-through cooling is the only option, one way of minimizing impacts is through the careful siting of new intake structures after consideration of the temporal and spatial distribution of organisms in the source water body. Identification and avoidance of spawning runs, egg-laying sites, hydrodynamically low-energy or sheltered areas where larvae may collect, juvenile nurseries, feeding areas, and overwintering refuges can greatly reduce potential impacts. To achieve this, some intakes may be resited to offshore locations to remove them from the most biologically productive habitats.

Intake structures raised slightly above the bottom will aid in keeping bottom-dwelling organisms from entering the intakes (but it will do little to deter planktonic or swimming organisms). Although the careful siting of intakes outside the most productive habitats will reduce impacts, there will always be some level of impacts due to CWIS unless some physical characteristic of the water (e.g., consistently low dissolved oxygen) prevents organisms from living there.

5.2 Operational Changes

Changes to operational procedures at the facility can also reduce impacts. For example, flow reductions may reduce impacts because entrainment and impingement rates are often directly correlated with flow rates. One way of reducing flow is to use variable-speed pumps (instead of on/off pumps) to reduce water intake during months when biological activity is highest. Another method is to operate in what is called "piggyback" mode, where the discharge from one unit is used to cool another unit. This cooling water recycling measure increases the risks to organisms that are initially entrained (because they may go through the cooling system several times) but decreases the overall amount of new cooling water needed, thereby decreasing impacts. Another operational procedure that can reduce impingement impacts is the continuous rotation of screens in front of the intake. Typically, screens are rotated only when excessive blockage of the screens (by vegetation, debris, or organisms) decreases intake rates beyond the threshold of condenser efficiency. Continuous rotation of screens during months when impingement is known or expected to be highest can reduce the amount of time that impinged organisms spend on the screens,

thereby increasing the probability of survival if there is an adequate return sluice.

5.3 Technological Upgrades

In addition to carefully planned CWIS siting and operational protocols, technological modifications and additions provide an opportunity for reducing impacts.

The upgrade to modified traveling screens with a well-designed fish return system can enhance survivorship of many impinged organisms. Improvements to traveling screens and fish return systems to minimize the physical stress on fish before they return to the receiving water body include modified fish buckets on the screens and low-pressure spray washes, variable speed drives for rotating the screens (3–30 feet/min), and separate fish and debris return troughs (Fig. 2). Furthermore, designing return sluices so that they deliver fish back into the receiving water body at all water levels (i.e., not allowing returned fish to drop on the shore at low water levels), and far enough away from the intake so as not to be reimpinged, may also increase survival rates of returned fish.

When biocides are necessary to reduce biofouling on intake screens, the location of biocide injection

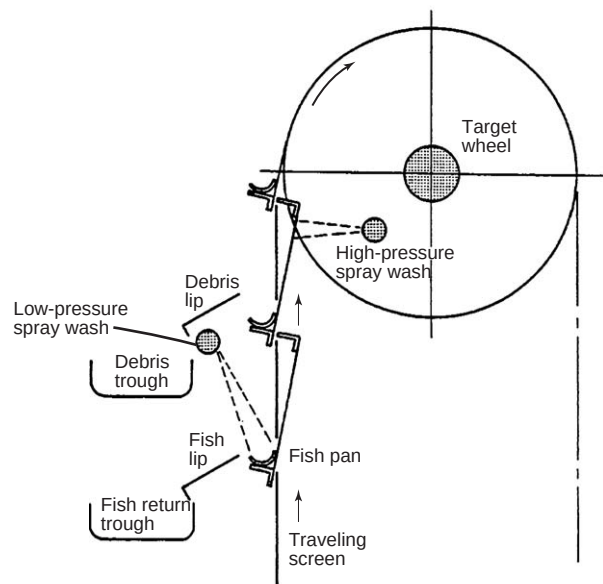


FIGURE 2 Schematic side view of a modified traveling screen with separate fish and debris troughs and a low-pressure spray to remove impinged fish. From Johnson, I. C. (2003). Fish return system efficacy and monitoring studies for JEA's Northside Generating Station. In "Proceedings: Symposium on Cooling Water Intake Technologies to Protect Aquatic Organisms." U.S. Environmental Protection Agency, Washington, DC.

can be important to the survival of impinged organisms. Positioning the biocide injectors so that they do not dose impinged organisms (i.e., aiming injectors toward the descending portion of the traveling screen) may help to increase survival rates. Furthermore, using targeted biocides (i.e., those that kill only biofouling organisms and not fish) will also increase the survival of impinged organisms.

Other technological upgrades include physical and behavioral barriers. Filter fabric curtains, fixed panel screens, or wedge wire screens placed in front of the CWIS create a physical barrier between organisms in the water body and the intake structure (Fig. 3). These barriers, if installed and maintained properly, reduce intake velocity and can prevent impingement and reduce entrainment (although the smallest eggs and larvae may still pass through the barrier and into an intake). Other systems that use angled screens or vertical panels placed at 90 degrees to the flow of water create flows that divert fish to a bypass on the outside of the screens that prevents them from entering the cooling system (Fig. 4). These systems have been shown to reduce impingement but not entrainment. In a similar way, localized turbulent flows generated at points near an intake can “sweep” fish away from intakes (Fig. 5). This method has been shown to be successful in preventing impingement of certain species of fish.

Placing the intake structure under gravel or behind a porous cobble dike is another upgrade that can potentially reduce entrainment and impingement if the intake velocity across the gravel or cobble is less than it would be across an unprotected intake. However, the

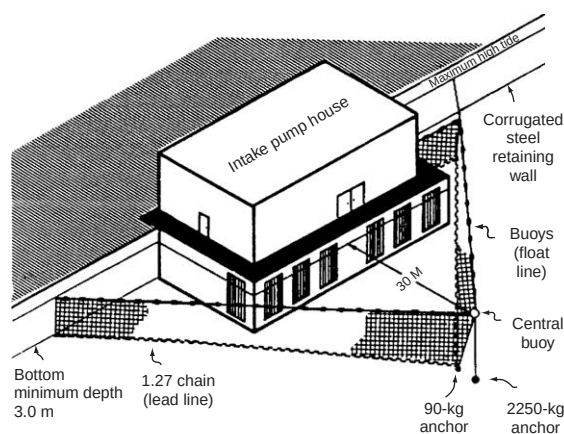


FIGURE 3 Schematic of a barrier net for reducing impingement at Bowline power plant in New York. From U.S. Environmental Protection Agency. (1994). “Preliminary Regulatory Development Section 316(b) of the Clean Water Act Background Paper Number 3: Cooling Water Intake Technologies.” EPA, Washington, DC.

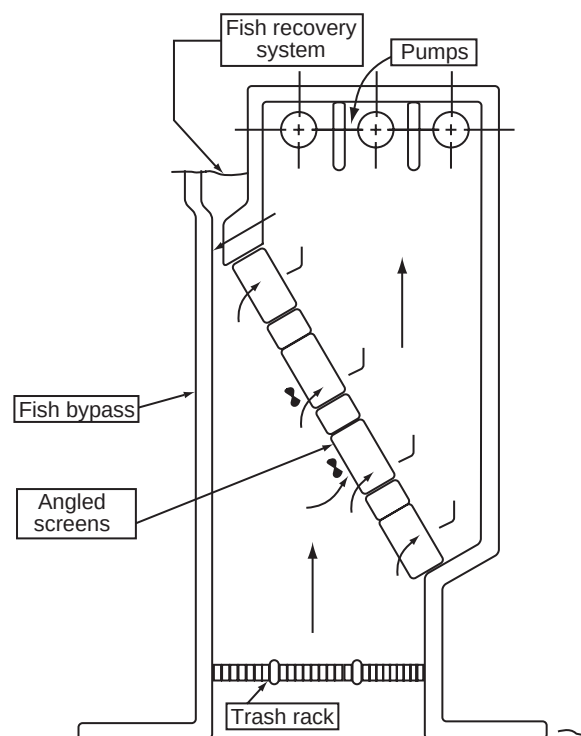


FIGURE 4 Schematic of angled screens located upstream from a cooling water intake to redirect fish toward a bypass around the intake structure. From U.S. Environmental Agency. (1994). “Preliminary Regulatory Development Section 316(b) of the Clean Water Act Background Paper Number 3: Cooling Water Intake Technologies.” EPA, Washington, DC.

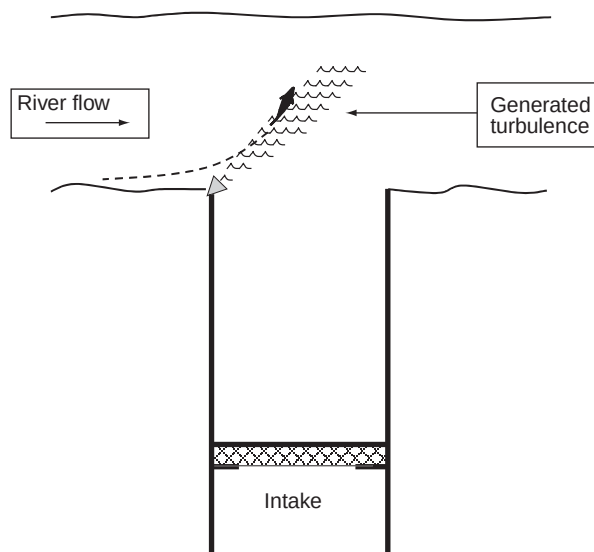


FIGURE 5 Diagram of how induced turbulent flows can sweep fish away from an intake structure. From Coutant, C., and Bevelheimer, M. (2003). Induced sweeping flows at CWIS for reducing fish impingement. In “Proceedings: Symposium on Cooling Water Intake Technologies to Protect Aquatic Organisms.” U.S. Environmental Protection Agency, Washington, DC.

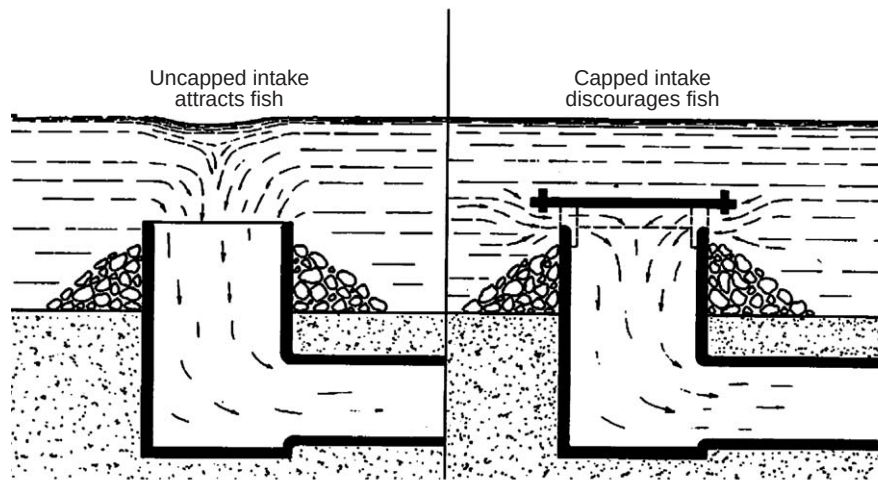


FIGURE 6 Schematic of the difference in intake velocity between capped and uncapped cooling water intakes. From U.S. Environmental Protection Agency. (1994). "Preliminary Regulatory Development Section 316(b) of the Clean Water Act Background Paper Number 3: Cooling Water Intake Technologies." EPA, Washington, DC.

efficacy of an under-gravel or porous dike barrier may decrease due to clogging by debris and sediment or fouling by colonizing aquatic organisms.

Other methods of impact reduction are behavioral barrier systems. One behavioral system that works on the premise that fish avoid rapid changes in horizontal flow is called a velocity cap. A velocity cap is placed on an offshore vertical intake and converts vertical flow into horizontal flow (Fig. 6). Like the panel and angled screen systems, this device has only been shown to reduce impingement but not entrainment impacts, and it is more effective for some species than for others. Acoustical or optical (strobe light or air bubble) deterrents have also been used as behavioral barriers to keep fish from approaching CWIS, but the efficacy of these techniques has been shown to be limited and species specific.

5.4 Biological Mitigation

An innovative but still unproven method for reducing impacts is biological mitigation. The theory behind biological mitigation is to accept that a certain number of losses will occur at a given facility and that these losses can be balanced by adding more individuals to the affected biological populations.

Some biological mitigation strategies propose to conserve or restore habitat so as to increase the abundance of certain life stages of affected populations. To facilitate this process, the EPA created a "habitat-based replacement cost" method of valuing

biological resources lost to entrainment and impingement. This valuation system identifies the costs of habitat restoration required to replace biological resources lost to entrainment and impingement. One of the strengths of this method is that it recognizes that all organisms, and not just those that are recreationally or commercially important, are valuable to an aquatic ecosystem. An intensive search for habitat restoration opportunities for each affected species and a system of prioritizing the options based on their expected benefits to the affected species is required. Once implemented, only intensive, long-term, and carefully controlled studies can determine whether habitat remediation has led to increases in the abundance of species life stages affected by power plants.

Another form of biological mitigation is restocking. In restocking plans, the sum total of equivalent adult fish entrained plus the number of adult fish that have been impinged is replaced by hatchery-raised fish of the same species. One of the challenges with this method is to ensure that the hatchery-raised fish have the same genetic makeup and ecologically important behavioral responses (e.g., habitat selection, prey capture, predator avoidance) as do the fish they are meant to replace. An additional consideration is that restocking often occurs without respect to restocking the same life stages that are experiencing the greatest impacts. For example, juvenile fish grown in a hatchery and released as mitigation for entrainment impacts will not have the same ecological role as will the larvae they are meant to replace.

Biological mitigation is a field that requires more research before its efficacy can be ascertained.

5.5 New Regulations to Reduce Impacts

At the regulatory level, the EPA has established requirements to reduce entrainment and impingement impacts at new facilities using once-through cooling. On December 18, 2001, it issued its “final rule” implementing Section 316(b) of the Clean Water Act for new facilities. The national requirements for new facilities establish a two-track system. Under track I, facilities using more than 10 million gallons of intake water per day must meet capacity, velocity, and capacity- and location-based proportional flow requirements. Facilities using more than 2 million but less than 10 million gallons of intake water per day do not have to meet the capacity limitation but must use technologies to reduce entrainment. More specifically, the track I standards require the following:

- Intake flow must be commensurate with that of a closed-cycle system (although facilities with 2–10 million gallons/day intake flow are exempt).
- The maximum through-screen intake velocity is 0.5 feet per second.
- If an intake is sited in a freshwater river or stream, flow must be less than 5% of the annual mean flow.
- If an intake is sited in a lake or reservoir, withdrawal from a cooling water intake must not disrupt the natural thermal stratification of the source water body.
- If an intake is sited in a tidal river or an estuary, withdrawal can be no greater than 1% of the volume of the water column within the area centered around the intake, with the diameter defined by the distance of one tidal excursion at mean low water.

Under track II, a facility has the option to demonstrate that alternative technologies or operations chosen by the manager of the facility will reduce entrainment and impingement mortality to a level comparable to that required under track I. Also under track II, flow levels are not limited to those

commensurate with closed-cycle cooling, and intake velocity does not need to be less than 0.5 feet per second; however, the other withdrawal constraints are the same. In addition, facility operators choosing track II need to undertake a comprehensive biological evaluation demonstrating that the chosen technologies and operations have an adverse environmental impact level commensurate with that achievable under track I. The EPA has been evaluating requirements to reduce entrainment and impingement effects at existing facilities that use cooling water in order to meet the goals of track II.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Electricity, Environmental Impacts of • Fisheries and Energy Use • Hydropower, Environmental Impact of • Thermal Pollution • Wetlands: Impacts of Energy Development in the Mississippi Delta

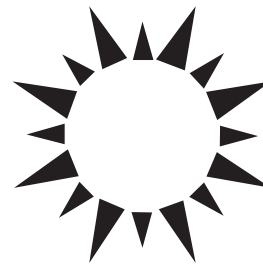
Further Reading

- Barnthouse, L. W., Klauda, R. J., Vaughan, D. S., and Kendall, R. L. (eds.). (1988). “Science, Law, and Hudson River Power Plants: A case study in environmental impact assessment monograph 4.” American Fisheries Society, Bethesda, MD.
- Boreman, J., Goodyear, C. P., and Christensen, S. W. (1981). An empirical methodology for estimating entrainment losses at power plants sited on estuaries. *Trans. Am. Fish. Soc.* **110**, 255–262.
- Environmental Protection Agency. (2001a). Efficacy of cooling water intake structure technologies. www.epa.gov/waterscience/316b/technical/ch5.pdf.
- Environmental Protection Agency. (2001b). Proposed Section 316(b) phase II existing facilities rule case study analysis. www.epa.gov/waterscience/316b/casestudy/.
- Van Winkle, W., Christensen, S. W., and Suffern, J. S. (1979). “Incorporation of Sublethal Effects and Indirect Mortality in Modeling Population-Level Impacts of a Stress, with an Example Involving Power-Plant Entrainment and Striped Bass,” ORNL/NUREG/TM-288. Oak Ridge National Laboratory, Oak Ridge, TN.

Entropy

CARLO BIANCIARDI and SERGIO ULGIATI

University of Siena
Siena, Italy



1. Introduction
2. Entropy as a Thermodynamic State Function
3. Statistical Meaning of Entropy and Boltzmann's H-Theorem
4. Entropy and Information Theory
5. The Free-Energy Concept
6. Conclusion

Glossary

energy Energy is usually defined as the capacity for doing mechanical work—that is, displacing an object from a position to another by means of the application of a force. This textbook definition is sometimes criticized for being reductive and not always easily applicable to processes other than purely physical ones. A broader definition for energy could be the ability to drive a system transformation, which clearly includes all kinds of physical, chemical, and biological transformations.

enthalpy Enthalpy, H , is a thermodynamic property expressing the total energy content of a system. It is defined as $H = U + pV$, where U is the internal energy of the system (any kind of energy form of its molecules) and pV is the system's energy in the form of expansion work (in fact pV , the product of pressure and volume, has the physical dimensions of energy). In a chemical reaction carried out in the atmosphere the pressure remains constant and the enthalpy of reaction (ΔH) is equal to $\Delta H = \Delta U + p\Delta V$. Enthalpy is a state function, because it only depends on the initial and final states of the system, no matter how the process develops.

macroscopic/microscopic Any system is composed of parts. Parts contribute to the functioning of the whole. Matter is composed of molecules, an economy is made up of production sectors and firms, for example. The dynamics of the whole system may be or appear to be different from the behavior of the component parts. A car may be stopping in a parking lot, yet its molecules possess a chaotic motion, which cannot be seen when we look at the car. A national economy may be in good health, on average, yet some sectors or local firms may face a difficult

survival. Macroscopic therefore concerns the system's dynamics, considered as a whole, while microscopic applies to the behavior of the lower scale components.

perfect gas A perfect or ideal gas is a gas the molecules of which are infinitely small and exert no force on each other. This means that their motion is completely chaotic and does not depend on any mutual interaction, but only on completely elastic collisions, which cause a transfer of energy from a molecule to another and a change of their direction and velocity.

probability The probability of an event is the ratio of the number of times the event occurs to a large number of trials that take place. The number of trials (all events that may occur, including those different from the ones fitting the request) must be very large for the probability to be reliable. As an example, assume we cast a die. If the die is perfectly constructed, the probability of a five is 1 out of 6 possible different castings. However, in a real case, we must cast it many times to ensure that a five occurs out of the other possible events in a 1:6 ratio.

reversible/irreversible A reversible change in a process is a change that can be reversed by an infinitesimal modification of any driving force acting on the system. The key word infinitesimal sharpens the everyday meaning of the word reversible as something that can change direction. In other words, the state of the system should always be infinitesimally close to the equilibrium state and the process be described as a succession of quasi-equilibrium states. A system is in equilibrium with its surroundings if an infinitesimal change in the conditions in opposite directions results in opposite changes in its state. It can be proved that a system does maximum work when it is working reversibly. Real processes are irreversible, therefore there is always a loss of potential work due to chemical and physical irreversibility.

self-organization All systems in nature self-organize—that is, use the available energy and matter inputs in order to achieve higher levels of structure and functioning under natural selection constraints. In so doing, they may maintain their present level of complexity or become more and more complex by also increasing the number of components, functions, and interactions.

statistic The science dealing with sampling and recording events, which occur in a different way in dependence of a given parameter (time, temperature, etc.). It estimates

events by recording and analyzing a given behavior verified by an appropriately large number of random samplings. The statistical analysis may lead to hypotheses about the behavior of a given set of events.

temperature Temperature is a property of an object that tells us the direction of the flow of heat. If the heat flows from A to B, we can say that the temperature of the body A is higher than that of the body B. In some way it measures the concentration of heat within matter and can be linked to the average kinetic energy of the molecules in a body. A thermometric scale (centigrade or Celsius scale of temperature) can be defined by dividing by 100 the temperature interval between the temperature of melting ice (0°C) and the temperature of boiling water (100°C) at sea level. Other existing scales depend on the reference parameters chosen as the zero level. A scale of absolute temperatures (Kelvin) defined as described in this article is linked to the Celsius temperature, t , by means of the relationship: $T\text{ (K)} = t\text{ (}^{\circ}\text{C)} + 273.15$. A unit absolute degree (K) coincides with a unit centigrade degree ($^{\circ}\text{C}$).

work The transfer of energy occurring when a force is applied to a body that is moving in such a way that the force has a component in the direction of the body motion. It is equal to the line integral of the force over the path taken by the body, $\int \mathbf{F} \times d\mathbf{s}$. Its meaning is different from the usual life work, often associated with activities not always involving visible motion. Physical work requires a body motion under the action of the force.

Entropy, a measure of energy and resource degradation, is one of the basic concepts of thermodynamics, widely used in almost all sciences, from pure to applied thermodynamics, from information theory to transmission technology, from ecology to economics. In spite of the mathematical difficulty of its definition and application for the nonspecialists, the word entropy is often used in the everyday life, where it assumes less constrained, looser meanings than in the scientific community, being generally associated to disorder, lack of organization, indefiniteness, and physical and social degradation. In this article, we first give the mathematical definition of entropy and discuss its meaning as a basic thermodynamic function in reversible, ideal processes. The relationship of the entropy concept with real irreversible processes, statistical thermodynamics, information theory, and other sciences is then reviewed.

1. INTRODUCTION

The development of thermodynamics as well as the introduction of the concept of entropy is deeply rooted into the fertile technological ground of the Industrial

Revolution, which took place in England between the beginning of the 18th and the second half of the 19th century. The development of steam-powered machines (first used to pump water out of coal mines and then to convert into work the heat generated by coal combustion) amplified both technological and purely scientific research, in an interplay of mutual reinforcement. The “miner’s friend,” an engine built by Thomas Savery in 1698, had a very low efficiency. Other more efficient engines (such as Newcomen’s engine in 1712) opened the way to Watt’s steam engine (1765), the efficiency of which was much higher than Newcomen’s engine. The Scottish technician James Watt had to solve several technological and theoretical problems before his machine could effectively replace the previous steam engines. However, at the beginning of 19th century, hundreds of Watt’s machines were used in Europe and constituted the basis for the significant increase of production that characterized the Industrial Revolution.

The conversion of energy from one form to another, and in particular the conversion of heat into work, has been investigated by many scientists and technicians in order to deeper understand the nature of heat toward increased efficiency. These studies yielded a general framework for energy conversion processes as well as a set of laws, the so-called laws of thermodynamics, describing the main principles underlying any energy transformation. It is worth noting that the investigation about thermal engines and the limits to the conversion of heat to mechanical work (second law, Carnot, 1824, “*Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance*”) preceded the statement of the first law of thermodynamics. This is because heat conversion to mechanical work (steam engine, second half of 18th century) generated first an attention to the existing constraints and then a much clearer understanding of the nature of heat and a quantitative equivalence between heat and mechanical work (middle of the 19th century).

Clausius then provided Carnot’s “*Réflexions*,” with a stronger mathematical basis and introduced the concept of entropy, S , in so developing a clear statement of the second law. The entropy concept is perhaps the most complex one in physical science, also because it was gradually charged by many meanings belonging not only to thermodynamics, but also to a variety of other fields and disciplines. A further definition of the entropy concept came from L. Boltzmann (1844–1906) in the second half of the 19th century, by means of the so-called H-theorem,

in which the statistical meaning of entropy was definitely assessed.

2. ENTROPY AS A THERMODYNAMIC STATE FUNCTION

A consequence of the first and second laws of thermodynamics is that the efficiency, η_R , of any reversible machine only depends on the two temperatures, T_1 and T_2 , between which the machine operates. In short, after adequately choosing the thermometric scale in order to have $T = t + 273.14$ (t is Celsius temperature), it is

$$Q_2/Q_1 = T_2/T_1 \quad (1)$$

and

$$\eta_R = (T_1 - T_2)/T_1 \quad (2)$$

independently on the fluid or the amount of heat exchanged. The temperature T is called absolute temperature and measured in absolute degrees (or Kelvin degrees, K). Since it does not depend on the particular thermometric substance used, it therefore coincides with the one obtained by using an ideal fluid as the thermometric substance.

Equation (1) can be rewritten as

$$Q_1/T_1 - Q_2/T_2 = 0 \quad (3a)$$

and generalized to the case of n sources in a reversible cycle, as

$$\sum Q_i/T_i = 0, \quad (3b)$$

where Q_i has a positive sign if heat is provided to the fluid and a negative sign if heat is released. Eq. (3b) can be further generalized to the case of infinitesimal amounts of heat δQ exchanged by each source in any reversible process with continuous temperature changes:

$$\int_{\text{cycle}} \delta Q/T = 0. \quad (4a)$$

For any *irreversible* process S , it would be

$$\int_{\text{cycle}} \delta Q/T < 0 \quad (4b)$$

because $\eta_s < \eta_R$

Let us now consider the most general case of a reversible cycle as the one shown in Fig. 1, with continuous temperature changes. It can always be simplified as the succession of two reversible transformations from A to B (along the pathway 1) and vice versa (along the pathway 2).

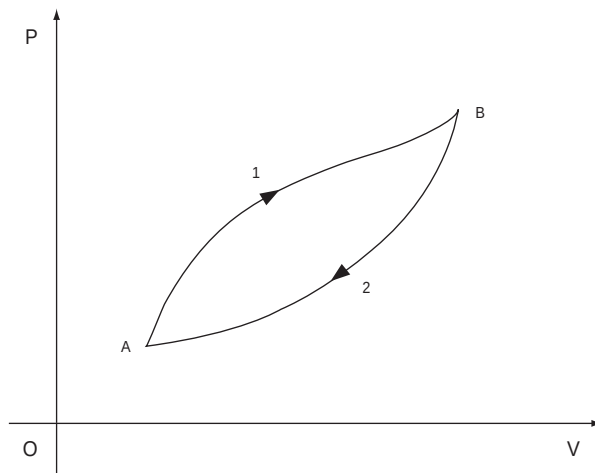


FIGURE 1 Diagram of a reversible cycle, with continuous temperature changes, simplified as the succession of two reversible transformations from A to B (along the pathway 1) and vice versa (along the pathway 2).

According to Eq. (4a), we may write

$$\int_A^B \delta Q/T|_1 + \int_B^A \delta Q/T|_2 = 0, \quad (5)$$

where the subscripts indicate the paths from A to B and vice versa. Due to the reversibility of the two transformations $A \rightarrow B$ and $B \rightarrow A$, we may also write

$$\int_A^B \delta Q/T|_1 = - \int_B^A \delta Q/T|_2 = \int_A^B \delta Q/T|_2. \quad (6)$$

This result is of paramount importance, because it demonstrates that the integral does not depend on the specific reversible transformation from A to B, but only on the actual states A and B. We are therefore able to define a new function only depending on the system state (such as the internal energy U).

Clausius (1850) called this function entropy, defined as follows:

Given an arbitrary initial system state O, an arbitrary final state A and an arbitrary reversible transformation from O to A, the value of the integral

$$S(A) = \int_O^A \delta Q/T + S(O) \quad (7)$$

calculated over any reversible transformation $O \rightarrow A$ is a function only of the states O and A. This value is called entropy.

Entropy is measured in cal/K or J/K. It clearly appears that $S(A)$ is defined but for an arbitrary constant $S(O)$, which has no practical utility, because we are always interested in entropy differences between states and not in absolute values. We can

therefore equate this constant to zero. For any reversible transformation from a state A to a further state B, the entropy variation dS will be

$$dS = S(B) - S(A) = \int_A^B \delta Q/T \quad (8)$$

independent on $S(O)$, the entropy of the arbitrary initial state O. Unlike dS , an exact differential independent on the path, δQ is path dependent and it is therefore indicated by means of the δ symbol. The expression at the right side of Eq. (8) is called the Clausius integral. As a consequence, in order to calculate the variation of entropy between the initial state A and the final state B of a system undergoing a real and irreversible process we must first consider the corresponding reversible transformation of the system between the same two states and calculate dS over this pathway. If we then consider a globally irreversible cycle (Fig. 2), made up with an irreversible transformation from A to B and a reversible transformation from B to A again, it must be according to Eq. (4b) and Eq. (8):

$$\int_A^B \delta Q/T|_{\text{irr}} + \int_B^A \delta Q/T|_{\text{rev}} < 0 \quad (9a)$$

$$\int_A^B \delta Q/T|_{\text{irr}} + S(A) - S(B) < 0 \quad (9b)$$

$$S(B) - S(A) > \int_A^B \delta Q/T|_{\text{irr}}. \quad (9c)$$

This famous inequality, due to Clausius and named after him, states that the variation of

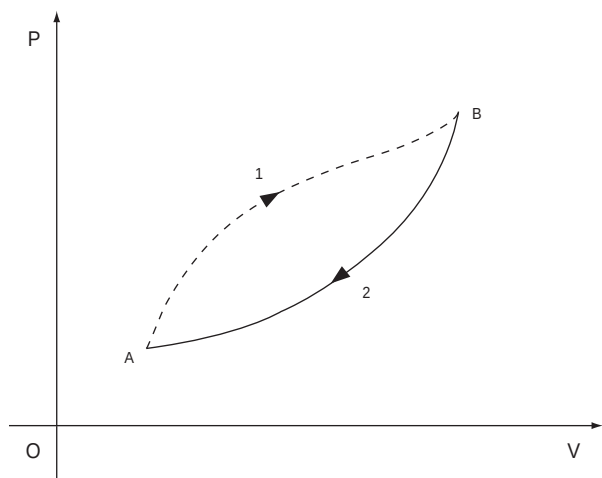


FIGURE 2 Diagram of a globally irreversible cycle, with continuous temperature changes, simplified as the succession of one irreversible transformation from A to B (along the pathway 1, dotted line) and one reversible transformation (along the pathway 2).

entropy between two states for an irreversible transformation is always higher than Clausius integral. From Eqs. (8) and (9c) we have that in a generic transformation,

$$S(B) - S(A) \geq \int_A^B \delta Q/T, \quad (10)$$

where the equality applies only to reversible processes. If a system is isolated (i.e., it has no exchange of energy [or matter] with the surrounding environment), we have

$$S(B) - S(A) \geq 0 \quad (11a)$$

$$S(B) \geq S(A) \quad (11b)$$

that is, “If an isolated system undergoes a transformation from an initial state A to a final state B, the entropy of the final state is never lower than the entropy of the initial state.” This statement is the so-called entropy law.

Therefore, when an isolated system reaches the state of maximum entropy consistent with the constraints imposed upon it, it cannot undergo any further transformation. The state characterized by the maximum entropy is the most stable equilibrium point of view, the increase of a system’s entropy measures the amount of energy no longer usable to support any further process evolution. When all the energy of a system becomes unusable (degraded heat with no temperature gradients to drive heat flows), no more system transformations are possible. In fact, heat cannot be further transformed into other forms of energy nor can it be transferred from a body to another within the system, because all bodies have the same temperature.

It is important to underline that the law of entropy states in an unequivocal way that in an isolated system all real transformations proceed irreversibly towards the state of maximum entropy. The whole universe is assumed to be an isolated system, and therefore we may expect it to reach the maximum entropy state in which its energy is completely converted into heat at the same temperature and no further transformation can be supported. This state would be the state of maximum equilibrium for the universe and has been defined as its thermal death.

Finally, unlike in classical mechanics, no time symmetry is observed in thermodynamics. While the movement of a point (i.e., a virtual object) may always occur from A to B and vice versa (time symmetry between t and $-t$ over the time axis), a real

process in an isolated system cannot reverse its evolution if this change is not accompanied by an increase of entropy.

In 1906, Nerst demonstrated that the entropy of any thermodynamic system at the absolute temperature of 0 K is equal to zero, whatever is the state of the system (Nerst's theorem). As a consequence of Nerst's statement, also known as the third law of thermodynamics, it is convenient to choose one of these states as the initial state O and to define the entropy of state A (at a temperature $T \neq 0$ K) as

$$S(A) = \int_O^A \delta Q/T, \quad (12)$$

where the arbitrary constant $S(O)$ of Eq. (7) is equal to 0, in order to be in agreement with Nerst's theorem, and where the integral must be calculated over any reversible transformation from O (at $T = 0$ K) to A (at $T \neq 0$ K).

It is worth noting that, at $T = 0$ K, it is impossible to have a mixture of finely ground substances. In fact, if so, it would be possible to diminish the entropy of the system by separating them out of the mixture. This yields an *equivalent* statement of the third law: The entropy of a pure substance is equal to zero at $T = 0$ K. As a consequence, the substances that are in the mixture will spontaneously separate from each other when their temperature drops in the proximity of $T = 0$ K.

3. STATISTICAL MEANING OF ENTROPY AND BOLTZMANN'S H-THEOREM

Dice players know very well that if we cast two dice it is much easier to get a seven than a twelve. This is because a seven can be obtained as $1 + 6$, $2 + 5$, $3 + 4$, $4 + 3$, $5 + 2$, and finally $6 + 1$, where the first number is the value obtained from the first die, and the second number is the value of the second die. There are therefore six ways to obtain a seven, and only one ($6 + 6$) to obtain a twelve. We may state that the macroscopic configuration seven has six microscopic states accomplishing it, and is therefore more probable than the configuration twelve.

A similar result can be achieved if we consider a box divided into two compartments (Fig. 3A). There are several ways of putting four balls in it. We may put four balls in its left compartment leaving the right compartment empty. Or we may do the

opposite. The two states are identical from the microscopic and statistical points of view. In fact, if we try to extract one ball from the box, in the dark, we have a 50% probability of putting our hand in the empty part and a 50% probability of putting our hand in the compartment with the four balls. The probability is the same for both compartments, which are therefore perfectly equivalent from a statistical point of view. Therefore, the macroscopic state "four balls in a compartment and no balls in another" can be achieved by means of two microscopic configurations perfectly equivalent.

Things are different if we want to distribute the four balls in both compartments, two in each one of them. We have now six possible ways for distributing our balls, each of them equivalent from a statistical point of view (Fig. 3B). The macroscopic state having two balls per compartment has more ways to occur (more microstates), which means that it is more probable than in Fig. 3A.

It is easy to understand now that a macroscopic configuration having three balls in a compartment and only one in the other is more probable than the two previously considered configurations (Fig. 3C) since it can be achieved by means of eight different microstates, all perfectly equivalent.

Let us now consider a macroscopic state of a perfect gas, defined by given values of the quantities P (pressure), V (volume), and T (temperature). It may result from a given number N of microstates—that is, a given number N of different positions and velocities of the billion molecules which compose it. In each microstate there will be different values of the position and the velocity of each molecule, but the global result of the behavior of all the molecules will be exactly that macrostate characterized by the measured values of P , V , and T . Of course, the larger the number of possible equivalent microstates N , the higher the probability of the corresponding macrostate. We know from the previous section that an isolated perfect gas tends to and will sooner or later achieve the state characterized by the maximum entropy consistent with the energy U of the gas. This state is also the macrostate characterized by the largest number of equivalent microstates. Ludwig Boltzmann (1844–1906) understood the strong link between the thermodynamic and the statistical interpretation of the entropy concept. In 1872 he published a milestone paper in which he introduced the basic equations and concepts of statistical thermodynamics.

As the first step, Boltzmann assumed that all the n molecules constituting the gas had the same velocity and divided the volume V into a very large number of

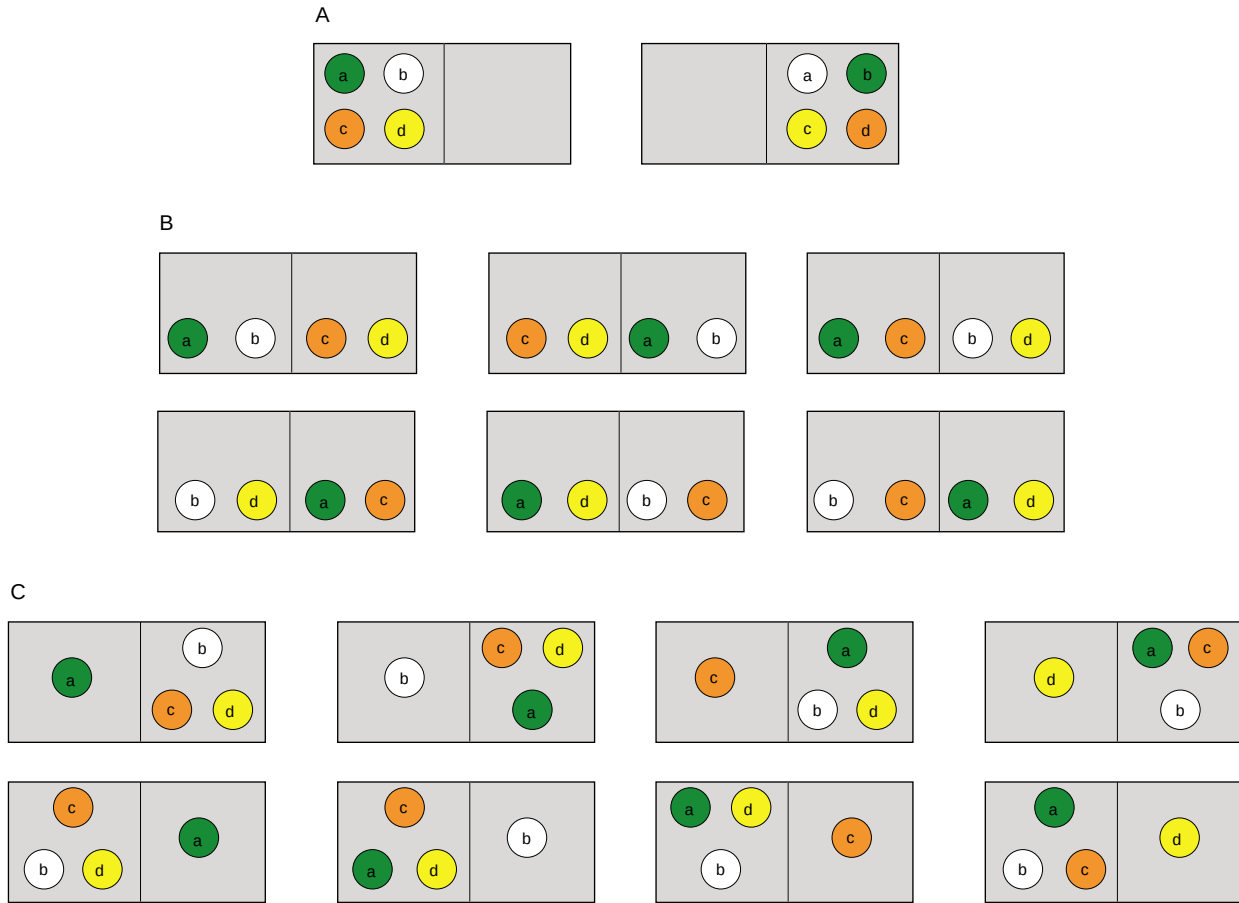


FIGURE 3 Macrostates and microstates in a system made up of four balls in two separated compartments. Each macrostate may correspond to different numbers of equivalent microstates. Figures A, B, and C show the different number of equivalent microstates that may correspond to a given macrostate, characterized by a measurable temperature, pressure, and volume.

cells, M , each having the smallest experimentally measurable size. It is then easy to demonstrate that the number of distinct microstates N corresponding to the macrostate of volume V is

$$N = \frac{n!}{n_1! \dots n_M!}, \quad (13)$$

where n_i is the number of molecules contained in the i -th cell and $n = \sum_i n_i$. Boltzmann also assumed that the entropy of the state is linked to the number N of microstates by the logarithmic expression

$$S = k \log N, \quad (14)$$

where $k = 1.38 \times 10^{-23}$ joule/K, the so-called Boltzmann's constant. The logarithm base used is arbitrary, since the conversion from a base to another one would simply require the introduction of a multiplicative constant. For a reason, which will clearly appear in the next section, it is preferable to use the logarithms with base two.

According to Nerst's theorem and to Eq. (14), it must be $N=1$ when $T=0$ K, in order to have $S=0$ —that is, at the absolute temperature $T=0$ K, a system's state is always composed by only one microstate.

By applying to Eqs. (13) and (14) the properties of natural logarithms and the Stirling formula for the mathematical expansion of $n!$, Boltzmann derived the expressions

$$S = k \log N = k \log \left(\frac{n!}{n_1! \dots n_M!} \right) = k H \quad (15a)$$

$$H = - \sum_i p_i \log p_i \quad \text{with} \quad \sum_i p_i = 1, \quad (15b)$$

where p_i is the probability to find one molecule in the i -th cell. It is not difficult to show that the maximum entropy state is that state for which all the p_i 's have the same value (i.e., that state in which all the cells are occupied by the same number of molecules). It can also be said that this is the most disordered

state, since the distribution of the molecules among the cells is the most uniform and therefore the one showing the lower level of organization.

As the second step, Boltzmann dealt with the most general case in which the positions and velocities of the molecules are all different. He introduced a distribution function, $f(x,y,z; v_x, v_y, v_z)$, to account for the different positions and velocities and expressed the p_i 's as functions of $f(x,y,z; v_x, v_y, v_z)$. He also transformed the summation $\sum$ of Eq. (15b) into an integral, due to the very high number of molecules:

$$H = - \int f \log f \, dx \, dy \, dz \, dv_x \, dv_y \, dv_z. \quad (16)$$

Boltzmann finally demonstrated in 1872 his well-known H-theorem—that is, that it must be

$$dH/dt \geq 0. \quad (17)$$

According to the Eq. (17), H (and therefore $S = kH$ as well) never decreases over time. Actually Boltzmann defined H without the minus sign in Eq. (16) and then demonstrated that $dH/dt \leq 0$. As a consequence, Eq. (15a) became $S = -kH$. We prefer, in agreement with other authors, the introduction of the minus sign in order to have H increasing over time according to Eq. (17). In so doing, H and S show exactly the same trend.

Boltzmann therefore restated by means of statistical considerations the law of entropy in isolated systems—that is, he gave a statistical reading of the second law of thermodynamics and the entropy concept. He also demonstrated that his velocity distribution function f coincided with the one hypothesized by Maxwell in 1859, from which the most probable state corresponding to the maximum entropy could be derived.

Boltzmann's seminal work generated a heated debate about mechanical reversibility and thermodynamic irreversibility. Several critics questioned that thermodynamic irreversibility could be described and explained in terms of mechanical reversibility. It is not the goal of this article to go into the details of this fascinating debate. The latter soon became a philosophical investigation about the mechanistic conception of reality as well as the so-called arrow of time (i.e., the spontaneous direction of natural processes as suggested by their irreversibility).

The most difficult part of this concept is trying to figure out two enormous numbers—that is, the number of molecules constituting the system (the ideal gas or any other) and the number of microstates contributing to a given macrostate of the system. We would like to underline that Boltzmann only stated

the statistical irreversibility of the evolution of an isolated system. After a sufficiently (infinitely) long time, the system will certainly achieve the most stable equilibrium state, characterized by the maximum entropy, although among the billion microstates contributing to a given macrostate in any instant, there is a percentually small (but different from zero) number of microstates capable of driving the system to a macrostate at lower entropy. This means that future lower entropy states are not impossible for the system, but only extremely unlikely.

4. ENTROPY AND INFORMATION THEORY

Information theory is relatively recent. It moved its first steps in 1948, when C. E. Shannon published his seminal paper, "A Mathematical Theory of Communication." This paper was the starting point of modern communication technologies, which have so deeply modified all aspects of our life. Shannon's problem was how to measure and transmit a message from a source (transmitter) to a recipient (receiver) through a transmission channel generally characterized by a background noise capable of altering the message itself. We will focus in this paper on the problem of information measure.

First of all, it is important to underline the fact that the concept of message is very broad. However, the set of messages that each source is able to transmit must be specified without any possible ambiguity. A letter of the alphabet, the ring of a bell, a word or a particular gesture are all examples of possible messages. Shannon considered the n elementary messages that a source could transmit (i.e., the basic alphabet of the source itself). He first assumed that these messages were equally probable (i.e., that the probability for each of the n options to be transmitted was $p = 1/n$). If two consecutive elementary messages are transmitted, the possible options (i.e., the number of possible different sequences of two messages out of n available ones) are n^2 . In general, if m messages are transmitted in a row, the possible different sequences become $N = n^m$ (i.e., the number of all the possible permutations of the n elementary messages in groups of m elements, also allowing for repetitions). At the same time, the information transmitted (i.e., the events or descriptions that characterize the source and that are not yet known by the recipient) should be twice, three times, ..., m times the one transmitted by means of only one elementary message. In any case, this

information should be somehow proportional to the number m of messages that make up the transmitted sequence of messages. Shannon assumed it was proportional to the logarithm of the possible options N , which easily yielded

$$\begin{aligned} I &= \log N = \log n^m = m \log n \\ &= -m \log 1/n = -m \log p. \end{aligned} \quad (18a)$$

The information conveyed by each elementary message is therefore

$$i = -\log p = \log 1/p. \quad (18b)$$

From Eq. (18b) it appears that the information conveyed by a message is proportional to the log of the reciprocal of its probability. If $p < 1$, it is $i > 0$. This also means that the less probable the message, the higher the information content transferred from the source to the recipient. If we already know that a given message will be transmitted ($p = 1$), it will not carry any additional information. In fact, we would not switch the TV on if we already knew which news are going to be announced. Instead, an unlikely and therefore unexpected message carries a significant piece of information. Finally, no information is carried by a message that cannot be transmitted ($p = 0$).

Shannon then generalized his equation to the case of messages showing different transmission probability. He assumed that the source could transmit n elementary messages in groups of m elements, with probabilities p_i , where $i = 1, \dots, n$. If m is large enough, the i -th message with probability p_i will appear $m p_i$ times and the total information carried by the sequence of m messages will be

$$I = - \sum_i m p_i \log p_i. \quad (19a)$$

As a consequence, the average information conveyed by each of the m messages will be

$$H = - \sum_i p_i \log p_i \quad \text{with} \quad \sum_i p_i = 1 \quad (19b)$$

Equation (19b) has the same mathematical form as Eq. (15b), thus indicating a surprising similarity between Boltzmann's entropy and Shannon's information. For this reason, the same symbol H is used and Shannon's expression is very often called entropy of a source. Actually, the entropy of the source should be indicated as in Eq. (15a), where k is Boltzmann's constant and H would be Shannon's expression for the average information of a message. The entropy H of a source, intended as the average information content of a given message, also indicates

the recipient's ignorance about the source, or in other words his a priori uncertainty about which message will be transmitted. This fact should not be surprising, since it is exactly by receiving the message that the uncertainty is removed.

In 1971, M. Tribus and E. C. McIrvine provided an interesting anecdote about Shannon, who is reported to have said:

My greatest concern was what to call it. I thought of calling it "information," but the word was overly used, so I decided to call it "uncertainty." When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

True or not, the story points out the links between the two concepts. Equation (19b) also allows the definition of a unit to measure information. Let us assume that the source releases information by means of a binary alphabet—that is, it is able to transmit only two elementary messages, both having a probability $p = \frac{1}{2}$. It is common habit to indicate these messages by means of the two digits, 0 and 1, which therefore become the basis of the so-called digital alphabet and transmissions. The average information carried by an elementary message is, in such a case, from Eq. (19b):

$$\begin{aligned} H_1 &= -1/2 \log 1/2 - 1/2 \log 1/2 \\ &= -1/2 \log_2 1/2 - 1/2 \log_2 1/2 \\ &= -1/2 \times (0 - 1) - 1/2 \times (0 - 1) \\ &= 1/2 + 1/2 = 1. \end{aligned} \quad (20)$$

This unit quantity is called a bit (acronym for binary digit), which indicates a pure number, with no physical dimensions. One bit of information is, by definition, the information carried by each of the two elementary messages released by a binary source with equal probability 1/2. The reason of the choice of logarithms in base two is now evident.

For the sake of completeness, we may also add that another very common unit is the byte (1 byte = 8 bit). Eight bits allow for the construction of $2^8 = 256$ different sets of elementary messages, or different 8-bit strings, to each of which we may for instance assign a given symbol on the keyboard of a personal computer.

4.1 Maxwell's Demon

The link between S and H was deeply investigated soon after the introduction of Shannon's theory. Each isolated system can be considered either a source of information or vice versa. For example, an isolated gas with totally thermal energy and uniform temperature is in a high-entropy, disordered state (from a thermodynamic point of view). It is also in a macrostate to which a high number of microstates may correspond and therefore it is a highly probable state (from a statistical point of view). If so, the uncertainty about the actually occurring microstate is huge and a large number of bits (much information) would be needed to ascertain it (from the point of view of Information theory).

A clear example of the link between entropy and information is the so-called Maxwell's demon, a problem that fascinated scientists for a long time. Assume that we have two rooms at the same temperature, due to the fact that in each room there is the same distribution of molecular velocities and kinetic energies. If we could separate the slowly moving molecules from the fast ones, the temperatures of the two rooms should vary. Maxwell assumed that this task could be performed by a demon quickly opening or closing a frictionless window in the separation wall, in order to pass fast (hot) molecules to a room and slow (cold) molecules to the other one (Fig. 4). The final effect would be the separation of hot from cold molecules and the generation of a temperature gradient (one room would be at higher temperature than the other one), without spending energy from outside, due to the absence of friction in the process. In so doing, the entropy of the system would decrease and the second law would be violated. This paradox was only solved in 1956 by L. Brillouin. He pointed out that the demon needed information about each approaching molecule, in order to decide about opening or not the window. He could, for instance, light them up. The process of generating information would cause an increase of the overall entropy of the system + environment, larger or equal to the entropy decrease of the system only.

From Eq. (15a) and the result of Eq. (20), we have

$$\begin{aligned} S &= k H = 1.38 \times 10^{-23} \text{ joule/K} \times 1 \text{ bit} \\ &= 1.38 \times 10^{-23} \text{ joule/K} \times 1 \text{ bit.} \end{aligned} \quad (21)$$

Therefore, 1 bit of information requires an entropy generation equal to 1.38×10^{-23} joule/K. This entropy cost of information is so low, due to the small value of k , that people thought information was free,

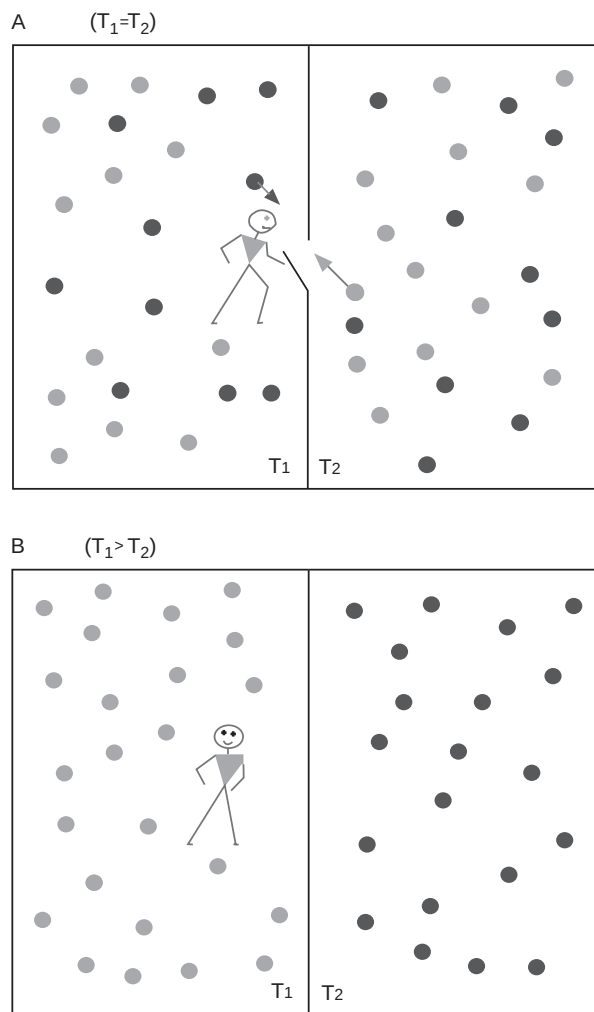


FIGURE 4 Maxwell demon. The demon seems to be able to violate the second law of thermodynamics, by selecting the molecules and sorting the hot (gray) from the cold ones (black). The two compartments are initially at the same temperature (A). After some time, all hot molecules migrated to the left compartment and cold ones to the right compartment. As a consequence of the selection, the temperature of the left side is higher than the temperature of the right side (B).

as for Maxwell's demon. Of course, if H is calculated by means of natural logarithms, then 1 bit corresponds to $k \ln 2$ entropy units.

5. THE FREE-ENERGY CONCEPT

What happens when entropy increases? What does it mean in practical terms? Lots of pages have been written, linking entropy increase with disorder, disorder with less organization, less organization with lower quality, lower quality with lower usefulness.

Sometimes links have been a little arbitrary, because things are different with isolated or open systems. How can we quantify these concepts?

We know from Section 2 that if an isolated system undergoes a transformation from a state A to a state B, its entropy never decreases. The entropy law states in an unequivocal way that in an isolated system all real transformations tend irreversibly towards a state of maximum entropy. This statement applies to any kind of energy conversion, not only to heat-to-work ones. An entropy increase means that a fraction of the energy provided to the system is transformed into degraded heat, no longer usable to support any further process evolution. The energy actually usable is less and less work can be done. Many feel confused by using increasing entropy to mean decreasing ability to do work, because we often look to increasing quantities with positive feelings in our mind. For the entropy concept it is the opposite. The chemist J. W. Gibbs (1839–1903) actually “solved” this problem by introducing a new quantity, which he called available energy, linked to the entropy concept by means of a fundamental equation named after him. Available or free energy was then recognized as being an important property not only in chemistry but also in every energy transformation, in which the dissipation of a gradient (temperature, pressure, altitude, concentration, etc.) yields a product or delivers work as well as releases some degraded energy or matter.

Several slightly different expressions exist for available energy, so that we will only refer to one of them, the so-called Gibbs energy, or Gibbs free energy, defined for reversible system transformations at constant pressure. This is the most frequent case, because the majority of processes on Earth occur at constant pressure.

5.1 Gibbs Energy

In general, the available energy is that fraction of a system's energy that is actually usable to drive a transformation process. It can be calculated by subtracting from the total system's energy the fraction that becomes unavailable due to the entropy increase. Assuming a process in which heat is transferred at constant pressure p , let us first define the Gibbs-energy function, G , as

$$G = H - TS, \quad (22a)$$

where

$$H = U + pV \quad (22b)$$

is the system's enthalpy, a measure of the total heat content of a system, while

$$U = Q + W \quad (22c)$$

is the well-known expression of the first law of thermodynamics. For a measurable change of the system's state at constant temperature T , from Eq. (22a) we have

$$\begin{aligned} G_f - G_i &= \Delta G = \Delta H - T\Delta S \\ &= (H_f - H_i) - T(S_f - S_i), \end{aligned} \quad (23)$$

which quantifies the amount of the available energy ΔG , after enthalpy and entropy changes have been accounted for. The subscripts i and f indicate the initial and final states.

This result shows the importance of the entropy concept also in practical applications. It indicates that the energy of a system is not always totally available to be transformed into work. Part of this energy is not converted but instead is released as high entropy waste (degraded heat), no longer able to support change of state of the system. The terms available and free can be now understood, as opposed to the bound, unavailable fraction of energy present in the entropy term of Eq. (23).

A process with $\Delta G < 0$ (i.e., a process whose initial state has higher available energy than the final state) may evolve spontaneously towards the latter, and deliver work. This result has a general validity and clearly shows the importance of the entropy-related, free energy concept.

Heat conversion into work and chemical reactions are clear examples of how the second law and the entropy concept work. In both cases we have spontaneous processes: (1) the heat flows from a source at higher temperature to a sink at lower temperature, and (2) the reaction proceeds from the higher to the lower free energy state. In both cases we have a process driven from a gradient of some physical quantity (respectively, a difference between two temperatures and a difference between two free energy states) and some work delivered to the outside at the expenses of the free energy of the system. In general terms, whenever a gradient of some physical quantity (altitude, pressure, temperature, chemical potential, etc.) exists, it can drive a transformation process. The free energy of the system is used whenever these spontaneous processes occur. The result is a lowered gradient, and an increased total entropy of the process + environment system, in addition to the work done.

6. CONCLUSION

The entropy concept is fascinating and rich of significance. At the same time, it appeared scaring and confusing to many. Many focused on the gloomy picture of increasing disorder and thermal death, characteristic of equilibrium thermodynamics and isolated systems. In this regard, entropy is used to point out and measure the decreasing availability of high quality resources (i.e., resources with low entropy and high free energy), the increase of pollution due to the release of waste, chemicals, and heat into the environment, the increase of social disorder due to degraded conditions of life in megacities all around the world, the “collapse” of the economy, and so on.

Others identify entropy as the style of nature: since the biosphere is an open system supported by a constant inflow of solar energy, its structures and life phenomena undergo a continuous process of self-organization. Geologic processes, atmospheric systems, ecosystems, and societies are interconnected through a series of infinitely different and changing relationships, each receiving energy and materials from the other, returning same, and acting through feedback mechanisms to self-organize the whole in a grand interplay of space, time, energy and information. During this self-organization process, entropy is produced and then released towards the outer space. Living structures do not violate the second law, but feed on input sources, which continuously supply low entropy material and energy (free energy) for system development.

While the first point of view calls for increased attention to avoid misuse of resources and prevent degradation of both natural and human environments (i.e., prevent the loss of stored information), the second point of view focuses on the creation of organization and new structures out of disordered materials, thanks to a flow of resources from outside. This point of view calls for adaptation to the style of nature, by recognizing the existence of oscillations (growth and descent) and resource constraints, within which many options and new patterns are however possible. Both points of view are interesting and stimulating and account for the richness of the entropy concept. The latter, first introduced in the field of energy conversion, acquired very soon citizenship in several other fields. We already mentioned the information theory in Section 4, but many other examples can be easily provided.

Environmentally concerned economists focused on the entropy tax—that is, on the fact that the development of economic systems is always dependent on the degradation of natural resources, the amount of which is limited. Some pointed out that the notion of entropy helps “to establish relationships between the economic system and the environmental system,” as noted by Faber *et al.* in 1983. In 1971, Nicholas Georgescu-Roegen underlined that “The Entropy Law...emerges as the most economic in nature of all natural laws.”

The ecologist Ramon Margalef introduced in 1968 a quantity, called ecological diversity, D , as a measure of the organization level of an ecosystem. This quantity is defined exactly in the same way as entropy, namely,

$$D = - \sum_i p_i \log p_i \quad (24)$$

where p_i is the probability for an organism to belong to a given species. The higher the ecological diversity, the lower the production of entropy per unit of biomass, because resources are better utilized and support the growth of the whole spectrum of ecosystem hierarchy. Instead, simplified ecosystems with fewer species (such as agricultural monocultures) do not use available resources and coproducts as effectively, resulting into a higher production of entropy per unit of time and biomass.

Finally, entropic styles have been adopted by several artists to express in their works their desire of novelty and their rejection of previous, static views on life. Some understood and meant entropy alternatively as disorder, indefiniteness, motion, rebellion—the opposite of the ordered and well-defined, Newtonian structures of Renaissance artists. In 1974, Rudolf Arnheim published an interesting essay on disorder and order in arts as well as on the need for complexity, which moves artists toward investigating new forms to express their creativity.

In the end, we would do a disservice to the entropy concept, by limiting its significance to the concept of disorder, which is, however, one of its meanings, or even to the practical application of the concept of free energy. Entropy, a word now very common in the everyday language, may be suggested as the other side on the coin of life itself, with its development, its richness and complexity. Life processes (such as photosynthesis) build organization, add structure, reassemble materials, upgrade energy, and create new information by also degrading input resources. Degraded resources

become again available for new life cycles, ultimately driven by the entropic degradation of solar energy.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • Energy in the History and Philosophy of Science • Entropy and the Economic Process • Exergy • Exergy Analysis of Energy Systems • Exergy Analysis of Waste Emissions • Exergy: Reference States and Balance Conditions • Thermodynamics and Economics, Overview • Thermodynamic Sciences, History of • Thermodynamics, Laws of

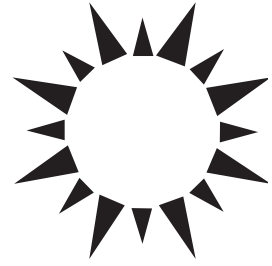
Further Reading

- Arnheim R. (1974). "Entropy and Art. An Essay on Disorder and Order." University of California Press, Berkeley, California.
- Boltzmann, L. (1872). Further studies on the Thermal Equilibrium of Gas Molecules (from *Sitzungsberichte der kaiserlichen Akademie der Wissenschaften*), Vienna.
- Brillouin, L. (1962). "Science and Information Theory." Academic Press, New York.
- Clausius, R. G. (1850). Ueber die bewegende Kraft der Wärme. *Annalen der Physik und Chemie* 79, 368–397, 500–524 [translated and excerpted in William Francis Magie, *A Source Book in Physics*, McGraw-Hill, New York, 1935].
- Faber, M., Niemes, H., and Stephan, G. (1983). "Entropy, Environment, and Resources. An Essay in Physico-Economics." Springer-Verlag, Berlin.
- Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard University Press, Cambridge, MA.
- Gibbs, J. W. (1931). "The Collected Works of J. Willard Gibbs." Longmans, Green and Co, London.
- Margalef, R. (1968). "Perspectives in Ecological Theory." The University of Chicago Press, Chicago.
- Maxwell, J. C. (1871). "Theory of Heat." Longmans, Green, & Co., London.
- Nernst, W. (1906). Ueber die Berechnung chemischer Gleichgewichte aus thermischen Messungen. *Nachr. Kgl. Ges. Wiss. Gött.* 1, 1–40.
- Prigogine, I. (1978). Time, structure, and fluctuations. *Science* 201, 777–785. Text of the Lecture delivered in Stockholm on 8 December 1977, when Prigogine was awarded the Nobel Prize in chemistry.
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J.* 27, 379–423.
- Tribus, M., and McIrvine, E. C. (1971). Energy and information. *Scientific American* 225(3), 179–184.

Entropy and the Economic Process

GABRIEL A. LOZADA

University of Utah
Salt Lake City, Utah, United States



-
1. Entropy
 2. The Laws of Thermodynamics
 3. Economic Implications of Entropy to Production Theory
 4. Entropy and Neoclassical Economic Analysis
 5. Interpretations and Misinterpretations of Entropy
 6. Conclusion

Glossary

entropy change The amount of heat absorbed by a material divided by the material's absolute temperature.

isolated system A collection of materials that exchanges neither matter nor energy with its surroundings.

macroeconomics The study of the economy as an aggregated whole, generally ignoring the behavior of individual industries or consumers and instead concentrating on economy-wide magnitudes such as unemployment, inflation, interest rates, the trade deficit, and growth of gross national product.

microeconomics The study of the behavior of one or a few firms, industries, or groups of consumers in an economy.

natural resources Exhaustible resources such as coal, renewable resources such as fish, and ecosystem capacities useful to humans such as the ability to cleanse dirty water.

neoclassical economics The school of economic thought which postulates that consumers act in a way that maximizes their well-being subject to their income and possibly other constraints, that firms act in a way that maximizes their profits subject to the technological constraints of production, and in general that economic agents act in a way that maximizes an objective function that can in principle be rigorously and completely mathematically described, subject to constraints that can also in principle be so described.

second law of thermodynamics Also known as the entropy law, it states that in an isolated system, entropy cannot decrease; it can either rise or remain constant.

neoclassical theory of value The theory that market prices are determined by the intersection of demand curves

and supply curves, which in turn arise from the behavior of self-interested atomistic economic agents; these "curves" may be more complicated than simple functions over Euclidian space.

To explain why a capitalist economy produces certain commodities and how the prices of those commodities are determined, a description must be given of consumers' desire and ability to buy the commodity and of how firms produce that commodity. The description of how firms produce a commodity is in many ways a description of the engineering opportunities open to the firm, although describing the allocation and supervision of labor is also important, as is the firm's access to diverse forms of financing. In describing the engineering opportunities open to the firm, economists usually have to work with a simplified version of all the potentially available technologies in order for the problem to be tractable. However, it is important for the simplifications to obey fundamental physical laws, or at least not flagrantly violate them, so economists studying production should have a basic knowledge of those laws. The writers who first emphasized the relevance of the laws of thermodynamics to production economics were Frederick Soddy (the 1921 Nobel laureate in chemistry), Nicholas Georgescu-Roegen, and Herman E. Daly. The second law of thermodynamics, called the entropy law, has been the focus of most of the work in this field.

1. ENTROPY

The first law of thermodynamics states that in an isolated system, energy is neither created nor destroyed. This law cannot explain why an ice cube placed on a sidewalk on a hot summer day always melts, because the first law would permit heat to flow from the ice cube as long as an equal amount of heat

flows to the air. Entropy is the concept used to explain why ice cubes on a hot sidewalk melt and in general why heat always flows from hotter objects to colder ones, as well as why spontaneous physical processes are spontaneous.

The definition of entropy is straightforward. Let T stand for absolute temperature, that is, temperature in degrees Kelvin, denoted K. (Kelvin temperature is 273.15 degrees higher than Celsius temperature.) Let Q stand for the flow of heat into a material (or system, which is a collection of materials). The metric unit of heat Q , as well as that work, is the joule, denoted 'J'; the English unit is the calorie or the BTU. Then the change in the entropy S of the material (or system) is defined to be

$$dS = dQ_{\text{rev}}/T, \quad (1)$$

where the subscript rev stands for a reversible process; physicists call a process reversible if it involves no dissipative effects such as friction, viscosity, inelasticity, electrical resistance, or magnetic hysteresis. If a system undergoes an irreversible process between an initial equilibrium state i and a final equilibrium state f , the entropy change of the system is equal to the integral from i to f of dQ/T , taken over any reversible path from i to f . No integral is taken over the original irreversible path.

Suppose one has an isolated system containing two bodies, one hot and one cold, placed in thermal contact. When a given quantity of heat, Q_0 , flows from the hotter body to the colder one, the change in entropy of the system is

$$\Delta S = \frac{-Q_0}{T_{\text{hot}}} + \frac{+Q_0}{T_{\text{cold}}}.$$

Since $T_{\text{hot}} > T_{\text{cold}}$, $\Delta S > 0$. If heat were to flow in the opposite direction, away from the colder body and toward the warmer body, then the Q_0 terms in the previous equation would change sign and ΔS would be negative. Because it is the former and not the latter which is always observed, 19th-century physicists postulated that in an isolated system, entropy never decreases. This is one form of the second law of thermodynamics.

2. THE LAWS OF THERMODYNAMICS

The laws of thermodynamics are as follows:

- *Zeroth law.* Two systems that are in thermal equilibrium with a third are in thermal equilibrium with each other.

- *First law.* During a process in which no heat is exchanged with the environment, the work done is only a function of the initial and final states of the system, not of the path. Furthermore, during any process, heat flow, Q , is equal to $U_f - U_i + W$ where U_f and U_i are the final and initial internal energies of the system and W is the net work done by the system.
- *Second law, Kelvin-Planck statement.* No process is possible whose sole result is the absorption of heat from a reservoir and the conversion of this heat into work.
- *Second law, Clausius statement.* No process is possible whose sole result is the transfer of heat from a cooler to a hotter body.
- *Second law, entropy statement.* In an isolated system, entropy is nondecreasing.
- *Third law, unattainability statement.* It is impossible to reach absolute zero by any finite number of processes.
- *Third law, Nerst-Simon statement.* In the limit as temperature goes to zero degrees Kelvin, the entropy change of any reaction is zero.

Proofs that the three statements of the second law are equivalent can be found in standard texts, as can proof that the two statements of the third law are equivalent.

3. ECONOMIC IMPLICATIONS OF ENTROPY TO PRODUCTION THEORY

3.1 Heat Engines

The first application of entropy was to "heat engines"; these are devices, such as internal combustion engines and power-plant steam engines (in both conventional and nuclear power plants), that use temperature differences to do their work. (A car's motor, for example, will not work in a room hotter than the spontaneous ignition temperature of gasoline.) The efficiency of a heat engine is defined to be its work output divided by its heat input. A surprising and important implication of the second law of thermodynamics is that the maximum efficiency of any heat engine operating between the temperatures T_{low} and T_{high} is $1 - (T_{\text{low}}/T_{\text{high}})$. This is less than 1 because T_{low} is strictly greater than zero due to the unattainability statement of the third law. A felicitous explanation of the maximum-efficiency

equation appears in Sir Oliver Lodge's *Energy*. Lodge also explains what it has to do with entropy:

If the quantity of heat energy Q is lost by a body at temperature T , and gained by a body at temperature T' , then the entropy lost is Q/T , and the [entropy] gained is Q/T' . The gain is greater than the loss, if T' is less than T . But if some of the heat is utilised,—converted into mechanical or other forms of energy by means of some kind of engine,—so that only Q' is imparted to the colder body (the work done corresponding to the difference $Q-Q'$) then it may be that $Q/T=Q'/T'$; and in that case the entropy of the system remains constant. This is the condition for a perfect engine conducting a reversible operation. Irreversible operations increase entropy or dissipate energy, for the lost availability cannot be recovered. The efficiency of an engine means the ratio between the portion of heat utilised and the heat taken from the source. It is $(Q-Q')/Q$; and we see now that in a perfect engine this would be $(T-T')/T$, because the Q 's and the T 's are then proportional. (pp. 39–40)

Lodge's "perfect" heat engine is called a Carnot engine; on a temperature-versus-entropy diagram, the cycle of this engine traces out a rectangle, called a Carnot cycle.

This second law limit on the maximum efficiency of heat engines is striking because it establishes a boundary beyond which technological progress will never cross. This has the salutary effect of throwing scientific doubt on the predictions of technological Panglossians who believe that technical progress is unbounded and can eventually remedy any problem facing industrial civilization. On the other hand, not all engines are heat engines (fuel cells, for example, are not), and the second law maximum-efficiency limitation only applies to heat engines.

All engines—and indeed all uses of energy—are, however, subject to another limitation due to the entropy law. Lord Kelvin's principle of the degradation of energy states that whenever an irreversible process takes place, the effect on the universe is the same as that which would be produced if a certain quantity of energy were converted from a form in which it was completely available for work into a form in which it is completely unavailable for work. This amount of energy is the temperature of the coldest heat reservoir at hand times the entropy change of the universe brought about by the irreversible process. As a consequence of this principle, it is never desirable to recycle energy (because accomplishing such recycling requires using up more energy than would be regenerated).

3.2 Chemical Processes

To discuss implications of entropy to chemical rather than mechanical processes—for example, if one

wished to analyze fuel cells—it is useful to rewrite Eq. (1) as

$$\Delta S = \int_i^f \frac{C_p(T)}{T} dT, \quad (2)$$

where i denotes some convenient initial reference state, f denotes the state of interest, and C_p denotes the heat capacity at constant pressure of the chemical substance. (The heat capacity of a substance, also known as its specific heat, is the amount of heat or work needed to raise the temperature of 1g of the substance by 1 degree Kelvin [or Celsius]. A substance's heat capacity changes slightly depending on its initial temperature.) In a laboratory, Eq. (2) can be approximated to any degree of accuracy by replacing the integral with a sum, as in

$$\Delta S \approx \sum_{j=i}^f \frac{C_p(T_j)}{T_j} (1K), \quad (3)$$

where the temperature range from T_i to T_f has been broken down into j parts. Using Eq. (3), the entropy of one gram of a substance is experimentally determined in the following way: Start with the substance at a convenient initial temperature and pressure. Measure the heat needed to raise its temperature from T_1 to T_2 , then divide this figure by T_2-T_1 to obtain the specific heat (at constant pressure). Divide this specific heat by the initial temperature; this is $C_p(T_1)/T_1$. Next measure the heat needed to raise the temperature from T_2 to T_3 , and divide that by T_3-T_2 , then divide that by the temperature; this is $C_p(T_2)/T_2$. Continue until T_f is reached. The change in entropy of the substance between the initial and final states is approximately equal to the sum of the $C_p(T_j)/T_j$ terms, times 1 degree Kelvin to adjust the units.

In this way, the entropy change of different materials can be experimentally determined. The results of such experiments are given—for example, in freshmen chemistry textbooks—in tables of entropies for many substances. For instance, the standard entropy of one mole of copper (63 grams) is 33 J/K = 8 cal/K, where "cal" denotes calories, and where standard (or absolute) entropies are defined with an initial condition of 0 K and one atmosphere pressure and a final condition of 300 K and one atmosphere pressure.

Armed with such knowledge, and the second law's principle that entropy in an isolated system never decreases, the equilibrium proportions of different chemicals reacting with each other can be determined merely by calculating, on paper, which proportions maximize the entropy of the mixture. In other words,

once entropy values have been tabulated, the equilibrium direction and extent of any chemical reaction can be determined without further laboratory work. There is a caveat: while the equilibrium direction and extent of any chemical reaction can be determined, not all chemical systems go to equilibrium; some get stuck in a metastable state in which an activation energy barrier blocks their path to the equilibrium. This barrier can be broken by a catalyst, which works by supplying the activation energy, which is then returned in full to the catalyst.

3.3 General Processes

The entropy law thus governs the efficiency of heat engines, the recycling of energy, and the direction and extent of chemical reactions. More generally, it sets the ground rules within which mechanical and chemical systems operate. Even more generally, it constitutes a fundamental principle of irreversibility in the following sense: Imagine a system that is isolated. By the second law, entropy always increases in such a system (it could remain the same, but only if there were no friction, only perfect machines, and so forth). Choose two dates t_1 and t_2 with $t_1 < t_2$. Then the entropy of the system at t_1 (call it S_1) is less than the entropy of the system at t_2 (call it S_2): $S_1 < S_2$. Then it is in principle impossible to reverse all of the changes that have happened in this isolated system between t_1 and t_2 —assuming the system remains isolated—because such a reversal would imply a decrease of entropy. This means that perpetual motion machines are impossible. It means that complete recycling of all a process's components is impossible; some fresh inputs will always be needed for the process to occur again. *A fortiori*, the economy will always require natural resource input of some sort from nature. The economy will never be a perpetual motion machine.

Nicholas Georgescu-Roegen wrote that “the Economic Process has a unidirectional irrevocable evolution” (from low to high entropy), and, in fact, all processes have such an evolution. The activities of an isolated mine, factory, or household—activities that ultimately turn valuable inputs into waste products—are all physically irrevocable. As Georgescu-Roegen and Herman Daly have emphasized, this replaces the typical freshman textbook conceptualization of the economy as a circular flow with an alternative conceptualization of the economy as entailing a one-way flow of matter and energy from usefulness to uselessness.

The type of economic growth occurring over the past four centuries has greatly accelerated this flow toward the ultimate maximum-entropy equilibrium from which no change could ever be made. In speeding up this final end, modern economic growth has produced not just more waste but also wonderful dissipative structures, which, as we learn from the work of Ilya Prigogine, can be the vehicles through which entropy is transformed from low to high. Dissipative structures include machines, human beings, and amazing new technologies like R. Dean Astumian's molecular motors, which function precisely because they are not in thermodynamic equilibrium and can use the resulting chaotic turbulence to do work.

Important as the conceptualization of the economy as a one-way process always dependent on fresh inputs from nature is, two caveats are in order. The first is that the word “isolated” is an important qualifier. A farm, for example, is not thermodynamically isolated from the sun. To construct a proper isolated system for agriculture, the sun has to be included as part of the system. The second caveat then comes into play: while a system as a whole cannot be returned to its previous state, a component of the system—such as a farm's soil—can be returned to its previous state as long as some other component of the system (such as the sun) suffers an increase in entropy sufficient to offset the decrease in entropy of the component that has returned to its previous state. In principle, this means it may be possible to perfectly recycle the materials of any production process that involves or could be made to involve the sun (until the sun dies).

Nicholas Georgescu-Roegen believed such perfect matter recycling to be impossible. Since such an impossibility does not follow from any of the scientifically accepted laws of thermodynamics, he proclaimed this impossibility as a new “fourth law of thermodynamics.” Ecologists often believe that nutrient recycling in ecosystems is 100% complete and thus dispute the fourth law's validity. By contrast, economics Nobel laureate Paul Samuelson, in a forward to a book about bioeconomics, wrote that Georgescu-Roegen's “Law of Inevitable Dissipation of Useful Concentrated Matter ... is good-sense certainty ...: only if all useful copper and gold and helium could be 100 per cent recycled would a *perpetuum mobile* be possible. If pigs could fly, if my Aunt Sally were a stagecoach, if ... and if” So the truth of the fourth law remains in dispute.

As an aside, not only does $t_1 < t_2$ in a closed system imply $S_1 \leq S_2$, but $S_1 < S_2$ in a closed system also

implies $t_1 < t_2$. The second statement has been used to explain the direction of time, “time’s arrow.” For example, it has been proposed that since we live in a universe whose expansion causes the universe’s entropy S to change, the reason we perceive time to increase in the direction it does is that that causes us to conclude that the universe’s S is increasing, in accordance with the second law of thermodynamics, not decreasing, which would violate the second law.

4. ENTROPY AND NEOCLASSICAL ECONOMIC ANALYSIS

Neoclassical microeconomic theorists have no difficulty incorporating scientific constraints into their work. Indeed, as far back as Tjalling Koopmans’ and Gerard Debreu’s work of the 1950s, general equilibrium theorists have explicitly incorporated an irreversibility assumption into their work, as well as a more basic assumption that there cannot be outputs without inputs. Unfortunately, general equilibrium analysis is conducted at a level of abstraction that makes it almost useless for policy applications. (For example, applied general equilibrium analysis of trade policy assumes full employment, which begs the question of free trade’s effects on employment.)

Neoclassical analysis of long-run economic growth has been conducted, not using microeconomic general equilibrium, but instead using a macroeconomic framework in which aggregate economic output is assumed to be a function $f(L, K)$ of the economy’s stock of labor L and physical capital K , or sometimes a function $f(L, K, R)$ including natural resources R . Inspired by the work of Robert U. Ayres, one notices that elementary mass balance (conservation of matter) is lacking in $f(L, K)$ —for where does K come from? Surely not from L , and most physical capital nowadays is not alive and able to reproduce itself! — and it is also lacking in $f(L, K, R)$ in the common formulation where K can be maintained at a constant or even increasing level while R goes to zero.

Neoclassical resource economists have constructed a useful body of microeconomic knowledge about how profit-maximizing firms exploit the environment, as well as the implications of different types of government regulation of that activity. However, these results usually do not treat the overall issue of natural resources in long-run growth.

Fundamentally, no economic technique will be able to definitively foretell the abundance and

accessibility of particular natural resources far in the future, nor the need for those resources in the future. Educated guesses are possible, but novelty, contingency, and lack of predetermination mean that it is not possible to predict the future. There is no basis to confidently assume that the future will resemble past centuries, in which technological progress made it possible and desirable to exploit vast, hitherto useless stocks of many different materials found in nature.

5. INTERPRETATIONS AND MISINTERPRETATIONS OF ENTROPY

Equations (1), (2), and (3) are completely straightforward in their literal meaning. However, they give no intuitive insight into what entropy is. Other famous formulas in physics, such as $F = ma$ or even $E = mc^2$, deal with concepts we have at least some everyday familiarity with or understanding of, such as force, mass, acceleration, energy, and the speed of light. By contrast, Q/T or C_P/T are remote and hard to grasp. This has led scientists to come up with interpretations or explanations of entropy. Some of these interpretations are extremely valuable to scientific specialists in the field. Others are unfortunate misinterpretations. Neither the useful ones nor the useless ones have much importance to economists, but their continuing great influence on the historical development of the field makes it important to understand them.

5.1 Disorder, Randomness, and Probability

One of the most surprising consequences of Eq. (1) comes from analyzing the mixing of two ideal gases in a container whose walls allow no heat to flow. (An ideal gas has molecules that do not interact with their neighbors.) At first sight, the formula textbooks give for the resulting change of entropy—which is

$$\Delta S = -R \left[n_1 \ln \frac{n_1}{n_1 + n_2} + n_2 \ln \frac{n_2}{n_1 + n_2} \right], \quad (4)$$

where n_1 and n_2 are the moles of the gases and R is the universal gas constant ($8.31 \text{ J/(K} \cdot \text{mole)}$)—seems wrong, because since no heat can flow into or out of the container, it seems from Eq. (1) that the system’s entropy cannot change. This doubt about Eq. (4)’s validity is incorrect. Entropy change is not $\int_i^f dQ/T$

(which is zero in this case) but rather $\int_i^f dQ/T$ over a reversible path. The mixing of two ideal gases is not reversible. The only way to calculate the entropy change accompanying the mixing of two ideal gases is to replace the original irreversible path linking the states before and after the mixing with a reversible path linking those same states, then performing the integration. Because entropy change is a function of the initial and final states only, entropy change along the new reversible path, which can be calculated using Eq. (1), must be the same as the entropy change along the original irreversible path. Along the reversible path, heat does flow, and the entropy change turns out to be Eq. (4), which is greater than zero.

The remarkable result that mixing two gases increases entropy has led observers to think that there is more to the idea of entropy than expressed by Eq. (1)—that entropy measures increasing mixing, or spatial disorder, or even some general notion of order per se. This is incorrect. Increasing entropy does not mean increasing disorder in any intuitive, layperson sense. For example, the spontaneous (i.e., entropy-increasing) evolution of an oil/water mixture is toward oil on top, water on the bottom, which is hardly increasing disorder according to any intuitive meaning of the word disorder.

Confusing the issue is that while increasing entropy does not mean increasing disorder in any intuitive, layperson sense, it does mean increasing disorder in an abstract, highly technical sense defined by the Boltzmann equation

$$S = k \ln W + S_0, \quad (5)$$

where k is Boltzmann's constant (1.38×10^{-23} J/K), where the constant S_0 is, by tradition but not by necessity, set equal to zero, and where W is defined by Eq. (6), where N is the number of particles in the system and n_i is the number of particles in energy level i :

$$W = \max_{n_0, n_1, \dots, n_k} \frac{N!}{n_0! n_1! \dots n_k!} \text{ such that} \quad (6)$$

$$\sum_i n_i = N \text{ and} \quad (7)$$

$$\sum_i n_i \cdot (\text{Energy level } i) = \text{Total system energy.} \quad (8)$$

(Note that $N!/(n_0! n_1! \dots n_k!)$ is the number of ways that N distinguishable objects can be put into groups of $n_0, n_1, \dots, n_k$ objects.)

While W defined in Eq. (6) is not a random variable and comes from combinatorics not statistics, Boltzmann christened it the thermodynamic probability, which has led observers to incorrectly think

that entropy has something to do with randomness. On the other hand, another interpretation of Boltzmann's ideas does involve randomness. In this interpretation, the maximization operator in Eq. (6) is dropped, and W becomes a random variable, which takes on the value $N!/(n_0! n_1! \dots n_k!)$ with probability

$$\frac{N!/(n_0! n_1! \dots n_k!)}{\sum_{n_0, n_1, \dots, n_k} N!/(n_0! n_1! \dots n_k!)}.$$

The problem with this interpretation is that it allows the second law to be violated with nonzero probability, and such a violation has never been observed.

Boltzmann's ideas have grown into the important science of statistical mechanics, by which the entropies of various substances are explained, from a molecular level; the older classical thermodynamics described how to measure entropy but could not explain why it should take on a particular value for a particular substance. For extremely simple physical systems, statistical mechanics does imply that lower entropy states are those which are intuitively more ordered and less random. Thus, one mole of a substance in the form of a simple solid has lower entropy than if that mole is in the form of a simple liquid, and that in turn has lower entropy than if that mole is in the form of a simple gas. Also, random mixtures of simple substances have higher entropy than if those substances are ordered into unmixed, pure components, as seen in Eq. (4). However, most substances are not simple in the sense necessary for this intuition to be valid, as the oil and water example showed. Treatments that ignore this caveat—such as Erwin Schrödinger's famous 1944 essay "What Is Life"—should be read with the greatest caution.

5.2 Absolute Entropy

Notice that Eq. (1) defines only changes in entropy; absolute amounts of entropy are not defined. It is sometimes asserted that, by contrast, in statistical mechanics absolute entropy is defined—equivalently, that the constant S_0 in Eq. (5) must be zero. Modern treatments of entropy and its physical meaning show that this is not true. This controversy has no bearing on laboratory entropy measurements, which always measure entropy differences anyway.

5.3 Thermal versus Configurational Entropy

Occasionally, scientists find it useful to distinguish the contribution to entropy change represented by Eq. (4)

from other contributions to entropy change. When this is done, the component of the entropy change given by Eq. (4) is called configurational entropy or the entropy of mixing, and the other components of the entropy change are together called thermal entropy. The choice of the term thermal entropy is unfortunate, because it can mislead one into thinking that configurational entropy has nothing to do with thermal changes and hence that configurational entropy has nothing to do with Eq. (1). Configurational entropy is actually just as much a component of Eq. (1) as thermal entropy is.

5.4 Entropy and Free Energy

To calculate the entropy change when an event occurs, the resulting entropy change of the system's surroundings must be added to the entropy change undergone by the system itself. Under conditions of constant pressure and temperature, the former is the energy released by the system into its environment (traditionally represented by $-\Delta H$) times $1/T$, and the latter is ΔS . Adding these two entropy changes together yields a total entropy change of $\Delta H/T - \Delta S = (1/T)(\Delta H - T\Delta S) = (1/T)\Delta G$, where ΔG is known as the Gibbs Free Energy. Tables of Gibbs Free Energies of common substances are readily available, and such a table is all one needs to determine how entropy changes when an event occurs under conditions of constant pressure and temperature. This makes Gibbs Free Energy extremely valuable for chemists, but it should not obscure the fact that entropy is the fundamental physical quantity and that Gibbs Free Energy is a derived quantity of no independent physical importance. Similarly, Helmholtz Free Energy $F = U - TS$ (where U is internal energy) is useful in situations where volume and temperature T are constant, but it is not a fundamental quantity either.

5.5 Information

An equation that looks like Eq. (5) with Eq. (6) appears in the field of information theory, with, of course, entirely different meanings for the symbols. This does not imply any connection between entropy and information, any more than the similarity between the formula for the sum of scores on two student examinations ($e_1 + e_2$) and the formula for the sum of the energy of two hotel elevators ($e_1 + e_2$) implies any connection between student examination scores and hotel elevators. The entropy of information theory, while no doubt quite useful to practi-

tioners of that field, has nothing to do with the second law of thermodynamics.

5.6 Minimum Energy Requirements

Near the end of Section 3.1 it was shown that the amount of energy E degraded during an irreversible process is the temperature of the coldest heat reservoir at hand (call it T_0) times the entropy change of the universe brought about by the irreversible process:

$$E = T_0 \Delta S. \quad (9)$$

Suppose we desire to obtain pure copper from oxide copper ores of the type often found in Africa. Because CuO is thermodynamically stable at room temperatures, breaking up CuO into copper and oxygen would cause an entropy decrease, so it cannot occur spontaneously. Some authors have calculated the amount of the entropy decrease, then used Eq. (9) to assert that T_0 times that amount is a minimum amount of energy necessary to liberate the copper from the CuO. (Their general proposition is that there is a definite theoretical limit to the minimum energy input to a process and that there is no substitute for this minimum energy.) There may be some sense in which this is true, but no amount of an energy commodity (such as coal or electricity) is needed to obtain copper from CuO. What is done is simply to introduce sulfuric acid into the environment of the copper oxide; a typical reaction is $\text{CuO} + \text{H}_2\text{SO}_4 \rightarrow \text{Cu}^{++} + \text{SO}_4^{--} + \text{H}_2\text{O}$, which increases entropy and is thus spontaneous. This example shows that there is no minimum "energy commodity" requirement for physical, chemical, and hence industrial processes.

5.7 Entropy and Energy

There is no such thing as low-entropy energy. Entropy is a characteristic of matter, and a substance's entropy changes when the substance reversibly absorbs or emits energy (or when an irreversible process involves the substance and a reversible process linking the same initial and final states would entail the substance's absorption or emission of energy); see Eq. (1).

5.8 Entropy Theory of Value

A glance at a table of standard entropies for different substances will show the error of believing that the lower an object's entropy, the higher its economic value. Similarly, each time one heats a kettle of water

in a kitchen, the water becomes more useful at the same time as its entropy increases. Adding impurities to iron to create steel simultaneously increases the entropy of the metal and increases its economic value. The entropy of oceanic pollutants is decreased by their concentration in the flesh of oysters, but this decrease of entropy makes the oysters unsafe to eat, thus less valuable to humans. So there is no general connection between low entropy and high economic value, nor between high entropy and low economic value. It is true that the process of production and consumption turns valuable inputs into useless waste, with the useless waste having, as a total assembly, higher entropy than the valuable inputs. However, this generalization only applies to assemblies as a whole, not to their component parts.

Furthermore, what is valuable will always be partially determined by psychological, historical, sociological, and in any case nonphysical traits unrelated to entropy; a personal computer made in 1987 was quite valuable then, but would be considered junk now, even if its physical traits are unchanged. Even a simple container of oxygen, whose entropy decreased when it was extracted from air, can be useful and valuable in one situation—helping hospital patients breathe—and be an unwelcome hazard in another situation, such as near a stove.

6. CONCLUSION

The entropy law implies that economic production will never be entirely self-contained or self-replicating, let alone self-expanding. Production will always require fresh inputs from nature, and nature's size, both of terrestrial stocks and of the rate of solar flow, is finite.

SEE ALSO THE FOLLOWING ARTICLES

Aggregation of Energy • Economic Systems and Energy, Conceptual Overview • Economic Thought,

History of Energy in • Entropy • Evolutionary Economics and Energy • Exergoeconomics • Exergy • Exergy Analysis of Energy Systems • Thermodynamics and Economics, Overview • Thermodynamic Sciences, History of • Thermodynamics, Laws of

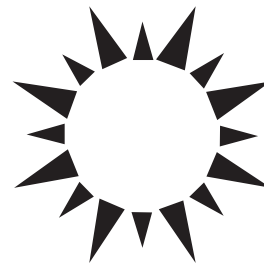
Further Reading

- Atkins, P. W. (1984). "The Second Law." Scientific American/W. H. Freeman, New York.
- Beard, T. R., and Lozada, G. A. (1999). "Economics, Entropy and the Environment: The Extraordinary Economics of Nicholas Georgescu-Roegen." Edward Elgar, Cheltenham, UK.
- Biswas, A. K., and Davenport, W. G. (1994). "Extractive Metallurgy of Copper." 3rd ed. Pergamon Press, Oxford.
- Daly, H. E. (1993). The steady-state economy: Toward a political economy of biophysical equilibrium and moral growth. In "Valuing the Earth: Economics, Ecology, Ethics" (H. E. Daly and K. N. Townsend, Eds.). The MIT Press, Cambridge, MA.
- Dugdale, J. S. (1996). "Entropy and Its Physical Meaning." Taylor & Francis, Bristol, PA.
- Faber, M., Proops, J., and Speck, S. (1999). "Capital and Time in Ecological Economics: Neo-Austrian Modelling." Edward Elgar, Cheltenham, UK.
- Gaskell, D. R. (1981). "Introduction to Metallurgical Thermodynamics." 2nd ed. Hemisphere, New York.
- Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard University Press, Cambridge.
- Goldstein, M., and Goldstein, I. F. (1993). "The Refrigerator and the Universe: Understanding the Laws of Energy." Harvard University Press, Cambridge, MA.
- Lozada, G. A. (1999). The role of entropy and energy in natural resource economics. In "Bioeconomics and Sustainability: Essays in Honor of Nicholas Georgescu-Roegen" (J. M. Goudy and K. Mayumi, Eds.). Edward Elgar, Hants, England.
- Toman, M. A., Pezzey, J., and Krautkraemer, J. (1995). Neoclassical economic growth theory and 'sustainability.' In "Handbook of Environmental Economics" (D. W. Bromley, Ed.). Basil Blackwell, Cambridge, MA.
- Wicken, J. S. (1987). "Evolution, Thermodynamics, and Information: Extending the Darwinian Program." Oxford University Press, New York.
- Zemansky, M. W. (1968). "Heat and Thermodynamics: An Intermediate Textbook." 5th ed. McGraw-Hill, New York.

Environmental Change and Energy

I. G. SIMMONS

University of Durham
Durham, United Kingdom



-
1. Human Access to Energy
 2. Conceptual Frameworks
 3. Historical Stages of Access to Energy
 4. Overview from Today's Perspectives

Glossary

agro-ecology Crop production as a component of an ecosystem that includes not only the desired species but also, for example, the weeds, water, and nutrient relations; soil characteristics; and human inputs.

cosmogenesis The creation of the whole cosmos in a single event. In some societies it has been seen as an annual renewal of nature.

developed countries Those nation-states with an industrial economy and generally high gross domestic product and commercial energy use per capita, though not necessarily evenly distributed.

extrasomatic Outside the body. In using a bow and arrow, the energy deployed is somatic but the technology used to channel it is extrasomatic.

heterotrophic Nourished from outside—that is, of all organisms incapable of photosynthesis, which constitute most living things except green plants.

intensification Delivery of more product per unit area per unit time, for example, double-cropping of rice or shelling with heavier ordnance.

less developed countries (LDCs) Those nation states with a low level of industrial development and usually a low consumption of commercial energy. Most have high rates of population growth. Sometimes referred to as low-income economies (LIEs).

moa The large flightless herbivorous bird of New Zealand, equivalent to the ostrich, emu, and cassowary. Extinct.

noosphere Concept of Soviet biologist V. I. Vernadsky that there would be a global envelope of thought equivalent to the lithosphere and biosphere.

organochlorine A class of chemical developed in and after World War II with biocidal properties; the most famous

is DDT. Without natural breakdown pathways, they have accumulated in ecosystems. Still used in LDCs.

overkill A popular but graphic term for the mass extinction of genera of animals. Originally conceived for late Pleistocene times but extended to the disappearance of the dinosaurs and held in reserve for a “nuclear winter.”

pastoralism A mode of life involving herds of domesticated animals pastured over large areas, with seasonal migrations that track the availability of water and animal food.

permanent cultivation The replacement of a preexisting ecosystem with a permanently manipulated agroecology in order to produce crops as frequently as climate and soil conditions allow.

photosynthesis The chemical process of the basis of life, in which the energy of the sun is transformed into the tissues of plants.

precautionary principle A commandment of environmental management in which actions are avoided if there is a probability that they will lead to environmental degradation. Like certain other Commandments, breaches are frequent.

savanna A regional-scale ecosystem of tall grasses with interspersed trees. It usually experiences fire at least once a year and in Africa may well have coevolved with the genus *Homo*.

shifting cultivation Crop production in which the land is allowed to revert to a wilder state once the soil is exhausted or weeding is too onerous. Much derided by colonial administrators but ecologically appropriate in times of low population growth.

teratogenic Of fetuses (usually but not exclusively applied to humans) that are deformed by the mother's ingestion of certain chemicals. Dioxins (of which some are more toxic than others, such as 2,3,4,7,8-TCDD) have been detected as a major cause.

1. HUMAN ACCESS TO ENERGY

In the planet's ecological webs, humans are heterotrophic: they cannot fix energy from sunlight like

plants. So in its 4-million year perspective, human history can be seen as a set of stages of access to energy sources. Early hominids relied on recently captured energy as edible plant and animal material, and at one stage (whose chronology is disputed but may have been as long ago as 0.5 my) gained control of fire at landscape scale. This reliance on solar energy as biomass is also characteristic of preindustrial or solar-based agriculture, but here the concentration of energy is enhanced by making it available in delineated fields or herded animals. Hence, many more people could be fed off a smaller area, especially if water control could be added to the technological repertoire. Once fossil fuels were exploitable for steam generation, solar power could be subsidized by energy from the stored photosynthesis as represented by coal, oil and natural gas. Energy concentration in these fuels was higher than more recent sources, and this trend is continued by the harnessing of the energy content of the nuclei of uranium and hydrogen.

2. CONCEPTUAL FRAMEWORKS

A number of concepts are embedded in this overview of human history. The first is that of a stage of simply garnering the usufruct in the form of plant and animal material that could be gathered, hunted, or scavenged. Most early human societies were probably mobile, so that these early species of the genus *Homo* could be said to have lived in an environmentally determined fashion. Thus, another concept is that of control and the rest of human history can be viewed as a desire to increase the degree of mastery over environmental components, of which energy has been arguably the most important. The first key step was regulation of fire away from the hearth and into the landscape, with the multiple benefits thus brought about, as amplified later in this article. Fire is also at the center of the developments of industrialization, when the combustion of coal released energy that could be controlled via a steam-using technology; its concentration was such that even mobile devices could be pioneered. Direction of the energy of the atomic nucleus has been achieved for fission but for fusion, only the awesome threat of the hydrogen bomb remains available.

Another major concept is that of technology. Apart from eating it as food, humans must access the usefulness of most energy sources via an extrasomatic device. They must learn to preserve and transport fire, for example; crop harvesting is enhanced by the use of an iron blade produced by

smelting metal ores with charcoal; the steamship carried goods, ideas, and troops from one end of the earth to the other. The energy made available to growing populations in the 19th century from the potato comes not only from the carbohydrates inherent in the tubers but from its better growth when encouraged by bagged fertilizers. Electricity is a flexible carrier of energy that has facilitated all kinds of environmental change, not least in the transmission of ideas about how to control the environment. In all these devices, the hope is that the group will improve their net energy yield—that is, the ratio of usable energy harvested to energy invested as labor, draft animals' food and the making of the technology itself. The major stages in human access to energy before the 20th century and the quantitative results are summarized in Table I.

Technology, however, is not simply a set of inventions. It is the product of, and embedded within, human societies. This means that the strict rationality and "scientific" qualities associated with it are not necessarily found in every application. To begin with, much technology was prescientific, in the sense that it was the product of practical-minded men (sic?), who happened on a better way of doing things. Only after the 17th century did the theory precede the technology in truly scientific procedures. (It has been argued that the steam engine, for example, was pioneered by practical men with no knowledge of the theories of physics.) Certain possibilities follow: it is feasible to reject technology on religious, economic or simply prejudicial grounds, as medieval Islam is said to have regarded the wheel

TABLE I
Energy Consumption through Time

Economy	Energy consumption (GJ cap day)	Sources
Staying alive only	0.0001	Plant and animal foods
Hunter-gatherers	2.0	Wild plant and animal foods plus fire
Preindustrial agriculture	20–40	Domesticated foods plus fire, charcoal
Eve of industrialization	70–100	Foods plus fire and plentiful fuels such as charcoal
Britain in 1800 CE	20	Foods plus wood and a little coal
Britain in 1900 CE	116	Foods plus developed coal-based economy

and sensible nations have regarded atomic weapons. Hence the distribution of energy technologies is likely to be uneven, as is shown by the access to commercial energy today, or by the way in which the adoption of fossil-fuel based technologies took off in the West after ca. 1750 AD but not elsewhere. So access to, and use of, energy in any society is mediated by technologies that are themselves subject to the organization of that society, with all its potentials for, especially, inequalities of distribution.

The relationship of all these kinds of energy use to environment has of course varied according to the intensity of available energy and the number (and density) of people, but a few historical generalizations can be made. The first is that the environment is a source of energy that may be altered as the energy is garnered. An animal population severely depleted by intensive hunting is an obvious example. A savanna fired frequently to drive out game will undergo species' shift as fire-intolerant plants are extirpated. A strip mine changes the ecology in a radius measured in tens of kilometers. Second, while many of these changes may be deliberate and desired by a society, there may well be side effects which are undesired and degradatory. The smoke from unregulated coal burning has been the source of much human ill health in cities since the 13th century, for instance; the fallout from the nuclear accident at Chernobyl in 1986 sterilized large areas of formerly productive land. The possibility of negative consequences of human actions in matters of energy access and use was pointed out in the myth of Prometheus, which retains a certain resonance today.

3. HISTORICAL STAGES OF ACCESS TO ENERGY

None of these stages, with exception of the last, can be given precise dates. Thus the chronology must be read as having wide margins for different places. Agriculture, to quote an obvious example, may have been dominant in southwest Asia by 8000 BCE but only arrived in Australia with European colonization in the 19th century. Coal was used to console Roman soldiers on Hadrian's Wall but is only a significant global influence from about 1800 AD/CE.

3.1 Hunters and Gatherers (4 MY–8000 BCE)

This stage encompasses (1) human biological evolution in the transitions from early African forms

of *Homo* to the anatomically modern human of *Homo sapiens* at ca 35 ky with a worldwide distribution and (2) cultural development through the various levels of the Palaeolithic and Mesolithic to hunter-gatherers described at and after European contact, usually described as near-recent groups. There are a few remaining groups but a way of life that comprised 100% of the human population in 10,000 BCE now accounts for something less than 0.001% and is shrinking. The term "hunting and gathering" also includes fishing and scavenging; it may also be written in the reverse to suggest the dominance of plant material in most of their life ways.

The environmental relations of hunter-gatherers are dominated by the practice of food collection rather than food production. This is not to say that there is no environmental manipulation in the cause of gaining nutrition or status, but that it is practiced in a context where there is (1) a low population density, (2) no conscious attempt to manipulate the genetics of biota with the exception of the dog, and (3) a mythological context of nonseparation from the natural world and of annual cosmogenesis. For many years, academic interpretation of these relationships postulated a "child of nature" construction in which hunter-gatherers lived in complete harmony with the natural world, taking little from it and exhibiting complete reverence. Yet the development of an ever more efficient technology was characteristic of that age, as it is of industrial and post-industrial times. Devices such as the bow, spear thrower, and blowpipe harnessed the energy of a group of human muscles to direct a missile down a narrow channel. Net energy returns were of the order of 1:5-15 and might rise to 1:40. Controlled fire seems to have been used even in wet places like the lowland tropical forests, !QJ; the Californian Coast redwoods, and the moorlands of England and Wales.

Some examples of hunter-gatherer life indeed fitted a Golden Age view. Work in the 1950s and 1960s on groups like the !Kung San Bushmen of the Kalahari seemed to show that with relatively little effort, a well-nourished population could persist even in an unforgiving environment. It is certainly true that given a large and preferably diverse territory, an abundant staple food (the case of some of the !Kung it was the mongongo nut), and a slow rate of population growth, hunter-gatherers could exert relatively little effect on their environment. Some indeed adopted conservationist attitudes to their resources. Nevertheless,

many historical examples have been shown to have brought about considerable environmental changes.

The most thoroughgoing of such manipulations has been the discovery that the introduction of humans into regions hitherto lacking their presence has resulted in the extinction of large groups of animals, with an emphasis on the loss of large (> 50 kg adult body weight) herbivore mammals and flightless birds. The initial scene for this revelation was the High Plains of North America in the late Pleistocene (12–11 ky) but the same phenomenon has been detected in islands in the Pacific and Indian Oceans at later dates, such as New Zealand/Aotearoa (900 CE), Hawaii (600 CE), and Madagascar (700 CE). By contrast, no such overkill phases have been noted for Africa, where humans evolved and spread gradually. Another relationship between hunters and animals sees massive kills but no long-term effects on the populations of the latter since they are so massive (or so rapid at reproduction) that humans at hunter-gatherer densities cannot bring about extinction. This seems to have been the case before European contact on the High Plains of North America. Here, as in the northern forests of Canada, such equilibria did not survive contact with the more energy-intensive economies, whether these were basically agricultural as with the Hudson's Bay Company and its fur traders or the railroads with their hungry and bored laborers. Such interrelationships have been found in recent food-collecting societies: the BaMbuti of the Zaire Basin have traded meat outside their forests to the detriment of duiker populations, for example. Some groups have internalized industrial technology, as with the Inuit, the rifle, and the snowmobile, for example, leading to controversy about their access to traditional prey such as the bowhead whale. Interestingly, no such contentions seem to have arisen over plant life, suggesting that the mythic role of dead animal flesh is deeply rooted in human societies: an aspect of control, perhaps.

At an early stage, hominid energy use must have been totally metabolic in nature, confined to that needed for growth, movement, and reproduction. As cultures became more complex, demands for additional energy grew, with any orientation toward meat consumption likely to cause more impacts than reliance on plant material. Notable technological developments seem oriented to the procurement of meat rather than plants, and in the great artistic monuments of the Upper Palaeolithic in the caves of France and Spain there are no paintings of roots and berries.

3.2 Pre-Industrial or Solar-Powered Agriculture (8000 BCE–1750 CE)

In energy terms, agriculture represents an *intensification* of access. To collect enough energy for survival, an average food collector needed some 26 km² of territory as a support system. Subsistence farmers with adequate water supplies can manage on 0.1 ha per head and achieve net energy returns of 1:20. Environmentally, this is achieved by breeding plants and animals in cultural images, by altering ecosystems so as to direct their energy flows at human societies and by coping with the unforeseen consequences of these alterations of the previous patterns. The actual biological channels through which these new patterns were directed varied in time and space. In southwest Asia in 9000 to 8000 BCE, wheat, barley, sheep, and goats were central; in southeast Asia at the same time it was rice above all; in the Americas somewhat later, the potato, tomato, maize, and llama were key elements. Yet the energetics had much in common, as had the environmental consequences.

Compared with hunter-gatherers, the energy is garnered closer to home. There is usually a permanent settlement and the food energy is produced so as to minimize expenditure on travel—hence the development of the field, where a crop is kept within bounds so that its maintenance and yield is under the eye of the farmer, be it wheat or cattle. Energy has to be invested in sowing, weeding, fencing, predator control, storage, and myriad other tasks so that the nearer at hand, the greater the probability of a positive energy balance at dinnertime. The domestication of animals meant that a seasonal round of tapping plants not edible by humans and turning them into milk, meat, blood, and skins could develop on the arid or mountainous fringes of cultivable areas.

Not surprisingly, humans enlisted help in the new diversity of tasks required by the novel economy. Draft animals are not really energy efficient in the sense that they consume food even when not working but they can deliver concentrated power at key times, such as spring ploughing or carrying metal ores to the smelter. They relieve humans of wearisome tasks and may furnish symbols of status. Equally important ameliorations of work came with the harnessing of solar energy in the forms of water and wind power. Perhaps the most important was the realization that precipitation (itself a consequence of solar energy distributions) could be stored and then released at times critical for crop growth or animal watering. Coupled with energy investment in, for

example, terracing, crop production was not only intensified but made more reliable. Since storage and distribution often meant that water had to be lifted, the application of wind power to the chore was another gain. The transformation of energy in wind and falling water to rotary motion also facilitated the technologies of the mill, useful for grinding corn and metal ores as well as lifting water. So as with the bow and arrow in the Upper Palaeolithic, a whole range of extrasomatic technology is brought into action to achieve both survival and wider cultural aims.

Food production is thus the core of preindustrial agriculture and at times of Malthusian pressure, virtually any environmental manipulation becomes acceptable in order to increase the quantity of crop. There are three main types of agroecology: shifting cultivation, permanent cultivation, and pastoralism. The first of these, sometimes regarded as a primitive system, is improved by conversion to permanence, but this is probably a colonialist view of a sophisticated piece of adaptation to forest and savanna environments. Its ecology focuses on the clearance of a patch of near-natural vegetation and the take-up of the nutrient capital into crops. Cleared vegetation may be burned to add to the available stock of nutrients. As fertility falls or weeds become tiresomely prevalent, the patch is abandoned and another opening replaces it. Revegetation then builds up the nutrient stock and smothers the weeds. The crop plants in fact much less efficient at energy capture than the original vegetation, but that loss is traded for the cultural acceptability of the food, herb, spice, and medicinal plants that replace it. Moreover, a cleared patch in a matrix of near-natural vegetation does not readily leak nutrients to runoff, so energy expenditure on adding manure or other sources of nutrients is limited compared with permanently cleared areas. The system is not easily intensified, however, and has to be transformed when population growth necessitates a return to a cultivable patch before the vegetation has restored the nutrient levels. Though generally thought of as a tropical system (especially in forests and savannas), it was used in the deciduous forests of Europe in the Neolithic (7000–5500 BCE) and as late as the early 19th century CE in the Boreal forests of Finland.

The historical energetics of food production are without doubt dominated by systems of permanent cultivation, found all over the world except in polar and subpolar regions. Along with sedentary settlement, they represent one of the truly revolutionary shifts in human-environment relations and the energy relationships reflect the new conditions. The basic

agroecology is that of the permanent replacement of the preexisting ecology with crop plants and animals and of progressive intensification of output as populations grow. The energy flow of the land is directed toward the crop biota and all else is labeled as weeds or pests and if possible killed or driven away. As with shifting cultivation, maintenance of nutrient levels is at the heart of the system, and both energy and materials have to be used in their upkeep. Cropping is, of course, a major factor in nutrient loss from the fields, but bare soil is vulnerable to losses from runoff and many systems have adopted strategies for minimizing such losses: cover plants are one such, but the most universal is the terrace, found in a broad belt from the equator to about 55°N and the equivalent southern latitudes.

To harvest the energy in a crop, there has to be considerable energy investment. At some time, there has to be breeding of the best strains of plants and animals for local conditions, with no necessary guarantee of success in prescientific times. More obviously, consider the average field in a preindustrial society. The soil has to be turned over for a start. A spade or digging stick may suffice, or draft animals may be harnessed to a plough with a wooden or metal-tipped ploughshare. The energy pathways can be immediately seen: human energy in the first and animal energy in the second. The animals have to be fed and may thus form part of the energy flows taken by the post-harvest crop. Or indeed they may compete for energy with the human population by needing special croplands: “horses eat people” was a proverb in early modern Hungary. If metal is used then there is an ancillary need for energy in smelting and working the iron. Sowing and weeding need investment, as does the fencing, which keeps the wrong animals out of the crop or the wolves from the sheep at night. Children are very useful for shepherding and bird scaring but have to be fed. Harvesting of plants is generally labor intensive, and only careful built storage will keep at least some of the rats out. Water management also requires energy investment, whether the problem is too much of it or too little. Lastly, many plant foods require cooking since the ability of the human gut to digest cellulose is restricted. Our species is better at processing animal tissues, but the thermodynamics of conversion of plant material to animal tissue ensure that meat in preindustrial societies is a luxury. Yet in many parts of the world, such systems were stable environmentally, nutritionally adequate and even provided surpluses.

On the fringes of settled areas, environments not conducive to permanent agriculture were often

found. Characteristically, these were semiarid in nature, with long periods without rainfall and no permanent surface drainage. Mountains form another category. These could nevertheless be tapped for human use via flocks of domesticated animals in the system known as pastoralism. The fundamental agroecology was the cropping of cellulose in plants inedible to humans by domesticates such as sheep, goats, and llamas (and many more species adapted to particular environments, such as the yak), whose sustainable products could be stored for later consumption and trade. Many dairy products have their origin in such systems and surplus animals might be sold off for meat (and skins) though eaten only rarely by the pastoral communities themselves, who, however, sometimes tapped blood from the live beasts as a source of protein. The energetics of this system are more like hunting and gathering than permanent cultivation in the sense that there has to be continual movement, in this case so as not to overcrop the vegetation. One form of the life way involves the movement up mountains in the summer. This transhumance was common in Europe until the 19th century, and a form of it persisted among people of Basque origin in Nevada until very recently. The energy flow of this system is generally completed by the selling off of most of the embedded energy in the animals and their products to sedentary societies, and there is no investment in nutrient levels. Once the pastures are visited often, then the photosynthetic efficiency declines as the species' mix shifts toward the arid tolerant and more bare ground is seen.

Fishing was often an efficient method of garnering energy since the investment came more or less free as river water, lakes, or wind in sails. Energy investment in construction of boats, weirs, and traps was generally long lasting: only nets required frequent renewal. Yet fish were usually an adjunct to diets: only in a very few places were agricultural societies dominated by their consumption, though as an addition to the palate of flavors they have long been popular.

One outcome of these systems was that they could often provide a surplus. This could be used to fuel, so to speak, population growth in the producing society. Alternatively, it could be mulcted off to feed a superior class who did not have to toil in order that they might eat. Instead they might effect environmental change in other directions. One direction might be that of developing certain nonagricultural skills full time: itinerant metalsmiths are postulated for prehistoric Europe, and presumably the astronomers of ancient Babylon did not grow their own barley. Likewise, the management of forests could

become a specialized occupation, with a society recognizing that the maintenance of its chief energy source apart from food was a key element in its survival. This was especially so in the great ages of empire building and trade in wooden ships, when ship timber became a key instrument of national policy for countries such as Great Britain and France. At a less material level, philosophers and university faculties might not have to bake their own bread, though interestingly monastic houses and rural clergy were usually required to be self-sufficient to some degree, even though the latter might be entitled to a tithe of local crops. The most visible product of diverting energy into the hands of an aristocracy (who in a fascinating piece of etymology in English wielded "power") was pleasure and conquest. The latter came by feeding large numbers of soldiers, usually by local communities and on an involuntary basis, diverting the energy stored after the last harvest into the army's day-to-day maintenance. The former can be seen in landscape relics such as hunting parks and gardens. Many a European city park was once the playground of the King or the Archbishop: St. James's and Hyde Parks in London had such a royal origin. The gardens of the Alhambra Palace in Granada were the work of the governors of the moorish province of Andalusia. Lastly, energy might be invested in the afterlife: only very productive irrigated cereal-growing provided enough surplus to feed the laborers who built the Giza pyramids in Egypt.

Above all, perhaps, surpluses made possible the city. Although small by today's standards, they represented concentrations of stored energy in the form of structures. They lived heterotrophically off imported energy as food and fuel, which they transformed into all kinds of wastes. (The basic ecology of today's cities is not very different.) Though gardens might produce food and water mills some energy, their citizens were mostly dependent on their hinterlands, for which a certain nostalgic reverence turns up as early as classical Rome. Some of the energy and matter was however turned into knowledge and stored not only in people's heads but in writing. It is difficult to see how the growth of abstract thought could have taken place without the concentrations made possible by the energy surpluses allowed by well-conducted agriculture. Even before the steamship, the control of wind power allowed the improvement of food production by its facilitation of transferring new crops to new places. The role of the potato in industrializing Europe followed on from its earlier spread out of South America, for example; the penetration of all kinds of European practices into

the Americas can scarcely be contemplated without the horse. So good energy trappers were taken around the world. It might be arguable that the major takeoff of human population growth usually dated to about 1650 CE was in part fuelled by improved energy capture and that better crops of all kinds were implicated in the improved nutrition. (It can be added that in some countries the use of energy to boil water to make tea and coffee has been suggested as a health-producing measure that added to survival rates. Maybe).

The theme of surplus energy produced by pre-industrial agriculture has one apogee: it fed the people who conceived and built the pyramids, the Gothic cathedrals, Angkor Wat, and the Ryoan-ji. Somewhat imperfectly, it must be admitted, Mozart got fed as a result.

3.3 Industrialism (1750–1950 CE)

This is sometimes called the age of fossil fuels. Without doubting the accuracy of the label, note should be made of the exploitation of these energy sources in the years before their full flowering. Natural gas, for example, was piped in hollow bamboos from natural seepages into Chinese towns in medieval times and lit the streets. Coal exports from the port of Newcastle in northeast England fuelled breweries and laundries in London in the 14th century CE. But all this was as nothing compared with the explosion of coal use (followed by oil in the late 19th century and natural gas in the mid-20th century) once the potential for combustion of these fuels to produce steam, kinetic energy of all kinds, heat, and electricity had been realized. Further, once past the mid-18th century, the technology might follow a science in that hypothesis testing and theory building preceded the construction of the machines.

The concentrated energy content of mineralized photosynthesis is central to the idea of surpluses. The environmental relations are characterized by the impacts of getting the fuels and then of using them via, as before, the mediation of technology. The key change from the previous era is the highly positive ratio of energy investment to energy yield at, for example, the pit head and the well head, with ratios of 1:20–30. Even when postextraction processing is complex, as with oil (though not with coal, the earliest fuel of this kind to be used on a massive scale), the ratio is still highly favorable. Only with the large heat losses encountered in the generation of electricity in conventional power stations do the ratios begin to drop toward preindustrial magnitudes.

The energy-environment story of the 18th and 19th centuries is that of the adoption of fossil fuels as the basis of transport, communications, and manufacturing in a set of regions that became what are now called the developed countries (or sometimes the north or the high-income economies [HIEs] or simply the rich) and the failure of other parts of the world to undergo the same transitions. They became the less developed countries (euphemized as the developing countries, or sometimes referred to as the south, the low-income economies [LIEs], or the poor). The reasons for both are complex and outside the scope of this work. The energy available in classical Greece, in 17th century Europe, and China in 1880 CE was about the same but the adoption of fossil fuels brought about radical change. The core areas in, say 1850 CE, were western and central Europe together with eastern North America. Soon to be added were eastern Europe and western Russia, Japan, a few parts of the northern shore of the Mediterranean, and other patches of North America. On the map, the sign of the times was the railroad. Those heartlands are still significant in the sense that they acted as core areas for whole nations to adopt enough of the new technologies to ensure their future access to fossil energy supplies, so the present-day situation in which access to energy (usually measured as per capita consumption of commercial energy) replicates roughly the precedence established in the 19th century. Though wood as a fuel has declined greatly in relative terms (its energy content is about 14 MJ/kg, compared with hard coal at 29 MJ/kg), in absolute terms its use is still rising due to population growth in the LIEs (Fig. 1).

The myriad ways in which these fuels are extracted and utilized exert considerable environmental impacts, and many books have been written about each of them. There is a basic ecology common to all in which the hydrocarbons are lifted from the earth, processed, and stored, and then transformed into their end uses. These latter may be direct (as heat, for example, to power boilers) or as embedded energy as in fertilizers and biocides. All kinds of machines mediate at every stage and millions of different items are manufactured. The environmental impact of some is obvious and direct: the bulldozer is a good instance. Others are more subtle: stored electricity makes possible battery-operated TVs in remote Asian forests whose people are then able to see HIE consumerism at work, or indeed at play or at prayer. The bulldozer is the culmination of earth-moving equipment that started with the steam crane; microelectronics carry on the tradition of the electric telegraph. In that sense,

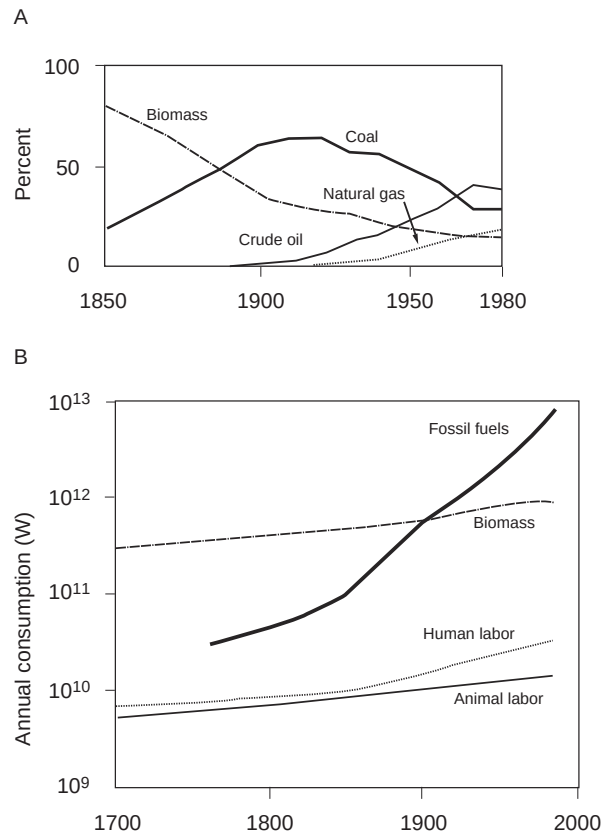


FIGURE 1 The transitions to fossil fuels. (A) shows the global proportions of primary energy for the period 1850–1980 CE and (B) the absolute amounts over a longer time period. Though biomass has declined in relative terms, the rising populations of LDCs (especially in Africa and Asia) have increased actual demand for wood fuel. From V. Smil (1991), “General Energetics.” Wiley-Interscience, New York and Chichester, Fig. 10.2 (part), page 185, and Fig. 14.6, page 303. Copyright © 1991. This material is used by permission of John Wiley & Sons, Inc.

the latter exemplifies the way in which energy as electricity above all carries ideas, which include those devoted alike to changing and protecting the environment. It has without doubt made possible the global coalescence of ideas which have on the one hand allowed corporations quickly to find states willing to accept GM crops but on the other hand pioneered agreements like the Kyoto protocols, which all civilized nations have agreed to ratify.

The historical effects of industrialization on food production are seen in many environmental changes in the 19th century. Many temperate grasslands were ploughed to grown cereals for export to the rapidly growing populations of industrializing regions. The demand for meat also allowed vegetation change in other grasslands as grazing densities increased. All such changes were facilitated by the takeover of areas

previously used by preindustrial people (nowadays labeled as “indigenous”), whether officially colonized (as in South Asia where “empty” lands were converted to tea and coffee by British colonial authorities) or internally reallocated, as on the High Plains of the United States. In the case of meat, the technology of refrigeration was critical, as was the reliable transport offered by railroad and steamship. Although to be more a feature of the post-1950 period, the disappearance of the seasons in the rich world began, as asparagus, strawberries, and flowers consumed transportation energy in their journeys from the warm and sunny to the still-emerging-from-winter shops of London, Paris, and New York. The loss of the local, as energy for transport plummeted in price, has its origins in the 19th century, though clearly not confined to that period. One short-lived counter-trend came about in the 1890s in Paris, where there was enough horse-dung to fertilize and heat an export trade in winter salad crops within the city boundaries.

The theme of increased material production can be carried beyond that of food. One result of immense environmental significance was the burgeoning appetite for material possessions, leading to environmental manipulations of all kinds. This was exacerbated by the craving for the new, so that reuse and recycling (a feature of poor societies everywhere) became far less common. Again, the introduction of electricity (as late as the post-1945 in many rural parts of Europe) impelled many authorities to speed up hydropower development, with its many environmental consequences. This is just one of many environmental outcomes of the extraction of energy sources: huge open-cast coal mines in central Europe and oil wells in Mesopotamian deserts were all visible before 1950. The overall impression is of self-amplifying systems in which the extraction of energy is followed by the development of technology to use it and this in turn spawns further environmental change. The motor vehicle is a case in point, where impacts include not only the emissions to the atmosphere but also the paving over of soils, with implications for runoff rates and flood levels on lower ground. Hence, all industrial societies are replete with embedded energy: obvious in the Chicago coal yards but less visible in the suit from Jermyn Street.

As in the previous energy era, opportunities for pleasure were increased. The form of travel now called tourism was made possible by the steamship and the railroad: before the 19th century, a pleasure visit from western Europe to Egypt was too great an

undertaking for almost everybody except the aristocracy. After about 1860, Britons might holiday in Switzerland, beginning a love-affair with the mountains denied them in their own small-scale islands, one which took them to the summit of Everest in 1953. (Actually it was a New Zealander and a Nepalese who first reached it, but imperial attitudes were still strong. The relevance of industrialization is in the specialized gear now used, and the fact that an expedition has had to be mounted to clear up the "litter" from such climbers including some bodies.) The ability to travel to relatively remote places led to conflicts over environmental change: the U.S. National Parks post-1945 slogan of "parks are for people" led to developments that people of later worldviews have found difficult to demolish, in all senses of the word. In a secondary frame, urban electricity made possible all kinds of entertainment, which may well help to inculcate environmental attitudes, whether these be of the exploitive or the conservative kind. Think of oil company advertisements on television or of documentaries on whales.

Only recently has the application of industrial energy to warfare been examined in an environmental context. At one level, there is the way in which calorie production must be stepped up, for soldiers need feeding well, as do the industrial workers who make their munitions. Horses, too, consume cereals and were important as late as 1945. So many grasslands are ploughed up to make way for wheat, oats, and rye crops, with downstream aftermaths for wildflower populations and soil losses. The opposite happens when relict landmines prevent land being used for agriculture: a World War II problem in Poland to this day. But the main impact of industrial warfare before 1950 was seen on the Western Front in World War I. This was in effect a huge concentration of energy use in the form of soldiers, horses, shells, and bullets. The environmental alterations were profound and almost total, as can be sensed from the paintings of John Nash, for example. Yet the land was quite quickly retransformed to agricultural use, and only a few consciously preserved

remnants of trench systems are left, though the whole of them can be seen from the air at the end of summer like an archaeological crop mark. The less static warfare of 1939 through 1945 left fewer long-term environmental changes in Europe but transformed the vegetation, soils, and indeed landforms of some Pacific islands.

If some summative expression of industrialization and environment is sought, then two phenomena stand out. The first is that enough carbon was emitted into the global atmosphere to start the rise in concentration of (especially) carbon dioxide, which underpins today's concerns about climatic change. The mid-18c is usually taken as the base-level of 280 ppm, to compare with recent values of 360 ppm. The second is the explosion of the city, parasitic on the rest of the world for energy in all forms and for oxygen. The carbon loading has been mostly caused by the rich, but the city's ecology is shared by LIE and HIE alike.

3.4 Post-Industrial Economies (post-1950 CE)

After 1950, most of the indicators of environmental change start to rise in magnitude. Graphs such as those in B.L. Turner's 1990 compilation show sharp upward trends in resource use, waste production, emissions to air and water, and land use transformations. The increases are seen in Table II, where some shift toward an increased proportion being devoted simply to movement can be detected. Most of these are outworkings of increased use of oil, whose ascendancy over coal in this period is unchallenged. So post-industrial economies are in large measure the old fossil-fuel-powered systems writ large in many ways (not least in the geopolitics of oil) but in some of its energy-environment relationships there is something different about the post-1950 world.

For once it is possible to date the genesis of the new world quite accurately: 2.30 p.m. on December 2, 1942, in Chicago, where Enrico Fermi conducted the

TABLE II
Energy Use in Industrial and Postindustrial Societies

Society	Total energy use (10 ³ kcal cap day)	Plant and animal (%)	Home and commerce (%)	Industry and agriculture (%)	Transportation (%)
Postindustrial	230	4.3	28.6	39.5	27.3
Industrial	77	9.0	41.5	31	18

first controlled fission of atomic nuclei and thus made possible both military and civilian nuclear power. (The energy content of pure ^{235}U is 5.8×10^5 MJ/kg, compared with oil at 43 MJ/kg.) The military applications have had some environmental effects, most notably in the fallout of radioactive material from atmospheric weapons testing before the ban in 1963. Levels of long-lived radionuclides (such as ^{137}Cs and ^{90}Sr) built up in high latitudes and have accumulated in the bodies of, for example, the Saami people. The potency of thermonuclear weapons was such that in the 1970s scientists constructed scenarios for a 5000 MT exchange that would produce at least 12 months of virtual darkness in the Northern Hemisphere ("nuclear winter"). This knowledge is sometimes credited with helping to bring about the end of the Cold War. The potency of radionuclides to affect human tissues on all timescales from the immediate to the 240,000-year half-life of ^{240}Pu has accompanied the development of civilian nuclear power where very strict international agreements control the discharges to air and water. In spite of these, nations like Norway and Ireland object vociferously to the marine releases of the UK plant at Sellafield on the coast of Cumbria. Of greater concern, however, is the possibility of accidents resulting in the meltdown of the core of a reactor, as has happened at Sellafield in 1957, at Three Mile Island (United States) in 1979 and at Chernobyl (Ukraine) in 1986. Only the latter was truly disastrous environmentally and to human health but the memory is very vivid all over Europe. Yet of even more extensive importance is the fate of high-level radioactive wastes. Nobody yet knows really what to do with them since they must be sequestered from the environment for about one-quarter million years, a commitment rather like that of the Netherlands in keeping out the sea. One further long-term consequence of this era has been the formulation of chemicals with long-term toxicities, which can accumulate in food chains. The organochlorine pesticides and dioxin are the best known examples with the former causing high avian mortality and the latter implicated in teratogenic fetuses after its distribution in Vietnam as a contaminant of herbicides used in the war of liberation in the 1960s and 1970s.

The adoption of nuclear power in HIEs has mostly been motivated by a fear that fossil fuels might run out, either physically, economically, or politically. The HIE lifestyle is now dependent on relatively cheap and totally constant energy supplies: to interrupt gasoline flow at the pumps for a couple of days is virtually to bring about insurrection, whereas

a coal miners' strike for 3 months has virtually no effect. In a similar vein, it takes little imagination to fashion scenarios of a western world without constant electricity. Since capital formation is now a matter of colored numbers on screens, then all kinds of projects to alter the environment might come to a standstill, and the standing of the small screen as a major source of authority in western society (especially if it shows a man (*sic*) in a white coat) would be greatly diminished—no more documentaries about global warming for a while at least.

Unlike the fossil fuel era, it is difficult to designate any particular piece of technology with direct environmental effects as characteristic only of the post-1950 period. In more ways, there has been more indirect effects through capital accumulation in ways that have spun off into environmental projects like the very large dams of the late 20th century planned for China, India, and possibly Turkey. One of the diagnostic features of the post-industrial society, however, is its appellation as a knowledge economy, and there is no quibbling with the fact that the environment is a big absorber of space on hard disks the world over. Cheap and ubiquitous energy makes possible the gathering of environmental data, its storage, and even its transmission in windowless, air-conditioned classrooms.

The present is always impossible to assess in historically oriented accounts and it may well be that the post-1950 phases is not really a new era but simply an intensification of the one that started in 1750. It could equally be the start of a transition to something totally new, in which case the energy-environment linkages would be a central feature along with one of their outgrowths already in place: the creation of a truly global network of instant communication (a noosphere in fact, as envisaged by Vernadsky in 1945) such as there has never been.

4. OVERVIEW FROM TODAY'S PERSPECTIVES

In the perspective of the last 10 ky, access to energy can be seen as having produced tensions in a world where access is highly differential (Table III). In a positive view, many developments have been facilitated, without which today's humans would not even be able to aspire to the cultural richness and diversity that have evolved. By contrast, access to energy has allowed impacts that if foreseen by some omniscient world power imbued with the precautionary principle would likely have been prohibited.

TABLE III
Energy Consumption at the End of the 20th Century, GJ Cap Year

Region	Commercial energy	Biofuels
World	60	3.7
N & C America	243	2.0
Europe	134	1.1
Africa	12	7.2

Looked at positively, access to energy has been one consequence of agriculture, which has brought about the possibility of adequate nutrition even for human populations of more than 6 billion: if they do not get it, then the causes are political, not to do with energy capture. In the wake of the shifts to more intensive sources of food than offered by hunting and gathering, a quite wonderful diversity of human cultures came about. (It is interesting that the more up-market the tourism, the further back along this spectrum people want to go, albeit only on vacation.) There are still elites who think that imperialism was a positive outcome of energy made available by gunpowder, the breech-loading rifle, the steamship, and by their latter-day equivalents. Travel has been one further outreach into which energy can be channeled. For humans, there is tourism and space travel (think of the rocket as an updated blowpipe), for instruments gathering data there is the satellite and the remote weather station on a buoy in the Pacific. To sum it all up, there is the city and its urbanized fringes, where the CBD stacks the decision makers in high towers (always a visible emblem of power, i.e., of personal access to energy: recollect as well the personal jet and the LDC leaders' Mercedes) and the suburbs spread out the expenditure of energy and money.

But change inevitably brings loss. Most of these have been irreversible, like the diminution of biodiversity caused by habitat change as, for example, agricultural intensification works toward monoculture or the demand for wood products devastates forests. In the period 1700–1980, forests and woodlands decreased globally by about 19%, grassland and pasture by about 1%, but croplands increased by 466%. Industrialized warfare is better at destruction of all kinds, including life in all its forms, than a few armies raging seasonally across selected terrain. The chemical industry has produced many quality-enhancing pharmaceuticals but has also thoughtlessly allowed out persistent substances that have engendered problems for wild species,

crops, and human health alike. One effect has also been to introduce uncertainty: models of climate change are in the end models and can be disregarded on those grounds by those wanting to emulate the moa rather than the ostrich. Though the multiplicity of environmental changes that are problematic for somebody or something is bewildering, the main two for humans can probably be summed up as meat and motors. The energy going into both produces all kinds of patterns that could be rearranged to the greater good of many more inhabitants of the planet.

If, though, there is only one allowable view of the history of energy and environment, it must be that access to the energy flows and stores of the planet, especially after the 18th century, has allowed the growth of the human population—from about 720 million in 1750 to 6.1 billion of the early 21st century. Whether regarded historically, culturally, technologically, or politically, that view fills the lens.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Energy, Overview • Development and Energy, Overview • Ecological Risk Assessment Applied to Energy Development • Energy Efficiency and Climate Change • Environmental Injustices of Energy Facilities • Industrial Agriculture, Energy Flows in • Sustainable Development: Basic Concepts and Application to Energy • Technology Innovation and Energy • Transitions in Energy Use

Further Reading

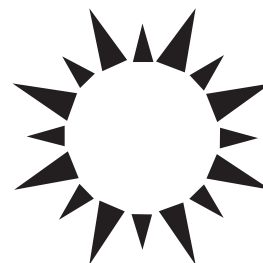
- Bormann, F. H. (1976). An inseparable linkage: Conservation of natural ecosystems and the conservation of fossil energy. *BioScience* 26, 754–760.
- Cottrell, F. (1955). “Energy and Society.” Greenwood Press, Westport, CN.
- Debeir, J. C., Deléage, J.-P., and Hémery, D. (1991). “In the Servitude of Power. Energy and Civilization through the Ages.” Zed Books, London.
- Hall, C. A. S., Cleveland, C. J., and Kaufmann, R. K. (1986). “Energy and Resource Quality. The Ecology of the Resource Process.” Wiley-Interscience, New York and Chichester.
- McNeill, J. R. (1991). “Something New Under the Sun: An Environmental History of the Twentieth-Century World.” Allen Lane, New York and London.
- Simmons, I. G. (1996). “Changing the Face of the Earth: Culture, Environment, History.” 2nd ed. Blackwell, Cambridge, MA, and Oxford.

- Smil, V. (1991). "General Energetics: Energy in the Biosphere and Civilization." Wiley-Interscience, New York and Chichester.
- Smil, V. (1993). "Global Ecology: Environmental Change and Social Flexibility." Routledge, London, and New York.
- Turner, B. L. (ed.). (1990). "The Earth as Transformed by Human Action." Cambridge University Press, Cambridge and New York.
- Vernadsky, V. I. (1945). The biosphere and noosphere. *Am. Scientist* 33, 1–12.

Environmental Gradients and Energy

CHARLES A. S. HALL

State University of New York
Syracuse, New York, United States



-
1. Introduction
 2. An Integrative Approach Based on Bioenergetics and Five Important Principles about Energy and Ecology
 3. Methodology for Measuring Energy Balance within N Gradients
 4. Energy Balance and Population Dynamics
 5. Summary

Glossary

abundance Number of a species found per area.

distribution Range of occurrence, as in brook trout are distributed in the Appalachian Mountains from Maine to Georgia.

gradient A gradual change in some condition from low to high. For example, there is a gradient in temperatures in the Luquillo Mountains of Puerto Rico from 22° at the top of the mountain to 30° at the bottom.

opportunity cost When an organism uses some of its surplus energy for one activity that energy is no longer available for some other activity.

population dynamics Changes in species' population abundance over time.

succession More or less orderly replacement of one species by another over time in an ecosystem following the relaxation of a disturbance.

The examination of an organism's energy costs and gains along environmental gradients, such as temperatures varying over a terrain, can help establish a central issue for all levels of ecology, which is what determines the distribution and abundance of species and populations.

1. INTRODUCTION

Population ecology is concerned with the distribution of individuals of a particular species over space and time, community ecology examines the distribution of species congregations over space and time, and ecosystem science considers materials transfer and other processes which in turn influences the collective distribution and abundance of organisms within defined environments over time. These issues are equally important to plant and animal ecologists, as well as to those who study behavior and other interactions.

1.1 Observed Patterns in the Distribution and Abundance of Biota

How are species distributed over space and time? There are several important generalities. Usually individuals of a species are more abundant near the middle of their geographic range and become progressively rarer toward the edges, a pattern that approaches a normal curve, or sometimes a skewed normal curve. This generality relates to the assumption that distribution and abundance of a species reflects the probability density distribution of the species' required environmental resources and biotic interactions. A corollary is that populations of species are often found in a greater variety of habitat types near the middle of their geographic range compared to the edges. A second major pattern is that in any given environment some species are relatively abundant, but most are relatively rare. Furthermore, species that are abundant in the center of their range tend to have widespread distributions, and the rarer species may be rare everywhere. Relative abundance also is influenced by differential survival and reproduction, although factors that

cause those differences are often not readily apparent and are rarely quantified.

A third important generality is that species or population abundances usually do not follow repeatable patterns over long time periods in a given environment, even though the number of species may persist. While examples of relatively constant populations certainly exist, such as *Spartina* grass in many coastal environments and brown trout in small streams, most populations that have been studied over long time periods seem to fluctuate widely in abundance, even when plotted on a log scale. The dynamics of populations tend to be, or at least appear, stochastic and should not be described or modeled by a deterministic or logistic equation. In most environments, populations are almost continually reduced by random environmental disturbances of the many abiotic (e.g., temperature, moisture, light, nutrients) and biotic (e.g., predation, competition, parasitism, dispersal) factors that influence population dynamics.

1.2 The Need for an Integrative Approach

Many investigators have remarked on the general lack of an integrative approach to assessing the many potentially important factors effecting organismal distribution and abundance, as well as the need for more comprehensive approaches of understanding the impact of more than one resource or factor. In fact, many studies have shown that all factors examined were clearly important in determining the distribution and abundance of the populations in question. However, the difficulties and costs associated with measurement of multiple factors in ecological studies are great, especially over the long time periods required to demonstrate interannual variation, which may be essential to inferences. The lack of multifactor empiricism in ecology is also rooted in the lack of a general, unifying theory that (1) integrates the myriad of biotic and abiotic factors that influence species distributions and abundances and (2) provides a mechanism for explaining how organisms or species identify resources and evaluate (or trade off) the costs associated with either obtaining them or adapting to alternatives.

1.3 Multiple Factors and Tradeoffs

Maguire examined the growth response of a phytoplankton as a function of two (or, in theory, more)

environmental gradients (e.g., variability in temperature and a nutrient). Maguire then examined how these resources changed over time as the season progressed and temperatures warmed while nutrients were reduced. By integrating the environmental changes in relation to concomittant changes in growth of that species over the year, or between years, he was able to show how abundances and distributions changed as environmental conditions waxed and waned. This seems substantially superior to assuming a constant rate of growth and a fixed carrying capacity, as is the case in the more generally used logistic models.

Many other studies have identified the tradeoffs associated with different resources or adaptations in relation to biotic distributions and abundances. Some very good examples pertain to plant life history strategies, where investments in chemical defense versus growth, fixation of carbon versus nitrogen, efficiency of water use versus carbon uptake, or growth rate versus structural integrity have been quantified. A particularly interesting example is one pertaining to nitrogen concentration of leaves. There is a remarkably linear relation between leaf nitrogen concentration and growth rates of plants. Aspen (*Populus tremuloides*) and other plants with high nitrogen concentrations have high growth rates, probably due to the presence of high levels of carbon fixation enzymes. However, leaves high in nitrogen are also more desirable food for many heterotrophs, from microbes to moose. But this means that high growth rates may also be associated with high loss rates through herbivory, and high rates of herbivory can induce plants to produce chemical defenses. Our point is that approaches emphasizing tradeoffs are important when trying to understand biotic distributions and population dynamics because essentially every resource can be obtained only at the expense of other resources. Smith and Huston proposed a general theory linking vegetation patterns to species-specific physiological and life history traits in relation to energetic constraints on the capture and use of resources. Their approach emphasized fundamental tradeoffs facing plants, such as CO₂ uptake versus water retention (at the level of stomates as well as the whole plant), root versus shoot allocations in relation to the costs of acquiring water or nutrients versus sunlight, and allocation of photosynthate to leaves versus the whole plant. Plants are very plastic with respect to these tradeoffs, both among and within species, by virtue of adaptations.

Any holistic scheme that attempts to provide a comprehensive approach must satisfy at least three

criteria. First, it must analyze, and give procedures for assessing, the tradeoffs that exist among more than two or three factors. Second, explicit explanations must be given as to why these tradeoffs should exist or of the mechanisms by which they operate relative to the observed distribution and abundance of a population or species. Third, they must start from first principles, an approach that tends to work well in other disciplines. Energy is the common currency of ecology and is necessary for any biophysical process to occur. Thus, energy should be fundamental in the development of an ecological theory that attempts to explain the distribution and abundance of animals and plants in relation to the myriad of ecological factors that impinge on all populations.

2. AN INTEGRATIVE APPROACH BASED ON BIOENERGETICS AND FIVE IMPORTANT PRINCIPLES ABOUT ENERGY AND ECOLOGY

The essence of the approach developed here is to integrate energy costs and gains along all appropriate environmental gradients to yield net energy gain in the form of biomass elaboration and reproduction. In this way, the observed distribution and abundance relationship may be explained in terms of the net energy balance or budget of the organism, population, or species. Our central tenet is that a sufficient, positive energy balance must be sustained through time and space such that a significant number of individuals reproduce effectively. If this does not occur, then eventually the population and ultimately the species will decline toward extinction or, through the process of natural selection, adapt some other energy strategy that will permit survival.

It is necessary first to identify five important principles about energy and ecology. First, as identified in many review articles, energy contributes to survival and reproduction of organisms in accordance with our contemporary understanding of basic evolutionary. We believe that the most appropriate way to express this relationship is energy per time, or power, as developed in many publications by Howard Odum and also others. In the words of Ware, "Optimal life history is concerned with power allocation." Energy is vital to organisms in that, other things being equal, the more energy an organism can sequester per unit time, the larger its energy reserves will be to help meet stress and

contingencies. Therefore, if an organism can maximize energy intake, it is more likely to reach age of reproduction rapidly and maximize the number of propagules it can propel into the future to perpetuate its genome. Orians and Solbrig summarize this view: "a plant capable of increasing photosynthesis within the constraints of its available resources and the physical environment should gain advantages in competition with other plants, defenses against herbivores, and should have more energy to devote to reproduction."

Second, an examination of an organism's ongoing energy balance allows a less circular definition of fitness than is possible through the more commonly used criteria of reproduction itself. Although at various times in an organism's life its ability to accumulate energy can be satiated, in nature this is very rare. This is especially true with respect to high-quality food items, whose availability is usually limited.

Third, energy balance is important from the perspective of how much energy an organism must divert from growth and reproduction to overcoming contingencies such as adverse environmental conditions. In general, organisms reproduce only from the energy surplus after meeting other contingencies in their life, including growth to a reproductive size. However, when conditions become so extreme that the organism itself has a high probability of death, all available resources often are put into reproduction.

Fourth, the importance of energy as it effects organisms encompasses far more than just the environmentally available or stored (i.e., trophic or elaborated) energy that is usually emphasized. It includes, among other things, metabolic energy drains imposed on organisms by nonoptimal environmental conditions or added stresses, thermal energy gained or lost by behavioral modifications and biotic interactions as discussed later. Most important, however, it includes the energy costs of respirational requirements as well as energy required for an organism to live in and to exploit particular environments, especially suboptimal environments. For example, a rainbow trout is isotonic to water of about one-quarter the salt content of seawater. In fresh or saltwater, that fish must use at least 30% more energy for basal metabolism simply for osmotic adjustments. This is probably one reason that estuaries are such important fish nurseries. Fish can grow better in environments where their basal metabolism is lower. Fifth, in order to grow and ultimately reproduce, organisms require resources and must invest stored energy reserves to get them.

Moreover, all resources vary in quantity and quality. The most important point about energy and its relation to other resources is not that everything can be reduced to energy (which is false) but rather that every material (and most nonmaterial) resource has an associated acquisition energy cost. Most resources become limiting in part because their acquisition may require too great an energy investment on the part of the organism. Thus, high-quality resources are often those that require less energy expenditure per unit of resource obtained. However, with sufficient energy reserves and as constrained by its genome, an organism can divert energy to the acquisition or mobilization of essentially whatever other resources are in short supply, even when their quality is low. Likewise, quality also reflects the value of the energy per kcal of heat content so that, for example, a kcal of food has more utility to an animal than a kcal of incident solar energy and food with a high nitrogen content may be much more useful. We define high-quality resources as those of highest utility that are available to the organism with relatively low energy investment.

Bioenergetics can provide a comprehensive and explicitly testable set of mechanisms for predicting where species will be found and for determining why they are not found in other locations, although it cannot account for conveyance of propagules to new sites. While it is true that organisms need resources other than energy—for example, various nutrients—these other resources can be examined from the perspective of their acquisition energy cost.

2.1 Net Energy Balance: The Elusive Integrator

Each factor that impinges on an organism has energy costs and, in many cases, potential energy gains. The life history strategy (phenology) of an organism is the outward expression of the genetic codification of past successful trials of an organism's energy return on investment. At each point in an organism's life it must invest energy for obtaining more energy while decreasing energy drains and converting tradeoffs about which contingency is most deserving of whatever energy reserves might be available. Tradeoffs occur because most energy gains are obtained only by increasing energy costs—that is, by increasing energy investments. Energy investments by plants to grow leaves versus roots is one obvious example. Another is the energy investment by animals in motility for the purpose of optimizing energy intake from food resources.

Life is a continuous balance of energy investments, each with potential gains but also real and potential costs. As one simple example, plants have energetic tradeoffs whereby the advantage gained by early reproduction tends to be at the cost of future reproduction. For another example, longer-lived but slower growing trees often put large energy investments into energy-intensive structural fibers that generally allow them to survive storms and other contingencies, but at the cost of reduced growth and reproduction. Similarly large, long-lived trees that are easy for insects to find tend to be more heavily defended with secondary compounds that operate to discourage insect feeding. Generally, plants that are exposed to high levels of infestations increase their investments in secondary compounds in subsequent years. In summary, this energy balance perspective derives from the fundamental energetics ideas of earlier investigators as detailed at the end of this entry and involves examination and integration of all pertinent energy costs and gains throughout the life history of the organism. Populations (or species) may be examined in the same context, if one accepts that the net energy balance of the population results from the cumulative energy balance of individuals within the population (or species). This gives a less circular and more mechanistic set of tools to assess tradeoffs facing an organism or population, and ultimately the relative success of its genome, than by using fitness (potential reproductive output) as a direct function of multiple environmental gradients. Ideally both approaches would give similar answers to the same question.

2.2 Energy Costs and Gains in Relation to Environmental Gradients

As noted, population densities tend to be higher in the middle of the species' geographic range, and we can extend that idea to say it is because physical, chemical, and biotic resource gradients are optimal there, which allow maximal energy profits. For plants, light, water, and nutrients are clearly important in this context, and for animals, food derived from photosynthate and (generally) oxygen are essential. For both, temperature, often considered the master regulator, is critical. The concept and data of gradient analysis, as developed initially by Ramensky, Gleason, and Whittaker, demonstrate this fundamental ecological pattern. In terms of our energy balance approach, individuals of a species will be distributed over space and time in a way that may

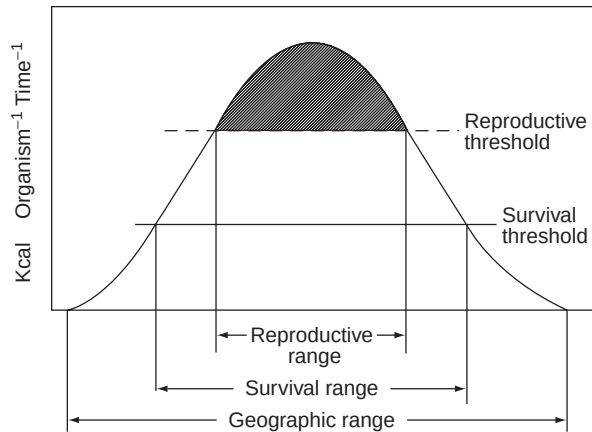


FIGURE 1 Net energy balance of a species over its geographic range. The horizontal solid line represents zero net energy gain for the species (survival threshold)—that is, conditions where energy gain is just sufficient to compensate for basal metabolic energy, plus the energy required to capture food and overcome stress. Below that line, organisms can exist only by consuming their own energy resources. The broken line indicates the larger amount of energy intake needed to generate sufficient energy surplus for reproduction (reproductive threshold). The hatched area of the curve is the only part of the range in which the species can exist in the long term. From Hall, C. A. S., Stanford, J., and Hauer, R. (1992). The distribution and abundance of organisms as a consequence of energy balances along multiple environmental gradients. *Oikos* 65, 377–390. Used with permission.

be idealized as a more or less normal curve representing energy return on investment along a gradient (Fig. 1). Two important points derive from Fig. 1. First, individuals near the center of the collective distribution enjoy a high energy profit, which allows a large net gain of energy, growth, and either numerous offspring per effort or frequent reproduction. Individuals on the margins of their physiological, territorial, or nutritional range with respect to the gradient are unable to generate a sufficient energy profit to produce population growth. Competition may also be more important at the margins of the range because it raises the cost of resource acquisition and creates more of an impact on individual energy gain in locations on the gradient where the net gain is already small. Hence, if the gradient is favorable and temporally stable, the species will be abundant and widespread, which is consistent with the first generality of observed patterns of distribution discussed earlier. The second point is that rare species are those whose particular morphologies and behaviors do not allow them to be abundant anywhere because they never make a sufficient energy profit. This is consistent with the second generality of observed patterns of distribution discussed previously.

This approach emphasizes that all abiotic variables that may be important for bioproduction of any species may be examined in the context of the energy costs and gains to that species. This is regardless of whether resource limitation is examined in the context of availability of the resource or the ability of the species to obtain the resource. For example, at very low nutrient concentrations, such as the 1 $\mu\text{g/l}$ of bioavailable phosphorus that commonly occurs in oligotrophic lakes (where phytoplankton production is often phosphorus limited), a 10 μm plankton cell is surrounded by roughly 100 atoms of phosphorus in the 10 μl of water adjacent to the cell. Even though it may seem that phosphorus atoms are relatively abundant, bioavailability of phosphorus atoms is more influenced by the large concentration gradient that exists between the inside and the outside of the cell membrane. Uptake of phosphorus by the cell to mediate growth requires energy expenditure (active transport). Therefore, nutrient uptake at low concentrations is influenced critically by light. A low concentration of nutrients acts as an energy drain to the phytoplankton, which in turn reduces the reproductive output of the plant and contributes to their low concentrations found where nutrients are low.

Finally, various biotic interactions may also be considered as gradients in our energy balance context. For example, natural selection operates to optimize the risks of predation versus the benefits of exposure to predators in such a way that net energy gains are optimized. Thus, it is possible to view predation energetically from a “Lloyd’s of London view.” In the old days of sailing ships, Lloyd’s knew that a certain proportion of the ships they insured would founder in any given year; but the company management also knew that if no ships left port, no profits would be forthcoming to the shipping industry (or to Lloyd’s). On the other hand, captains who loaded too much sail trying to maximize profits usually foundered (or otherwise lost more than they gained) and were selected against. As an analogous example, Gilliam and Fraser found that the risk taken by minnows could be manipulated by changing the energetic return of a foray into predator-infested waters, and that the behavior of minnows in accepting risk was predictable from an energetic model of energy gains and costs. Likewise, it costs energy to avoid predation, including both the direct costs of avoiding food rich areas that contain predators or of manufacturing defenses; there are also the indirect costs of decreased motive efficiency. Using similar reasoning, competition (both inter- and

intraspecific) usually reduces the availability of a resource for one or both of the competitors, and hence may reduce the range or move the optimum location of an organism along the resource gradient. Hence, competition for resources may reduce the net energy profits, unless the organism is a very good competitor, in which case net gains may indeed be enhanced.

2.3 Some Examples of Energy Tradeoffs in Response to Resource Gradients

An important aspect of any analysis of energy balance in relation to resource gradients is that there are opportunity costs to being a particular species or genome. This simply refers to the fact that energy investment in life history strategies that produce prosperity at one place along an environmental gradient (opportunity costs) precludes adaptation to other conditions. Thus, inherent in the concept of phenological energy balance in response to an environmental gradient is the concept of tradeoffs. At any given time, the energy available to adjust to contingencies or to reproduce effectively usually is limited and must be invested wisely.

Often the relations defining costs and gains (tradeoffs) along a gradient are nonlinear. Such nonlinearities, at least in principle, can produce many of the classic hump-shaped growth curves along a gradient found in the literature. In the case of the growth of a particular aquatic insect or fish as a function of temperature, foraging rates increase, often linearly at least up to a point. But respiration increases more rapidly as temperature rises, because of its Q₁₀-type response (Fig. 2A). More complicated relations can exist, especially at the extremes of the gradient, but even so the maximum net gain is normally near the middle of the gradient, even for homeotherms including humans. The net energy return to an organism is the difference between energy gains and energy costs. Thus, the observed growth curve versus temperature relation can be explained as the difference between energy gains and costs along a temperature gradient (Fig. 2). A clear tradeoff exists between temperature and growth for many organisms. Vannote and Sweeney elegantly showed how this idea applied to distribution and fecundity of different species of mayflies.

Bunce and his colleagues measured photosynthesis and respiration of eight plant species that had different distributions over the elevational gradient of moisture on a mountain in Arizona and calculated

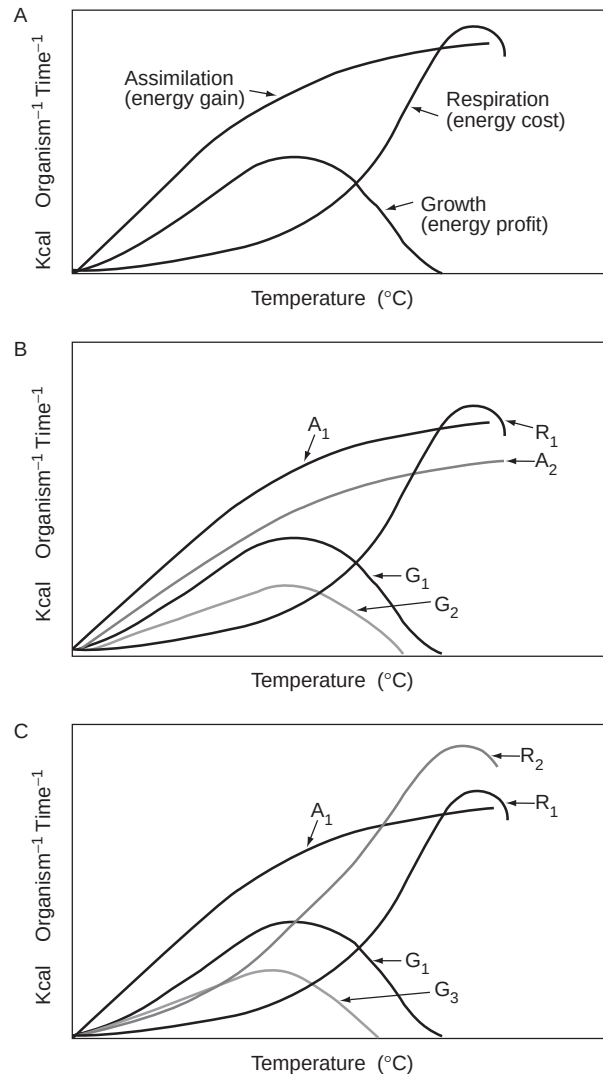


FIGURE 2 (A) Growth of a poikilotherm in response to an environmental gradient (in this case temperature). We hypothesize that net energy profit, which equals growth plus reproduction, is the result of changing energy costs and gains along the gradient. The commonly observed bell-shaped growth curve is the net difference between metabolic energy expenditures (respiration) and gains (assimilation), realized as individual growth (energy profit). (B) Factors that reduce energy assimilation (curve A₁ changes to curve A₂) will result in a decrease in individual growth rate (G₁ to G₂). Decreased quantity or quality of food assimilated may be the result of abiotic factors (e.g., nutrient availability, disturbance) or biotic factors (e.g., competition or predation pressure). (C) Factors that result in increased respiration (R₁ increases to R₂) will also decrease individual growth rate (G₁ to G₂). Increased respiration may be the result of changing abiotic conditions (e.g., caused by disturbance) or manifestation of biotic interactions (e.g., increased search time for food due to competition, increased activity caused by predation pressure). From Hall, C. A. S., Stanford, J., and Hauer, R. (1992). *Oikos* 65, 377–390. Used with permission.

the carbon balance of each species as a function of moisture and hence elevation. They found that the lower limit of each of the species was approximately 200 m above the zero carbon balance elevation predicted from their field measurements. In other words, each species would not be able to make a significant carbon, and hence energy, profit very much beyond the measured limits of their range. The actual mechanisms for drought tolerance or avoidance were complex, but the net effect was well predicted by their model. Thus, tradeoffs along the moisture gradient were apparent between species. Some had mechanisms, such as a relatively large area of open stomates, favoring high carbon gain caused by high rates of photosynthesis, but they were also prone to high water losses. Other species were more conserving of water but had lower carbon gains.

Trout in a river are principally drift feeders. If they are able to maintain position in rapidly flowing water, the current brings more food to them, but the energy costs of swimming rapidly to maintain position may be too great to maintain a positive energy balance in the long term. Bachman developed a general theory of the energy costs and gains of trout feeding in waters of different velocities. Dominant or fastest-growing fish were found approximately where the difference between the costs and gains were greatest. Thus, dominant trout grew faster where the current was optimal relative to food delivery (energy gains), swimming speed, and duration (energy costs). Young steelhead trout were also distributed along the velocity gradient of a stream approximately as predicted from this model. Brook trout change the area they defend, and their antagonistic displays, as a function of changes in water velocity in a way that maximized the energy efficiency with which they defended their territories. The same kind of result held for pelagic fish and their swimming speeds. Young bleak (a British minnow) voluntarily swam at velocities (107 mm s^{-1}), similar to velocities that theoretically optimized their net energy gains (111 m s^{-1}). In a more general sense, organisms that defend larger territories get more food for their young but must expend more energy in that defense, resulting in a bell-shaped net return curve.

2.4 Molecular and Cellular Tradeoffs

While it is not difficult to find organisms responding to gradients, there is not a well-developed or recognized procedure for *why* organisms should do well at certain parts of gradients and not well at others, or for why this should affect energy gains

and costs. There are probably many reasons, depending on the gradient considered, but an important point is that many enzymes controlling growth, maturation, and reproduction tend to have great specificity. Moreover, this specificity is greatly influenced by the environment, especially temperature. Enzymes exhibit thermal optima that correspond closely to body temperatures in the case of homeotherms or to the preferred habitat temperatures in poikilotherms. Organisms tend to seek temperatures where their particular enzyme variants are most effective; therefore, the distributions of closely related species may be explained by variable enzymatic responses within an environmental gradient. At temperatures below the thermal optima, enzymes function slowly. This is due partially to low levels of kinetic energy and partially to the constraints on flexibility because of the specificity of temperature-mediated activities, such as folding, at the active site. Above the thermal optima, the weak chemical bonds that are responsible for formation of transition state structures, and ultimately catalysis, are weakened and break. Various physiological mechanisms have evolved to broaden the functional range of temperatures and lower Q_{10} values. Clearly, however, there are limits to the functional range of a given protein molecule. In addition, the number of isozymes with different thermal optima available to an organism necessarily is limited to the number of alleles for that enzyme, a maximum of two in the case of diploids (assuming no genetic duplication). There are exceptions. Rainbow trout (*Oncorhynchus mykiss*), a fish with a relatively broad thermal tolerance among salmonids, are tetraploid. As a consequence they have twice as many alleles to generate enzymes as other trouts. Presumably this allows various races of the rainbow to be adapted more specifically to warm or cool water.

It is probably efficient for organisms to have enzymes with a relatively narrow temperature range for metabolic activity, because enzymes that operate well at a given temperature appear to operate less well at other temperatures. This has been demonstrated empirically for different subpopulations of the killifish (*Fundulus heteroclitus*). Each subpopulation of the fish had an optimal operating temperature where metabolic efficiency was maximized. Moreover, an associated optimal temperature for maximum kinetic performance of lactate dehydrogenase B was observed, which related to swimming efficiency and hatching success. Such tradeoffs along gradients are also important for the structural

components of cells. Membrane lipids, for example, are subject to functional alteration at different temperatures.

2.5 Extension to N Gradients

Not all the gradients affect an organism's welfare equally, because some, such as temperature and food quantity and quality, have a large effect on metabolism whereas others, such as trace metals, often have minimal effect. But the cumulative effects will be reflected in the organism's integrated net energy balance in relation to each relevant gradient. For example, a tree species may grow progressively well or poorly from the bottom of a mountain to the top, due to differences in temperature. Likewise, it may do well or poorly as a function of sunlight or moisture gradients from one side of the mountain to the other. Shelford introduced and Hutchinson developed the idea that the realized niche of an organism could be determined empirically by plotting the boundaries of that organism's existence as a function of the N factors that influenced the distribution of the organism. The resulting N -dimensional hyperspace is the definition of the range (and sometimes niche) of that organism. A species is the aggregate genetic code, and its phenotype is the organism that can generate a sufficient energy profit to reproduce at some place along each possible environmental gradient. Since it apparently does not pay to exist along the entirety of each environmental gradient, natural selection of each species is focused on the part of each gradient that allows a sufficient profit for the species to survive and reproduce. This may be a strong factor favoring speciation. Widespread species are those that have evolved mechanisms for making significant energy profits within a large portion of the hyperspace created by all gradients that impinge on the phenology of the species.

Hutchinson's concept of n -dimensional hyperspace can be extended such that all relevant resource gradients, and the energy costs and gains for an organism to live along each of those gradients are integrated graphically and mathematically (Fig. 3). The well-being, growth, and eventual reproductive success of the organism is defined by the interpretation of energy costs and gains on all gradients that influence the organism's phenology. For purposes of illustration, these energy profit gradients can be envisioned as a Chinese egg set of N dimensions, with the exterior egg representing the limits of the organism's range. Where individuals

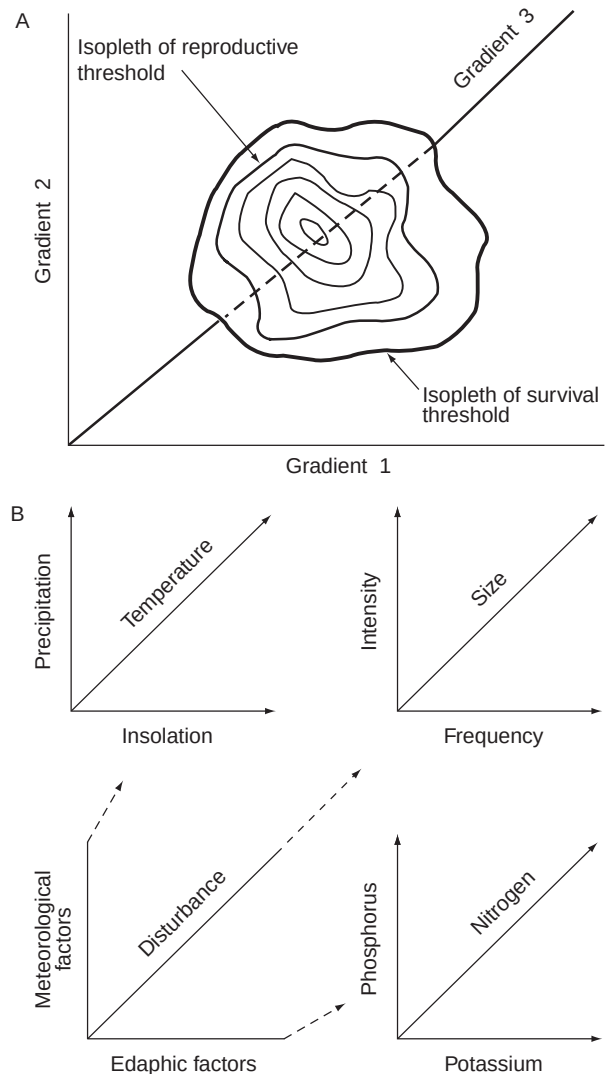


FIGURE 3 (A) Isopleths of net energy profits, defined as energy gains minus costs, of a species as a function of three environmental gradients. Net energy profits and, hence, species distribution and abundance respond to the influence of each environmental gradient involved in the energy balance of the species. Smaller energy profits therefore are shown as isopleths occurring successively outward from the point of optimum profit in relation to these three gradients. Although drawn hypothetically here in just three dimensions, in reality energy profits are influenced by many (N) environmental gradients and should be visualized in the context of an N -dimensional hyperspace (see text). (B) One way to display the effects of more than three gradients simultaneously. In this case, the figure is showing nine gradients (three each: edaphic, disturbance, and meteorological) thought to influence the distribution and growth of plants and animals in the Luquillo Mountains of Puerto Rico. Edaphic gradients include soil nutrient (e.g., N, P, K) concentrations, disturbance gradients include intensity, size, and duration of hurricanes, landslides, and tree-falls, and meteorological gradients include the relation between precipitation, temperature, and solar insolation at different altitudes in the mountains. From Hall, C. A. S., Stanford, J., and Hauer, R. (1992). *Oikos* 65, 377–390. Used with permission.

of a species range beyond where they can survive for long periods, the next interior egg represents the survival range, and so on. Reproduction takes place within some interior egg, which represents an even greater optimization within the hyperspace. Maximum reproduction takes place at a center of gravity representing simultaneously the midrange of all impinging gradients. Thus, many gradients can be represented and analyzed as in Fig. 3.

2.6 Niches, Succession, and Competition

In the context of this energy balance perspective, a Gaussian niche is the volume of all possible energy profits for a species within the hyperspace of all gradients to which the organism might be exposed. A Hutchinsonian niche is the volume of all realized energy profits for a species and is at least partly dependent on the geographical history of the species. Succession, then, may be caused in large part by the changing energy profit regimes of the environment in response to the presence or absence of various species and the changes produced on that site by various species over time. It is not clear whether species generally affect other species by any other mechanism than by changing the quality and hence energy cost of resources. For example, trees influence understory plants by shading or lowering available nutrients. The conditions favoring a positive energy balance for particular species change over successional time.

Competition often has been invoked as the principal mechanism that causes the distribution patterns of organisms, but that view seems to be increasingly called into question. The data of Whittaker, for example, show plant species distributed along elevational gradients in more or less random patterns, rather than in avoidance patterns that would be more consistent with competition. This indicates that the species are not causing each other to move "away" from one another on the gradient. Although few ecologists would argue with the existence of competition in nature, there is a very large difference between establishing its presence and attributing to this one factor the fundamental role in structuring communities. Where competition is demonstrated to be operational, it can be factored easily into the energy balance analysis promoted herein. As noted earlier, exploitive or interference competition can alter the availability of a resource and impact the energy balance of the competitors. For example, light-intolerant shrubs are aided by the

shading of other plants (exploitation) or one species may produce a substance that is toxic or stressful to another (interference). Such influences can be treated as additional gradients (axes) that impinge on the organism's energy balance. If enough axes are examined, the concept of competition may not be as relevant in explaining the displacement of different species along resource gradients. Each species simply does not have the morphological, behavioral, or, especially, physiological (biochemical) wherewithal to provide an energy balance that allows it to exist along the full length of all important gradients.

3. METHODOLOGY FOR MEASURING ENERGY BALANCE WITHIN N GRADIENTS

Species-specific responses of individuals at locations within many environmental gradients result in a complex pattern of energy costs and gains. This seemingly intractable pattern can be analyzed and predicted by using simulation modeling, if the basic input data for each gradient exist. First, an appropriate time loop is generated. Then, the environmental gradients are established mathematically according to historical or simulated patterns. Geographical specificity can be included; for example, procedures have been developed for projecting meteorological information from a single station to a montane landscape. Next, the individuals are located mathematically along each relevant environmental gradient for each time unit (month or day or hour) as a function of the species' known preferences or behavior in response to each gradient. Data on the energy costs and gains are needed for each gradient as per Fig. 2. Then, all costs and gains are summed across gradients for the time period of interest or as the various state variables (gradients) are changed.

4. ENERGY BALANCE AND POPULATION DYNAMICS

A final point is that the multifactorial energy balance approach given here should be useful in attempting to understand population dynamics. Very simply, the net energy balance of a species changes as environmental conditions change. Temporal variations are inherent in resource gradients, and phenological energy balances will likewise vary. Therefore,

populations will be dynamic in relation to net energy profits invested in offspring and the energy balance of those offspring as environmental conditions become more or less favorable over space and time (Fig. 4). This replaces the dogma that survival is the only determinant of future population levels and leads to mechanistic, explicitly testable hypotheses that can be applied to any individual or population. One good example of where the energy balance approach might be useful in producing a more realistic explanation of year-to-year variation in abundance of a species is in regard to recruitment

variability and the dynamics of commercial marine fisheries. For years, fisheries managers have focused harvest goals on density-dependent (logistic) responses in considering controls on stock-recruitment dynamics. However, in the multispecies, environmentally complex world of most exploited marine fisheries, the implications of the analysis in Fig. 4 seems to be much more appropriate based on the rationale of this paper.

5. SUMMARY

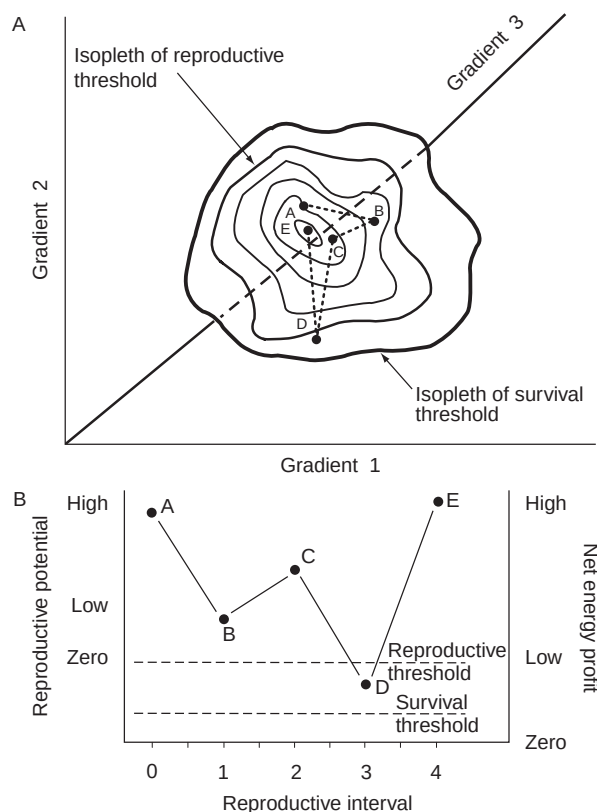


FIGURE 4 (A) Net energy balance (isopleths) of a population existing at A–E different positions within the N -(in this case three dimensional) hyperspace determined by important ecological gradients that influence growth throughout the life history. (B) Temporal reproductive potential of the population represented in Fig. 4A. In this example, the cumulative energy gains of individuals exceed losses and the population survives, although investments in reproduction are variable over time. For genomes that die after a single reproductive effort and do not have multiple cohorts (e.g., many insects), population extinction would occur at D. For organisms that live many years (e.g., most commercially important fishes), such year-to-year fluctuations in reproduction and similar changes in death rates lead to the population dynamics that we observe. From Hall, C. A. S., Stanford, J., and Hauer, R. (1992). *Oikos* 65, 377–390. Used with permission.

This article began with a discussion of generally observed patterns of biotic distributions. It concludes that these patterns can be explained as a consequence of energy balance along multiple environmental gradients due to species-specific energy costs and gains (energy balance) in response to the many (Hutchinsonian N -dimensional) environmental or resource gradients. Competition, predation, and other biotic interactions operate principally by increasing energy costs to the species and can be included in our energy balance methodology as additional environmental gradients of energy costs. Persistence of a population and, ultimately, the species in a given locality will occur only where energy return on investment allows a significant energy profit in the form of propagules. At any given time, energy investment in physiological, behavioral, or morphological adaptations that optimize abundance at one point within the N -dimensional hyperspace of environmental gradients precludes the investment of the same energy elsewhere. Within the spatial and temporal range where the species can reproduce, abundance is related to the stochastic nature of the important environmental forcing functions because energy balance is ultimately controlled by the sum of organismal responses to all environmental influences. Populations tend to be most abundant where their genome allows the optimization of energy profits. Therefore, population dynamics reflect the cumulative effect of environmentally induced shifts in organismal energy balance.

Holistic energy perspectives in the past have been complicated and controversial and often based on very theoretical considerations. But when an energy perspective is integrated with the vast bulk of appropriate physiological and ecological analysis, as done here, the energy approach can generate an elegant and powerful syntheses. The goal here is to

show that energy analysis can be used to integrate formerly disparate components of ecology using an approach that is straightforward enough to require little in the way of going beyond most traditional views in biology. If the energy balance theory presented here is robust, it has the potential of synthesizing ecology at many levels.

Acknowledgments

This article was derived from Hall, C. A. S., Stanford, J., and Hauer, R. (1992). The distribution and abundance of organisms as a consequence of energy balances along multiple environmental gradients. *Oikos* 65, 377–390.

SEE ALSO THE FOLLOWING ARTICLES

Ecological Footprints and Energy • Ecological Risk Assessment Applied to Energy Development • Ecosystem Health: Energy Indicators • Ecosystems and Energy: History and Overview • Environmental Change and Energy

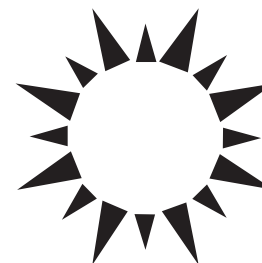
Further Reading

- Bunce, J. A., Chabot, B. F., and Miller, L. (1979). Role of annual leaf carbon balance in the distribution of plant species along an elevational gradient. *Bot. Gaz.* 140, 288–294.
- Fry, F. E. J. (1971). The effect of environmental factors on the physiology of fish. In “Fish Physiology” (W. S. Hoar and D. J. Randall, Ed.), Vol. VI, pp. 1–87. Academic Press, New York.
- Gilliam, J. F., and Fraser, D. F. (1987). Habitat selection under predation hazard: test of a model with foraging minnows. *Ecology* 68, 1856–1862.
- Hochachka, P. W., and Somero, G. N. (1984). “Biochemical adaptation.” Princeton University Press, Princeton, NJ.
- Hutchinson, G. E. (1965). The Niche: An abstractly inhabited hypervolume. In “The Ecological Theory and the Evolutionary Play,” pp. 26–78. Yale University Press, New Haven, CT.
- Krebs, C. J. (1985). “Ecology: The Experimental Analysis of Distribution and Abundance.” 3rd ed. Harper & Row, New York.
- Odum, H. T., and Pinkerton, R. C. (1955). Times speed regulator, the optimum efficiency for maximum output in physical and biological systems. *Am. Sci.* 43, 331–343.
- Shelford, V. E. (1951). Fluctuations of forest animal populations in east central Illinois. *Ecol. Monogr.* 21, 183–213.
- Vannote, R. L., and Sweeney, B. W. (1980). Geographic analysis of thermal equilibria: A conceptual model for evaluating the effect of natural and modified thermal regimes on aquatic insect communities. *Am. Nat.* 115, 667–695.
- Ware, D. M. (1982). Power and evolutionary fitness in teleosts. *Can. J. Fish. Aquat. Sci.* 39, 3–13.

Environmental Injustices of Energy Facilities

NANCY IRWIN MAXWELL

Boston University School of Public Health
Boston, Massachusetts, United States



-
1. Coal and Coke Production
 2. Petroleum Production
 3. Coal- and Oil-Fired Electric Power Plants
 4. The Nuclear Fuel Cycle
 5. Large Hydroelectric Dams
 6. Home Use of Biomass Fuels

Glossary

biomass fuels Any plant or animal matter used as fuel, including animal dung, crop residues, wood, charcoal, or peat.

respirable particulates Airborne particles with diameter less than 10 microns.

vectorborne disease A disease that affects more than one host species and that is transmitted by one host (the vector) to other host species. Many vectors are insects or arthropods; mosquitoes, for example, are the vector that transmits malaria.

working level (WL) One WL is defined as any combination of radon progeny in one liter of air that ultimately releases 1.3×10^5 MeV of potential alpha energy.

working level month (WLM) One WLM is equal to 170 WLs (that is, exposure to an average of 1 WL for a working month of 170 hours).

The energy sector relies on a broad range of facilities to extract and distribute resources, produce energy, and handle wastes. All these facilities benefit those who use the energy produced. But many facilities also create environmental pollution and have negative impacts on human health—impacts that may fall on populations who do not benefit from the energy produced. In some cases the environmental exposures or health impacts of energy facilities, like those of other industrial facilities, fall disproportionately on certain geographic regions or nations, due to the

concentration of resources or facilities in certain locations. In other situations, disproportionate burdens may fall on particular subpopulations, including disenfranchised groups such as low-socioeconomic status communities, communities or workers of color, less developed nations, women or men (male workers, for example), children or other age groups, future generations, or susceptible populations. This article describes such environmental injustices, or inequities, associated with energy facilities. It begins with energy production activities related to fossil fuels (coal and oil) and electric power generated from these fuels, and then the article considers in turn the nuclear fuel cycle, large hydroelectric dams, and home use of biomass fuels. International statistics on energy production and consumption are taken from the World Energy Council.

1. COAL AND COKE PRODUCTION

Nearly all the world's coal production—and its environmental and health consequences—is concentrated in Europe, North America, and Asia. Asia accounted for 38% of 1999 global production (China 24%, India 7%). The United States and Europe each produced 23% of the world's coal. The global distribution of coal consumption mirrors that of production rather closely.

1.1 Environmental Impacts of Underground and Surface Mining of Coal

Coal mining traditionally took place underground, with men and equipment following seams of coal deep within mountains. Since the 1960s, surface mining methods have become more common, and today they account for more than half of the coal

produced in the United States. In this approach, also known as strip mining, oversized bulldozers remove the land surface until coal is exposed, producing large amounts of waste rock—and dust—in the process.

Historically, destruction of forests often accompanied underground coal mining in the United States. The usual practice in the 1800s was to cut down all trees with commercial value on the overlying land before beginning to mine. Much of the timber was used for props in the coal mine shafts and for railroad ties as track was laid to carry away the coal.

Both underground mining and surface mining of coal cause acid mine drainage. This problem occurs when the mineral iron disulfide is exposed to water and air, leading to the formation of sulfuric acid. Large volumes of acidic liquid drain into local streams, killing fish as well as aquatic plants.

An extreme form of strip mining now being practiced in the Appalachian region of the United States is known as mountaintop removal. As the name suggests, several hundred feet are sheared off the tops of mountains that are close to one another, exposing the coal, which is removed. The waste rock is dumped into valleys so that a set of neighboring mountains in effect become a single plateau. The resulting valley fill buries streams and may become unstable as it becomes saturated with water; its impacts on local groundwater transport are unknown.

1.2 Community Burdens of Coal Mining

Coal production in the United States is highly concentrated in the Appalachian region. The leading coal-producing states, West Virginia and Kentucky, together accounted for 43% of the value of U.S. coal produced in 1997, as reported by the U.S. Economic Census. Pennsylvania and Virginia together accounted for an additional 17%, bringing the total for the Appalachian region to 61% of the value of U.S. coal production. Wyoming, Colorado, and Alabama together account for another 16%. The leading coal states, West Virginia and Kentucky, are also two of the poorest states in the United States, ranking 50th and 44th, respectively, on median household income in 1999, and 4th and 7th on the proportion of their population who lived in poverty (18% and 16%).

In Appalachian communities where mountaintop removal is under way, residents endure dynamite blasts, falling rock, and dust. Property values decline, and many residents move away. Studies of communities near ordinary surface mining in the United Kingdom indicate that particulate concentrations in

air are somewhat elevated in these areas, and children in mining communities make somewhat more visits to their doctors for respiratory symptoms such as wheezing and coughing.

On a global scale, community impacts of coal mining in less developed nations are likely to be more pronounced than in the United States and the United Kingdom, which can afford more stringent regulatory controls. For example, the study of surface mining of coal in the United Kingdom reported mean concentrations of respirable particulates (particulates less than 10 microns in diameter, known as PM_{10}) ranging from 14.4 to 22.3 $\mu\text{g}/\text{m}^3$ (micrograms particulates per cubic meter of air) in nearby communities. A study of surface mining of coal in one of India's major coal fields reported concentrations of respirable particulates of roughly 70 $\mu\text{g}/\text{m}^3$ in nearby residential areas, as well as concentrations of total suspended particulates (i.e., including larger particulates) of 200 to 500 $\mu\text{g}/\text{m}^3$ in these areas.

1.3 Impacts to Workers

Communities also bear the burden of disease that occurs among their workers. Coal miners, traditionally men, are at high risk of physical injury from machinery and from strenuous, sometimes repetitive, physical labor. Many coal miners suffer hearing loss from the noise of machinery. Coal workers' pneumoconiosis, also known as black lung disease, is a chronic lung disease caused by exposure to coal dust. As the lungs become progressively scarred with continuing exposure, they become stiffer and lung function is impaired. There is no cure, and the disease is sometimes fatal.

Some coal is converted to coke for use as fuel in blast furnaces in the steel industry. Coke itself is made in a coke oven, where gaseous components of the coal are driven off using extreme heat, leaving the coke as a residue. Coke oven workers, especially those working in undesirable jobs at the top of the oven where fumes are strongest, are exposed to high levels of chemicals known as coal tar pitch volatiles.

A classic 1975 occupational health study by Mazumdar and colleagues, which focused on lung cancer risk to coke oven workers, gives a window onto historical racial inequities in this work setting. The study estimated the men's cumulative exposures to coal tar pitch volatiles based on their tenure in various job categories. As shown in Fig. 1, the proportion of nonwhite workers increases steadily from lower to higher categories of cumulative exposure. This pattern reflects the historical practice

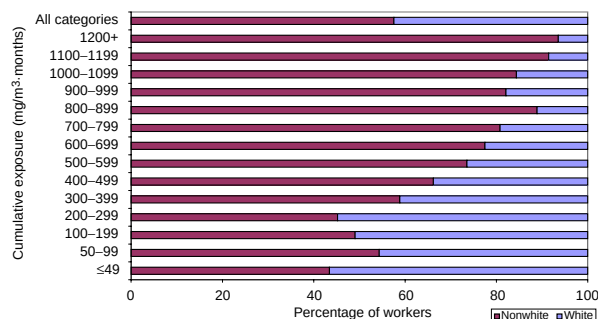


FIGURE 1 White and nonwhite coke oven workers, by exposure category. From Mazumdar *et al.* (1975).

of differentially assigning African-American workers to the high-risk “topside” jobs, while white workers were more likely to be assigned safer positions away from the top of the coke oven.

2. PETROLEUM PRODUCTION

The regions that consume the largest share of the world’s petroleum production are not those that produce the largest shares. North America is the largest consumer (30% of global consumption) and is a major producing region (19%), but not the foremost. Similarly, proportional consumption outstrips production in Asia and Europe. The Middle East, by contrast, produced 30% of the world’s oil and natural gas in 1999 but consumed a mere 9%.

The environmental and health impacts of oil production can be substantial; the compensating benefits that accrue to producing regions are not only access to the energy produced, but also financial gain from oil sales and political power gained through controlling resources and production.

Where foreign interests control oil production, the host country may bear the environmental health impacts while a multinational corporation realizes the financial gain. Such a situation now exists in Nigeria, which produced about 30% of Africa’s oil in 1999. British and U.S. oil companies, in joint ventures with the Nigerian state oil company, have done most of the oil exploration and extraction in the country. Regional populations within Nigeria have bitterly protested the exploitation of local oil resources, with profits departing from the region and environmental contamination left behind.

2.1 Oil Refineries

Most chemical releases from oil refineries are to air. In the U.S. Environmental Protection Agency’s

(EPA’s) Toxic Release Inventory data for 2000, ammonia and sulfuric acid were released in the largest amounts; also prominent were volatile aromatic hydrocarbons that are found in both crude oil and petroleum products (e.g., benzene, toluene, xylene), as well as *n*-hexane and the gasoline additive methyl *tert*-butyl ether.

Like crude oil resources, oil refineries are geographically concentrated, so their environmental and health impacts fall on local populations. Within the United States, the 1997 U.S. Economic Census identifies 244 petroleum refineries nationwide (excluding small establishments). Three states—Texas, California, and Louisiana—host more than 40% of those refineries, representing more than half the total value of U.S. petroleum shipments (Table I).

Petroleum facilities may also be highly concentrated on a more local scale. In Louisiana, for example, data from the U.S. EPA’s Toxics Release Inventory show that the 29 reporting petroleum facilities are located in just 19 of the state’s 525 five-digit zip codes—mostly near Baton Rouge, in Calcasieu and Caddo parishes on the Texas border, and in parishes near New Orleans. Of all the chemical releases from these facilities, half are from a single ExxonMobil refinery in East Baton Rouge, whose releases are almost five times greater than those of the second-ranking refinery. As shown in Table II, this immense refinery is located in a zip code whose population is 89% nonwhite—about three times higher than the state average. The third-ranked refinery in Louisiana, located in Caddo parish, is in a zip code whose population is 90% nonwhite. The stretch of the Mississippi River from Baton Rouge to New Orleans, known locally as “cancer alley,” is home not only to several refineries but also to a much larger number of petrochemical factories. Activists have long protested the concentration of these petrochemical industries in poor, nonwhite communities.

2.2 Oil Storage Facilities

The U.S. EPA’s Toxics Release Inventory reported releases from 474 petroleum bulk terminals in 2000. Not surprisingly, some of the largest of these transport hubs are located at ports. For example, five facilities, representing 7% of releases nationwide, are coastal facilities in Greater Boston, and five facilities, representing 4% of releases nationwide, are located in Los Angeles County. The chemicals released in the largest amounts by weight are some of the same volatile organic chemicals released by oil refineries, though the total releases from bulk terminals amount to only about 5% of those from refineries.

TABLE I

Geographic Concentration of Oil Refineries in the United States

Location	Number of refineries	Percentage of U.S. refineries	Value of shipments (\$1000)	Percentage of value of shipments
United States	244	—	157,934,872	—
Texas	49	20	44,412,926	28
California	31	13	19,296,030	12
Louisiana	25	10	25,795,152	16
Top three states	105	43	89,504,108	57

Source: U.S. Economic Census (1997).

TABLE II

Minority and Poverty Status of Louisiana Zip Codes Hosting Three Refineries with Largest Environmental Releases

Location		Total on- and off-site reported releases (lbs)	Percentage nonwhite	Percentage in Poverty
Louisiana (statewide)		10,420,308	32	20
Louisiana (median of 525 zip codes)		—	27	20
Zip Code	Parish			
70805	East Baton Rouge	5,169,845	89	34
70669	Calcasieu	1,084,032	16	15
71109	Caddo	835,951	90	33

Source: U.S. EPA, Toxics Release Inventory (2000); U.S. Census (2000).

Petroleum bulk terminals, as large industrial installations, are often located in lower income neighborhoods. Four of the five Boston-area “tank farms,” for example, are located in some of the most poverty-stricken communities in the eastern part of the state. In addition to routine chemical releases and the potential for spills or attacks, communities that host petroleum bulk facilities endure the exhaust, noise, and accident hazards of the many trucks that roll through their neighborhoods each day.

3. COAL- AND OIL-FIRED ELECTRIC POWER PLANTS

Inequitable environmental impacts of burning fossil fuels in electric power plants include local air pollution burdens on disadvantaged populations, regional and international effects of acid deposition, and climate change on a worldwide scale.

3.1 Disparate Impacts of Local Air Pollution

Coal- and oil-fired electric power plants contribute to local air pollution, especially particulates, sulfur

dioxide, and nitrogen oxides. In the United States, industrial sources of air pollution tend to be located in lower-income urban areas, and so local impacts often fall most heavily on poor and minority populations. The prevalence of asthma, which may be exacerbated by exposure to air pollutants, has been rising in the United States in recent decades and is particularly elevated among inner-city children.

3.2 Transboundary Acid Deposition

Through a series of complex chemical reactions in the atmosphere, sulfur dioxide and nitrogen oxides emitted from coal- and oil-fired power plants may be converted to sulfuric acid and nitric acid, respectively. These chemicals may be transported hundreds of miles by wind and ultimately be dissolved in precipitation—rain, snow, or fog—and deposited far from their point of origin. Such long-distance transport was brought about by the installation of tall smokestacks in the 1970s; intended to alleviate local pollution, the tall stacks placed air pollution within reach of higher level air currents.

As a result, acid deposition as an environmental impact is very often inequitable, since the receiving area is typically far from the source of the pollution.

For example, acid deposition in the northeastern United States and parts of eastern Canada is due largely to sources in the Midwestern United States. Similarly, burning of fossil fuels in Germany and the United Kingdom contributes to acid deposition in Scandinavia, and coal-fired plants in China were the original source of acid deposition in Japan. Acid deposition can acidify drinking water sources, potentially mobilizing toxic metals such as lead from the sediments of drinking water reservoirs or from drinking water distribution pipes. Acid deposition also harms aquatic life and damages buildings and statues.

3.3 Inequities Related to Global Climate Change

The very designation of “developed” and “less developed” nations identifies an enormous disparity in material security and public health status between these groups of countries. It is inherently unfair that much of the economic growth in the developed nations occurred before the impacts of carbon fuels on global climate were appreciated, whereas the less developed nations must take account of these impacts as they seek economic growth. The dilemma of defining an equitable approach to controlling emissions will not easily be resolved.

Some of the anticipated impacts of global climate change are expected to fall more heavily on less developed countries simply as a matter of geography. For example, warming climates will increase the ranges of mosquitoes and other vectors of infectious disease, and this impact will be felt most strongly in parts of the globe where these diseases, such as malaria, are already endemic. Similarly, much larger numbers of people in the developing world live in low-lying areas that will be threatened by rising sea levels. Small island nations from the Pacific to the Caribbean are experiencing increased temperatures and are uniquely vulnerable to a rising sea level.

The Intergovernmental Panel on Climate Change (IPCC) in its report *Climate Change 2001: Impacts, Adaptation, and Vulnerability* projects that many areas in the developing world are more vulnerable to the effects of global climate change, whereas impacts will generally be more moderate, or more mixed, in North America and Europe. In the more arid regions of Africa, for example, global climate change is expected to exacerbate water scarcity and increase desertification. Food insecurity, already a problem in much of the continent, is expected to worsen due to water stress in agricultural areas and impacts of climate change on fisheries. Increases in tempera-

ture and rainfall will increase the range of vector-borne diseases and will also make it more difficult to maintain sanitation, with a resulting increase in waterborne illness. Increased food insecurity and infectious disease are also projected for areas in Asia. In parts of Asia that are already vulnerable to typhoons or cyclones, droughts, and floods, such extreme weather events are projected to increase. IPCC projects that with a 100 cm rise in sea level, 5763 km² in India would be submerged, putting a population of 7.1 million at risk. Under the same scenario, Bangladesh could lose 29,846 km² of land area, putting 14.8 million people at risk (13.5% of the country's population). Even a sea-level rise of 45 cm is projected to displace 5.5 million in Bangladesh.

The IPCC projects that global climate change will bring more difficult conditions for the southern and Arctic areas of Europe than for other areas of the continent. Pressures on water supply will increase. Crop yields in northern Europe are projected to increase somewhat, though yields in southern Europe are likely to be somewhat reduced. The major direct risks to human health are anticipated from episodes of extreme heat, especially in urban areas, from some increase in vector-borne disease, and from flooding. For North America, a mix of positive and negative impacts is projected. More extreme weather events are anticipated, such as episodes of extreme heat or cold, storm surges, and floods. Overall agricultural production is not threatened under moderate climate change scenarios, though some regions will gain and others will lose in agricultural productivity, and greater warming could cause more disruption of farming. Some increase in vector-borne diseases is also projected. The strong social, economic, and public health infrastructures in Europe and North America are expected to mitigate effects somewhat.

In addition to placing an inequitable burden on developing nations that are less responsible for the problem and less able to manage its impacts, climate change will differentially affect future generations in all regions of the world. More than any other environmental health problem, climate change is global in the scale of its impacts and uncertain in the scope and nature of its intergenerational effects.

4. THE NUCLEAR FUEL CYCLE

Like other energy resources, uranium reserves are unevenly distributed around the globe. The United

States and Canada have produced the most uranium over the long term. U.S. production has fallen off dramatically since the 1980s, but Canada remains the leading producer of recent years. Historically South Africa and Germany have also been major producers of uranium; today Australia, Niger, Namibia, and several countries of the former Soviet Union are producing increasing shares of global uranium output. In contrast, about half of the world's nuclear generating capacity is currently concentrated in Europe and about one-third in North America (mostly in the United States); Japan is the largest user of nuclear power in Asia.

4.1 Environmental Health Hazards of the Nuclear Fuel Cycle

The nuclear fuel cycle, with its unusual suite of environmental health hazards, raises issues of environmental inequity distinct from those of fossil fuels. Nuclear power plants do not produce air pollution in the normal sense of the term, nor do they contribute to global climate change. The hazards associated with nuclear power are those of ionizing radiation, and include the potential for catastrophic reactor failure as well as the unique challenges of disposing of wastes that remain radioactive for thousands or millions of years. Until the emergence of global climate change as a threat to human health, nuclear power was unique in shifting to future generations much of the burden of health risk associated with today's energy production.

Short-term exposure to high levels of ionizing radiation, as in a nuclear accident, may be rapidly fatal or may cause radiation sickness—a syndrome that includes effects on the blood-forming system, the gastrointestinal system, and the central nervous system. Exposure to lower levels of ionizing radiation over a period of time leads to an increased risk of leukemia and of other cancers including cancers of the thyroid, bone, lung, and breast. Birth defects including microcephaly, growth retardation, and mental retardation have also been associated with low-level exposure to ionizing radiation *in utero*.

4.2 Distribution of Nuclear Energy's Benefits and Risks

The distribution of benefits and public health risks associated with nuclear power is inherently inequitable, simply in geographic terms. The “front

end” of the nuclear fuel cycle (uranium mining and milling, and fabrication of reactor fuel) offers no benefit to local populations, other than employment, and workers in uranium mines and mills are exposed to elevated concentrations of radon gas. When radon gas is inhaled, exposure to its radioactive breakdown products, known as radon progeny or radon daughters, increases lung cancer risk. People living in nearby communities are also exposed to radioactive dust from the freestanding piles of waste rock (called tailings) left near mines and mills. In some communities, tailings have been used as construction materials for roads or even homes, resulting in higher radiation exposures to residents.

The hazards associated with mining and milling are usually located far from the people who benefit from the electric power produced by a nuclear reactor. Those populations, often in densely populated areas, do, however, bear the uncertain risks associated with routine low-level radioactive emissions and would be hardest hit in the event of a severe accident—though the incident at Chernobyl demonstrated how geographically widespread such impacts can be. At the “back end” of the nuclear fuel cycle, specially constructed waste repositories are ideally located in isolated areas, which would typically be far from the beneficiaries of the power plants that produced the wastes. This strategy reduces the size of the population at risk, but shifts the burden of risk to people who got no benefit from nuclear power. In the absence of such a repository, spent fuel may be held for long periods at the power plants, as currently in the United States, or disposed of in a less secure manner.

In the United States, most uranium mines and mills are in the West—predominantly in Arizona, Colorado, New Mexico, Utah, and Wyoming. The 103 nuclear power plants operating in the United States in 2003, by contrast, are mostly located east of Kansas and Nebraska. Indeed, if a map of uranium mill tailings piles were overlaid on a map of nuclear power plants, the tailings piles would drop neatly into an area not served by nuclear power. Yucca Mountain, in the southernmost corner of Nevada, is the long-planned site for permanent disposal of U.S. nuclear power plant wastes. The wastes are to reach this remote location by rail lines that crisscross densely populated areas that benefit from nuclear power, regions with no nuclear activities at all, and areas already heavily affected by uranium mining and milling.

4.3 Profile: Differential Impacts on Native Americans of the Nuclear Fuel Cycle in the United States

4.3.1 Early Uranium Mining in the United States

The Four Corners region—where Colorado, New Mexico, Arizona, and Utah meet at right angles—is home to rich reserves of carnotite, an ore containing oxides of both uranium and vanadium. The Navajo Nation is here, located mostly in Arizona but extending into New Mexico and Utah. Carnotite was first mined on a small scale in the early 1900s for the orange color it gave to ceramic pottery glazes. During World War I, radium was in demand for painting instrument dials, and between the world wars vanadium—used to harden steel—was much sought after. But uranium mining began in earnest in the early 1940s with the start of the Manhattan Project and the loss of Czechoslovakian uranium reserves to the Nazi invasion of that country. It became a true boom industry in the late 1940s when the U.S. Atomic Energy Commission (AEC) began to ratchet up development of atomic bombs. For national security reasons, the AEC was at this time the sole buyer of the area's uranium, though mining operations were run by private companies. The AEC would soon begin to press the development of nuclear power as well, further increasing the demand for uranium.

When the mining peaked in 1955–1956, there were hundreds of active uranium mines on the Navajo reservation, as well as several mills where the ore was crushed and the uranium extracted. Jobs were scarce on the reservation, and by the time the boom ended, some 2500 to 3000 Navajos had worked in the mines, either on the reservation or in nearby areas. Smaller numbers of Hispanic men and non-Hispanic white men also worked in the mines. These early mines were known for their poor working conditions. Until the early 1960s, there was no ventilation and no protective gear. Large-scale uranium mining in the region ceased in the late 1960s.

Miners' exposure to radon progeny, which emit radiation as alpha particles, is measured in units called working levels (WL). The WL is a measure of the concentration of potential alpha particles in air. (One WL is defined as any combination of radon progeny in one liter of air that ultimately releases 1.3×10^5 MeV of potential alpha energy.) A working level month (WLM) is equal to 170 WLs (that is, exposure to an average of 1 WL for a working month of 170 hours). The federal exposure limit for miners was set in 1971 at 4 WLMs per year.

4.3.2 Evolving Understanding of the Health Risks of Uranium Mining

As early as 1879, a lung disease called *Bergkrankheit* ("mountain illness") was described in uranium miners in the Erz Mountains in Europe. By the time of World War I, lung cancer had been identified as an important cause of death in these miners. Uranium ore had been known for some time to be radioactive, and radon gas had also been identified. However, scientific confusion over whether and how radon might cause lung cancer continued until the role of radon progeny was elucidated in the early 1950s. Still, the link between occupation and illness was clear, and both Czechoslovakia and Germany had begun to compensate lung cancer in uranium miners as an occupational disease by the early 1930s.

The U.S. Public Health Service (PHS) undertook the first study of radon exposure and lung disease in U.S. uranium miners, many of whom were Navajo, through medical surveys during the 1950s. Both the health agency (the PHS) and the regulatory body (the AEC) have been strongly criticized for failing to obtain informed consent from the miners before using them as research subjects and for recording miners' exposures and health outcomes without providing ventilation and protective gear to reduce exposures already known to be harmful.

More than four decades of published research now documents the experience of the early U.S. uranium miners as this cohort has aged, accumulating illnesses and deaths. Exposure to radon progeny has been found to be associated with increased risk of both malignant and nonmalignant lung disease. Further, because the ore in this region is found in formations of quartz sandstone, workers are exposed to silica dust and are at risk for silicosis. Smoking enhances the lung cancer risk of radon exposure—that is, the lung cancer risk of smoking and radon exposure combined is greater than the sum of their separate risks.

In 1990, Congress passed the Radiation Exposure Compensation Act (RECA). Miners who worked during the years that the U.S. government was the sole purchaser of uranium (1942–1971) may qualify for one-time compensation in the amount of \$100,000 for primary lung cancer and certain nonmalignant respiratory diseases including pulmonary fibrosis, pneumoconiosis, and silicosis. To qualify, a miner (or his survivors) must provide documentation of a designated medical condition and must also be able to document a certain total radon exposure.

4.3.3 Differential Impacts of Uranium Mining on Navajo Miners

Because most U.S. uranium mines were located on or near the Navajo Nation, the Navajos were vastly overrepresented among the miners, compared to their representation in the U.S. population as a whole. Navajos were more likely to work underground than other miners, partly because they were less likely to hold supervisory positions. It appears that in some small Navajo-run mines, working conditions were even worse than in the larger corporate mines. Some Navajos did not speak English, inhibiting their access to any health information that might have been available to the miners. Thus the burden of the government's ongoing failure to regulate working conditions in mines, and of its conduct of research violating accepted ethical standards, fell heavily on the Navajo.

Although early publications on the U.S. uranium mining cohort included some data on mortality among Navajos, a full analysis of the lung cancer risk among Navajo uranium miners was not published until 1984. This landmark study, led by Jonathan Samet of the New Mexico Tumor Registry, identified all cases of lung cancer among Navajo men in New Mexico that had been identified by the tumor registry and compared the men's work histories to those of a control group of Navajo men not diagnosed with lung cancer. Because every Navajo with lung cancer had a history of uranium mining, it was not possible to calculate the relative risk for this exposure (with a zero denominator, the relative risk is infinity), but the lower bound risk was statistically estimated to be a 14-fold increase. That is, the study found that a history of uranium mining carried at least a 14-fold increased risk of lung cancer, and probably a much higher risk.

A second milestone in documenting the differential impacts of the uranium era on Navajo miners came in a 1997 publication by Douglas Mapel and others at the University of New Mexico comparing the *nonmalignant* lung disease experience of Native American, Hispanic, and non-Hispanic white uranium miners. The study clearly documented some expected contextual patterns. Nearly two-thirds of the miners (63.4%) were Native American, 24.7% were Hispanic, and 11.9% were non-Hispanic white. A higher proportion of Native American miners had worked only in *underground* uranium mining (84.3% versus 59.4% Hispanic and 62.4% non-Hispanic white), and a lower proportion of Native American miners had worked only in *aboveground* uranium mining (3.7% versus 7.5% Hispanic and

9.3% non-Hispanic white). The study group reflected the well-known differences in smoking patterns—a much larger percentage of Navajos in the study had never smoked (72.9% versus 34.6% Hispanic and 22.2% non-Hispanic white). Further, those Navajos who did smoke had smoked a much smaller number of cigarettes (Hispanic and non-Hispanic white miners had accumulated more than five times as many “pack-years” of cigarettes smoked). Statistical analysis revealed that this difference in smoking behavior had an important consequence: although all three groups of miners experienced obstructive lung disease, among Hispanic and non-Hispanic white miners this condition was largely attributable to smoking, whereas among the Navajo (most of whom had never smoked), obstructive disease was mostly attributable to uranium mining. Other research has documented a similar effect for lung cancer, identifying a situation—perhaps unique in the annals of public health—in which most of the lung cancer in a population is attributable to a single occupational exposure.

To receive compensation under RECA for non-malignant lung conditions, such as obstructive pulmonary disease or restrictive pulmonary disease, miners must meet certain diagnostic criteria—for example, lung function must be reduced by a certain percentage below normal on a standard lung function test. RECA specified the standards to be used, and they were based on the general U.S. population of non-Hispanic whites. Mapel and colleagues showed that 24% of Navajo miners who would be classified as having restrictive lung disease on the basis of a Native American-specific formula were excluded using the RECA standard. That is, some Navajo miners have been denied compensation under RECA although ethnic-specific criteria indicated that they were sufficiently impaired to warrant compensation. Amendments to the RECA in 2000 adjusted some eligibility criteria.

4.3.4 Differential Impacts of Uranium Mining on Navajo Communities

The risks of large-scale uranium mining were also brought to the Navajo at home, because much of the uranium ore so highly valued by the broader U.S. society was located within the boundaries of the Navajo Nation. The U.S. government, which held reservation lands in trust for Native Americans, freely exploited those resources, with profound impacts to the natural environment and risks to the health of local populations. Especially in the early decades, uranium was mined disproportionately on

native lands. The sense of injustice felt by Native Americans from this experience is reflected in the use of terms such as “internal colonialism,” to describe U.S. government exploitation of uranium resources on Native American lands, and “national sacrifice area,” to describe the environmental devastation that resulted.

Because groundwater seeped into both underground shaft mines and the deeper open-pit mines, the water had to be continuously pumped out so that operations could continue. This “dewatering” released large volumes of water. A single site might dump tens of thousands of gallons of water per day into local streams—water contaminated with radionuclides and heavy metals. Uranium mills, many located on or near reservation land, left behind enormous piles of tailings, an estimated 140 million tons altogether in seven western states. These heaps of pulverized waste rock are an ongoing source of radioactive dust. In 1979, an earthen dam collapsed, releasing more than 1000 tons of tailings sludge and almost 100 million gallons of water that were carried downstream into a Navajo community. Even the introduction of vents in mines sometimes resulted in community exposures, if vents were placed in populated areas above mines. All these burdens fell on communities that already lived with more poverty, poorer health, and a lower life expectancy than the general U.S. population.

4.3.5 Potential Siting of Nuclear Waste Facilities on Native American Lands

Since the 1980s, the federal government has planned to build a permanent repository for high-level radioactive wastes from nuclear power plants at Yucca Mountain, Nevada, but the expected opening year for the facility has receded steadily to 2010. As the permanent repository has been delayed, pressure has built for an interim storage facility (known as monitored retrievable storage, or MRS), so that spent fuel can be moved from storage sites at power plants. During the 1990s, the U.S. Department of Energy actively sought communities and Native American tribes to host such an MRS facility. The department worked to create an interactive process through which safety provisions and financial or social benefits would be negotiated such that the community would be satisfied that the potential negative impacts of the site had been offset. Though dozens of communities and tribes began the process, extended negotiations were held only with Native American constituencies—the Mescalero Apache of New Mexico and the Skull Valley Band of the

Goshute in Utah. Both negotiations ultimately foundered. The Skull Valley Band has since entered negotiations with a private utility to host an MRS, resulting in legal battles that are not yet resolved. Although these Native American groups have participated voluntarily in siting negotiations, the process has nevertheless raised the question of an environmental injustice against disadvantaged populations whose economic need made them vulnerable.

5. LARGE HYDROELECTRIC DAMS

Although hydropower is a renewable energy technology and is often seen as relatively benign in terms of pollution, its facilities may have very different impacts on different populations.

5.1 Hydropower and Environmental Equity

Hydropower is a well-developed technology that is relatively accessible to developing nations. At present, five nations—the United States, Canada, China, Brazil, and the Russian Federation in that order—are global leaders in hydropower capacity, together representing almost 40% of the world’s operating capacity. Dam projects now under construction worldwide, including the massive Three Gorges Dam in China, are expected to increase current global operating capacity by 16%. Taking account of capacity currently under construction or in the planning stages, China and Brazil are expected in coming years to push the United States and Canada to the third and fourth ranks, and India is expected to move up into sixth place behind the Russian Federation.

Thus, at a global scale, hydropower is becoming more equitably distributed between richer and poorer nations. On the smallest scale, too, when the benefits and burdens of small dams are shared by the same people, hydropower’s impacts may be equitable. But to a great extent the very large dam projects that emerged in the latter 20th century and continue into the 21st deliver benefits and burdens to different populations, creating persistent inequities.

Hydropower is a relatively inexpensive form of electric power after the initial investment in construction, so the principal benefit of a large dam is cheap electricity for some population. Large dams may offer other benefits related to management of a waterway, such as flood control, improved navigation, or increased opportunities for irrigation. On the

other hand, new reservoirs may displace large numbers of people, often disenfranchised populations, and there may be detrimental downstream effects on fishing or farming. Finally, those located downstream of a dam are at risk in the event of a catastrophic failure, though this is generally considered an unlikely event.

Although this discussion focuses on the outward environmental impacts of hydropower facilities on different demographic groups, a historical episode from the occupational setting deserves mention. In 1930–1931, the 3-mile Hawk’s Nest Tunnel was bored through a mountain in West Virginia to divert water from its natural riverbed to a hydropower plant being built for an industrial facility. The rock was high in silica, and dust was thick in the tunnel, but no measurements were taken, and attempts to ventilate the tunnel or damp down the dust were minimal. Although chronic silicosis was a recognized occupational illness in miners at the time, it was the fast-paced construction of the Hawk’s Nest Tunnel that revealed *acute* silicosis, an illness that killed many workers within months of their tunnel labors. The workers, many of whom were brought in from outside the county for the tunnel project, were about two-thirds African American. Of the white laborers, 44% worked inside the tunnel at any time, and 22% worked *only* inside the tunnel (Fig. 2). By contrast, among the African-American laborers, fully 70% worked at some point inside the tunnel, and 35% worked *only* inside the tunnel. Thus the episode is noteworthy not only for revealing the acute form of silicosis but also for the striking racial inequities in workforce exposures to the silica-laden dust of the tunnel.

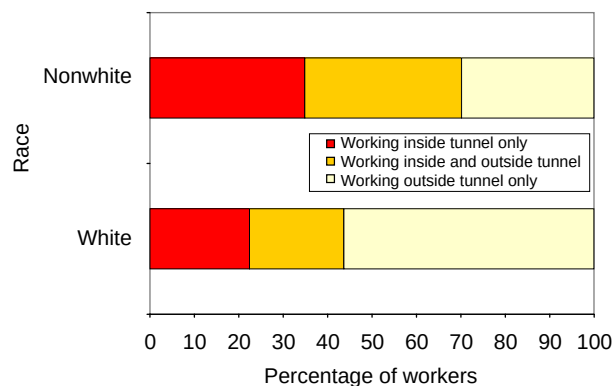


FIGURE 2 Exposure of white and nonwhite workers in building the Hawk’s Nest Tunnel. Data from Cherniak, M., *et al.* (1986). “The Hawk’s Nest Incident.” Yale University Press, New Haven, Connecticut.

5.2 Profile: The Three Gorges Dam on the Yangtze River in China

The Three Gorges Dam now being built in Hubei province in China will be the largest dam ever constructed. The high dam is to create a narrow reservoir 600 kilometers (370 miles) long filling the Qutang, Wu, and Xiling Gorges of the Yangtze River. Construction began in 1994; by 1997 a diversion dam was completed so that work could begin on the main dam. Flooding and power generation began in 2003, and the project is to be completed in 2010.

5.2.1 Benefits

The dam is expected to produce 18,000 MW of electric power, and some 15 million people in five provinces in Central China will get inexpensive electricity from the project. By replacing coal-burning power plants, the dam is also expected to result in cleaner air in the Yangtze Valley.

In addition to its energy benefits, the dam is intended to control flooding, a major problem on the Yangtze River—although many tributaries enter the river below the dam site. It is also expected to smooth navigation of the river above the dam, though critics argue that locks needed to carry ships around the dam may stall river traffic.

5.2.2 Displacement of Residents

The Three Gorges Dam will flood nearly 20 cities, hundreds of towns, and more than a thousand villages. Estimates of the number of people to be displaced by the dam range from 1.3 million to 1.9 million. To date, more than 600,000 have been moved from their homes. Some 30,000 hectares (74,000 acres) of farmland will be flooded, and perhaps half of the people who will ultimately be displaced are farmers.

Because so many people must be moved within such a short time, the competition for resources is intense. Many of the displaced face hunger or food insecurity, occurring in the context of a disrupted social fabric. Although the government has compensated those displaced by the dam project, the compensation has been inadequate and many of the displaced become migrants. Resentment against the government is widespread, as is resentment against the displaced populations in the areas that receive them.

Because the Three Gorges Dam is such a high-profile project, institutional pressure to continue with it is intense, despite any problems or challenges that may occur. In this context, the large-scale removal and resettlement of people creates troubling

potential for human rights abuses, including punishment of those who refuse to leave—or who leave but later return to their homes.

A survey of displaced persons by Li Heming and colleagues gave insight into the experience of the dispossessed. The resettling of farmers offers special challenges because of their intense attachment to the land. Many of those who have refused to leave, or who have returned, are farmers. For the most part, the resettlement program has sought to move rural residents to other rural locations not too far distant from their original homes. But farmers have been given less land than they owned in the area to be flooded, and land of lower quality—for example, with poor soil on a steep slope—or suited for growing different crops from the ones they are familiar with.

Other displaced residents, urban and rural, have been moved to urban settings far away from the Three Gorges Area. Here housing costs often outstrip the compensation provided. Some of the displaced are not given jobs in their new locations, and some of those who are given jobs lose them. The resistance to living in a distant location is strong. The most successful relocations appear to take place when elderly persons, who are no longer working, move in with relatives. Many displaced persons believe that the project has been tainted by corruption.

5.2.3 Other Detrimental Impacts

The Three Gorges Dam will destroy the habitat of a number of indigenous species, including some found only in China and a species of sturgeon found only in the Yangtze river. The dam will flood three gorges famous for their natural beauty, along with the many antiquities located there. Some critics of the project have suggested that upstream pollution sources will degrade the quality of the water in the reservoir. Downstream concerns include a loss of fertility in farmlands that are no longer regularly flooded by the river and loss of wells because groundwater will not be replenished.

Finally, there is some concern about a catastrophic failure of the dam in the event of an earthquake, especially if engineering decisions or construction quality have been compromised. Some critics even fear that the sheer weight of the reservoir could trigger an earthquake along a fault line that lies under it.

6. HOME USE OF BIOMASS FUELS

A total of 1.4 billion tons of fuelwood was produced worldwide in 1999, and a roughly equal amount is

consumed. Global production and use of fuelwood is concentrated in Asia (46%) and Africa (30%). India alone accounts for 14% of global fuelwood production; India, China, and Indonesia together account for almost one-third of the world's production.

The use of wood for fuel declined rapidly in the developed world with the increasing availability of oil and natural gas after World War II. In the United States, wood was used to heat almost one-quarter of all homes in 1940, but only 1% in 1970 (Fig. 3). Wood heat enjoyed something of a resurgence in the 1970s and 1980s, especially among middle- to upper-income households in New England and in the Northwest, but fell off again in the 1990s.

Global peat production amounted to only about 1% by weight of fuelwood production in 1999. Peat remains predominantly a fuel of Europe, where 95% of the world's peat was produced. Finland accounts for 37% of the world's peat production, Belarus and the Russian Federation 15% each, and Ireland 14%.

6.1 Health Effects of Domestic Woodsmoke in Developed Countries

Woodsmoke's constituents form a profile of products of combustion, including respirable particulates less than 10 microns in diameter (PM₁₀), carbon monoxide, nitrogen oxides, and sulfur oxides, polycyclic aromatic hydrocarbons (PAHs), formaldehyde, and various volatile organic compounds (VOCs). Exposure to woodsmoke or to its constituent respiratory irritants (e.g., particulate matter, nitrogen dioxide, sulfur dioxide) has been associated with inhibition of the cilia that line the respiratory tract and sweep foreign objects up and out; constriction of airways due to swelling of the membranes that line them

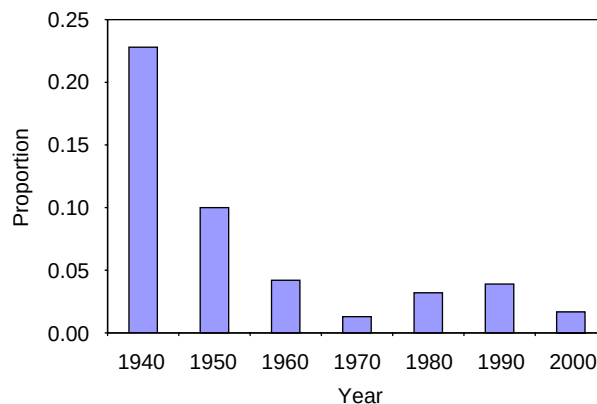


FIGURE 3 Proportion of U.S. housing units using wood for heat, 1940–2000. Data from 2000 U.S. Census.

(chronic airway resistance); a decrease in the volume of air that can be deliberately expelled from the lungs after inhaling (vital capacity); an increase in such respiratory symptoms as coughing, wheezing, and shortness of breath; and increased risk of infectious respiratory disease in children. Formaldehyde is a respiratory irritant that exacerbates asthma and causes headaches; in addition, chronic exposure is associated with increased risk of nasal cancer. Some PAHs have also been shown to be carcinogenic or to have immunosuppressive effects.

6.2 Profile: Effects of Home Heating and Cooking in Developing Nations

It is estimated that approximately half of the world's population—around 3 billion people—use biomass fuels to heat their homes and cook their foods. In rural locations in developing countries, the great majority of households use biomass fuels to meet domestic needs. Biomass fuels are derived from plants or animals, including not only wood and peat, but also animal dung and crop residues. Among these fuels, animal dung and crop residues are the dirtiest to burn but also least expensive; wood and charcoal burn somewhat more cleanly but cost more. Oil, gas, and electricity are the cleanest and most expensive fuels, out of reach of much of the world's population.

Biomass fuels are typically burned in primitive stoves with very incomplete combustion, in an open fire, or even using more than one stove, or some combination of open fire and stove. Depending on the time of year, the type of fuel and stove being used, and the nature of the structure, pollutant concentrations in indoor air may be many times those found in homes in the developed world. The U.S. EPA sets a standard for respirable particulates (particulate matter less than 10 microns in diameter, or PM_{10}) of $50 \mu\text{g}/\text{m}^3$ as an annual average and $150 \mu\text{g}/\text{m}^3$ as a daily average. (These small particles penetrate deeply into the lung; in fact, particles less than 2.5 microns in diameter penetrate more deeply still.) In a recent study of health impacts of wood stoves in Quebec, Canada, 9% of the homes examined had 24-hour average concentrations greater than $50 \mu\text{g}/\text{m}^3$; 67% had 24-hour average concentrations less than $11 \mu\text{g}/\text{m}^3$. These measurements contrast sharply with those taken in homes in developing countries where biomass fuels are used for heating and cooking. Here, researchers report typical 24-hour average concentrations in homes ranging from 200 to

$5000 \mu\text{g}/\text{m}^3$; and peak concentrations of more than $50,000 \mu\text{g}/\text{m}^3$ recorded near stoves.

6.2.1 Assessing Exposures to Domestic Biomass Fuels

An individual's cumulative exposure to the products of combustion depends on a number of factors in addition to the type of fuel and the combustion efficiency of the stove. For example, the fire may be burning brightly or smoldering. The use of the stove is affected by season and time of day. The degree to which pollutants are evenly dispersed through the house is affected by whether the cooking and living areas are in the same room or are separate from one another. Ventilation varies widely across homes.

However, research by Majid Ezzati and others has shown that individual patterns of behavior are the most important determinants of exposure. How much time is spent in the cooking area? How much time is spent bending over the stove or fire, adjusting the fuel or stirring the food? Studies have shown that women aged 16 to 50 and girls aged 6 to 15 spend more time in the cooking area, and more time engaged in high-exposure activities related to cooking, than do men or boys in the same age groups. As a result, women's cumulative exposures to indoor air pollution are higher than men's, and the proportion of their exposure that comes from cooking is also higher. Similar gender differences in exposure are not documented for children under 6, who have not yet taken on many household duties. However, young children's exposures are higher than those of older boys or of men aged 16 to 50, reflecting time spent in their mother's company.

6.2.2 The Global Burden of Disease from Domestic Use of Biomass Fuels

Exposure to indoor air pollution in developing countries has been most clearly associated with increased risk of acute lower respiratory infection, middle ear infection (otitis media), and chronic obstructive pulmonary disease. An estimated 1.5 to 2 million deaths each year in developing countries are due to indoor air pollution from the burning of biomass fuels, with perhaps 1 million of these due to acute lower respiratory infections in children under age 5. In sum, the burden of disease from domestic burning of biomass fuels rests most heavily on disenfranchised groups: the world's poorest countries, lower income populations within those countries, women, and young children.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Public Health: Emerging Infectious Diseases • Economic Geography of Energy • Environmental Change and Energy • Equity and Distribution in Energy Policy • Geopolitics of Energy • Global Energy Use: Status and Trends • Public Reaction to Nuclear Power Siting and Disposal • Women and Energy: Issues in Developing Nations

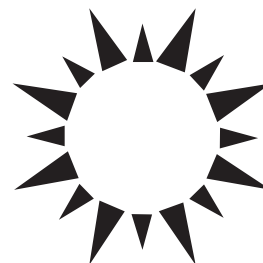
Further Reading

- Ezzati, M., and Kammen, D. M. (2002). The health impacts of exposure to indoor air pollution from solid fuels in developing countries: knowledge, gaps, and data needs. *Environmental Health Perspectives* 110(11), 1057–1068.
- Heming, L. (2001). Reservoir resettlement in China: past experience and the Three Gorges Dam. *Geographical J.* 167(3), 195–212.
- Intergovernmental Panel on Climate Change. Technical Summary: Climate Change 2001: Impacts, Adaptation, and Vulnerability. A report of Working Group II of the Intergovernmental Panel on Climate Change.
- Mapel, D. W., Coultas, D. B., James, D. S., Hunt, W. C., Stidley, C. A., and Gililand, F. D. (1997). Ethnic differences in the prevalence of nonmalignant respiratory disease among uranium miners. *Am. J. Public Health* 87(5), 833–838.
- Mazumdar, S., Redmond, C., Sellecito, W., and Sussman, N. (1975). An epidemiological study of exposures to coal tar pitch volatiles among coke oven workers. *J. Air Pollution Control Assn.* 25, 382–389.
- Samet, J. M., Kutvirt, D., Waxweiler, R. J., and Key, C. R. (1984). Uranium mining and lung cancer in Navajo men. *New England J. Medicine.* 310(23), 1481–1484.
- World Energy Council. Survey of Energy Resources (1999). www.worldenergy.org.

Environmental Kuznets Curve

DAVID I. STERN

Rensselaer Polytechnic Institute
Troy, New York, United States



-
1. Introduction
 2. Theoretical Background
 3. Econometric Framework
 4. Results of EKC Studies
 5. Theoretical Critique of the EKC
 6. Econometric Critique of the EKC
 7. Other Evidence
 8. Decomposing Emissions
 9. Conclusions

Glossary

cointegration Classical regression analysis assumes that the error term in a regression is normally and independently distributed. An extreme violation of this condition is when the error term is a random walk. Such a regression is termed a spurious regression and the results are not reliable. It usually indicates that a required variable with random walk behavior has been omitted from the model or that the variables in the model follow random walks and are not related to each other.

consistent estimation May show bias and inefficiency, but both are reduced toward zero as the sample size increases.

developed economies Countries with high income levels; typically have an income per capita greater than U.S. \$10,000, life expectancy greater than 70 years, and literacy rates greater than 95%.

developing economies Countries with low- and middle-income levels; equivalent to the term "Third World."

econometrics The specialized branch of mathematical statistics applied to the analysis of economic data.

economic growth An increase in economy-wide economic production usually measured by an increase in gross domestic product; also, the process of the economy growing over time.

environmental degradation Any human-caused deterioration in environmental quality.

externality When a production or consumption activity has unintended damaging effects, such as pollution, on

other firms or individuals and no compensation is paid by the producer or consumer to the affected parties (negative externality); when activities have beneficial effects for others, such as freely available research results, and the recipients do not pay the generator of the effect (positive externality).

heteroskedasticity When the variance of the error term in a regression model varies systematically across observations or cases.

income elasticity The percentage increase in some variable when income increases by 1%.

income per capita Usually computed as the gross national product of a country for one year divided by the total population.

integrated variable A random walk is an integrated variable—the current value is equal to the previous period's value plus a random shock. This means that the effect of these shocks on the future does not decline over time. Rather, the current value of the variable is an integration of all past shocks.

omitted variables bias Occurs when variables that should be in a regression model are omitted and they are correlated with those variables that are included.

panel data A data set that includes observations of a number of individuals, countries, firms, and the like over a number of time periods such as months or years.

simultaneity bias When there is mutual feedback between two variables and ordinary regression analysis cannot provide consistent estimates of model parameters.

sustainable development If the current average well-being of people could be maintained into the indefinite future.

technological change The invention and introduction of new methods of production as well as new products.

The environmental Kuznets curve is a hypothesized relationship among various indicators of environmental degradation and income per capita. During the early stages of economic growth, degradation and pollution increase, but beyond some level of income per capita (which will vary for different indicators) the trend reverses, so that at high income levels economic growth leads to environmental improvement. This implies that the environmental

impact indicator is an inverted U-shaped function of income per capita. Typically, the logarithm of the indicator is modeled as a quadratic function of the logarithm of income.

1. INTRODUCTION

An example of an estimated environmental Kuznets curve (EKC) is shown in Fig. 1. The EKC is named for Simon Kuznets, who hypothesized that income inequality first rises and then falls as economic development proceeds. Emissions of various pollutants, such as carbon dioxide, sulfur, and nitrogen oxides, are tightly coupled to the use of energy. Hence, the EKC is a model of the relationship among energy use, economic growth, and the environment.

The EKC is an essentially empirical phenomenon, but most of the EKC literature is statistically weak. It is very easy to do bad econometrics, and the history of the EKC exemplifies what can go wrong. The EKC idea rose to prominence because few paid sufficient attention to the relevant diagnostic statistics. Little or no attention has been paid to the statistical properties of the data used such as serial dependence and random walk trends in time series, and few tests of model adequacy have been carried out or presented. However, one of the main purposes of doing econometrics is to test which apparent relationships are valid and which are spurious correlations.

When we do take such statistics into account and use appropriate techniques, we find that the EKC does not exist. Instead, we get a more realistic view of the effect of economic growth and technological changes on environmental quality. It seems that most indica-

tors of environmental degradation are monotonically rising in income, although the "income elasticity" is less than 1.0 and is not a simple function of income alone. Time-related effects, intended to model technological change common to all countries, reduce environmental impacts in countries at all levels of income. However, in rapidly growing middle-income countries, the scale effect, which increases pollution and other degradation, overwhelms the time effect. In wealthy countries, growth is slower and pollution reduction efforts can overcome the scale effect. This is the origin of the apparent EKC effect.

The econometric results are supported by recent evidence that, in fact, pollution problems are being addressed and remedied in developing economies.

This article follows the development of the EKC concept in approximately chronological order. Sections II and III review in more detail the theory behind the EKC and the econometric methods used in EKC studies. Sections IV to VI review some EKC analyses and their critiques. Section VII discusses the recent evidence from Dasgupta and colleagues and others that has changed the way in which we view the EKC. The final two sections discuss an alternative approach, decomposition of emissions, and summarize the findings.

2. THEORETICAL BACKGROUND

The EKC concept emerged during the early 1990s with Grossman and Krueger's pathbreaking study of the potential impacts of the North American Free Trade Agreement (NAFTA) and Shafik and Bandyopadhyay's background study for the *World Development Report 1992*. However, the idea that economic growth is necessary for environmental quality to be maintained or improved is an essential part of the sustainable development argument promulgated by the World Commission on Environment and Development in *Our Common Future*.

The EKC theme was popularized by the International Bank for Reconstruction and Development's (IBRD) *World Development Report 1992*, which argued that "the view that greater economic activity inevitably hurts the environment is based on static assumptions about technology, tastes, and environmental investments" and that "as incomes rise, the demand for improvements in environmental quality will increase, as will the resources available for investment." Others have expounded this position even more forcefully, with Beckerman claiming that "there is clear evidence that, although economic

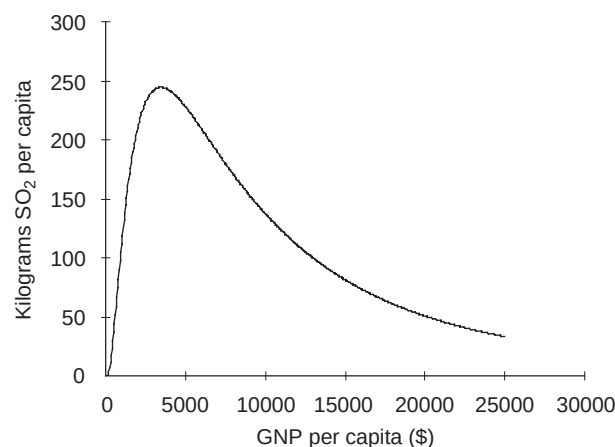


FIGURE 1 Environmental Kuznets curve for sulfur emissions. Data from Panayotou (1993) and Stern *et al.* (1996).

growth usually leads to environmental degradation in the early stages of the process, in the end the best—and probably the only—way to attain a decent environment in most countries is to become rich.” However, the EKC has never been shown to apply to all pollutants or environmental impacts, and recent evidence challenges the notion of the EKC in general. The remainder of this section discusses the economic factors that drive changes in environmental impacts and that may be responsible for rising or declining environmental degradation over the course of economic development.

If there were no change in the structure or technology of the economy, pure growth in the scale of the economy would result in a proportional growth in pollution and other environmental impacts. This is called the scale effect. The traditional view that economic development and environmental quality are conflicting goals reflects the scale effect alone. Panayotou, a proponent of the EKC hypothesis, argued,

At higher levels of development, structural change towards information-intensive industries and services, coupled with increased environmental awareness, enforcement of environmental regulations, better technology, and higher environmental expenditures, result in leveling off and gradual decline of environmental degradation.

Therefore, at one level, the EKC is explained by the following “proximate factors”:

1. Scale of production implies expanding production, with the mix of products produced, the mix of production inputs used, and the state of technology all held constant.
2. Different industries have different pollution intensities, and the output mix typically changes over the course of economic development.
3. Changes in input mix involve the substitution of less environmentally damaging inputs to production for more damaging inputs and vice versa.
4. Improvements in the state of technology involve changes in both (a) production efficiency in terms of using less, *ceteris paribus*, of the polluting inputs per unit of output and (b) emissions-specific changes in process that result in less pollutant being emitted per unit of input.

These proximate variables may, in turn, be driven by changes in underlying variables such as environmental regulation, awareness, and education. A number of articles have developed theoretical models about how preferences and technology might interact to result in different time paths of environmental

quality. The various studies make different simplifying assumptions about the economy. Most of these studies can generate an inverted U-shape curve of pollution intensity, but there is no inevitability about this. The result depends on the assumptions made and the value of particular parameters. It seems fairly easy to develop models that generate EKCs under appropriate assumptions, but none of these theoretical models has been empirically tested. Furthermore, if in fact the EKC for emissions is monotonic as recent evidence suggests, the ability of a model to produce an inverted U-shaped curve is not necessarily a desirable property.

3. ECONOMETRIC FRAMEWORK

The earliest EKCs were simple quadratic functions of the levels of income. However, economic activity inevitably implies the use of resources, and by the laws of thermodynamics, use of resources inevitably implies the production of waste. Regressions that allow levels of indicators to become zero or negative are inappropriate except in the case of deforestation where afforestation can occur. This restriction can be applied by using a logarithmic dependent variable. The standard EKC regression model is

$$\ln(E/P)_{it} = \alpha_i + \gamma_t + \beta_1 \ln(\text{GDP}/P)_{it} + \beta_2 [\ln(\text{GDP}/P)_{it}]^2 + \varepsilon_{it}, \quad (1)$$

where E is emissions, P is population, GDP is gross domestic product, ε is a random error term, and $\ln$ indicates natural logarithms. The first two terms on the right-hand side are intercept parameters that vary across countries or regions i and years t . The assumption is that although the level of emissions per capita may differ over countries at any particular income level, the income elasticity is the same in all countries at a given income level. The time-specific intercepts are intended to account for time-varying omitted variables and stochastic shocks that are common to all countries.

The “turning point” level of income, where emissions or concentrations are at a maximum, can be found using the following formula:

$$\tau = \exp[-\beta_1/(2\beta_2)]. \quad (2)$$

The model is usually estimated with panel data. Most studies attempt to estimate both the fixed effects and random effects models. The fixed effects model treats the α_i and γ_t as regression parameters, whereas the random effects model treats them as components of the random disturbance. If the effects

α_i and γ_t and the explanatory variables are correlated, the random effects model cannot be estimated consistently. Only the fixed effects model can be estimated consistently. A Hausman test can be used to test for inconsistency in the random effects estimate by comparing the fixed effects and random effects slope parameters. A significant difference indicates that the random effects model is estimated inconsistently due to correlation between the explanatory variables and the error components. Assuming that there are no other statistical problems, the fixed effects model can be estimated consistently, but the estimated parameters are conditional on the country and time effects in the selected sample of data. Therefore, they cannot be used to extrapolate to other samples of data. This means that an EKC estimated with fixed effects using only developed country data might say little about the future behavior of developing countries. Many studies compute the Hausman statistic and, finding that the random effects model cannot be estimated consistently, estimate the fixed effects model. But few have pondered the deeper implications of the failure of this test.

Tests for integrated variables designed for use with panel data find that sulfur and carbon emissions and GDP per capita are integrated variables. This means that we can rely only on regression results that exhibit the cointegration property. If EKC regressions do not cointegrate, the estimates will be spurious. Very few studies have reported any diagnostic statistics for integration of the variables or cointegration of the regressions, and so it is unclear what we can infer from the majority of EKC studies.

4. RESULTS OF EKC STUDIES

The key features differentiating the broad variety of EKC studies can be displayed by reviewing a few of the early studies and examining a single indicator (sulfur) in more detail. This section also discusses EKCs for total energy use, seen by some as a proxy indicator for all environmental impacts.

The early EKC studies appeared to indicate that local pollutants, such as smoke and sulfur emissions, were more likely to display an inverted U-shaped relation with income than were global impacts such as carbon dioxide. This picture fits environmental economics theory; local impacts are internalized in a single economy or region and are likely to give rise to environmental policies to correct the externalities on

pollutees before such policies are applied to globally externalized problems.

Grossman and Krueger produced the first EKC study as part of a study of the potential environmental impacts of NAFTA. They estimated EKCs for SO₂, dark matter (fine smoke), and suspended particles (SPM) using the GEMS data set. This data set is a panel of ambient measurements from a number of locations in cities around the world. Each regression involved a cubic function in levels (not logarithms) of purchasing power parity adjusted (PPP) per capita GDP and various site-related variables, a time trend, and a trade intensity variable. The turning points for SO₂ and dark matter are at approximately \$4000 to \$5000, whereas the concentration of suspended particles appeared to decline even at low income levels.

Shafik and Bandyopadhyay's study was particularly influential in that the results were used in the *World Development Report 1992*. They estimated EKCs for 10 different indicators using three different functional forms. Lack of clean water and lack of urban sanitation were found to decline uniformly with increasing income and over time. Deforestation regressions showed no relation between income and deforestation. River quality tended to worsen with increasing income. However, local air pollutant concentrations conformed to the EKC hypothesis, with turning points between \$3000 and \$4000. Finally, both municipal waste and carbon emissions per capita increased unambiguously with rising income.

Selden and Song estimated EKCs for four emissions series—SO₂, NO_x, SPM, and CO—using longitudinal data from World Resources. The data were primarily from developed countries. The estimated turning points all were very high compared with those in the two earlier studies. For the fixed effects version of their model, the estimated turning points were as follows (converted to 1990 U.S. dollars using the U.S. GDP implicit price deflator): SO₂, \$10,391; NO_x, \$13,383; SPM, \$12,275; and CO, \$7,114. This study showed that the turning point for emissions was likely to be higher than that for ambient concentrations. During the initial stages of economic development, urban and industrial development tends to become more concentrated in a smaller number of cities that also have rising central population densities, with the reverse happening during the later stages of development. So, it is possible for peak ambient pollution concentrations to fall as income rises, even if total national emissions are rising.

Table I summarizes several studies of sulfur emissions and concentrations, listed in order of estimated

TABLE I
Sulfur EKC Studies

Author(s)	Turning point (1990 U.S. dollars)	Emissions or concentrations	Purchasing power parity	Additional variables	Data source for sulfur	Time period	Countries/ cities
Panayotou (1993)	3,137	Emissions	No	—	Own estimates	1987–1988	55 developed and developing countries
Shafik and Bandyopadhyay (1992)	4,379	Concentrations	Yes	Time trend, locational dummies	GEMS	1972–1988	47 cities in 31 countries
Torras and Boyce (1998)	4,641	Concentrations	Yes	Income inequality, literacy, political and civil rights, urbanization, locational dummies	GEMS	1977–1991	Unknown number of cities in 42 countries
Grossman and Krueger (1994)	4,772–5,965	Concentrations	No	Locational dummies, population density, trend	GEMS	1977, 1982, 1988	Up to 52 cities in up to 32 countries
Panayotou (1997)	5,965	Concentrations	No	Population density, policy variables	GEMS	1982–1984	Cities in 30 developed and developing countries
Cole <i>et al.</i> (1997)	8,232	Emissions	Yes	Country dummy, technology level	OECD	1970–1992	11 OECD countries
Selden and Song (1994)	10,391–10,620	Emissions	Yes	Population density	WRI: primarily OECD source	1979–1987	22 OECD and 8 developing countries
Kaufmann <i>et al.</i> (1997)	14,730	Concentrations	Yes	GDP/Area, steel exports/GDP	UN	1974–1989	13 developed and 10 developing countries
List and Gallet (1999)	22,675	Emissions	N/A	—	U.S. EPA	1929–1994	U.S. states
Stern and Common (2001)	101,166	Emissions	Yes	Time and country effects	ASL	1960–1990	73 developed and developing countries

income turning points. On the whole, the emissions-based studies have higher turning points than do the concentrations-based studies. Among the emissions-based estimates, both Selden and Song and Cole and colleagues used databases that are dominated by, or consist solely of, emissions from Organization for Economic and Cooperative Development (OECD) countries. Their estimated turning points were \$10,391 and \$8,232, respectively. List and Gallet

used data for 1929 to 1994 for the 50 U.S. states. Their estimated turning point is the second highest in Table I. Income per capita in their sample ranged from \$1,162 to \$22,462 (in 1987 U.S. dollars). This is a wider range of income levels than was found in the OECD-based panels for recent decades. This suggests that including more low-income data points in the sample might yield a higher turning point. Stern and Common estimated the turning point at more

than \$100,000. They used an emissions database produced for the U.S. Department of Energy by ASL that covers a greater range of income levels and includes more data points than do any of the other sulfur EKC studies.

The recent studies that use more representative samples of data found that there is a monotonic relation between sulfur emissions and income, just as there is between carbon dioxide and income. Interestingly, an estimate of a carbon EKC for a panel data set of OECD countries found an inverted U-shaped EKC in the sample as a whole. The turning point was at only 54% of maximal GDP in the sample. A study using a spline regression implied a within-sample turning point for carbon for high-income countries. All of these studies suggest that the differences in turning points that have been found for various pollutants may be due, at least in part, to the different samples used.

In an attempt to capture all environmental impacts of whatever type, a number of researchers estimated EKCs for total energy use. In each case, they found that energy use per capita increases monotonically with income per capita. This result does not preclude the possibility that energy intensity (i.e., energy used per dollar of GDP produced) declines with rising income or even follows an inverted U-shaped path.

A number of studies built on the basic EKC model by introducing additional explanatory variables intended to model underlying or proximate factors such as "political freedom," output structure, and trade. On the whole, the included variables turned out to be significant at traditional significance levels. However, testing different variables individually is subject to the problem of potential omitted variables bias. Furthermore, these studies did not report cointegration statistics that might indicate whether omitted variables bias was likely to be a problem or not. Therefore, it is not really clear what can be inferred from this body of work.

The only robust conclusions from the EKC literature appear to be that concentrations of pollutants may decline from middle-income levels and that emissions tend to be monotonic in income. Emissions may decline over time in countries at many different levels of development. Given the likely poor statistical properties of most EKC models, it is hard to come to any conclusions about the roles of other additional variables such as trade. Too few quality studies of other indicators apart from air pollution have been conducted to come to any firm conclusions about those impacts as well.

5. THEORETICAL CRITIQUE OF THE EKC

A number of critical surveys of the EKC literature have been published. This section discusses the criticisms that were made of the EKC on theoretical, rather than methodological, grounds.

The key criticism of Arrow and colleagues and others was that the EKC model, as presented in the *World Development Report 1992* and elsewhere, assumes that there is no feedback from environmental damage to economic production given that income is assumed to be an exogenous variable. The assumption is that environmental damage does not reduce economic activity sufficiently to stop the growth process and that any irreversibility is not severe enough to reduce the level of income in the future. In other words, there is an assumption that the economy is sustainable. However, if higher levels of economic activity are not sustainable, attempting to grow fast during the early stages of development, when environmental degradation is rising, may prove to be counterproductive.

It is clear that the levels of many pollutants per unit of output in specific processes have declined in developed countries over time with technological innovations and increasingly stringent environmental regulations. However, the mix of effluent has shifted from sulfur and nitrogen oxides to carbon dioxide and solid waste, so that aggregate waste is still high and per capita waste might not have declined. Economic activity is inevitably environmentally disruptive in some way. Satisfying the material needs of people requires the use and disturbance of energy flows and materials. Therefore, an effort to reduce some environmental impacts might just aggravate other problems. Estimation of EKCs for total energy use is an attempt to capture environmental impact regardless of its nature.

Various critics argued that if there was an EKC-type relationship, it might be partly or largely a result of the effects of trade on the distribution of polluting industries. The Heckscher–Ohlin trade theory, the central theory of trade in modern economics, suggests that, under free trade, developing countries would specialize in the production of goods that are intensive in the production inputs they are endowed with in relative abundance: labor and natural resources. The developed countries would specialize in human capital and manufactured capital-intensive activities. Part of the reductions in environmental degradation levels in developed countries and part of the increases in environmental degradation levels in

middle-income countries may reflect this specialization. Environmental regulation in developed countries might further encourage polluting activities to gravitate toward developing countries.

These effects would exaggerate any apparent decline in pollution intensity with rising income along the EKC. In our finite world, the poor countries of today would be unable to find other countries from which to import resource-intensive products as they themselves become wealthy. When the poorer countries apply similar levels of environmental regulation, they would face the more difficult task of abating these activities rather than outsourcing them to other countries. There are no clear solutions to the impact of trade on pollution from the empirical EKC literature.

Some early EKC studies showed that a number of indicators—SO₂ emissions, NO_x, and deforestation—peak at income levels around the current world mean per capita income. A cursory glance at the available econometric estimates might have led one to believe that, given likely future levels of mean income per capita, environmental degradation should decline from this point onward. However, income is not normally distributed but rather very skewed, with much larger numbers of people below mean income per capita than above it. Therefore, it is median income, rather than mean income, that is the relevant variable. Both Selden and Song and Stern and colleagues performed simulations that, assuming that the EKC relationship is valid, showed that global environmental degradation was set to rise for a long time to come. More recent estimates showed that the turning point is higher; therefore, there should be no room for confusion on this issue.

6. *ECONOMETRIC CRITIQUE OF THE EKC*

Econometric criticisms of the EKC concern four main issues: heteroskedasticity, simultaneity, omitted variables bias, and cointegration issues.

Heteroskedasticity may be important in the context of cross-sectional regressions of grouped data. Data for a country are a sum or mean of observations for all of the regions, firms, or households within that country. Smaller residuals may be associated with countries with higher total GDPs and populations. Taking heteroskedasticity into account in the estimation stage seems to significantly improve the goodness of fit of globally aggregated fitted emissions to actual emissions.

Simultaneity may arise because increasing energy use drives increasing income, but increasing income may also increase the demand for energy in consumption activities. In addition, if environmental degradation is sufficiently severe, it may reduce national income, whereas the scale effect implies that increased income causes increased environmental degradation. A Hausman-type statistical test for regressor exogeneity can be used to directly address the simultaneity issue. No evidence of simultaneity has been found using such a test. In any case, simultaneity bias is less serious in models involving integrated variables than in the traditional stationary econometric model. The Granger causality test can be used to test the direction of causality between two variables. In tests of causality between CO₂ emissions and income in various individual countries and regions, the overall pattern that emerges is that causality runs from income to emissions or that there is no significant relationship in developing countries, whereas causality runs from emissions to income in developed countries. However, in each case, the relationship is positive, so that there is no EKC-type effect.

Three lines of evidence suggest that the EKC is an incomplete model and that estimates of the EKC in levels can suffer from significant omitted variables bias: (1) differences between the parameters of the random effects and fixed effects models tested using the Hausman test, (2) differences between the estimated coefficients in different subsamples, and (3) tests for serial correlation. Hausman-type test statistics show a significant difference in the parameter estimates for estimated random effects and fixed effects models in the majority of EKC studies. This indicates that the regressors—the level and square of the logarithm of income per capita—are correlated with the country effects and time effects, indicating that the regressors are likely correlated with omitted variables and that the regression coefficients are biased. As expected, given the Hausman test results, parameter estimates turn out to be dependent on the sample used, with the global and developing country estimates of emissions EKCs for sulfur and carbon showing a turning point at extremely high income levels and the developed country estimates showing a within-sample turning point. In the cases where serial correlation in the residuals was tested, a very high level was found, indicating misspecification in terms of either omitted variables or missing dynamics.

Tests for integrated variables designed for use with panel data indicate that sulfur emissions per capita, carbon emissions per capita, and GDP per capita are

integrated. This means that we can rely only on regression results that exhibit the cointegration property. Results for cointegration are less clear-cut than those for integration in the individual time series. For sulfur emissions in 74 countries from 1960 to 1990, approximately half of the individual country EKC regressions cointegrate, but many of these have parameters with "incorrect signs." Some panel cointegration tests indicate cointegration in all countries, and some accept the non-cointegration hypothesis. But even when cointegration is found, the forms of the EKC relationship vary radically across countries, with many countries having U-shaped EKCs. A common cointegrating vector in all countries is strongly rejected.

7. OTHER EVIDENCE

Dasgupta and colleagues present evidence that environmental improvements are possible in developing countries and that peak levels of environmental degradation will be lower than those in countries that developed earlier. They present data that show declines in various pollutants in developing countries over time. They show that although regulation of pollution increases with income, the greatest increases occur from low- to middle-income levels. In addition, diminishing returns to increased regulation would be expected, although better enforcement at higher income levels would also be expected. There is also informal or decentralized regulation in developing countries. Furthermore, liberalization of developing economies over the past two decades has encouraged more efficient use of inputs and less subsidization of environmentally damaging activities. Multinational companies respond to investor and consumer pressure in their home countries and raise standards in the countries in which they invest. Furthermore, better methods of regulating pollution, such as market instruments, are having an impact even in developing countries. Better information on pollution is available, encouraging government to regulate and empowering local communities. Those who argue that there is no regulatory capacity in developing countries would seem to be wrong.

Much of Dasgupta and colleagues' evidence is from China. Other researchers of environmental and economic developments in China came to similar conclusions. For example, China is adopting European Union standards for pollution emissions from cars with an approximately 8- to 10-year lag. Clearly, China's income per capita is far more than

10 years behind that of Western Europe. Furthermore, China has reduced sulfur emissions and even carbon emissions during recent years. Ambient air pollution has been reduced in several major cities, and with the exception of some encouragement of road transport, the government is making a sustained effort in the direction of environmentally friendly policies and sustainable development.

8. DECOMPOSING EMISSIONS

As an alternative to the EKC, an increasing number of studies carry out decompositions of emissions into the proximate sources of emissions changes described in section II. The usual approach is to use index numbers and detailed sectoral information on fuel use, production, emissions, and so on that, unfortunately, is unavailable for most countries. Econometric decomposition models may require less detailed data. Decomposition models represent the relations between energy use and the environment more explicitly than do EKC models.

The conclusion from the studies conducted so far is that the main means by which emissions of pollutants can be reduced is by time-related technique effects, particularly those directed specifically at emissions reduction, although productivity growth or declining energy intensity has a role to play, particularly in the case of carbon emissions where specific emissions reduction technologies do not yet exist. Estimates of emissions-specific reductions for sulfur range from approximately 20% globally over a 20-year period to 50 to 60% in West Germany and The Netherlands during the 1980s alone.

Although the contributions of structural change in the output mix of the economy and shifts in fuel composition may be important in some countries and at some times, their average effect seems less important quantitatively. Those studies that include developing countries find that technological changes are occurring in both developing and developed countries. Innovations may first be adopted preferentially in higher income countries but seem to be adopted in developing countries with relatively short lags. This is seen, for example, regarding lead in gasoline, where most developed countries had substantially reduced the average lead content of gasoline by the early 1990s but many poorer countries also had low lead contents. Lead content was much more variable at low income levels than at high income levels.

9. CONCLUSIONS

The evidence presented in this article shows that the statistical analysis on which the EKC is based is not robust. There is little evidence for a common inverted U-shaped pathway that countries follow as their incomes rise. There may be an inverted U-shaped relation between urban ambient concentrations of some pollutants and income, although this should be tested with more rigorous time series or panel data methods. It seems unlikely that the EKC is a complete model of emissions or concentrations.

The true form of the emissions-income relationship is likely to be monotonic, but the curve shifts down over time. Some evidence shows that a particular innovation is likely to be adopted preferentially in high-income countries first with a short lag before it is adopted in the majority of poorer countries. However, emissions may be declining simultaneously in low- and high-income countries over time, *ceteris paribus*, although the particular innovations typically adopted at any one time could be different in various countries.

It seems that structural factors on both the input and output sides do play a role in modifying the gross scale effect, although on the whole they are less influential than time-related effects. The income elasticity of emissions is likely to be less than 1.0 but not negative in wealthy countries, as proposed by the EKC hypothesis.

In slower growing economies, emissions-reducing technological change can overcome the scale effect of rising income per capita on emissions. As a result, substantial reductions in sulfur emissions per capita were observed in many developed countries during the past few decades. In faster growing middle-income economies, the effects of rising income overwhelmed the contribution of technological change in reducing emissions.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Economic Growth and Energy • Economic Growth, Liberalization, and the Environment • Energy Ladder in Developing Nations • Environmental Change and Energy • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

- Arrow, K., Bolin, B., Costanza, R., Dasgupta, P., Folke, C., Holling, C. S., Jansson, B.-O., Levin, S., Mäler, K.-G., Perrings, C., and Pimentel, D. (1995). Economic growth, carrying capacity, and the environment. *Science* 268, 520–521.
- Beckerman, W. (1992). Economic growth and the environment: Whose growth? Whose environment? *World Dev.* 20, 481–496.
- Cole, M. A., Rayner, A. J., and Bates, J. M. (1997). The environmental kuznets curve: An empirical analysis. *Environ. Dev. Econ.* 2, 401–416.
- Dasgupta, S., Laplante, B., Wang, H., and Wheeler, D. (2002). Confronting the environmental kuznets curve. *J. Econ. Perspect.* 16, 147–168.
- Grossman, G. M., and Krueger, A. B. (1994). Environmental impacts of a North American Free Trade Agreement. In “The U.S.–Mexico Free Trade Agreement” (P. Garber, Ed.). MIT Press, Cambridge, MA.
- International Bank for Reconstruction and Development. (1992). World Development Report 1992: Development and the Environment. Oxford University Press, New York.
- Kaufmann, R. K., Davidsdottir, B., Garnham, S., and Pauly, P. (1997). The determinants of atmospheric SO₂ concentrations: Reconsidering the environmental Kuznets curve. *Ecol. Econ.* 25, 209–220.
- List, J. A., and Gallet, C. A. (1999). The environmental Kuznets curve: Does one size fit all? *Ecol. Econ.* 31, 409–424.
- Panayotou, T. (1993). “Empirical Tests and Policy Analysis of Environmental Degradation at Different Stages of Economic Development, working paper WP238.” International Labor Office, Geneva, Switzerland.
- Panayotou, T. (1997). Demystifying the environmental Kuznets curve: Turning a black box into a policy tool. *Environ. Dev. Econ.* 2, 465–484.
- Perman, R., and Stern, D. I. (2003). Evidence from panel unit root and cointegration tests that the environmental Kuznets curve does not exist. *Austr. J. Agric. Resource Econ.* 47, 325–347.
- Selden, T. M., and Song, D. (1994). Environmental quality and development: Is there a Kuznets curve for air pollution? *J. Environ. Econ. Environ. Mgmt.* 27, 147–162.
- Shafik, N., and Bandyopadhyay, S. (1992). “Economic Growth and Environmental Quality: Time Series and Cross-Country Evidence,” background paper for World Development Report 1992. The World Bank, Washington, DC.
- Stern, D. I. (1998). Progress on the environmental Kuznets curve? *Environ. Dev. Econ.* 3, 173–196.
- Stern, D. I. (2002). Explaining changes in global sulfur emissions: An econometric decomposition approach. *Ecol. Econ.* 42, 201–220.
- Stern, D. I., and Common, M. S. (2001). Is there an environmental Kuznets curve for sulfur? *J. Environ. Econ. Environ. Mgmt.* 41, 162–178.
- Stern, D. I., Common, M. S., and Barbier, E. B. (1996). Economic growth and environmental degradation: The environmental Kuznets curve and sustainable development. *World Dev.* 24, 1151–1160.
- Torras, M., and Boyce, J. K. (1998). Income, inequality, and pollution: A reassessment of the environmental Kuznets curve. *Ecol. Econ.* 25, 147–160.
- World Commission on Environment and Development. (1987). “Our Common Future.” Oxford University Press, Oxford, UK.

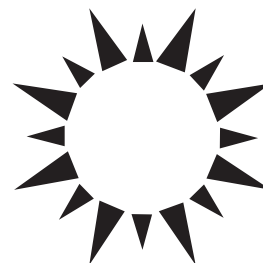
Equity and Distribution in Energy Policy

ADAM ROSE

The Pennsylvania State University
University Park, Pennsylvania, United States

SNORRE KVERNDOKK

Ragnar Frisch Centre for Economic Research
Oslo, Norway



-
1. Introduction
 2. Welfare Analytics
 3. Illustrative Examples
 4. Conclusion

Glossary

competitive equilibrium The point, defined in prices and quantities, at which demand for goods and services equals the supply. All producers and consumers are so small that they cannot influence the prices in the market.

equity A normative concept related to the distribution of rights and goods. Equity seems to be a broader concept than fairness, while justice is sometimes taken to be an umbrella term that incorporates all dimensions of evaluation besides efficiency.

global justice The solution to a comprehensive redistributive scheme of society-wide justice, where global does not necessarily mean international.

libertarianism A theory for the organization of society emphasizing private property rights and individual freedom. The baseline is that individual freedom prevails except where others may be harmed.

market failure When the markets fail to allocate resources efficiently. Reasons for failure are imperfect competition, natural monopolies, externalities, and public goods. The existence of market failures requires market intervention in the form of supplementing, supplanting, or strengthening the market.

Pareto efficiency A situation characterized by no waste of resources. As long as resources can be allocated so that we experience an increase in production/utility of one firm/individual without increasing other firm's/individual's production/utility, we experience a Pareto improvement. An allocation is Pareto optimal if nobody

can be made better off without reducing the well-being/production of others.

social welfare function A representation of the welfare of the society based on individual objectives. The most widely used analytical version is a function of utility of the individuals, in terms of both absolute and relative levels.

utilitarianism A theory for the organization of society. The utilitarian aim is to distribute goods so as to maximize the total utility of members of the society, where "goods" are interpreted broadly to include economic goods, rights, freedom, and political power.

utility possibility frontier A graphical representation of the maximum utility combinations two individuals or groups can attain, given the efficient allocations of finite resources in society.

welfare economics A normative theory for allocating scarce resources and distributing output and income.

The ideal goals of energy policy are to achieve efficiency in production and consumption, as well as an acceptable distribution, of energy goods and services. Efficiency is achieved without policy intervention if there are no imperfections in the market for energy and if all prices reflect the true value of resources. However, it is often necessary to correct prices (e.g., by taxes or subsidies) and to ensure competition (e.g., by anti-trust laws, trade liberalization, or deregulation). Still, even efficient markets can have undesirable distributional consequences for households and producers. A policy that leads to an efficiency gain may not be overall welfare improving if some are made worse off, or it may be viewed as inequitable if it widens the gap between rich and poor. Thus, the distribution of energy goods is also an important part of energy policy.

1. INTRODUCTION

1.1 Overview

Given the prominence of energy in many realms of society, as well as the several unique considerations surrounding its extraction, refinement, transportation, and end use, energy has long been the focus of policy making. This has ranged from minor tweaking of market allocation through taxes and subsidies to outright government control. Energy policy began with concern about market failure due to ordinary and natural monopoly but is now increasingly oriented to reducing externalities, primarily environmental pollution. It has oscillated between *laissez-faire*, government intervention, and deregulation. Allocation issues have dominated but distributional issues have long been important. Moreover, nearly all policies tended to improve the allocation of energy resources also have significant implications for distribution and its assessment in terms of equity or fairness.

This section first briefly summarizes the range of root causes of energy policy and how they relate to distribution and equity. It then offers some basic definitions of equity. Section 2 provides an overview of welfare economics, the theoretical underpinnings for energy policy, and provides some illustrations of key concepts. Section 3 provides some major examples of distribution and equity in energy policy. They are typical but not necessarily representative of the broad range of issues. Rather, they represent the authors' research interests. Finally, Section 4 provides a brief summary.

1.2 Pervasiveness of Energy Policy

1.2.1 Basic Issues

A combination of factors make energy policy pervasive throughout the world. All of these factors have important direct or indirect distributional implications. Energy is a basic necessity for individual, business, and national survival everywhere. In many places, energy is a major commodity of local production and external trade. All stages of the energy cycle are subject to market failures, which typically require policy intervention. Large rents accrue from energy, making it a ripe target, including for purposes of redistribution. Energy is not evenly distributed on a geographic basis, so there is a spatial mismatch between sources and uses. Many energy resources are finite and concern is mounting over the legacy we leave to our

descendants. Energy, especially renewable energy, is a key to sustainable economic development, which appears to require much policy guidance. Energy is also the focus of major technological breakthroughs that often require large R&D funding. Finally, energy has become a major political tool and even a weapon.

1.2.2 Relationship to Distribution and Equity

Because energy is a necessity, there is concern that all consumers have access to it and at an affordable price. This has led to price controls or subsidies for lower income groups in many countries. The survival of business customers, however, is often left to the market, but there are some exceptions for small businesses or those in lagging regions. Other cases in which distributional considerations affect access to energy include the aftermath of an earthquake or terrorist attack, or in the face of an electricity shortage during a war or other crisis, where customers are prioritized for rationing. Interestingly, residential customers often receive higher priority than many businesses.

Several countries, and portions of many others, are highly dependent on energy for their livelihood directly or indirectly. Some, such as Appalachia, are relatively poor, while others such as Texas or the Middle East, are relatively wealthy. In essence, every major energy producing region has an exportable surplus that serves as a stimulus to its economy, though the competitiveness of the relevant energy market has significant effects on how much income is derived from this activity and how it is shared between the various factors of production. The situation has also been especially ripe for protectionist policies, especially at the national level, that often favor one segment of society over another. Environmental, technology, and trade policies have and will continue to cause major shifts in the relative economic vitality of these regions.

The uneven geographic distribution of energy has been a root cause of concentration of energy production, political tensions, and redistributional issues. This is most vivid in the international oil arena and several portions of the world where oil has been used as a bargaining chip or even a weapon. A massive redistribution of wealth has taken place in the more obvious case of the OPEC cartel, but also in the more subtle ways because comparative advantage is highly dependent on relative resource endowments.

One of the most prominent factors that require energy policy tension relates to market failure, the

inability of markets to allocate resources efficiently without policy intervention (including market strengthening). All four major categories of failure discussed below have strong implications for the distributional consequences of energy policy. With respect to externalities, defined as a by-products of activities not reflected in the market price, fossil energy use is probably the greatest single source of air pollution, and the production and transportation of fossil energy are major sources of water pollution. Moreover, these types of pollution have transboundary features, which means one geographic area that benefits from energy production or use imposes problems on other regions. Corrective measures, such as pollution taxes or tradeable permits, involve large sums of money. The imposition of a tax may be on a producer, but through economic interdependencies and tax shifting, its incidence is often spread much more broadly. The relative progressiveness and regressiveness of these policy instruments is a long-standing issue in terms of socioeconomic entities (income classes, racial/ethnic groups, age groups). At international levels, this has become a serious issue in terms of burden-sharing between industrialized and developing countries.

The major consideration in the area of public goods is what is often referred to as the subset of common property resources, where the good is nondivisible but does have nonexcludability/nonappropriability (of payment) characteristics. One example of this subset may be oil and gas fields. Conventional markets lead to degenerate competition in terms of over-production and plummeting prices. Solutions such as unitization involve difficult decisions on the distribution of rents.

Ordinary monopoly, or market power in general, is one of the most visible market failures in the energy area. This includes U.S. oil companies, state-owned hydropower in many countries, and the world crude oil market. Not only does this cause a misallocation of resources but also a transfer of wealth in the direction of the monopolist. In fact, many public policies (e.g., excess profits taxes) stem from distributional rather than efficiency concerns of this type of market failure.

Natural monopolies arise because scale economies and other factors imply the efficiency of single providers. In addition to the conundrum of attempting to find efficient prices in this context, there is the concern that deviations from competitive levels will come at the expense of consumers. Over the past decade, there has been a reassessment of the extent to which natural monopolies do in fact exist. Deregulation or restructuring have in some cases been

poorly designed and have had dramatic distributional consequences, most vividly illustrated by the crisis in California.

Not only does the market often fail to allocate resources efficiently but even more often it fails to distribute them, on the proceeds from them, fairly. Essentially, the market is blind to equity, and extra-market means are usually necessary to steer it in the right direction.

Large rents (profits) accrue from energy production making it a prime target for policy machinations to capture them. These rents are often viewed as indicating monopoly profits or as unfair; however, in many cases, they relate to scarcity rents, returns to high risk in mineral extraction, or to superior quality deposits. The existence of large rents has led to aggressive policies or outright regulation. Also, in the United States and other countries, vast amounts of energy resources are located on public lands, so that the distributional issues are in the forefront. Public officials must often make hard choices between constituencies. This typically has to be accomplished in the absence of market signals, though an increasing number of market institutions have been developed to address the issue (e.g., bonus-lease bidding).

Many energy resources are finite, and there is concern about utilizing them today and leaving fewer or higher cost deposits for future generations (intergenerational equity). This has also caused energy policy to become intertwined with sustainability issues, which have both normative and positive (e.g., public participation) distributional consequences. This has also shifted the emphasis to renewable energy as discussed later.

Energy is key to economic development, especially of the sustainable type. Market forces alone will not promote development, so this is another, and very broad sweeping, area of policy, replete with distributional consequences and equity considerations. Economic development is not simply pursued to raise the absolute standard of living but also to help close the economic gap between countries.

The form of energy has major distributional implications. Fossil fuels are geographically concentrated and subject to natural monopoly considerations, especially in their transportation. Renewable energy, such as solar, wind, dry-rock geothermal, and biomass, are relatively ubiquitous. Hence, they have the potential for decentralizing energy production and even eliminating a great deal of the transportation/transmission needs. These alternate energy resources thus have the ability to lessen the impact

of OPEC and large producers of energy in general and thus to redistribute wealth.

Energy is the subject of major technological advances. This arises because fuels and conversion lend themselves to technical refinement. Major breakthroughs can cause major geographic and socioeconomic redistributions. Much of the research is undertaken or subsidized by government. Issues about concerning industry favoritism. A “level playing field” is absent here, as well as in most areas of energy policy.

Energy is a political tool. Wealth has always been a major basis for political power, and the concentration of energy deposits, and the nature of energy refining and transportation, have been a major source of strength and tension. The Arab oil embargo is a vivid example, and many contend that the Gulf War, as well as the invasion of Iraq, have been motivated by the politics of oil. The situation is complicated in many parts of the world as a battle between the haves and have-nots or as a remnant of colonialism, one of the ultimate distributional issues.

1.3 Definition of Equity

The concepts of efficiency and equity are central in policy analyses. While Pareto optimality, a situation characterized by no waste of resources, is an accepted efficiency principle, there is not a consensus on a best equity principle. As a result, most economic analyses, also those related to energy, have concentrated on efficiency problems, and equity often plays a secondary role in economic policy making, even if it will always play an important role as a principle of all social interactions.

Distinguishing between the related concepts of equity, fairness, and justice may in general be difficult, and we do not attempt to make extensive distinctions. However, fairness often has a specific meaning in economic theory and philosophical theories of justice. One example is to define fairness as nonenvy. Equity seems to be a broader concept than fairness, while justice is sometimes taken to be an umbrella term, incorporating all dimensions of evaluation besides efficiency.

2. WELFARE ANALYTICS

2.1 Welfare Maximization

One branch of mainstream economics studies equity by formalizing the concept of a social welfare

function (SWF), in which it was possible to separate efficiency and equity considerations though tradeoffs between the two objectives still prevail. We can illustrate this by a simple diagram containing two major constructs. First is a utility possibility frontier, or UPF, which traces the maximum combination (m) of the utility (U) two individuals or groups (1,2) can attain, given all the efficient allocations of finite resources in society (see curve $U_m^1 U_m^2$ in Fig. 1). Interestingly, all the points on the UPF are Pareto optimal, (i.e., one person cannot be made better off without another person being made worse off). Thus, points such as C and D in Fig. 1 are not Pareto optimal. Note that there are an infinite number of Pareto optimal points on the UPF. We can be assured, under ideal conditions, that the market will lead to a point on the UPF, thus illustrating the first fundamental theorem of welfare economics. But from the standpoint of efficiency alone, the points on the UPF cannot be ranked.

However, if we introduce equity, then we have a way of preferring some to others. In fact, that is the essence of the definition of a social welfare function—a complete ranking of all states (resource allocations and product or income distributions) of the economy. Although many alternative equity concepts exist, the one most prevalently used in welfare economics is “vertical” equity, which holds that equity increases as utility (or income) disparities between individuals decrease. Generally, equity is

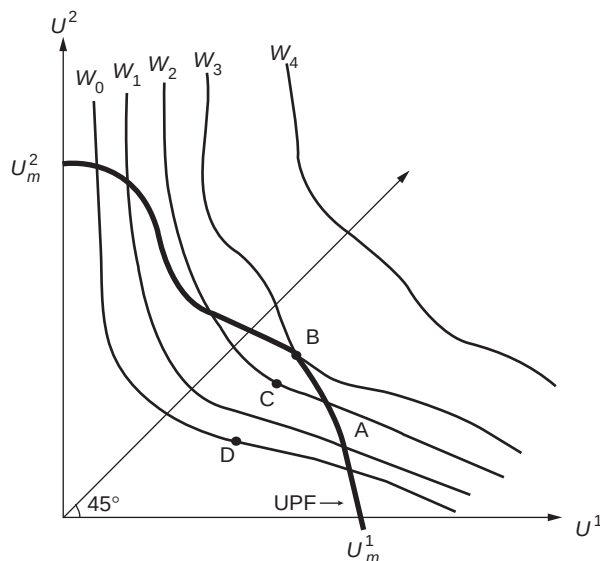


FIGURE 1 Welfare maximization and efficiency-equity trade-offs.

enhanced as joint utility allocations in Fig. 1 move closer to the 45° line.

A social welfare function that embodies both efficiency and equity can be superimposed on Fig. 1. The welfare contours (W_0, W_1, W_2, W_3, W_4) are analogous to indifference curves (i.e., they are horizontal slices of a three-dimensional surface). The shape of these contours reflects equity; the more bowed toward the center, the greater the emphasis on equity.

The welfare optimum is given by point B. Again, there is no guarantee that a market economy will lead to this outcome, and, in fact, given the infinity of possible Pareto optimal outcomes, it is likely that it will not. Moreover, while theoretically society can move from a competitive equilibrium that is not a maximum welfare outcome, such as A, to point B through the use of various types of transfers, there may be some limitations in practice, as discussed in the following section.

For some, one major justification for downplaying equity is the efficiency-equity tradeoff. This refers to the possibility that transfers to promote equity will create a work disincentive to the transferring party (usually the well-to-do), with the ensuing inefficiency limiting society to a lower level of welfare even when utilities (or incomes) are equal or made more equal. Still, this is an empirical question, though most studies indicate the disincentive effect is rather small.

We can use Fig. 1 to illustrate this potential problem. Previously we noted the ideal transfer would move society from point A to point B. However, if the disincentive effect exists, and is weak, the move would be from point A to point C, whereby there is an efficiency-equity tradeoff but an unequivocal welfare gain. But it is also possible that the disincentive is strong and will result in a move from point A to point D, whereby overall welfare is even lower than in the original state. The threat of moving to point D, and even the possibility of moving to point C (note the discussion of benefit cost analysis below), causes many economists to be nervous about pursuing equity goals. We can be more certain of achieving equity with the minimum of efficiency loss if we utilize lump sum transfers, or if we avoid price-distorting policies. The difficulty is that there are few lump sum transfers (i.e., transfers that are not based on effort).

2.2 Social Justice Theories

So far we have discussed distribution within the framework of a social welfare function, which is the

mainstream approach in economic theory. These welfaristic theories (meaning theories that focus on welfare outcomes) are in the tradition of utilitarianism. We will now take a brief look at this theory, and also consider some nonwelfaristic approaches.

A theory of justice is, as theories of rationality and ethics, a normative theory. Such a theory has two aspects. First, it will regulate individual rights (and duties); second, it will propose or evaluate a distribution of goods (and burdens). Different theories of justice may weight these two aspects differently. They are, for instance given equal weight by utilitarians, while John Rawls gives political freedom and rights a lexicographical priority over economic distribution.

In the following, we will consider philosophical theories of justice (or global justice theories)—that is, theories that are centrally designed for the whole society and are intended to compensate people for various sorts of bad luck that may result in low levels of income. This means that we will not consider theories for decentralized distribution decisions (or local justice theories), such as who shall perform military service, who shall receive organs for transplantation, and so on, but theories that provides suggestions to society-wide problems such as income distributions or the social security system. These theories are also relevant for energy policy, as we will see in Section 3.

2.2.1 Utilitarianism

The utilitarian aim is to distribute goods so as to maximize the total utility of members of the society, where “goods” are interpreted broadly to include economic goods, rights, freedom, and political power. A utilitarian welfare function is usually defined as the sum of the utility of all the members (i.e., all individuals have an equal weight). One problem with this function as well as for other forms of SWFs is the interpersonal comparison of utility, which is not a straightforward issue. We will, however, not examine these problems in this article.

Even though utilitarianism does not explicitly address equity, its welfare maximization objective does have distributional implications. Different versions of the theory exist. One important distinction is the difference between act-utilitarianism and rule-utilitarianism. The former enjoins us to perform the act that will maximize the utility on every single occasion, while the latter tells us to act so that we will maximize total utility over time. These two forms of utilitarianism may conflict, as many actions

provide immediate pleasure but long-run costs. For example, a high level of fossil fuel consumption today may maximize current welfare, but due to the accumulation of greenhouse gases it will cause a reduction of maximum attainable welfare in the future.

2.2.2 Rawlsian Theory

The theory of John Rawls can basically be seen as a critique of utilitarianism. According to Rawls, utilitarianism has no respect for individuals. A person is not seen as valuable and worth protecting on his or her own right. A classical but grotesque example is if you kill a healthy person and use his or her organs to save the life of five other human beings. This may increase the utility of the society, but it does not respect individual rights. Rawls argues that a theory of justice should respect the individuals as ends in themselves.

The methodological starting point of Rawls theory is the original position. Most persons would argue for what is just or unjust depending on their own position in the society. Therefore, as a starting point to decide the basic structure of the society, which according to Rawls is the primary subject of justice, we have to think about a hypothetical and idealized world where all individuals sit behind a veil of ignorance; the original position. They do not know their abilities, sex, race, or position in the society. All they know is that they are going to live in the society about which they are going to decide the basic structure. In this hypothetical situation, Rawls argues that they will agree on certain principles:

First principle. Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.

Second principle. Social and economic inequalities are to be arranged so that they are both (1) to the greatest benefit of the least advantaged and (2) attached to office and positions open to all under conditions of fair equality of opportunities.

The first principle, the principle of greatest equal liberty, is about individual rights. As this principle has a lexicographic priority, none of the basic liberties should be traded against material advancement. However, it is the first part of the second principle, the difference principle, that is mostly referred to. One common misinterpretation in economics is that Rawls argues for maximizing the utility of the least advantaged. However, Rawls theory is not utilitarian, rather an alternative theory

and Rawls argues for distributing “primary goods”—that is, goods everybody needs to realize his or her plans of life, independent of what the plan is. While the first principle distributes one subset of the total set of primary goods, the basic liberties, the difference principle distributes another subset including wealth, income, power, and authority. The last part of the second principle, the principle of fair equality of opportunity, requires that we go beyond formal equality of opportunity to insure that persons with similar skills, abilities, and motivation enjoy equal opportunities. Finally it should be noted that Rawls’s theory is a theory of intragenerational distributive justice, and the principles cannot easily be extended to intergenerational concerns.

One element of Rawls’s theory of justice, which may be particularly relevant when discussing energy policy, is what he defines to be morally relevant and morally arbitrary. The distribution of goods should not be based on morally arbitrary components. To illustrate, consider two elements that are often claimed to contribute to economic growth. One is the human-cooperative activity, and the other is the natural component. The second element is components present on the earth’s surface, which are not due to human production. One example is natural resources or assets that are arbitrarily distributed on the earth’s surface. In the literature, these are compared to the distribution of natural endowments, defined as such contingencies that are inescapable and present from birth, and may contribute to material advancement, or economic growth because of the implications for human cooperative activity. Examples are nationality, genetic endowments, race, gender, and social class. According to Rawls, natural endowments are “neither just or unjust; nor is it unjust that men are born into society at any particular position. These are simply natural facts. What is just or unjust is the way that institutions deal with these facts.” The distribution of natural assets and endowments are, in general, assumed to be morally arbitrary and unacceptable as standards for distribution, since they are beyond the control of human beings. One interpretation is that the geographical distribution of natural resources like oil and gas are simply facts, in the same way as individual talent is a fact. This distribution is, therefore, not just or unjust. However, the income generated from these resources should be considered when discussing justice. Thus, this may be an argument for taxing this income as a part of a distributional policy.

2.2.3 *Libertarian Theory*

Probably the most widely used decision rule in economic theory is the Pareto criterion, which is most often applied to efficiency and political feasibility (e.g., in bargaining theory). It has also been applied to equity, for instance by a common-sense conception of justice. In libertarian theory, the baseline is that individual freedom prevails except where others may be harmed. Thus, this ethic is in the same line as Pareto superiority saying that there is an improvement in welfare if one or more persons are made better off due to change in resource use as long as the other persons are at least as well off as before. A strict interpretation of libertarian theory is that an act is only immoral if anyone is worse off because of it.

Robert Nozick's theory is the best-known statement of libertarian thought. The theory can be summarized in three principles: justice in appropriation, justice in transfer, and justice in rectification. A distribution of goods is just if it is the end result of an unbroken chain of just transfers, beginning from a just original appropriation. If these conditions are not satisfied, justice in rectification requires that we should establish the distribution that would have occurred if all unjust links in the chain had been replaced by just ones.

The first principle is essentially a finder's keepers principle, where the basic idea is that anyone has the right to appropriate, exploit, and enjoy the fruits of any unowned piece of nature. The principle of just transfers says that the outcome of any voluntary transaction between two or more individuals is just. It is assumed that there is no coercion. If individuals agree on a contract that will benefit all, there is no reason to stop the contract. The only reason to invalidate the contract is if anyone uses its power to make the nonagreement state worse for other parties than it would otherwise have been. The last principle is probably the main weakness of the theory, as identifying the point in time where the earliest violation occurred and thereafter the counterfactual chain of just transfers may be rather indeterminate.

2.3 Link to Energy Policy

Despite the prevalence of distributional and equity considerations in energy policy, much less attention has been paid to them in contrast to the attention to efficiency, policy instrument design, and various industrial organization topics. Several reasons exist for this situation but paramount is the fact that distribution nearly always, and equity by its very

essence, are in the realm of normative economics. Energy policy, of course, is also inherently normative, but most analysts feel more comfortable promoting an objective such as economic efficiency, for which the definition has almost universal consensus. Moreover, most interpretations of efficiency call for making all parties better off (or at least not making anyone worse off). On the other hand, there is no consensus on the best definition of equity, as illustrated by the different theories described earlier. Furthermore, the promotion of this objective (in a static context, as opposed to a case like economic development) is often viewed as a zero-sum game, where improvements for one entity must emanate from another one.

Of course, in the real world, few policies do not at least have some adverse effect on some parties. However, this is finessed by benefit-cost analysis (BCA), which evaluates only the overall net gain (and ignores the fact that some incur net cost). In essence, BCA stipulates that benefits and costs be evaluated equally no matter to whom they accrue. Some attempts have been made over the years, however, to introduce distributional issues into BCA by suggesting winners compensate losers and by assigning weights to individuals. However, when choosing a SWF for doing a BCA analysis, ethical views and other specific assumptions are made at the same time. Thus, another approach is to provide sufficient welfare indicators (i.e., information that can be used as input in a BCA analysis). Instead of providing information that is the output from a certain BCA analysis, the analyst can provide factual information about different population groups (such as their social state, size, or characteristics, as well as their changes in income and access to public goods).

As will be shown later, in the primary theoretical underpinnings of economic policy, welfare economics, equity receives considerable attention. Moreover, direct redistribution is viewed as a sound and even ideal, policy option to promote equity, while not adversely affecting efficiency. The classic example is the debate over natural gas well-head price ceilings in the United States. In 1953, a U.S. court ruled that natural gas extraction had monopoly elements and imposed price caps at the wellhead. Opposition to this policy increased over the next two decades because of the bureaucratic backlog of legitimate cost adjustments and instances of major gas shortages. The counter-argument, however, began to hinge less on the monopoly condition and more on concern that, if wellhead activity were deregulated, the ensuing fly-up of gas prices would have an

especially adverse effect on the poor. Thus, the debate hinged on the relative merits of two approaches: (1) maintaining price caps to make low-cost natural gas available to the needy, while distorting market signals potentially causing shortages, and (2) deregulating gas prices and providing some form of tax relief to low income groups. Natural gas wellhead prices were in fact deregulated beginning in 1978. No major legislation was passed to provide subsidies or tax relief for low income groups, though attempts were made to lessen the potential fly-up by a slow phasing in of deregulation, and hence market price increases. Fortunately, large price increases did not take place, primarily due to the drop in world oil prices and economic recession of the early 1980s. This is an example showing that efficiency and equity can be in conflict, though external circumstances can modify the tension. Also, ideal policies do not always take place (an example of political failure analogous to market failure), and distributional issues may have a lower priority than overall efficiency.

3. ILLUSTRATIVE EXAMPLES

3.1 Carbon Taxes

A number of studies have examined the income distribution impacts of carbon taxes or carbon emission permits. The complexity and international nature of the issue is one reason, an increasing number of analyses are performed in a general equilibrium context.

Gbadebo Oladosu and Adam Rose begin an examination of this issue by summarizing the three special features most emphasized to distinguish the impacts of these policies in contrast to the incidence of taxes in general. First, although the initial focus is on a few but very prominent sectors that emit carbon (coal/oil/gas extraction, transportation, and refining), the fundamental role of these products, however, means that carbon reduction policies will eventually ripple throughout the economy, with possibly surprising outcomes. This is one of the major reasons computable general equilibrium models are used (i.e., numerical simulation models based on behavioral responses by producers and consumers to price signals in interconnected markets).

Second, fossil energy products and most energy-intensive processed goods (food, housing, automobiles) are necessities, making it relatively more difficult to substitute away from them. Spending on necessities is inversely related to income and,

hence, all other things being equal, carbon taxes would lean toward being regressive in partial equilibrium terms.

Third, unlike most existing taxes, carbon taxes are not aimed primarily at raising revenue. Moreover, they do not create a distortion in the price system but are intended to correct one. The factors have important implications for the disposition of carbon tax revenues (or revenues from the auction of carbon emission permits), including the possibility of using carbon tax revenues for tax relief that promises to reduce the distortionary nature of the preexisting tax system. This revenue recycling can take a number of forms (reductions in personal income taxes, corporation income taxes, etc.), with different distributional impacts (both regressive and progressive). Again, however, the final impacts of these alternatives are not *a priori* obvious when one allows for general equilibrium considerations.

Overall, a large number of other factors, both unique to carbon taxation and applicable to tax policy in general, can have a major bearing on the relative unevenness of impacts. It is also important to note several factors that affect the size of the aggregate impact, since it will also have a bearing on the degree to which the baseline income distribution changes. Of course, the size of the aggregate impact can affect the distribution of impacts in highly nonlinear models or where such factors as income elasticities (i.e., the percentage change in demand following from a 1% increase in income of demand vary strongly across income groups). Major factors include the following:

1. The energy-intensity of the economy. The higher this factor, the larger the overall impact and the more profound income inequalities of impacts can become in relation to the baseline.
2. The magnitude of the carbon tax or emission permit price, for reasons similar to factor 1.
3. The unit upon which the tax is based (e.g., energy equivalent, carbon emissions, or carbon content). This bears on the relative bluntness or precision of the policy and hence its cost-effectiveness and overall impacts.
4. The narrowness or breadth of products or entities on which the tax is imposed. For example, a tax on energy as a whole may be regressive but a tax on motor fuels may be progressive.
5. The point of initial imposition of the tax (i.e., upstream suppliers of energy or downstream users of energy). This is related to whether a tax on production or consumption is specified. Progressivity

is likely to be greater if imposed upstream and if tax shifting is difficult.

6. The ability to shift the tax forward onto customers or backward onto factors of production. In terms of the latter, shifts on to labor are likely to be more regressive than shifts onto capital.

7. The extent of factor mobility, as in the regressivity of transitional effects on coal miners having to find jobs in other industries.

8. The extent to which general equilibrium effects are taken into account. While this typically exacerbates the level of impacts, it mutes their unevenness as they are spread out over the economy as a whole.

9. The extent of production/income distribution/consumption interactions. For example, a large decrease in coal production may have a disproportionate effect on income of high-wage unionized miners, but the decrease in their consumption may be for products that are characterized by a predominant number of low-wage earners.

10. The extent to which dynamic effects are taken into account (e.g., with respect to savings and investment). The income distribution has an effect on economic growth, which in turn affects future distribution. Here progressivity is often thought to have a detrimental effect on future growth, though the effect on future income distribution is ambiguous.

11. The use of annual income versus lifetime income as a reference base (e.g., as proxied by consumption). The latter is the more appropriate measure given the long-run nature of the issue.

12. The extent to which demographic considerations pertaining to household composition are taken into account; related to this is the demarcation of income groups, especially at the highest and lowest levels.

13. The type of revenue recycling (including lump-sum transfers) and in contrast to alternatives such as budget deficit reduction and individual and corporate tax relief. The latter is usually considered the most regressive.

14. Asset market considerations, such as the extent to which financial returns are affected. This factor bears primarily on higher income groups.

15. The degree to which the impacts result in unemployment. Those already in lower income groups are less able to withstand the shocks of both temporary and long-term unemployment.

16. Basic parameters and assumptions of the analytical model (especially price elasticities of demand and supply, elasticities of substitution with respect to inputs and imports, market structure, etc.).

For example, the higher the elasticities of substitution between energy and other factors, the less the negative impacts. Also, the greater the variation in price and income elasticities of demand, the greater the potential progressivity or regressivity.

If we try to summarize studies of carbon taxes on households, we find that the distributional effects of a carbon tax can often be regressive unless the tax revenues are used either directly or indirectly in favor of the low-income groups. Consequently, there seems to be a conflict between efficiency and equity goals, although some analyses are less conclusive. Due to the many causal factors involved and the relative lack of analyses on this topic, it is difficult to draw conclusions about country differences. The same applies to the impact of model type or size of the carbon tax.

3.2 OPEC and Environmental Policy

A carbon tax also has distributional consequences for producers of energy. One important issue in the international context is how a carbon tax affects the producers of fossil fuels. In the ongoing international climate negotiations, the possible impacts on fossil fuel producers and producing countries have played an important role, and both coal producers and OPEC countries have shown a very reluctant position in the negotiations.

A study by Elin Berg, Snorre Kverndokk, and Knut Einar Rosendahl shows the impacts on fossil fuel producers and especially oil producers, as well as the distribution of burden between producers and consumers, of a global carbon tax. This is analyzed by an intertemporal equilibrium model for fossil fuels, where the dynamic aspects of resource extraction is taken into account. According to their results, a global carbon tax of \$10 per barrel will hurt OPEC more than other producers. The reason is that the cartel finds it optimal to reduce its production in order to maintain a high price level when the tax is introduced. Hence, the crude oil price is almost unchanged initially, and, in the first 40 years, the tax is almost completely shifted to the consumers. As the producer price of oil grows slower compared to the reference case, non-OPEC accelerates its oil production. Thus, while the market structure is important in explaining OPEC's behavior in the model, the dynamic aspect explains the behavior of the fringe. The oil wealth of non-OPEC is reduced by merely 8%, whereas OPEC's wealth is reduced by more than 20%. Thus, if we do not consider the environmental

benefits, the main losers under the carbon tax regime will be the consumers and OPEC, while the tax revenue is collected by the governments in the consuming countries.

The importance of market power is demonstrated by the results of a competitive model—that is, a model where all consumers and producers behave competitively (no market power). Here low-cost producers (OPEC countries) supply significantly more initially than in the model with market power, which implies a much lower oil price, making oil production unprofitable for high-cost producers (non-OPEC) in the first period. In this situation, a carbon tax will have a relatively much higher impact on the non-OPEC oil wealth compared to that noted earlier (i.e., about 40%), as the low-cost producers will no longer act to maintain the price. However, this reduction is less dramatic for non-OPEC than the original change in market structure, which reduces the oil wealth by more than 70%. For the OPEC countries the relative loss due to the carbon tax is about the same as mentioned previously.

The results may explain why OPEC countries are reluctant to introduce global carbon taxes or other restrictions on carbon emissions. In contrast, other oil producers seem to have less to fear if such an agreement were reached, and this seems to be partly consistent with their positions toward carbon restrictions. Therefore, economic factors may be important in explaining attitudes of oil producers in global warming negotiations.

3.3 Distribution of Tradable Emission Permits

Several alternative criteria have been put forth for the interregional and international analysis of the equity implications of environmental policy. Nine equity principles, a general operational rule emanating from each, and a corresponding rule applicable to the allocation of tradable emission permits at the regional level are presented in Table I.

Of course, there are other levels of application of equity criteria, including sectors and households. In some ways, these units are superior to regions, because they address the issue of relative impacts in terms of welfare of the individual (the basic unit in a democracy). Two reasons are offered for orienting the analysis to the regional level, however. First, are limitations on the availability of data for production and household delineations according to welfare

levels (e.g., income brackets or profit rates). Second, political realities indicate that regions are important. This is reinforced by the “regional” character of legislative bodies and several major economic sectors (e.g., oil and gas production, steel), which have translated into regional influence in previous environmental policy making, such as U.S. acid rain legislation. Several other important regional dimensions of the issue relating to natural and social science considerations.

To be operational, an equity criterion must be applied to a reference base, essentially a metric or index, which itself has no ethical content. Consider the following six permit allocation formulas (consisting of an equity criterion and an associated reference base):

1. Sovereignty (emissions based)
2. Egalitarian (population based)
3. Economic activity (GRP based)
4. Ability to pay (inverse-GRP based)
5. Sovereignty (energy-use based)
6. Sovereignty (energy-production based)

Note that there is not a one-to-one correspondence between criteria and reference bases. For example, the sovereignty criterion can be implemented according to more than one reference base, and gross regional product (GRP) can serve as a reference base for more than one criterion. Reference base distinctions are important for several reasons. First, as in the case of three bases applied to the sovereignty criterion, they can reflect alternative energy policy positions. For example, Formula 5 simulates a downstream administered program on fossil energy end users, while Formula 6 simulates an upstream program on fossil energy producers. Most important, the welfare implications of various reference bases for the same equity criterion will differ as well.

The results of the application of the six equity formulas to the sharing of the Kyoto protocol target within United States by Adam Rose and Zhong Zhang are presented in Fig. 2. Some interesting results arise.

First, for some of the formulas (e.g., Sovereignty/Energy Use), there is very little variation of welfare gains across regions. Second, for some regions (e.g., Southwest, Mid-Atlantic, New England), there is little difference across equity formulas. For some regions, the equity formula does matter greatly (e.g., south central, north central). Note also the results for some equity criteria differ greatly from their application in the international domain. For example, the egalitarian (per capita) criterion leads

TABLE I
Alternative Equity Criteria for Greenhouse Gas Mitigation Policy

Criterion	Basic definition	General operational rule	Operational rule for CO ₂ permits
Allocation based			
Sovereignty	All regions have an equal right to pollute and to be protected from pollution	Cut back emissions in a proportional manner across all regions	Distribute permits in proportion to emissions, energy-use, land area, etc.
Egalitarian	All people have an equal right to pollute or to be protected from pollution	Allow emissions in proportion to population	Distribute permits in proportion to population
Ability to pay	Mitigation costs should vary directly with regional economic well-being	Richer regions should shoulder a higher proportion of gross cost of abatement ^a	Distribute permits inversely to GRP
Econ Activity	All regions should be allowed to maintain their standard of living	Richer regions should not be penalized	Distribute permits in proportion to GRP
Outcome based			
Horizontal	All regions should be treated equally in terms of changes in welfare	Equalize net welfare change across regions (net loss as proportion of GRP equal for each region) ^b	Distribute permits to equalize net welfare change (net loss as proportion of GRP equal for each region) ^b
Vertical	Welfare gains should vary inversely with regional economic well-being; welfare losses should vary directly with GRP	Progressively share net welfare change across regions (net loss proportions directly correlated with per capita GRP) ^b	Progressively distribute permits (net loss proportions directly correlated with per capita GRP) ^b
Compensation	No region should be made worse off	Compensate net losing regions	Distribute permits so no region suffers a net loss of welfare
Process based			
Rawl's Maximum	The welfare of the worst-off region should be maximized	Maximize the net benefit to the poorest regions	Distribute largest proportion of net welfare gain to poorest regions
Consensus	The negotiation prices is fair	Seek a political solution promoting stability	Distribute permits in a manner that satisfies the (power weighted) majority of regions

^aGross cost refers to mitigation cost only and does not include benefits or permit transactions.

^bNet welfare change is equal to the sum of mitigation costs + permit sales revenues – permit purchase expenditures.

to the relatively greatest cost burden being incurred by one of the regions of the United States with the lowest per capita income (south central states). This is just the opposite of the result in the international domain, where countries such as China and India stand to gain the most from application of this criterion.

4. CONCLUSION

Energy is a necessity in both consumption and production. Thus, energy policies may have large consequences on income and profit. The ideal goals of this policy are to achieve efficiency, as well as an acceptable distribution of energy goods and services. However, even efficient markets can have undesirable

distributional consequences for households and producers. A policy that leads to an efficiency gain may not be overall welfare improving if some are made worse off, or it may be viewed as inequitable if it widens the gap between rich and poor. This makes the distribution of energy goods an important part of energy policy.

This article briefly summarizes the range of root causes of energy policy and how they relate to distribution and equity, and it provides an overview of welfare economics and social justice theories, the theoretical underpinnings for energy policy. It also provides some illustrations. As our survey shows, there is no single best equity principle nor one that is universally agreed upon. Therefore, it is difficult to develop unambiguously good policy proposals with respect to the distributional effects of energy policy

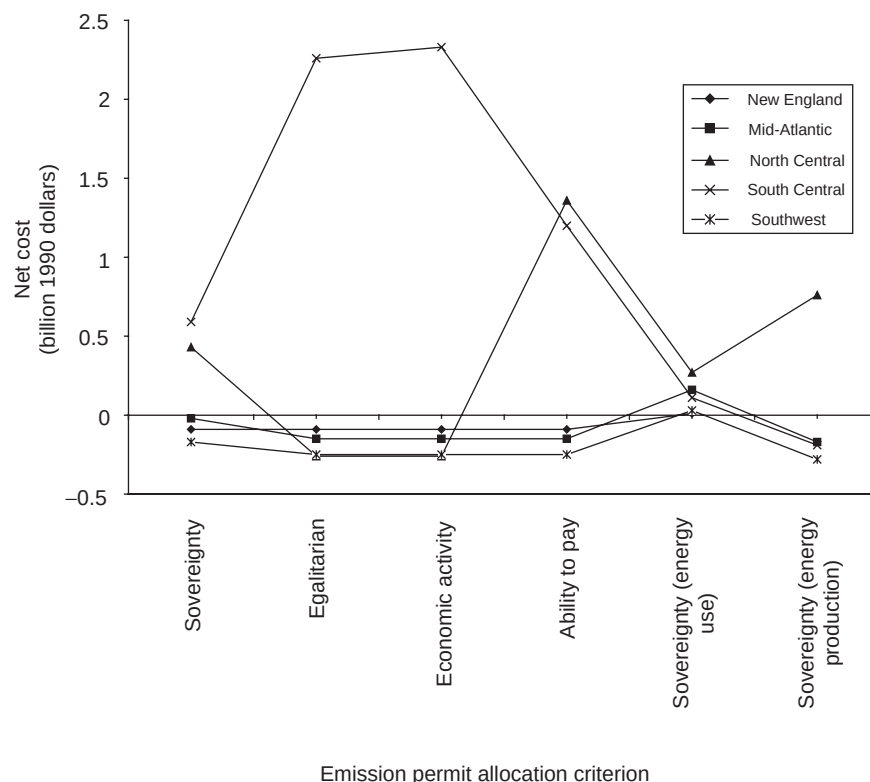


FIGURE 2 Net cost of meeting U.S. Kyoto protocol commitments in selected regions under alternative permit allocation criteria.

in particular. However, policymakers should be aware of the distributional effects before a policy is implemented in general.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, Energy Policy in • Cost-Benefit Analysis Applied to Energy • Development and Energy, Overview • Economic Growth, Liberalization, and the Environment • Electric Power Reform: Social and Environmental Issues • Energy Ladder in Developing Nations • Environmental Injustices of Energy Facilities

Further Reading

- Adelman, M. A. (1972). "The World Petroleum Market." Johns Hopkins University Press, Baltimore, MD.
- Berg, E., Kverndokk, S., and Rosendahl, K. E. (1997). Market power, international CO₂ taxation and petroleum wealth. *Energy J.* 18(4), 33–71.
- Brennan, T., Palmer, K., Kopp, R., Krupnick, A., Stagliano, V., and Burtraw, D. (1996). "A Shock to the System: Restructuring

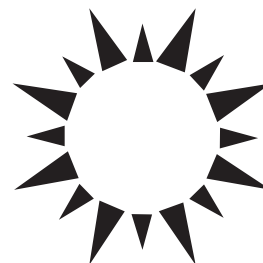
- American's Electricity Industry." Resources for the Future, Washington, DC.
- Bye, B., Kverndokk, S., and Rosendahl, K. E. (2002). Mitigation costs, distributional effects, and ancillary benefits of carbon policies in the Nordic countries, the UK, and Ireland. *Mitigation and Adaptation Strategies for Global Change* 7(4), 339–366.
- Daniziger, S., Haveman, R., and Plotnick, R. (1986). How income transfer programs affect work, savings, and income distribution. *J. Economic Literature* 19, 1–36.
- Gordon, R. L. (1985). Energy policy issues. In "Economics of the Mineral Industries" (W. Vogely, ed.), 4th ed. AIME, New York.
- Goulder, L. (1995). Effects of carbon taxes in an economy with prior tax distortions: An intertemporal general equilibrium analysis. *J. Environmental Economics and Management* 29, 271–295.
- Harrison, D. (1994). "The Distributive Effects of Economic Instruments for Environmental Policy." OECD, Paris.
- McAvoy, P., and Pindyck, R. (1975). "The Economics of the Natural Gas Shortage (1960–1980)." North-Holland, Amsterdam.
- Munasinghe, M. (1993). "Energy Policy Analysis and Modeling." World Bank, Washington, DC.
- Oladosu, G., and Rose, A. (2003). "The Income Distribution Impacts of Climate Change Mitigation Policy in the Susquehanna River Basin Economy." Paper presented at the 14th Annual Conference of the European Association of Environmental and Resource Economics, Bilbao, Spain.
- Rawls, J. (1971). "A Theory of Justice." Oxford University Press, Oxford.

- Roemer, J. (1996). "Theories of Distributive Justice." Harvard University Press, Cambridge, MA.
- Rose, A., Stevens, B., and Davis, G. (1988). "Natural Resource Policy and Income Distribution." Johns Hopkins University Press, Baltimore, MD.
- Rose, A., and Zhang, Z. X. (2003). Interregional burden-sharing of greenhouse gas mitigation in the United States. *Mitigation and Adaptation Strategies for Global Change* (forthcoming).

Ethanol Fuel

CHARLES E. WYMAN

Dartmouth College
Hanover, New Hampshire, United States



1. Use as a Fuel
2. Ethanol Production Approaches
3. Production from Sugar Crops
4. Production from Starch Crops
5. Production from Cellulosic Biomass
6. Directions and Opportunities for World Ethanol Production
7. Conclusions

Glossary

anhydrous ethanol Essentially pure ethanol containing very little, if any, water.

biomass Plant matter of recent (nongeologic) origin or materials derived therefrom.

cellulase A combination of enzymes that catalyze the reaction of water with cellulose to release shorter chains and ultimately soluble glucose sugar.

cellulose A polymer of up to about 10,000 glucose molecules joined by beta 1-4 linkages in long, straight cellulose chains that can align parallel to one another and join by hydrogen bonds in crystalline regions that typically contain much of the cellulose. About 40 to 50% of the structural portions of plants is cellulose.

cellulosic biomass Biomass composed primarily of inedible fibrous plant parts having cellulose as a prominent component. Examples include grass, wood, and agricultural and forestry residues. Also known as lignocellulosic biomass.

corn A grain composed of about 70% starch, 10 to 11% crude protein, 4.5 to 6.0% oil, 6% hemicellulose, 2 to 3% cellulose, 1% lignin, and 1% ash.

E85 A mixture of 85% ethanol with 15% gasoline.

E95 A mixture of 95% ethanol with 5% gasoline.

enzymes Proteins originating from living cells and capable of catalyzing chemical changes in organic substances.

ethanol A chemical compound with the formula C_2H_5OH that is widely used in beverages, solvents, cosmetics, and other commercial products and is also an excellent transportation fuel. Ethanol is readily metabolized and biodegradable and does not persist in the environment.

hemicellulose An amorphous and noncrystalline polymer usually containing arabinose, galactose, glucose, mannose, and xylose sugars and typically including lower amounts of other compounds such as hexuronic acids and methoxyl, acetyl, and free carboxyl groups. Hemicellulose represents about 25 to 30% of the overall dry weight of cellulosic biomass and works with lignin to cement the cellulose fibers into a composite material of great strength.

hydrolysis The reaction of water with a compound to form other compounds.

hydrous ethanol Often a mixture of 95.5% ethanol and 4.5% water as used in Brazil.

lignin A complex phenylpropanoic acid polymer that is chemically linked with hemicellulose to cement cellulose together. Lignin represents about 15 to 20% of the dry weight of cellulosic biomass and is not fermentable at an appreciable rate.

starch A polymer of glucose molecules joined predominately by alpha 1-4 linkages.

sugar A family of simple, often sweet, compounds consisting of carbon, hydrogen, and oxygen obtained in the juice of many plants and particularly from sugarcane and sugar beets. Also refers specifically to sucrose, a particular type of sugar. Other sugars typically contained in biomass include arabinose, galactose, glucose, mannose, and xylose.

Ethanol provides a valuable liquid fuel alternative to a transportation sector that is disproportionately dependent on one politically volatile and finite energy source: petroleum. In 2000, the United States consumed almost 100 quads of energy (a quad is 10^{15} BTUs or 1 EJ), with the largest single fraction, almost 40%, coming from petroleum. Furthermore, while the domestic supply and demand for all other energy sources are nearly balanced in the United States, close to two-thirds of petroleum is imported, with unstable regions of the world controlling the largest sources, and imports continue to increase as domestic well productivity drops. This high dependence on imported oil results in strategic vulnerability to price and availability disruptions such as

occurred in the “energy crises” of the 1970s. Imported petroleum also accounts for about 45% of the net trade deficit of the United States and was the single largest contributor at \$59.2 billion in 1999. Transportation uses over two-thirds of petroleum consumed in the United States and is almost totally dependent on oil, with over 96% of transportation energy from this one source. Furthermore, vehicle miles and total numbers of vehicles continue to rise, increasing petroleum demand more. Thus, abundant alternative domestic energy sources for transportation are vital. Although other countries may not have the high per capita transportation energy use of the United States, many suffer from similar oil-related strategic and economic vulnerabilities.

Our transportation fuel situation presents other important challenges. First, petroleum is a finite resource that cannot be sustained indefinitely, and its price must increase and availability decrease over time. Petroleum also represents a valuable resource for chemicals and materials that benefit from the unique complexity of petroleum far better than burning it to carbon dioxide and water. Additionally, burning petroleum contributes about 43% of the carbon dioxide released into the atmosphere each year in the United States, and this large quantity makes carbon dioxide the most potent greenhouse gas that could cause global climate change. Furthermore, because of its high dependence on petroleum, transportation contributes about 33% of the carbon dioxide released in the United States annually, more than for residential, commercial, or industrial applications. The United States is the largest contributor to release of greenhouse gases accounting for about 25% of the world total, and new fuels that dramatically reduce such emissions are essential to change this situation. Better transportation fuels and utilization systems can also substantially improve urban air quality and reduce environmental effects of handling and consuming fuels.

Ethanol has favorable fuel, resource, and environmental attributes as a transportation fuel. Ethanol is a liquid fuel that readily integrates with our infrastructure, is widely accepted for blending with gasoline, and can be used in nearly pure form. Large quantities of this fuel are now made from starch in the United States and sugar in Brazil as well as lesser amounts in other countries. Ethanol can also be made from abundant, sustainable sources of low-cost cellulosic biomass such as agricultural and forestry residues, portions of municipal solid waste, and herbaceous and woody crops that can dramatically reduce petroleum consumption and greenhouse gas

emissions. For the latter, costs could potentially be reduced to be competitive as a pure fuel. This article summarizes some of the key attributes of this beneficial fuel and outlines how it can be produced from these resources.

1. USE AS A FUEL

1.1 Background

Ethanol is also known as ethyl alcohol or fermentation alcohol. Although ethanol is often referred to as just “alcohol,” particularly as a beverage, the term “alcohol” technically refers to a broader family of compounds with the chemical formula ROH in which R contains one or more alkyl groups containing carbon and hydrogen atoms and OH refers to the hydroxyl group containing oxygen and hydrogen atoms, respectively. Ethanol is a member of the alcohol family and has the chemical formula C_2H_5OH in which C, H, and O refer to carbon, hydrogen, and oxygen atoms, in that order.

Ethanol is a colorless, clear liquid that looks like water and is completely miscible with it. Ethanol has a somewhat sweet flavor when diluted with water; a more pungent, burning taste when concentrated; and an agreeable ether-like odor. It is widely used in medicines, lotions, tonics, colognes, rubbing compounds, and solvents and also for organic synthesis. Beverage ethanol dates back to ancient times, with Sumerians and Babylonians making beer before 6000 BC. It was also among the earliest compounds investigated with the study of ethanol from distillation of wine likely in 11th-century Italy.

1.2 Transportation Fuel in Internal Combustion Engines

Ethanol is more volatile than water, flammable, burns with a light blue flame, and has excellent fuel properties for spark ignition internal combustion engines. Its research octane number (RON) is 109, its motor octane number (MON) is 90, and its average octane is 99 while the latter for regular gasoline is typically about 88. The latent heat of vaporization of 663.4 kJ/liter (2,380 Btu/gal) is higher than that of gasoline at about 250 kJ/liter (900 Btu/gal). The energy content of ethanol is 21.2 mega joules/liter (76,000 Btu/gallon), as the lower heating value, about two-thirds that of gasoline at about 30.1 mega joules/liter (115,000 Btu/gallon). Thus, somewhat more ethanol is needed to achieve the same

range as gasoline. On the other hand, engine efficiency can be improved by running at higher compression ratios that take advantage of the better fuel properties of pure ethanol, and ethanol can achieve about 75% or more of the range of gasoline on a volume basis for an optimized ethanol engine. The boiling point of ethanol is 78°C (172°F) while that of gasoline is about 27°C (81°F), and the Reid vapor pressure (closed vessel pressure at 38°C (100°F)) for ethanol is only 16 kilopascals (2.3 psi) versus about 71 kilopascals (10.3 psi) for gasoline. Thus, evaporative emissions that cause smog formation are less for pure ethanol than for gasoline, but because ethanol is a polar molecule that resembles water more than gasoline, low levels of ethanol blended with gasoline increase vapor pressure. Although vapor pressure drops as more ethanol is added, U.S. vehicle manufacturers do not warranty conventional vehicles with these blends, and fuel providers have been reluctant to reduce the vapor pressure of gasoline to compensate. The lower vapor pressure of pure ethanol also means that starting is more difficult in cold weather. Ethanol has a very low toxicity, particularly in comparison to other fuels, and is readily biodegradable in water and soils, reducing penetration of plumes from leaks and consequences of spills compared to MTBE and petroleum-based fuels.

In Brazil, ethanol, mostly from cane sugar, is produced as either anhydrous ethanol that contains 99.6% ethanol and 0.4% water for use in 22% blends with gasoline or as hydrous ethanol containing 95.5% by volume ethanol and 4.5% water that is burned directly as a pure fuel. Brazil originally promoted ethanol use through creating a state-owned company, Petrobras, which guaranteed sales to companies and sold the fuel to customers at a discount relative to gasoline. Economic incentives were also offered to ethanol producers. As a result, annual production of hydrous ethanol fueled vehicles climbed from only 3114 in 1979 to more than 697,000 at the peak in 1986. Hydrous ethanol use peaked at about 10.88 billion liters (2.88 billion gallons) in 1989 with another 1.72 billion liters (0.46 billion gallons) added to gasoline as 22% ethanol blends during that year, as shown in Fig. 1. However, incentives were then reduced, and hydrous vehicle manufacture virtually ceased by 1999 due to sugar pricing and ethanol supply problems. Although production of some hydrous ethanol vehicles resumed recently and the Brazilian government approved increasing ethanol content in blends to 26% for gasoline and 3% for diesel, Fig. 1 shows that total ethanol production slipped to about 11.15

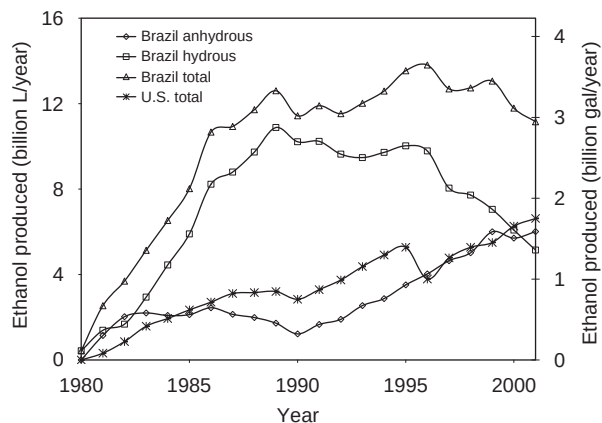


FIGURE 1 Annual production of fuel ethanol in Brazil and the United States since 1980. Ethanol is made from cane sugar in Brazil and from starch crops (mostly corn) in the United States. In Brazil, anhydrous ethanol is sold as 22% blends with gasoline, while hydrous ethanol, containing about 4.5% water, is used as a pure fuel. Almost all the fuel ethanol sold in the United States is used as a 10% blend in gasoline.

billion liters (2.95 billion gallons) in 2001 with only about 5.14 billion liters (1.36 billion gallons) used as hydrous fuel as ethanol vehicles have been replaced with those that use anhydrous/gasoline mixtures.

In the United States, fuel ethanol production grew from virtually nothing in 1980 to about 6.63 billion liters (1.75 billion gallons) by 2001, as shown in Fig. 1. Most of this ethanol is from corn starch and is used as 10% ethanol blends, reducing gasoline consumption and extending petroleum supplies. Ethanol also improves octane, decreasing the need for other octane boosters such as benzene, which are toxic and often carcinogenic, and oxygen in ethanol reduces emissions of unburned hydrocarbons and carbon monoxide, particularly for older vehicles that tend to burn rich.

Special vehicles are sold in the United States that can burn any fuel containing from 0 to 85% ethanol in gasoline, the latter being designated as E85, with 15% gasoline being used to promote cold starting. Only 172 of these vehicles were on the road in 1992, but the number rose to 48,022 in 2001, with 13,342 net new vehicles added that year. E85 engines consumed about 26.2 million liters (6.9 million gallons) of ethanol in 2001. These flexible fueled vehicles have in line sensors that measure ethanol levels and adjust the air-fuel ratio to maintain proper combustion. Although the compression ratio cannot be adjusted to fully capitalize on the favorable fuel properties of ethanol in high-level blends as with hydrous ethanol in Brazil, the technology ensures

fuel availability. Catalytic converters are needed to minimize tailpipe emissions of aldehydes, the primary pollutant resulting from incomplete ethanol combustion, and NO_x formation could increase for high-compression engines designed to maximize fuel utilization efficiency.

Ethanol has been used in compression ignition engines, but the lubrication properties and cetane of ethanol are not matched to conventional diesel technology. Thus, the engines must be modified or additives used to improve compatibility. On the other hand, adding ethanol to diesel fuel reduces particulate emissions, a key need, and formulations have been developed to stabilize dispersion of ethanol in diesel fuel.

1.3 Fuel Cells

Fuel cells are attractive because of low emissions and good efficiencies. Fuel cells react hydrogen with oxygen to make water at controlled conditions that minimize formation of pollutants such as NO_x and produce no unburned hydrocarbon emissions. In 2002 nearly every major automaker had large teams working on vehicle fuel cells, with most focused on operation with either gasoline or hydrogen. Because storing hydrogen onboard vehicles is challenging, using a liquid fuel such as ethanol as a hydrogen source is highly desirable but more complex to implement. Reformer technologies being developed for gasoline in fuel cells would generally be capable of using ethanol, but experience is limited on releasing hydrogen of acceptable quality from either one. Reforming methanol to hydrogen for fuel cells in buses and cars was developed and demonstrated in the United States, Germany, and Japan before environmental concerns about methanol limited this option. Reforming ethanol requires more steps, and this technology has been demonstrated in the laboratory but not onboard vehicles. Ethanol can also be partially oxidized to hydrogen, carbon monoxide, and carbon dioxide followed by reaction of carbon monoxide to hydrogen, and a partial oxidation reactor coupled to a shift reactor and a polymer electrolyte membrane (PEM) fuel cell followed by heat recovery achieved an efficiency of 43% with undetectable emissions of hydrocarbons, carbon monoxide, and nitrous oxides. Thus, ethanol fuel cells are very promising, but considerable research and development would enhance the readiness of this not widely studied technology.

2. ETHANOL PRODUCTION APPROACHES

Ethanol is manufactured by two primary paths: (1) fermentation of sugars derived from sugar, starch, or cellulosic materials or (2) reaction of ethylene with water. The former is favored for production of fuel. The latter has been used to make industrial grade ethanol for solvents, cosmetics, medicines, and so on, but purification of fermentation ethanol is displacing ethylene-derived ethanol for these applications. Some new processes are under development for fermentation or catalytic conversion of synthesis gas (carbon monoxide and hydrogen) from wood, natural gas, or coal, but synthesis gas from sustainable resources such as biomass is expensive. This article focuses on fermentation ethanol because it can potentially meet fuel cost and volume needs for large-scale fuel production, while ethylene is too expensive and supplies would likely be stretched.

3. PRODUCTION FROM SUGAR CROPS

3.1 Processing Sugar to Ethanol

Brazil produced about 18.5 million metric tonnes (20.4 tons) of processed sugar in the 2001/2002 harvest, with approximately 9.45 million metric tonnes (10.4 tons) used domestically and the rest exported. Brazilian sugar is mostly derived from sugarcane, a drought-tolerant tropical and subtropical crop containing about 12 to 17% sugars (90% sucrose, 10% glucose) and 68 to 72% moisture. Brazil grew about 272 million metric tonnes (300 million tons) of sugarcane in the 2001/2002 season, making it second to India in world cane production that year. In 1975, Brazil adopted a Pro-Alcool Program to convert sugar to ethanol to reduce dependence on petroleum imports that were damaging the economy, and Fig. 2 summarizes the sugar extraction and fermentation approach. In dedicated ethanol plants, cane juice is heated to about 105 to 110°C (221 to 230°F) to reduce microbial contamination and then fermented. In combined ethanol-sugar facilities, sucrose is removed by crystallization and the remaining molasses containing up to about 65% sugars is fermented. In either approach, sugar, $\text{C}_6\text{H}_{12}\text{O}_6$, is readily fermented to ethanol, $\text{C}_2\text{H}_5\text{OH}$, by common yeast such as *Saccharomyces cerevisiae* according to a reaction carried out for

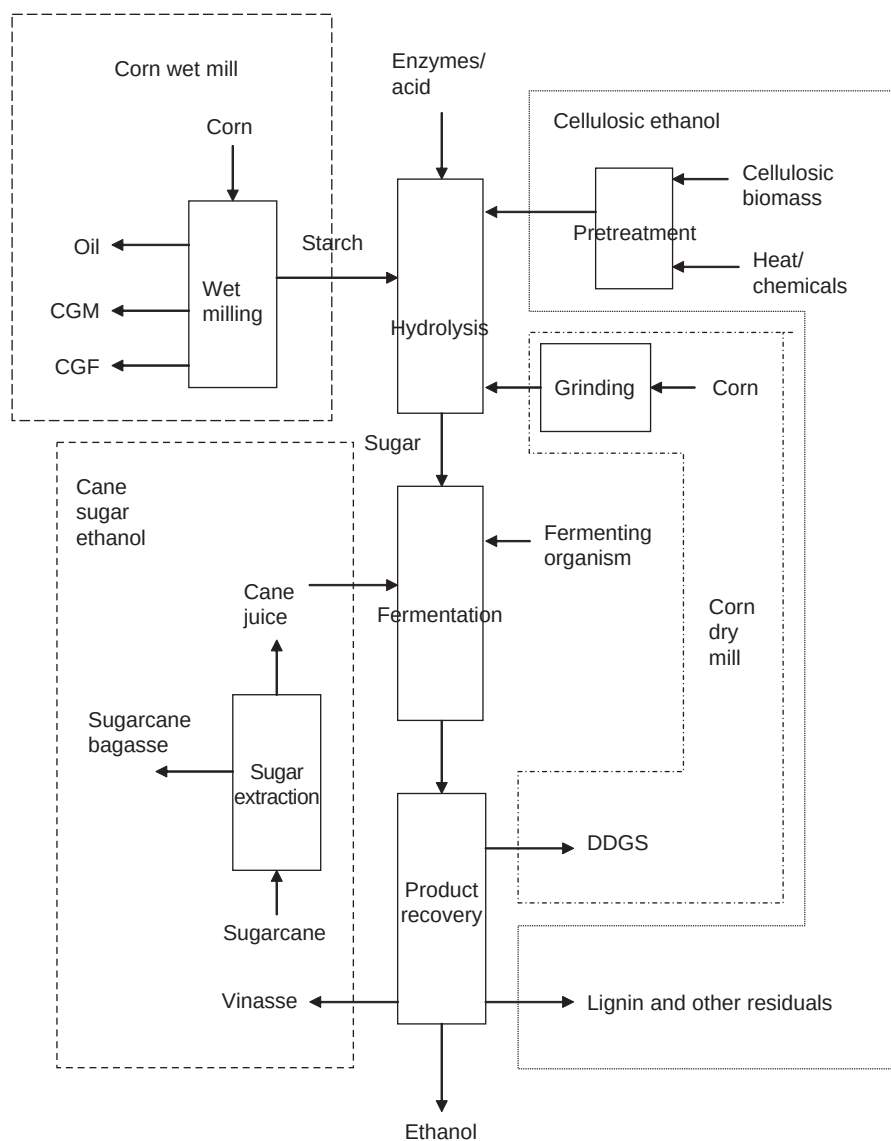


FIGURE 2 A simplified process flow diagram for production of ethanol from cane sugar, corn, and cellulosic biomass. All have similar fermentation and ethanol recovery operations but use different approaches to release sugars and generate different coproducts. Sugar can be directly extracted from sugarcane, and the residual bagasse is used as a boiler fuel to provide much of the energy for the extraction and ethanol production and recovery operations. In a corn dry mill, corn is ground, and enzymes and heat are added to hydrolyze starch to sugars for conversion to ethanol, while the oil, protein, and fiber in corn are recovered after fermentation as an animal feed known as DDGS. Wet mills first fractionate corn to separate corn oil, corn gluten meal (CGM), and corn gluten feed (CGF) to capture value for food and animal feed, and the starch can then be hydrolyzed to sugars for fermentation to ethanol. For cellulosic biomass, heat and acids or enzymes hydrolyze the hemicellulose and cellulose portions to release sugars that can be fermented to ethanol, and the lignin and other remaining fractions can be burned to provide all the process heat and electricity for the conversion step with excess electricity left to export.

millennia:



in which CO_2 is carbon dioxide released during this process. About 70% of Brazilian ethanol facilities, often called distilleries, use batch operations in which

sugar, nutrients, and other ingredients are added at the beginning of the fermentation or fed batch processes in which one or more of these ingredients are added as fermentation proceeds. However, continuous processes in which ingredients are constantly added and product removed from

fermentation vessels are also used. Yeast is typically concentrated and separated from the fermentation effluent, washed with sulfuric acid to reduce bacterial contamination, and recycled to fermentors, resulting in high cell densities that shorten fermentations to 6 to 10 h at temperatures of about 33 to 35°C with high ethanol yields (90 to 92% of theoretical) and concentrations (8 to 11% on a volume basis). According to Eq. (1), no more than 51.1 kg of ethanol can be obtained from fermentation of 100 kg of sugar, with yeast achieving about 90 to 92% of this. Thus, about 460 to 470 kg (920 to 940 pounds) or 580 to 590 liters (139 to 142 gallons) of ethanol result from a metric tonne (ton) of sugar. Following yeast removal, the fermentation broth containing ethanol, water, nutrients, and some yeast passes to distillation to make 95.5% “hydrous” ethanol, which can be taken to 99.6% “anhydrous” ethanol in a dehydration section, if desired. The remaining dissolved solids, called vinasse, are used as a fertilizer and in irrigation. About 5170 liters of ethanol could be obtained from a hectare of land in Brazil in 1996, and 13.9 billion liters (3.7 billion gallons) of ethanol in addition to 13.5 million metric tonnes (14.9 million tons) of sugar were made from 273 million metric tonnes (300 million tons) of sugarcane that year.

3.2 Costs

The costs of making ethanol in Brazil are estimated to be about \$0.006 to 0.007/liter (\$0.023 to 0.026/gallon) for labor, \$0.004 to 0.006/liter (\$0.015 to 0.023/gallon) for maintenance, \$0.002/liter (\$0.008/gallon) for chemicals, \$0.002 to 0.003/liter (\$0.008 to 0.011/gallon) for energy, \$0.004/liter (\$0.015/gallon) for other items, and \$0.127 to 0.134/liter (\$0.481 to 0.507/gallon) for sugar. Thus, the total cash cost of ethanol is on the order of \$0.167 to 0.185/liter (\$0.632 to 0.700/gallon). The overall capital investment is about \$0.52/annual liter (\$2.00/gallon) of capacity for a process that operates for 150 days/year making about 240,000 liters (900,000 gallons) of ethanol per day. Allowing another \$0.04 to 0.06/liter (\$0.15 to 0.23/gallon) for interest on capital, capital recovery, and other fixed costs, the total cost including return on capital as applied in Brazil is about \$0.21 to 0.25/liter (\$0.79 to 0.95/gallon) at the plant gate without transportation, distribution, taxes, and other costs. More on-stream time would greatly reduce the unit capital cost charged to ethanol. Hydrous costs about 7% less than anhydrous ethanol because less energy

and capital are needed to meet final product specifications. Costs in Brazil dropped at about a 30% rate with each doubling of experience through 1990 and continue to decline at about a 10% rate due to learning curve effects resulting from commercial experience. Opportunities have been identified to reduce costs by about 23%, mostly by reducing the intensity of growing cane but by also improving conversion operations.

3.3 Energy Balance

About 197 to 222 mega joules of fossil energy are used for providing seeds, labor, power for agricultural operations, agricultural equipment, fertilizers, lime, herbicides, insecticides, and cane transportation to grow a metric tonne of sugarcane (169,000 to 191,000 Btu/ton). However, much of the energy for sugar extraction and fermentation is provided by burning the bagasse left after processing, but some fossil energy is used for electricity generation, chemicals and lubricants manufacture, erection of buildings, and manufacture of equipment, a total of about 40 to 70 mega joules/metric tonne (34,000 to 60,000 Btu/ton) of cane. Overall, about 237 to 292 mega joules of fossil energy are consumed (204,000 to 251,000 Btu/ton) compared to 1707 to 1941 mega joules of energy in the ethanol from a metric tonne (1.470×10^6 to 1.669×10^6 Btu/ton). Based on these values, the energy balance ratio of ethanol energy output to fossil fuel input is between about 5.9 to 8.2, not including the 175 to 328 mega joules/metric tonne of cane (150,000 to 282,000 Btu/ton) left from burning bagasse after sugar has been extracted, a very favorable energy balance.

3.4 Greenhouse Gas Emissions

In simple terms, accumulation of methane, carbon dioxide, nitrous oxide, and other gases that trap infrared radiation are predicted to cause global climate change, and carbon dioxide has the greatest impact because of the huge quantities released when fossil fuels are burned. Burning coal, petroleum, and natural gas, in order of decreasing impact, release previously trapped carbon into the atmosphere, resulting in a net accumulation of carbon dioxide. On the other hand, if fuels are made with renewable biomass such as wood, carbon dioxide released during processing and burning can be recaptured by photosynthesis when more biomass is grown to replace that harvested. While biomass is carbon dioxide neutral, fossil inputs to plant, grow, harvest,

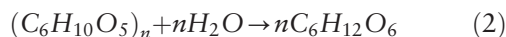
transport, convert, and distribute biomass and its products contribute to greenhouse gas accumulation and must be considered.

Burning bagasse can provide 100% of the thermal energy and 92% of the electricity for processing of sugar to ethanol, and less than about 8% comes from fossil sources. Based on Brazilian experience that 1.2 liters of ethanol can replace 1.0 liters of gasoline, the net contribution to carbon dioxide accumulation by vehicles running on hydrous ethanol is less than about 10% of gasoline, and substituting surplus bagasse or other renewables for fossil fuel to generate electricity could improve the balance further. However, intensive sugarcane agricultural practices raise serious concerns about its sustainability, and low and no-till and other practices will be needed for cane sugar to become more viable.

4. PRODUCTION FROM STARCH CROPS

4.1 Starch Processing

Corn grain contains about 70% starch, 10 to 11% crude protein, 4.5 to 6.0% oil, 6% hemicellulose, 2 to 3% cellulose, 1% lignin, and 1% ash and is the dominant choice for current fuel ethanol production by dry or wet milling operations in the United States. In a typical dry mill, grain is milled to a powder, heated with water addition to about 85°C, mixed with alpha-amylase enzyme, held for up to an hour, heated further to 110 to 150°C to liquefy the starch and reduce bacteria levels, cooled back to about 85°C, and held with more alpha amylase for about 1 h. Then, the stream is cooled further, and gluco-amylase enzyme added to complete conversion to sugars known as dextrose. This overall saccharification operation occurs by the hydrolysis reaction:



in which $C_6H_{12}O_6$ is a glucose sugar molecule formed when the alpha bonds linking n units of $C_6H_{10}O_5$ in long chains of starch are broken and combined with n molecules of water, H_2O . Yeast then ferment glucose to ethanol and carbon dioxide, just as in Eq. (1). The fermentation can be in batch operations held until the fermentation is complete in about 48 h or by continuous processes with ongoing sugar addition and takeoff of fermented broth known as beer. The beer containing about 12 to 14% or more ethanol is typically distilled to around 95% purity, close to the azeotropic composition at

which further enrichment by ordinary distillation ceases, and this mixture is dehydrated with molecular sieves to anhydrous ethanol. The solids containing the remaining protein, oil, and fiber are dried to a 27% protein product known as distillers dried grains with solubles (DDGS) or just distillers dried grains (DDG), depending on whether process syrup is combined with the solids or not, and used in animal feeds. The carbon dioxide can be sold for carbonated beverages, dry ice, and other applications. Overall, about 460 to 490 liters (110 to 118 gallons) of ethanol weighing 365 to 390 kg (730 to 780 lbs), 357 to 335 kg (714 to 670 lbs) of DDGS, and 350 to 370 kg (700 to 740 lbs) of carbon dioxide result from each metric tonne (per ton in parentheses) of corn processed—close to equal masses of each coproduct. Figure 2 presents a simplified block flow diagram for dry milling.

Although the greatest number of facilities are dry mills run by smaller producers, more ethanol is made in larger wet milling operations generally owned by larger companies. In wet mills, corn is first steeped and milled to separate the starch, protein, fiber, and germ, again as summarized in Fig. 2. Corn oil is removed from the germ for food, and the remaining germ meal is combined with the fiber and water from steeping as a 21% protein animal feed known as corn gluten feed. A portion of the corn protein is sold as corn gluten meal, a 60% protein product added to chicken and other animal feeds. Starch can be used directly for industrial uses and food or broken down by enzymes as for dry milling to make dextrose that can be enzymatically converted to high fructose corn syrup for a beverage sweetener or fermented to ethanol. One metric tonne (1 ton in parentheses) of corn can yield approximately 34 kg (67 lbs) of oil, 52 kg (105 lbs) of corn gluten meal, 283 kg (567 lbs) of corn gluten feed, 470 liters (112 gallons) or 370 kg (740 lbs) of ethanol, and 354 kg (708 lbs) of carbon dioxide. Again, all corn components are utilized, with protein sold for animal feed, the major market for corn.

4.2 Costs

Although ethanol costs tend to be site specific and business sensitive, recent surveys provide a useful picture. The average cost of all feedstocks for dry and wet mills was about \$0.23/liter (\$0.87/gallon) of ethanol with dry mills varying from \$0.25/liter (\$0.94/gallon) for small plants to \$0.21/liter (\$0.79/gallon) for medium plants and \$0.22/liter (\$0.82/gallon) for large facilities and wet mills being about \$0.024/liter (\$0.09/gallon) greater than the overall

average, apparently because of the greater quantities of corn required. The multiple wet mill coproducts provided revenue of \$0.11/liter (\$0.43/gallon) of ethanol versus about \$0.07/liter (\$0.28/gallon) for dry mills. The difference between the cost of corn and coproduct revenues, termed the net cost of corn, is about \$0.14/liter (\$0.53/gallon) for dry mills and \$0.13/liter (\$0.48/gallon) for wet mills but varied between \$0.095/liter (\$0.36/gallon) and \$0.24/liter (\$0.91/gallon) for wet mills and \$0.12/liter (\$0.44/gallon) and \$0.25/liter (\$0.95/gallon) for dry mills over the ten years ending in 2000.

Other costs are also incurred. Averages for electricity and fuel were about \$0.030/liter (\$0.112/gallon) for wet mills and \$0.035/liter (\$0.131/gallon) for dry mills for an overall weighted average of \$0.031/liter (\$0.118/gallon). Operating labor, repairs, and maintenance were lower for dry mills at about \$0.029/liter (\$0.109/gallon) compared to wet mills at about \$0.033/liter (\$0.124/gallon). Yeast, enzymes, chemicals, and other ingredients contributed another \$0.024/liter (\$0.090/gallon) for dry mills and \$0.030/liter (\$0.114/gallon) for wet mills. Averages for management, administration, insurance, and taxes were similar at \$0.009/liter (\$0.037/gallon) for dry mills and \$0.010/liter (\$0.038/gallon) for wet mills. All other costs averaged about \$0.019/liter (\$0.072/gallon) for wet and \$0.013/liter (\$0.051/gallon) for dry mills. Thus, total cash costs for other than corn were about \$0.12/liter (\$0.46/gallon) for wet mills and \$0.11/liter (\$0.42/gallon) for wet mills, and combined with net feedstock costs, the average cost is about 0.25/liter (\$0.94/gallon), with a less than 1% advantage to wet mills. This gate cost does not include transportation, distribution, taxes, and other costs to the consumer.

The difference between ethanol revenues and cash costs of production must be weighed against capital costs to determine return on investment. Capital costs are estimated to be between \$0.283/annual liter (\$1.07/annual gallon) to \$0.631/annual liter of capacity (\$2.39/gallon), varying with facility type, size, and technology. Wet mills tend to cost more than dry mills, and larger dry mills generally cost less than smaller ones per volume of ethanol made because of economies of scale. Capital and operating costs have dropped over time through learning curve effects and research.

Several considerations apply to starch ethanol costs, many of which apply to ethanol from sugar. Costs and coproduct revenues vary with location and economic climate, and many costs components do not always work together. For example, the value of

ethanol usually drops with gasoline prices. In addition, higher corn or sugar prices or softer coproduct markets increase the net feedstock cost. In this vein, increasing corn ethanol production would eventually drive up corn prices and drive down coproduct selling prices, making further capacity increases uneconomic. On the other hand, technology improvements enhance profitability.

Tax incentives in the United States promote blending with gasoline and were essential to initiating ethanol production in the late 1970s and early 1980s in response to petroleum shortages and price spikes. Over time, this excise tax exclusion dropped to \$0.0140/liter (\$0.053/gallon) of 10% blended fuel, reducing ethanol costs by \$0.140/liter (\$0.53/gallon) of ethanol used. Alternatively, the tax on blended gasoline is reduced by \$0.0110/liter (\$0.0416/gallon) for 7.7% ethanol and \$0.0081 (\$0.0308/gallon) for gasoline containing 5.7% ethanol. This exemption is not paid to farmers but is an avoided tax for blenders of ethanol from any form of biomass with gasoline. Blenders could apply an income tax credit instead of the excise tax exemption. The exemption is scheduled to be further cut to \$0.0137/liter (\$0.052/gallon) of 10% blended fuel in 2003, \$0.0135/liter (\$0.051/gallon) in 2005, and eliminated on December 31, 2007. Many states, particularly in the Midwest, also provide incentives to encourage ethanol use. Although subject to controversy, tax and other assistance policies are vital for introducing new technology, particularly for commodity markets, and have supported emergence of new industries such as the large petrochemical sector that grew out of World War II.

4.3 Energy Balance

Incorrect and outdated data are periodically used to spur controversy about whether making corn ethanol consumes more or less energy than contained in the product. However, up-to-date facts show continual improvements, with reductions in energy demand as the technology matures resulting in favorable energy benefits. Accounting for the variability of corn productivity over time and with location, an input of 2789.3 mega joules/dry metric tonne (57,476 Btu/bushel or 2,415,100 Btu/dry ton) of corn was determined for all the energy required to grow and harvest corn and haul it to storage. The energy to transport corn from storage to ethanol production was estimated to be about 292.3 mega joules/dry metric tonne (6020 Btu/bushel or 253,000 Btu/dry ton) or 690.6 J/liter (2.263 Btu/gallon) of ethanol made. Recent surveys showed that about 13.594 mega

joules of heat and electricity are needed to make one liter (48,772 Btu/gallon) of ethanol via dry mills and about 15.118 mega joules/liter (54,239 Btu/gallon) for wet mills. However, new plants generally achieve on the order of 11 mega joules/liter (40,000 Btu/gallon) or less. The survey showed about 460 and 467 liter of ethanol/dry metric tonne (2.64 and 2.68 gallons/bushel or 111 and 113 gallons/dry ton) were derived from corn in dry and wet mills, respectively. Finally, about 442.6 kilojoules/liter (1,588 Btu/gallon) of ethanol was required to distribute ethanol to end users. Combining these estimates, dry and wet mills required about 20.75 and 22.16 mega joules/liter (74,447 and 79,503 Btu/gallon) of ethanol, respectively, with the average weighted by number of facilities being about 21.525 mega joules/liter (77,228 Btu/gallon).

Some energy should be assigned to coproducts from corn milling, with distribution by energy inputs, energy content, relative market values, relative weight, and the amount of energy needed to replace them analyzed. However, the latter based on the energy to produce equivalent feed and food as from coproducts is often favored, with growing soybeans an appropriate alternative. On that basis, the final estimate is a net energy displacement of about 6.307, 5.369, and 5.883 mega joules/liter (22,629, 19,262, and 21,105 Btu/gallon) for dry, wet, and weighted average facilities, respectively, when ethanol is made from corn compared to the higher heating value for ethanol of 23.402 mega joules/liter (83,961 Btu/gallon).

This analysis means that one volume of ethanol generates about 37%, 30%, and 34% more energy than needed to manufacture it by dry, wet, and the overall weighted average of production facilities, respectively. Furthermore, because about 6.34 units of liquid fuel energy result from 1.0 energy unit of petroleum, corn ethanol provides even greater advantages for displacing imported liquid fuels, a major concern for economic and political stability in the United States.

4.4 Greenhouse Gas Emissions

The positive energy balances suggest that corn ethanol has a lower impact on global climate change than gasoline. Again, if the greenhouse gas emissions that would result if soybeans were grown to meet markets served by corn ethanol coproducts are subtracted, burning ethanol from starch crops as a mixture of 85% ethanol with 15% gasoline (E85) reduces carbon dioxide emissions by about 40% and

emissions of all greenhouse gases including methane and nitrous oxides by about 20 to 30% relative to pure gasoline. Because a significant portion of greenhouse gases comes from the gasoline in E85, greater reductions would be possible if technology were developed to use hydrous ethanol as in Brazil. If corn stover or other renewable energy sources were substituted for fossil fuels to power dry and wet mills as in Brazil, the greenhouse gas benefits of corn ethanol would be better yet.

5. PRODUCTION FROM CELLULOSIC BIOMASS

5.1 Cellulosic Biomass

Ethanol can also be made from sugars contained in the structural portion of plants. In particular, about 40 to 50% is cellulose—up to about 10,000 glucose molecules joined by beta 1-4 linkages as long linear molecules that can align parallel to one another and join by hydrogen bonding in crystalline regions that typically contain much of the cellulose. Another sugar polymer known as hemicellulose represents about 25 to 30% of the dry plant mass; unlike cellulose, hemicellulose is amorphous. Hemicellulose is typically made up of the five sugars arabinose, galactose, glucose, mannose, and xylose, along with lower amounts of other compounds including hexuronic acids and methoxyl, acetyl, and free carboxyl groups. Lignin represents about another 15 to 20% by dry weight and is a complex phenylpropanoic acid polymer. Lignin and hemicellulose are joined by covalent bonds and cement cellulose fibers into microfibril bundles. Plants also contain lesser amounts of minerals, free sugars, proteins, and other components that vary with species. Together, this composite of cellulose, hemicellulose, lignin, and other generally lesser ingredients is called lignocellulosic, or just cellulosic, biomass and has great strength to support the aerial portion of plants.

Cellulosic biomass is abundant with prominent examples including agricultural residues (e.g., corn stover, sugarcane bagasse), forestry residues (e.g., forest thinnings), industrial wastes (e.g., paper sludge), major portions of municipal solid waste (e.g., waste paper), herbaceous plants (e.g., switchgrass), and woody species (e.g., poplar). In addition to its abundance, cellulosic biomass can be grown with low energy inputs and is comparable in cost to

petroleum. For example, biomass costing about \$44/dry metric tonne (\$40/ton) is competitive with petroleum at about \$6.00/barrel based on equal mass or about \$12.70/barrel based on equivalent energy content. Some cellulosic wastes such as rice straw can be obtained for lower prices, and in some cases, such as paper sludge, users would be paid to dispose of them (i.e., tipping fee). Waste conversion to ethanol can also help solve serious disposal problems.

5.2 Processing of Cellulosic Biomass

As outlined in Fig. 2, the overall process for making cellulosic ethanol is conceptually similar to processing corn but with distinctions that have important energy balance and greenhouse gas emission implications. Sulfuric and other acids can convert both hemicellulose and cellulose into fermentable sugars. Alternatively, enzymes known as cellulase hydrolyze cellulose to glucose while a collection of enzymes loosely termed hemicellulase could breakdown hemicellulose. The remaining lignin cannot be converted to sugars but can provide boiler fuel for process heat and electricity or be reacted into coproducts.

Biological conversion of cellulose to glucose by cellulase and fermentation of glucose to ethanol realize high yields vital to success, and the advances of the evolving field of modern biotechnology promise significantly lower costs. In this case, mills or other devices reduce the feed material size, facilitating transfer of heat and chemicals later, and the material is pretreated to expose cellulose to cellulase to realize high glucose yields. Currently, pretreatment is often by dilute acid hemicellulose hydrolysis because good yields of sugars are realized from hemicellulose during pretreatment and from cellulose in the subsequent enzymatic hydrolysis and acid costs are relatively low. In this approach, cellulose is reacted in 0.5 to 1.0% sulfuric acid at about 160 to 190°C for approximately 10 minutes, releasing most hemicellulose sugars by a hydrolysis reaction similar to Eq. (2) but including up to five different sugars as well as other compounds in the cellulosic matrix. The contents are flash cooled to stop the reaction as the temperature quickly drops to about 100°C and also to remove some acetic acid, furfural, and other fermentation inhibitors that are released from biomass or produced by degradation reactions during pretreatment. Next, the solid and liquid fractions are separated and ion exchange employed in some designs to remove a portion of acetic acid and much of the sulfuric acid from the liquid. Lime is added to neutralize the remaining

sulfuric acid and then increase the pH to about 10 in an “overliming” operation that removes toxics to downstream biological steps.

A small portion of the solids alone or including treated liquid is fed to make cellulase by an organism such as the fungus *Trichoderma reesei*. The effluent is then combined with the bulk of the pretreated solid to hydrolyze cellulose to glucose, and an organism such as yeast is added to convert the glucose to ethanol soon after release in a simultaneous saccharification and fermentation or SSF approach. Because SSF rapidly converts glucose and the two glucose molecule cellobiose to ethanol, the concentration of both of these powerful inhibitors of cellulase activity are reduced. Thus, the rates, yields, and concentrations for ethanol production are all improved via the SSF approach compared to sequential hydrolysis and fermentation even though the temperature has to be reduced from optimum levels for hydrolysis alone to accommodate less temperature tolerant yeast. The SSF configuration also eliminates about half of the fermentation vessels, cutting costs, and ethanol accumulation reduces the likelihood of successful invasion by unwanted organisms that thrive if sugar accumulates. Alternatively, conditioned liquid hemicellulose hydrolyzate can be added to SSF in an approach labeled simultaneous saccharification and cofermentation (SSCF) if the organism used ferments hemicellulose sugars and glucose to ethanol. Of course, hemicellulose sugar fermentation can be performed in a separate vessel by such an organism, albeit at the expense of more equipment and somewhat higher costs.

Distillation concentrates approximately 5% fermentation beer streams to about 95% ethanol, and this azeotropic mixture is often passed to molecular sieves to recover a nearly 100% ethanol product suitable for blending with gasoline or use as a pure anhydrous fuel. For use as a pure fuel without gasoline addition, the 5% water might be left in a la Brazilian experience. The solids and solubles remaining following SSF/SSCF pass out in the distillation column bottoms, are concentrated, and then are burned to generate steam that can provide all of the process heat and electricity with some excess electricity left to export. Water from the system is treated by anaerobic digestion, and the resulting methane is burned for steam generation.

5.3 Technical and Economic Progress

The United States and other countries have funded research to improve cellulosic biomass conversion to

ethanol, and despite modest and sporadic funding, the total estimated cost at the plant gate including capital charges dropped from about \$1.05 to 1.32/liter (\$4.00 to 5.00/gallon) of ethanol in 1979 to approximately \$0.29 to 0.32/liter (\$1.10 to 1.20/gallon) now, a value competitive with corn ethanol. For a facility making about 197.6 million liters/year (52.2 million gallons/year) of ethanol, the capital costs were projected to be about \$234 million, the feedstock cost about \$19.3 million/year, the other variable operating costs about \$5.7 million/year, and the fixed operating costs about \$7.5 million/year. Such cost projections depend strongly on location, estimation methods, available data, key performance parameter assumptions, technology applied, and financing arrangements and vary considerably. Furthermore, although corn and sugar conversion costs are based on commercial experience, no cellulosic facilities are yet operational. Thus, while useful for guiding research, cellulosic ethanol costs are approximate indicators of actual costs. However, the fact that several firms are working to commercialize the first cellulosic ethanol facilities confirms the economic merits, and their biggest challenge is financing the substantial capital cost for first-of-a-kind technology in light of its perceived risk.

Progress on two major fronts has fed the cost reductions outlined. First, important improvements have been made in overcoming the recalcitrance of biomass. Dilute acid pretreatment has been optimized to increase sugar yields from both the hemicellulose and cellulose fractions, and better cellulase formulations have achieved higher yields with lower doses. SSF improved economics substantially, with enzymes or organisms that utilize cellobiose, a polymer of two glucose molecules, proving particularly important by reducing the concentration of this powerful cellulase inhibitor.

The second hurdle, overcoming the diversity of sugars, was a major breakthrough vital to economic viability. Native organisms cannot ferment all five sugars typically contained in cellulosic biomass to ethanol with high yields, a necessity to competitiveness. However, in the late 1980s, researchers genetically engineered bacteria to make ethanol from all the sugars at close to theoretical yields by inserting two genes from the bacteria *Zymomonas mobilis* into other host bacteria to produce the enzymes pyruvate decarboxylase and alcohol dehydrogenase that transform pyruvate, an intercellular compound in the host, to ethanol instead of the products ordinarily made. This accomplishment increases revenue while reducing costs for disposal of unused sugars.

5.4 Energy Balance

Burning solids and concentrated solubles after distillation, solids from aerobic treatment, and methane from anaerobic treatment can generate enough steam to meet all process heat and power demands. In addition, about 10.9 MW of surplus power is left for sale for a currently envisioned ethanol plant fed 2000 metric tonnes/day (2,205 tons/day) of cellulosic biomass, resulting in about 0.4 energy units of ethanol per unit of energy in the biomass feed. The amount of ethanol will improve as yields are increased through learning curve effects and better technology, and power exports to the grid will increase as (1) cellulosic ethanol technology improves to use less heat and power and (2) more efficient cycles are used for making electricity from steam. The key is that no fossil fuel is needed to convert cellulosic biomass into ethanol, and significant amounts of excess electricity are left for sale into a market compatible with large-scale fuel production.

It has been estimated that about 6 to 15% of the ethanol fuel value is required to grow biomass, another 1 to 4% to transport it to the conversion facility, and 1 to 5% for fuel distribution. If all of these energy needs are met from fossil sources, about 4.0 to 6.7 times as much ethanol energy is released as fossil fuel used for the overall cellulosic ethanol process, but the ratio increases to about 4.3 to 7.2 if the thermal value of exported electricity is included. More comprehensive analyses project that current cellulosic ethanol technology provides a net energy output from about 16.7 to 21.2 mega joules/liter (60,000 to 76,000 Btu/gallon) of ethanol compared to its lower heating value of 21.2 mega joules/liter (76,000 Btu/gallon), depending on the feedstock employed. Overall, these low fossil inputs result from the low-energy demands to supply cellulose, generation of all heat and power for conversion from lignin and other biomass residues, and to a lesser extent, export of excess power.

Given its favorable energy balance, cellulosic ethanol could be substituted for gasoline or diesel fuel to harvest and transport cellulose to the process and distribute ethanol to market. In addition, methane produced by anaerobic digestion of process residues could be converted into ammonia to reduce fossil fuel use for growing biomass. The result would be that virtually no fossil fuels are needed to grow, convert, and transport ethanol when it is made from cellulose while both fuel and power are produced, and the ratio of energy outputs to fossil energy inputs

would become very large. Advances in ethanol production and power generation would improve this picture.

A recent study showed that fossil fuel inputs for cellulosic ethanol were lower than for a large number of options including gasoline, other petroleum-based fuels, natural gas-based fuels, and electrolysis hydrogen. Yet the total energy required to make cellulosic ethanol was toward the high end. This apparent contradiction is readily explained because burning lignin and other residues totally meets the large process energy demand without fossil fuels. The fact that most of the more efficient technologies in the study are practiced commercially while cellulosic ethanol is not suggests that significant improvements will evolve for cellulosic ethanol through the commercial learning curve, and the likely outcome is that the energy balances will improve dramatically, increasing the energy displacement benefits far more.

5.5 Greenhouse Gas Emissions

As shown for ethanol from sugar and corn, carbon dioxide generation by burning fossil fuels has a large impact on greenhouse gas emissions, and the low fossil energy inputs for cellulosic ethanol production are expected to have important implications for global climate change. In fact, studies have shown that using 85% cellulosic ethanol with gasoline, E85, reduces GHG emissions by about 65% compared to straight gasoline, and that running E95 in vehicles would decrease GHG emissions by about 75 to 78%. A large part of these emissions comes from the gasoline, and even greater reductions in GHGs of about 85% or more are possible if pure ethanol were used. Advances in both ethanol production and power-generation technologies should improve yields, reduce energy inputs, and generate more electricity, reducing GHG emissions even more.

5.6 Some Other Technologies for Processing Cellulosic Biomass

A favorable pretreatment must produce highly digestible cellulose, recover hemicellulose sugars with high yields, have a low capital cost, require limited energy inputs, minimize degradation, use low-cost chemicals, minimize problems with chemical additives, and integrate well with the overall process. Many studies with physical pretreatments such as size reduction and radiation and biological pretreatments with various fungi have been unsuc-

cessful, and only a few chemical pretreatments other than dilute sulfuric acid currently promise to be cost-effective. In the ammonia fiber explosion (AFEX) process, ammonia added to biomass reduces cellulose crystallinity, opens up the structure for enzymes, can be mostly recovered and recycled, provides fermentation nutrients, and produces a readily fermentable hydrolyzate without post treatment. A near neutral pH approach favors soluble oligomer release from hemicellulose to keep formation of sugar monomers that can degrade low. Ammonia circulation through biomass removes lignin, improving cellulose hydrolysis yields. Lime also removes lignin, but the reaction is slow. Dilute nitric acid reduces materials of construction costs compared to sulfuric but is itself more expensive. More research is needed to better understand the performance and costs of these and other options that promise to meet the demanding criteria for effective pretreatment.

Acids can be used to breakdown cellulose. Concentrated acids realize high yields vital to economic success, but the cost of required acid recovery operations must be significantly reduced. Substantial experience exists for dilute acid hydrolysis, but yields are not high unless practical counter-current flow configurations can be developed.

5.7 Opportunities for Further Improvements

The pretreatment step and combined biological operations for cellulase production, cellulose hydrolysis, and sugar fermentation are the most expensive in the currently defined process. *Trichoderma reesei* and other fungi generally used for manufacturing cellulase arose out of World War II research to combat decay of cotton gear in the tropics. Commercial cellulase is estimated to cost over \$0.26/liter (\$1.00/gallon) of ethanol produced, and making it on site is estimated to cost about \$0.092/liter (\$0.35/gallon). Although now expensive, cellulase offers near theoretical glucose yields from cellulose, and the evolving tools of modern biotechnology promise reduced costs by improving its effectiveness in hydrolyzing cellulose and increasing the rate of making enzyme. Better pretreatment could also reduce cellulase use, and development of continuous processes could cut equipment idle time and eliminate expensive propagation systems.

Although subsidies are needed for ethanol at current costs to compete for blending with gasoline, the cost of cellulosic biomass is competitive with

petroleum, and ethanol promises to be competitive as a pure fuel without subsidies if yields are increased, energy used more efficiently, operating costs reduced, and capital costs brought in line with those for corn processing. Less energy for size reduction, lower cost materials of construction, lower chemical use, and reduced process energy while maintaining or enhancing yields are all important pretreatment targets. Combining the operations for cellulase production, cellulose hydrolysis, and fermentation in a continuous operation now referred to as Consolidated BioProcessing (CBP) promises to dramatically reduce operating and capital costs for biological steps. Together such improvements would drop total cash and capital recovery costs to about \$0.13/liter (\$0.50/gallon) based on performance judged to be achievable for mature technology or \$0.09/liter (\$0.34/gallon) if near theoretical performance results. Thus, research targeted at advanced pretreatment and biological process integration could lead to a cost-competitive sustainable fuel with major benefits. Alternative uses for lignin, minerals, organic acids, and other components in cellulosic feedstocks and chemicals made from the low-cost sugars could increase profitability.

Enhanced cellulosic crop productivity will reduce land requirements for a given fuel production, promote economies of scale, and reduce feedstock costs. Greater impact for a given land use can be achieved with more efficient vehicles (e.g., fuel cells, hybrid vehicles). Because feed for livestock is grown on most agricultural land, recovering protein from fast growing crops such as switchgrass also used to produce ethanol could support making both fuel and food.

Several steps in cellulosic conversion are well positioned for commercial applications. Recombinant organisms are effective now for conversion of all five sugars in cellulosic biomass to ethanol. Although erroneous information may suggest otherwise, ethanol purification is not expensive or energy intensive. Rather, attention should be focused on significantly advancing pretreatment and cellulose conversion to make cellulosic ethanol costs competitive.

6. DIRECTIONS AND OPPORTUNITIES FOR WORLD ETHANOL PRODUCTION

Brazil is the leading producer of fuel ethanol in the world and is increasing the ethanol in blends from 22 to 26% and promoting 3% ethanol in diesel fuel.

However, replacement of ethanol vehicles with vehicles that burn low level blends will cut ethanol consumption unless large-scale purchase of hydrous ethanol fueled vehicles resumes.

Ethanol use is growing in the United States. The Clean Air Act Amendments of 1990 require 2% oxygen in gasoline to reduce vehicle tailpipe emissions of unburned hydrocarbons and carbon monoxide, and about 16.0 billion liters (4.2 billion gallons) of methyl tertiary butyl ether (MTBE) and 6.63 billion liters (1.75 billion gallons) of ethanol were used as oxygenates in 2001 to meet these requirements. Because MTBE works its way into groundwater from underground storage tanks, is a persistent contaminant, has a turpentine-like odor and solvent-like taste, and is a suspected carcinogen, U.S. legislation mandates phase out of MTBE. Ethanol does not have these problematic features and is the leading alternative. A renewable fuels standard (RFS) that bans MTBE, eliminates the oxygen requirement in gasoline, enhances clean air rules, and establishes a minimum renewable fuels standard that would grow to 5 billion gallons by 2012 appears likely. Passage of this or related legislation could triple ethanol use after 10 years and result in the United States overtaking Brazil as the world's largest producer of fuel ethanol. Enhanced ethanol production would reduce petroleum imports, the trade deficit, and greenhouse gas emissions, and would create domestic employment.

Other countries have begun or are about to begin ethanol production. Canada produced about 238 million liters (62.9 million gallons) of ethanol in 2001, with plans to more than double that output soon. Canada provides a federal excise tax exemption of \$0.10/liter (\$0.38/gallon) for the ethanol portion of blended fuels. Although most Canadian ethanol is from wheat and corn starch, Canada produces about 17 million liters (4.5 million gallons) of ethanol from a wood processing waste, spent sulphite liquor, but most of it is sold for industrial uses.

France has been the largest ethanol producing country in Europe, making about 115 million liters (30 million gallons) per year in 2000, mostly from beet sugar for conversion to ethyl tertiary butyl ether (ETBE) for use as an oxygenate in gasoline. However, Spain now has two grain plants with a total capacity of about 325 million liters (86 million gallons) per year, and a third plant planned to reach 526 million liters (139 million gallons) in capacity. Sweden made about 25 million liters (6.7 million gallons) of ethanol in 2000. Recently, the European Commission encouraged increased use of ethanol and biodiesel for

transportation with the goal of substituting 20% of diesel and gasoline by alternative fuels that reduce greenhouse gas emissions by 2020.

China planned to introduce 380 million liters (100 million gallons) of ethanol production capacity in 2001 and achieve a total annual ethanol production capacity of about 2.0 billion liters (534 million gallons) by late 2002. Mostly grain is used to relieve structural surpluses while also reducing oil imports, and blending ethanol will help address air pollution problems in major cities. India has a major program in progress to use some of their large sugar stocks to make ethanol for 5 and 10% blends with gasoline and plans for more than 30 facilities making up to about 380 million liters (100 million gallons) of fuel ethanol annually by the end of 2002. In addition, Thailand is pursuing production of ethanol from cane sugar and tapioca, corn, and rice starch with a near term goal of achieving an annual capacity of about 500 to 730 million liters (130 to 190 million gallons).

This expansion of ethanol production creates an immense opportunity for ethanol from cellulosic biomass. Cellulosic ethanol processes can be the low-cost producer of ethanol, and the transition to cellulose will become even more attractive as large-scale ethanol production from grains puts upward pressure on grain prices and likely reduces the selling price of coproducts. However, the perceived risk and higher capital costs of first-of-a-kind facilities present serious challenges to implementing this important technology. Advanced technologies that realize the low-cost potential of cellulosic feedstocks could lower capital investments and improve returns but delay implementation. Coupling development of advanced technologies with a commercial learning curve effect would accelerate the emergence of cellulosic ethanol as a competitive pure fuel without subsidies.

7. CONCLUSIONS

Considerable experience has been gained with low-level ethanol blends (10 and 22% ethanol in gasoline) and nearly pure (95.5% ethanol with the rest water and 85% with gasoline) in spark-ignition engines in Brazil and the United States that demonstrates its excellent fuel properties, and ethanol use is spreading. Ethanol can also be added to diesel fuel with additives and could undoubtedly provide hydrogen for fuel cells. Environmental attributes such as biodegradability and low toxicity avoid many problems with gasoline and MTBE.

When made from domestic biomass, ethanol provides an important alternative for a transportation sector almost totally dependent on economically, politically, and strategically vulnerable petroleum imports that continually climb. Furthermore, use of such resources can dramatically reduce emissions of greenhouse gases that trap radiation and cause global climate change.

Brazil led ethanol use as a pure fuel and for blending with gasoline, and a substantial fleet of dedicated ethanol vehicles evolved. However, world sugar demands and the phase down of incentives impacted pricing and availability, and dedicated vehicle production virtually stopped, hurting ethanol consumption. Because residues are burned to provide process heat and electricity, limited fossil fuel is needed to convert cane sugar into ethanol, and net emissions of carbon dioxide are much less than for gasoline. The cost of Brazilian cane sugar ethanol has dropped significantly through learning curve effects, but sugar prices and cultivation practices likely limit competitiveness with gasoline for large-scale uses.

In the United States, ethanol is produced from starch from corn and other crops, and the remaining components are used for feed and food to realize value from all of the grain. Energy consumption has dropped dramatically over time via technology advances and commercial learning curve effects, and corn ethanol now generates about one-third more energy than used to make it. Also, corn ethanol replaces about six times its energy content in petroleum, substantially reducing oil imports. Overall, the result is a reduction of about 40% in carbon dioxide emissions and 20 to 30% in total greenhouse gas emissions for E85 fuel relative to burning just gasoline with even greater reductions possible for pure corn ethanol use and powering the process by renewable resources. Further cost reductions would allow greater impact without tax incentives but become challenging if corn costs rise and coproduct prices drop as corn ethanol production expands.

Cellulosic biomass is abundant and competitive in price with petroleum. Costs for processing this recalcitrant, heterogeneous feedstock to ethanol have been reduced considerably, and companies are striving to introduce cellulosic technology commercially. Because burning lignin and other residues can provide the energy for conversion with electricity left for export and fossil energy requirements to provide biomass feed and distribute ethanol are low, at least four to more than seven times as much energy is contained in products than fossil energy inputs, and cellulosic ethanol would have less than 10 to

15% of the greenhouse gas emissions of gasoline for current technology. Once commercialized, learning-curve effects are expected to dramatically improve costs and environmental benefits, making it vital to successfully introduce cellulosic ethanol technologies soon. Costs and environmental attributes would further benefit through development of advanced technologies that reduce total production costs to compete with gasoline without subsidies and novel technologies that recover protein and other valuable cellulosic components. Ultimately, high productivity cellulosic crops would reduce land use and costs, while efficient transportation technologies would bring fuel demand in line with sustainable supplies. However, substantially more research is needed on advanced technologies to achieve these goals, and attention must be focused on overcoming perceived risk and capital investment challenges that now inhibit commercializing first-of-a-kind cellulosic ethanol processes before their major benefits can result.

Acknowledgments

This work was made possible through the support of the Thayer School of Engineering at Dartmouth College, for which the author is grateful. The Feldberg Library at Dartmouth College and particularly Karen Sluzenski are recognized for rapidly gathering valuable data. Hosein Shapouri of the United States Department of Agriculture and Michael Wang at Argonne National Laboratory provided important cost and energy balance information, and Luiz Carlos Correa Carvalho from UNICA in Brazil furnished Brazilian ethanol facts. The author appreciates input from several experts, including Jeffrey Bentley, Michael Gaylor, Donald Johnson, Lee Lynd, Jack Saddler, and Hosein Shapouri, who provided helpful suggestions. The National Renewable Energy Laboratory is commended for readily available, well-documented, detailed reports on cellulosic ethanol processes and economics.

SEE ALSO THE FOLLOWING ARTICLES

Biomass, Chemicals from • Biomass Combustion • Biomass for Renewable Energy and Fuels • Biomass Gasification • Biomass: Impact on Carbon Cycle and Greenhouse Gas Emissions • Biomass Resource Assessment • Greenhouse Gas Emissions, Alternative Scenarios of • Transportation Fuel Alternatives for Highway Vehicles

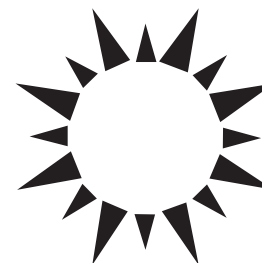
Further Reading

- General Motors Corporation, Argonne National Laboratory, BP, ExxonMobil, and Shell. (2001). "Well-to-Wheel Energy Use and Greenhouse Gas Emissions of Advanced Fuel/Vehicle Systems: A North American Analysis, Volume 2, Main Report," Argonne National Laboratory, Report No. ANL/ES/RP-104528, Argonne, Illinois, June.
- Goldemberg, J., Monaco, L. C., and Macedo, I. C. (1993). The Brazilian fuel-alcohol program. In "Fuels and Electricity from Renewable Resources" (R. H. Williams, T. B. Johansson, H. Kelly, and A. K. N. Reddy, Eds.), United Nations Solar Energy Group for Environment and Development (SEGED), 1992 UN Conference on Environment and Development in Brazil, Chapter 20, pp. 841–863.
- Lynd, L. R. (1996). Overview and evaluation of fuel ethanol from cellulosic biomass: Technology, economics, the environment, and policy. *Ann. Rev. Energy Environ.* 21, 403–465.
- Lynd, L. R., Cushman, J. H., Nichols, R. J., and Wyman, C. E. (1991). Fuel ethanol from cellulosic biomass. *Science* 251, 1318–1323.
- Moreira, J. R., and Goldemberg, J. (1999). The alcohol program. *Energy Policy* 27, 229–245.
- Shapouri, H., Duffield, J. A., and Wang, M. (2002). "The Energy Balance of Corn Ethanol: An Update," U. S. Department of Agriculture, Agricultural Economic Report No. 813, Washington, DC.
- Shapouri, H., Gallagher, P., and Graboski, M. (2002). "USDA's 1998 Ethanol Cost-of-Production Survey," U. S. Department of Agriculture, Agricultural Economic Report No. 808, Washington, DC.
- Wang, M., Saricks, C., and Santini, D. (1999, January). "Effects of Fuel Ethanol Use on Fuel-Cycle Energy and Greenhouse Gas Emissions," ANL/ESD-38, Center for Transportation Research, Argonne National Laboratory, Argonne, IL.
- Wheals, A. E., Basso, L. C., Alves, D. M. G., and Amorim, H. V. (1999). Fuel ethanol after 25 years. *TIBTECH* 17, 482–487.
- Wooley, R., Ruth, M., Glassner, D., and Sheehan, J. (1999). Process design and costing of bioethanol technology: A tool for determining the status and direction of research and development. *Biotechnol. Progress* 15, 794–803.
- Wooley, R., Ruth, M., Sheehan, J., Ibsen, K., Majdeski, H., and Galvez, A. (1999, July). "Lignocellulosic Biomass to Ethanol Process Design and Economics Utilizing Co-Current Dilute Acid Prehydrolysis and Enzymatic Hydrolysis: Current and Futuristic Scenarios," National Renewable Energy Laboratory, Report NREL/TP-580-26157 Golden Colorado.
- Wyman, C. E. (1999). Biomass ethanol: Technical progress, opportunities, and commercial challenges. *Ann. Rev. Energy Environ.* 24, 189–226.
- Wyman, C. E. (Ed.) (1996). "Handbook on Bioethanol: Production and Utilization." Applied Energy Technology Series, Taylor and Francis.
- Wyman, C. E., Hinman, N. D., Bain, R. L., and Stevens, D. J. (1993). Ethanol and methanol from cellulosic biomass. In "Fuels and Electricity from Renewable Resources" R. H. Williams, T. B. Johansson, H. Kelly, and A. K. N. Reddy, Eds.), United Nations Solar Energy Group for Environment and Development (SEGED), 1992 UN Conference on Environment and Development in Brazil, Chapter 21, pp. 865–924.

European Union Energy Policy

FELIX C. MATTHES

Oeko Institute, Institute for Applied Ecology
Berlin, Germany



-
1. European Union
 2. Areas of Regulation Concerning Energy Policy
 3. Energy Policy Structures
 4. Energy Policy, Competition, and The Single Market
 5. Energy, Environment, and Climate Protection Policy
 6. Nuclear Energy and Energy Policy
 7. Energy Research Policy
 8. Security of Energy Supply
 9. Energy Policy and EU Enlargement
 10. Energy Policy and External Relations

Glossary

acquis communautaire The entire body of European laws.

This includes all the treaties, regulations, and directives passed by the European institutions as well as judgments laid down by the European Court of Justice.

European Community (EC) The generic term for three different European communities: the European Coal and Steel Community (Treaty of Paris, 1951), the European Atomic Energy Community (Treaty of Rome, 1957), and the European Economic Community (Treaty of Rome, 1957). The EC formed the umbrella organization for these three communities following the signing of the fusion treaty in 1967.

European Economic Area (EEA) An agreement in 1992 between the European Community and three of the four members of the European Free Trade Association (i.e., Norway, Iceland, and Liechtenstein) to form a free trade area in all markets but agriculture; previously called the European Economic Space.

European Free Trade Association (EFTA) An organization founded in 1961; EFTA aims at supporting free trade among its members and increasing the liberalization of trade on a global basis, but particularly within Western Europe. Member states are Iceland, Liechtenstein, Norway, and Switzerland.

European Union (EU) The collective term for the European Community, the Common Foreign and Security Policy, and cooperation in the areas of justice and home

affairs formed by the European Union Treaty (Maastricht Treaty, 1992).

European Union burden sharing The sharing out of emissions allowances among the 15 member states under the EU umbrella. The agreement on burden sharing was reached in June 1998 and is legally binding as part of the EU's instrument of ratification of the Kyoto Protocol.

single market (internal market) An economic area without internal restrictions, in which the four fundamental freedoms (free movement of goods, persons, services, and capital) can be realized. The Community's internal market project ("Europe 1992") was initiated during the 1980s and is the main reason for the huge movement toward integration at the end of the 1980s and beginning of the 1990s.

The European Union (EU) is one of the largest economic entities in the world. Although there is a significant and fast integration of many relevant policies of the EU member states, the development of a common EU energy policy in its different dimensions (single market, environment, security of energy supply, research, external relations, etc.) is a complicated process. Furthermore, the enlargement of the European Union creates new challenges for energy policy and the energy policy arena in Europe.

1. EUROPEAN UNION

With 15 states, a population of approximately 375 million, and economic output of roughly 8500 billion euros (€8500 billion), the European Union comprises one of the largest economic areas in the world. In 2004, 10 more states, primarily in Central and Eastern Europe, will join the European Union. The population of the EU will then increase to approximately 450 million, although the economic output will rise to only about €9000 billion as a result of the enlargement. The economic output is expected to rise to approximately €13,000 billion by

2020, although the population is predicted to decrease slightly for demographic reasons.

The primary energy consumption of the current 15 member states of the European Union (EU-15) is about 1450 million metric tons of oil equivalents; this has increased by ~1% annually in the past decade (see Table I). Oil is the largest source of primary energy in the EU-15, accounting for over 40% of energy consumption, followed by natural gas, which accounts for over 20%, and coal, accounting for ~15%. The dependency of the EU on energy imports has increased steadily in the past few decades. Over 70% of oil consumption in the EU is covered by imports, and natural gas

and coal imports account for roughly half of the total demand.

Sources of energy in the EU that can be used at competitive costs are very unevenly distributed. The main countries that have large natural gas reserves are the United Kingdom, The Netherlands, and Denmark; it is expected that production in all three countries will decrease. Out of all the member states of the EU-15, only the United Kingdom, Germany, and Spain have large coal reserves, and coal production can only be sustained with massive subsidies. Substantial lignite volumes are produced in Germany and Greece (Table I). Nuclear power plays a very different role in each of the EU member

TABLE I

Production and Consumption of Fossil Fuels in the European Union^a

	Hard coal	Lignite	Crude oil	
Member state	(× 10 ³ metric tons) ^b			Natural gas (TJ) ^c
Austria	—	1.3 → (97%)	1.1 ↓ (13%)	71.3 ↑ (23%)
Belgium	—	—	—	0.1 → (0%)
Denmark	—	—	17.8 ↑ (221%)	344.8 ↑ (167%)
Finland	—	5.6 ↓ (84%)	—	—
France	2.0 ↓ (15%)	0.3 ↓ (84%)	1.7 ↓ (2%)	69.6 ↓ (4%)
Germany	30.4 ↓ (58%)	175.4 → (95%)	3.2 ↓ (3%)	735.0 ↑ (22%)
Greece	—	67.0 ↑ (99%)	0.3 ↓ (1%)	2.0 ↓ (2%)
Ireland	—	—	—	44.6 ↓ (28%)
Italy	—	—	4.6 ↓ (5%)	633.7 ↓ (24%)
Luxembourg	—	—	—	—
Netherlands	—	—	2.4 ↓ (4%)	2414.6 ↓ (150%)
Portugal	—	—	—	—
Spain	14.5 ↓ (40%)	8.7 → (101%)	0.2 ↓ (0%)	6.9 ↓ (1%)
Sweden	—	—	—	—
United Kingdom	32.5 ↓ (52%)	—	126.2 ↑ (147%)	4542.6 ↑ (112%)

^a Production for 2000/2001 and consumption trends for the past decade.

^b The consumption trend is indicated by the arrow position; the percentage of gross domestic consumption is shown in parenthesis.

^c Gross calorific value.

states. Only eight member states of the EU-15 operate nuclear power stations, and nuclear power accounts for ~33% of the total electricity volume overall within the EU-15, a proportion that remained more or less constant throughout the 1990s. The output of nuclear power stations has risen from ~117 to 125 gigawatts (GW) since 1990.

The framework for energy policy in the European Union is therefore based on a common economic area of considerable size, on the one hand, and on highly varied (1) energy industry structures in relation to primary energy volumes, (2) dependency on energy imports, and (3) orientation to specific energy sources, on the other hand. Ultimately, the basic conditions for European Union energy policy will change substantially as a result of the addition of new members in 2004.

2. AREAS OF REGULATION CONCERNING ENERGY POLICY

Since the beginnings of the European communities, the energy sector has formed a constituent element. Two of the three European treaties (European Coal and Steel Community and European Atomic Energy Community) particularly relate to the energy sector or to specific segments of the energy industry, even if the main reasons for concluding the treaties were primarily connected to security policy rather than to energy policy. The creation of the Single European Market, which was initiated in 1986 and includes the energy sector, lent a new quality to the energy policy and energy industry dimensions of the European Community (EC). However, the European Union has no direct jurisdiction over a European energy policy, and energy policy is the sole responsibility of the member states. Despite several attempts, all efforts so far to incorporate an energy chapter, which could be used to establish EU jurisdiction over energy policy, into the European Community Treaty have failed. Nevertheless, many of the areas of regulation that concern the energy sector directly or indirectly are influenced considerably by the European Union, even though, in practice, policy integration beyond the fundamental jurisdiction of the EU has been furthered in the different areas to a varying extent. The development of the European Single Market is very advanced. The liberalization of the electricity and gas markets is providing increasing dynamism for the energy sector. The development of trans-European networks belongs to the areas of energy policy that

are explicitly within the jurisdiction of the EU. In the area of competition policy, the strict aid regulations concerning the reduction of state subsidies have, in particular, had a considerable effect on energy policy.

Integration has made relatively substantial progress in traditional environmental policy regulation areas. Above all, regulations for reducing conventional pollutant emissions are relevant to the energy sector in this respect. A common policy on climate protection is in the initial stages, but is becoming increasingly dynamic. In the area of nuclear energy, there are specific Community regulations that concern the area of research and development, the provision of beneficial financing, and the ownership and control of fissionable materials, whereas regulations concerning nuclear safety are strictly within the jurisdiction of the member states. A whole range of Community approaches to policy concern the area of research, development, and demonstration, albeit with very different equipment, deadlines and continuity.

The dialogue with energy-producing states is gaining in importance in the European Union's external relations. The strong relevance to energy policy of many EU regulations in related policy areas (single-market policy, competition policy, environment policy, etc.), despite the EU's lack of energy policy jurisdiction, leads to the situation that the principle of unanimity is claimed for policies and measures that have a significant effect on the energy sector (e.g., some fields of environmental policy). In reality, and above all in the 2004 new-member process, the veto option made possible with this principle is only of secondary importance.

3. ENERGY POLICY STRUCTURES

The European Commission has a key role in European energy policy. The Commission has the sole right of initiative for the legislation process and has a multitude of implementation, arbitration, and monitoring functions and the like. The Commission's Directorate-Generals for Energy and Transport, Competition, the Internal Market, and Environment are particularly relevant to energy policy. The governments of the member states participate in the decision-making processes through the Council of Ministers of the EU. The course of important compromises in Community policies is set here. The Council of Ministers for Energy, the Council of Ministers for the Environment, and the Council of Ministers for Finance are the main decision-making bodies important for energy policy.

The role of the European Parliament, which is directly elected by EU citizens, has become much more significant over the past few years. For most, but not all, legislation processes in the EU, the European Parliament must be involved on a graded basis. This involvement may take the form of codecision, assent, or hearing. For activities concerning energy policy, the codecision process is most important: three readings take place in both the Parliament and the Council of Ministers and a Conciliation Committee procedure may be initiated if necessary. The most important committees for energy policy in the European Parliament are the Committee on Industry, External Trade, Research, and Energy, as well as the Committee on the Environment, Public Health, and Consumer Policy.

European Union law arises from the founding treaties of the Community (primary law) and from secondary law, i.e., the legal instruments produced by the institutions of the Community. European Union regulations are wholly binding and are directly valid in each member state. European Union directives are binding for each member state with respect to the result to be achieved, and the directives must be translated into domestic law. However, the member states do have some freedom to choose the form and methods. The Court of Justice of the European Communities has gained in importance for energy policy. It checks the legal instruments of the EU and the member states to ensure that they are consistent with the treaties, and it decides on the requests of national courts concerning the interpretation or application of European Union law.

4. ENERGY POLICY, COMPETITION, AND THE SINGLE MARKET

The development of the European single market for energy forms a crucial basic condition for European energy policy. On the basis of the directives for liberalization of the electricity and gas sector that came into effect in 1997, the member states had to open their electricity and gas markets to competition by 1999 and 2000, respectively, and had to meet specific minimum requirements for regulation (decartelization, network access provisions). However, the minimum requirements for opening the markets (33% through 2003) and for network access transparency (possibility of negotiated network access), as well as the unbundling requirements for the companies, remained low.

Although the creation of a European single market was a priority during the liberalization, the electricity and gas markets essentially remained national markets, despite the fact that they were opened for competition. Energy prices became particularly dynamic, even on the national markets, and dropped considerably, especially at wholesale level. Even 4 years after the markets opened, changes in supplier are the exception rather than the rule in most member states, both for large customer groups and other consumer groups. The directives were revised in 2003 after the various member states found liberalization to be sluggish in reality. Accordingly, all member states must amend their national legislation by 2007 such that the electricity and gas markets are fully open to competition. The minimum requirements for companies (more rigorous rules for unbundling of generation, transmission, distribution, and sales), for regulation of the electricity and gas markets, and for consumer information are being intensified. The conditions for cross-border energy supply within the European Union are regulated directly by a corresponding European Union regulation. In addition, the EU has identified a range of infrastructure projects for electricity and natural gas, with which network bottlenecks are to be eliminated and cross-border trade intensified.

In contrast to the grid-bound final customer markets, the energy supply companies have become Europe oriented relatively quickly. A large number of company transactions in which the major European companies (EDF Energy, E.ON Energie, RWE Solutions, Vattenfall, etc.) played a particularly active role have been performed within and between the member states. These companies have stopped pursuing a large number of activities in the nonsupply industry and are instead pursuing massive growth strategies in the core business of energy through mergers and acquisitions. Non-European supply companies are rarely involved. Due to differences in timing of implementation of the market opening process, and also to a number of national conditions, the process within the European Union has involved in part a strong increase in market concentration. Trends toward oligopolies, however, are emerging on some electricity and gas supply submarkets in Europe. Expansion of the dominant market position of some large European suppliers was also evident during the privatization process in Central and Eastern Europe and the Baltic states.

State aid for the energy sector is strictly forbidden in the European Union, unless specific assent has been given. This has particular significance for state-

subsidized coal production in the EU, which, at €6.2 billion annually, still reaches a substantial level; German coal production is awarded the largest proportion (over 65%) of these subsidies. In the past, coal subsidies have steadily been reduced, primarily due to pressure from the European Commission, and they will continue to decline in the future. The reduction of regulations specific to energy sources in the European single market is also evident from the fact that one of the EU founding treaties—the treaty establishing the European Community for Coal and Steel (ECCS)—was not extended when it expired after 50 years on 23 July 2002. After it expired, coal and steel gained the status of normal industrial products. Via a special regulation of the Council, degressive aid for securing resources, for reducing production activity, and for extraordinary financial burdens (inherited burdens) has been declared permissible, at least until 2007, without it having to be approved in individual cases. If state aid is to be used for the production of renewable energy, it is regulated by Community guidelines governing state aid for environmental protection.

Discussions within the European Commission on the admissibility of so-called national energy bases (in other words, specific proportions of primary energy consumption for which advanced subsidy options could be made available to the member states) have been very controversial.

5. ENERGY, ENVIRONMENT, AND CLIMATE PROTECTION POLICY

Based on the Community's jurisdiction over environmental protection, the European Union is implementing a whole range of policies that have considerable effects on the energy sector. The range of policies extends from EU measures with immediate effect, to directives involving limited degrees of freedom for the member states when they are implemented, to the specification of goals that can be realized by the member states in any way they wish.

Climate protection is one of the strategic priorities of the EU in relation to sustainable development. A Community strategy on climate protection is inevitable because the EU has ratified the Kyoto Protocol and has therefore undertaken a double commitment. On the one hand, the EU, as an independent party to the agreement, has promised to reduce greenhouse gas emissions by 8%, from 1990 levels, in the

initial period of 2008–2012; on the other hand, the member states have undertaken varied commitments as part of EU burden sharing, ranging from promises to make reductions (Luxembourg, 28%; Germany and Denmark, 21%; United Kingdom, 12.5%) to achieving growth restrictions (Spain, +15%; Greece, +25%; Portugal, +27%). For this purpose, the EU has developed a European Climate Change Program (ECCP), which comprises a number of measures that are also relevant to the energy sector. They include a directive on the promotion of electricity produced from renewable sources (2001), a directive on the promotion of cogeneration (2003), a directive on energy performance of buildings (2003), and a directive on energy services (planned for 2004). In contrast to the directives on liberalization, these directives generally contain only basic restrictions and definitions, and in some cases indicate target goals for expanding specific technologies. For example, in the field of renewable energy, the target is to increase the contribution of renewable energy to total primary energy consumption from 6 to 12% by 2010 and to increase its contribution to electricity consumption from 11 to 22%. Because implementation of the directives allows member states considerable freedom, it cannot be assumed that the targets will actually be achieved. However, more binding specifications result from a range of other measures. These include, for example, Commission directives for labeling electric appliances with energy efficiency data (1999), for adopting the Energy Star system for office equipment from the United States (2001), and for creating minimum requirements for the energy efficiency of electrical devices (refrigeration equipment, 1996; washing machines, tumble driers, ovens, water heaters, television sets, video equipment, electric motors, pumps, air compressors, etc., per agreements negotiated with industry at the end of the 1990s).

The initiatives pursued in the European Climate Protection Programme and in an independent Community strategy for promoting energy efficiency serve to realize the goal of improving energy productivity in the EU-15 annually by an additional 1% compared with the usual levels. However, the most prominent basic conditions for European Union energy policy result from an EU-wide emissions trading scheme for greenhouse gases, the introduction of which is planned for 2005. CO₂ emissions should be restricted for all industrial plants with a thermal input capacity greater than 20 MW and made flexible using a Europe-wide system of tradable certificates. In addition to a range of Europe-wide

specifications (participating sectors, coverage of carbon dioxide only, incorporation of the flexible mechanisms of the Kyoto Protocol, Joint Implementation and Clean Development Mechanism, free allocation of most of the emission allowances at first, etc.), the member states have considerable freedom to shape the system, particularly when distributing the emission rights to the different plants on the basis of the national allocation plans.

Considerable progress has been made in the reduction of conventional pollutant emissions in the energy sector. Sulfur dioxide emissions from power plants in the EU-15 dropped by almost 60% in the 1990s, and the corresponding nitrogen oxide emissions decreased by roughly 25%. Particularly with the directive on the limitation of atmospheric emissions of certain pollutants from large combustion plants (2001), similar demands have been established at EU level for ecological standards for power plants in terms of conventional pollutants, and hence a consistent ecological framework has been created for competition in the field of thermal power plants. New power plants must meet strict standards, and existing plants must observe defined emission ceilings and targets. The directive on national emission ceilings for certain atmospheric pollutants (2001), which sets ambitious reduction targets for the entire spectrum of conventional atmospheric pollutant emissions within the respective member states, forms a framework beyond the energy industry in the strictest sense, but still has an important formative influence on the energy sector. One of the main targets of this directive is to halve, within a reasonable period, the number of areas in which the critical load of acidifying substances is exceeded.

6. NUCLEAR ENERGY AND ENERGY POLICY

Nuclear energy plays an important role in the energy supply of the European Union. Eight of the fifteen EU member states operate a total of 145 nuclear reactors with a net output of roughly 120,000 MW, which corresponds to about one-fifth of the entire power plant capacity in the EU-15. The most recent nuclear power plant in the EU-15 was commissioned in France in 1999.

In the EU member states that operate nuclear power plants, the proportion to power supply contributed by nuclear energy varies. In The Netherlands, ~4% of all electricity is generated by nuclear

power plants, and in Finland, Germany, Spain, and the United Kingdom, the contribution is between 26 and 33%; in Sweden, Belgium, and France, it is 47, 58, and 75%, respectively. Three member states (Belgium, Germany, and Sweden) pursue an active policy of withdrawing from nuclear energy. Two member states (Finland and France) pursue a policy of setting up new nuclear power stations, although an actual project is underway in Finland alone. In the remaining three countries (The Netherlands, Spain, and the United Kingdom), it is expected from today's perspective that the use of nuclear energy will be phased out. Out of the 10 new EU member states, five operate nuclear power plants, which are predominantly of Russian design.

The treaty on establishing the European Atomic Energy Community (Euratom Treaty), which was signed in 1957 and is valid for an indefinite period, created the basis for a range of specific regulations on the use of nuclear energy; unlike other areas of regulation in the EU, these regulations should actually be part of secondary legislation. First, these special regulations relate to specific problems in the use of nuclear energy. This includes monitoring nuclear fuel in the EU and a range of other regulations on radiation protection. It must be emphasized in this respect that the nuclear fuels used in the EU are officially the property of the Community and the nuclear power operators have only rights of use. Second, there is a whole range of regulations concerning the promotion of nuclear energy. They include the promotion of research specifically concerning the use of nuclear energy, as well as special financing possibilities for nuclear energy projects. Funds for research into nuclear fission and nuclear fusion amounted to between €1.2 billion and €1.3 billion for each 4- or 5-year Euratom research program (see Section 7) in the past few years. In the framework of Euratom loans, the EU provides investments for nuclear projects, the (very favorable) refinancing conditions of which are based on those for public borrowers. In the period 1977–2002, a total of €3.2 billion was available through Euratom loans, which were also provided for nuclear projects in Central and Eastern European states, as well as in the states of the former Soviet Union, in order to commission new nuclear power plants, among other things. In 2002, the European Commission presented a proposal to increase the Euratom loan ceiling by a further €2 billion, which led to heated debates on the legitimization of Euratom loans, e.g., before the background of market distortions in favor of nuclear energy. Third, other regulations focus on the

Community control of fuel procurement. Based on the original assumption that diversification is a necessity due to a shortage of fuel, nuclear fuel procurement is predominantly handled by a separate institution, the Euratom Supply Agency, and strict checks are carried out to ensure adherence to specific import quotas (for example, from Russia). This has led to intervention in private contracts of individual power plant operators. Finally, the legislation procedures in the Euratom Treaty differ greatly from those in other EU regulations. In particular, this affects the participation of the European Parliament, whose opportunities for involvement and influence in the context of Euratom fall considerably short of the procedures provided for in the EU treaty.

With regard to nuclear energy, the concentration of binding regulations at the EU level is comparatively low, except in the fields of proliferation and health-related radiation protection. However, in the important and—in terms of the common market and competition provisions—cost-related regulation areas of disposal and liability, the regulations of the member states dominate. Therefore, the service specifications, implementation periods, and reserve variants that are relevant to fuel and plant disposal vary greatly among the EU member states, leading to significant differences in the disposal cost burdens of nuclear power generation, in addition to varying implementation efficiency. Companies operating nuclear power plants in EU member states must also make liability provisions to a widely varying extent, which is also connected to differing cost burdens. With regard to safety standards, the lack of coordinated regulations, particularly in the context of EU enlargement, has proved to be a major failing. Against this background, attempts have been made since 1999 to develop a common approach to reactor safety and corresponding regulation procedures in the Western European Nuclear Regulators Association (WENRA); however, the objective of establishing coordinated and binding safety standards for the use of nuclear power has not yet been achieved.

7. ENERGY RESEARCH POLICY

The European Union promotes energy research through specific programs. A distinction is made between nuclear and nonnuclear research in this respect. Nonnuclear energy research is implemented through the EU's general framework program for research. The sixth and most recent EU Framework Programme for Research and Technological

Development (2002–2006) has a total budget of €17.5 billion, €2.12 billion of which is earmarked for sustainable development, global changes, and ecosystems. This also includes the field of sustainable energy systems, for which €810 million has been appropriated. Compared with the corresponding research field in the fifth Euratom framework program for research (1998–2002), this constitutes a reduction in funds (€1.042 billion was employed for nonnuclear energy research previously). The fourth Euratom framework program for research (1995–1998) had a budget of €1.03 billion in the field of research, development, and demonstration concerning nonnuclear energy technologies.

Renewable energy systems (including integration into energy supply systems), energy saving technologies, hydrogen and fuel cells, CO₂ separation and dumping, and socioeconomic research in the energy sector are promoted above all in this regard. Fossil energy generation technologies play a much smaller role in the current research program than in previous promotion programs. Research specifically for the coal sector was promoted until 2002 through the research and development programs of the European Community for Coal and Steel (ECCS). It is stipulated in the agreements on the expiry of the ECCS in July 2002 that income from the remaining assets of the ECCS will be placed in a research fund for coal and steel; 27.2% of this is earmarked exclusively for the coal sector.

The promotion of nuclear energy research is implemented in a special program in the scope of the Euratom Treaty. A total of €1.23 billion is provided for in the sixth Euratom research program for the period 2002 to 2006. The majority of the budget (€750 million) is earmarked for fusion research, the International Thermonuclear Experimental Reactor (ITER) project being the focus of the research. Research in the field of radioactive waste, radiation protection, plant safety, and new reactor designs is supported by the remaining funds. The budget for the promotion of research by Euratom has remained largely constant in comparison with previous promotion programs. In the fifth Euratom framework program for research (1998–2002), a total of €1.26 billion was employed, whereas the fourth Euratom framework program for research (1995–1998) had €1.336 billion.

Beyond research, the EU establishes framework programs lasting several years for supporting the field of energy research. The total budget for the energy promotion program from 1998 to 2002 was €175 million, the majority of which was earmarked

for energy saving (SAVE program; €66 million) and renewable energy (ALTENER program €77 million). The "Intelligent Energy for Europe" program (2003–2006) has a total budget of €215 million, which is to be used primarily in the areas of energy efficiency (continuation of SAVE), renewable energy (continuation of ALTENER), corresponding activities in the international field (COOPENER), and energy matters in the transport sector (STEER).

8. SECURITY OF ENERGY SUPPLY

The major oil price fluctuations in 1999/2000 led to discussions on the security of energy supply in the European Union being conducted with renewed intensity. The European Commission has drawn up a well-heeded discussion framework for this, as presented in the Green Paper, "Towards a European Strategy for the Security of Energy Supply." The particular challenges of energy supply security for the first two to three decades of the 21st century are a result of the EU's increasing dependency on imports (from the current 50% to an estimated 70% in 2030), the necessary reinvestment in the European energy industry on a large-scale, and increased environmental and climate protection requirements.

The main features of a long-term European Union strategy on energy policy are as follows:

- An increase in energy efficiency on the demand side of the market is to form an important core of future EU energy policy, because an increased energy supply from the EU member states cannot be developed under acceptable conditions.
- The precedence of climate-friendly energy sources, such as new and renewable energy sources for generating electricity and heat as well as for the transport sector; a further discussion on the use of nuclear energy in Europe is also required for this purpose in the view of the European Commission.
- The coordinated increase of strategic stocks of crude oil and natural gas.
- The development of new sources for energy imports and new import channels for crude oil and natural gas.

Even though the debate initiated by the European Commission on security of energy supply in the EU is clearly driven by a whole range of ulterior motives (fight for more jurisdiction for EU institutions over energy policy, attempts to reevaluate nuclear energy, etc.), and even though, despite an extensive discussion

process, the Green Paper has continued to be merely discussion input from the European Commission (and other EU institutions, such as the European Parliament, have not become involved), the debate on energy supply security for EU energy policy will form an important framework in the future.

9. ENERGY POLICY AND EU ENLARGEMENT

Essential changes to the basic political conditions for energy policy in Europe will emerge from enlargement of the European Union in 2004. In addition to the accession of 10 states in mid-2004 (Malta, Cyprus, the Czech Republic, Hungary, Poland, Slovakia, Slovenia, Estonia, Latvia, and Lithuania), negotiations are continuing with other states (Bulgaria and Romania). The acceptance of Turkey into the EU will remain a highly controversial issue in the next few years. Other potential applicant states are those in the Western Balkans. The number of EU member states will therefore rise from 15 (EU-15) to 25 (EU-25) in the short to medium term. In the medium to long term, the number of member states may increase to over 30 (EU-30+). In addition, the European Free Trade Association (EFTA) regulations and bilateral regulations with the non-EU states of Norway and Switzerland also form a specific European framework for energy policy. The decision-making structures in a European Union with 25, 30, or more member states will (have to) change considerably. Majority decisions and a clearer allocation of tasks will play a larger role if the capacity of the Community and member states to act is to be retained and expanded. At the same time, competition for available funds will increase.

The European Union's enlargement to the east is the enlargement with the greatest economic divide compared with previous enlargement phases. In the four enlargement phases up to 1995, the economic strength of the new member states was comparable to, or even greater than, that of the European Union. The states to be incorporated into the EU in 2004 have an average gross domestic product (GDP) level per resident of only ~20% of that of the EU today. The adjustment of the economic strength of the accession states to the level of the EU requires strong growth processes in these countries on the one hand, and a large rise in the cost efficiency requirements of EU policies on the other hand. With respect to their energy markets, the

future EU member states face a threefold challenge that none of the current EU members had to confront. They must (1) liberalize energy prices and the energy markets in a relatively short space of time and at the same time must (2) renovate and modernize their plants, and (3) they have much less economic strength and purchasing power compared to the EU as it is today.

The restructuring of the energy sector is progressing to a varying degree in the different states and is generally also closely associated with the privatization of the energy industries. The motives for privatization are varied, but are often mainly due to the generation of income as well as the necessity and expectation of a considerable influx of capital from international investors. In the privatization procedures that have taken place so far, the companies that were already strongest in the EU were particularly successful, so privatizations in the EU accession states have already contributed to a further concentration of companies and to a substantial increase in the market power of individual energy suppliers in the European energy industry. This results in competition-related and liberalization-related challenges for forming the basic conditions of the energy markets.

In many of the accession states, energy efficiency is much lower compared to the current level in the EU-15. Although the primary energy consumption per resident in the EU-15 is mostly (much) higher than it is in the accession states, energy productivity (expressed as primary energy consumption per unit GDP on the basis of purchasing power parities) is poorer in most of the accession states, compared to the EU-15 average, by a factor of roughly two. The degree of macroeconomic stabilization in the accession states will, in particular, play a large role in the creation of stable basic conditions for investment, particularly in the area of energy efficiency. Economic stability and a high degree of legal security are central requirements for corresponding investment. In view of the restricted purchasing power and rising energy prices, primarily for households and small consumers as a result of liberalization, increasing energy efficiency also has a strong social element.

The situation concerning emissions corresponds to the initial condition of energy productivity. There is a much higher specific output of the main pollutants (carbon dioxide and sulfur dioxide) in most states than there is in the EU-15. This means substantial exertions and expenditure are necessary for the states to observe the EU environmental

standards quickly; in part, the air pollution control measures have already progressed relatively far, however.

In principle, at the date of accession, the entire body of European legislation (*acquis communautaire*) will be applicable to the accession countries, obliging them to fulfill the specific requirements of each piece of legislation. Against this background, EU enlargement creates for all applicant countries a process, in the course of which significant progress has to be achieved before official accession. The interaction of this basic requirement with economic restructuring has led to significant improvements in economic efficiency and in the quality of the environment. For example, total sulfur dioxide emissions from power plants in the Czech Republic have been reduced by 88% in the period from 1990 to 2000. Relevant reductions can also be observed for greenhouse gas emissions. Due to the provisions of the Kyoto Protocol, however, greenhouse gas reductions of the accession countries cannot be taken into account in compliance on the part of the European Union until 2012. The primary energy structure of an expanded EU will not change much on the whole as a result of enlargement. Nevertheless, in some of the future member states, substantial moves toward mineral oil and natural gas are to be expected. There are additional challenges for some countries that have a very specific energy industry structure.

Poland, the largest producer of coal in Europe today, will become a member of the EU. Although coal production has almost halved since the end of the 1980s, Polish coal production in 2001, at ~100 million t, was 31% higher than the entire coal production in the EU. Due to the unfavorable geological conditions, Polish coal extraction has been classified as being just as uncompetitive as Western European production, so the massive reduction in subsidies deemed necessary by the EU will result in a considerable drop in production. The discussion on coal subsidies will therefore become more explosive in terms of European politics. Among the new and potential EU member states, some countries are extensive lignite producers (e.g., Poland, Czech Republic, Hungary, and Bulgaria). Because only Germany, Greece, and Spain (although the latter is cutting back) lay claim to significant lignite production in the EU, the role of lignite will change in the short, medium, and long term in an enlarged EU.

Substantial growth in oil and natural gas consumption is expected in most accession states in the

next few years. On the one hand, this places considerable demands on the inclusion of these states in the stockpiling regime of the International Energy Agency (IEA). On the other hand, the natural gas markets in most of the states applying for EU membership are still strongly oriented to procuring natural gas from Russia. The diversification of natural gas supplies for states in Central and Eastern Europe will therefore form an important factor in energy policy in the medium term.

With the enlargement of the EU, the Community is faced with specific safety problems concerning substantial nuclear power plant capacity (mostly Soviet-designed plants), even though nuclear power is a less significant energy source in most of the accession states than it is in the EU today. Because no generally binding safety standards apply at the EU level, only a comparison with the practices and regulations applicable in all member states was used as a measure for demands for high safety standards in the field of nuclear power in the EU accession process. An enlarged EU will then be confronted with a whole range of reactors that, in the general view, do not satisfy the requirements and practices commonly accepted in the EU and have (serious) design faults. Therefore, in the course of EU accession, the closure of reactors has been agreed with some states, although the implementation of this agreement has not yet been fully guaranteed. Other reactors need to be retrofitted urgently, but there is a dispute about whether and to what extent “acceptable safety levels” can be achieved through retrofitting measures, particularly in light of the fact that construction weaknesses and quality shortcomings have arisen when plants have been commissioned. Many questions concerning the ultimate disposal of highly radioactive waste and the financing of nuclear plant closures have not been answered at all by representatives of most of the new EU member states.

District heating accounts for a large proportion of the total energy supply (up to 25%) in most of the Central and Eastern European accession states, and in the Baltic accession states. This figure is much higher than the EU-15 average ($\sim 2\%$), and is also higher than the figure in states with a high proportion of district heating (Denmark, 15%; Finland, $\sim 11\%$; Germany, 4%). The renovation and modernization of district heating systems and the introduction of highly efficient combined heat and power (CHP) production is not only an area for action concerning the energy industry in the Eastern European EU applicant states, but also has far-reaching social effects.

10. ENERGY POLICY AND EXTERNAL RELATIONS

Energy policy is a feature in the external relations of the European Union to an increasing extent. This is mainly because of two processes. First, the EU set the target of creating a European single market that includes other countries, in addition to the EU member states. Second, there is also a need for stronger cooperation with some energy-exporting regions in light of the EU enlargement, and also due to developments on the global energy markets in view of energy supply security. An essential determinant for the EU's relations with its immediate neighbor states is the aim of developing network infrastructures in such a way that at least 10% of the respective domestic electricity consumption will be covered by cross-border supply.

The European Union is also trying to influence the competition structures in neighbor states. For example, the bargaining cartel of Norwegian gas producers for natural gas supplies in the EU was broken up due to pressure from the European Union. Establishment of comparable market access and specification of equal standards (especially in the field of nuclear energy) are at the forefront of single-market activities in relation to third-party states.

The EU has institutionalized three fields of external relations and has also defined further areas of particular interest in relationship to Russia, the Mediterranean countries, and southeast Europe. In October 2000, the EU agreed on a regular European Union–Russia Energy Dialogue proposal to develop a strategic energy partnership. Because Russia is the largest single partner of the EU for energy supplies, long-term natural gas supply agreements, the establishment of legal and investment security for energy projects, trade in nuclear materials, and environment and climate protection issues (with the Kyoto Protocol to the Framework Convention on Climate Change in mind) are at the forefront. Previously, in 1997, the European–Mediterranean Energy Forum was established in order to increase cooperation with a total of 12 Mediterranean countries (Egypt, Israel, Jordan, Lebanon, Malta, Morocco, the Palestinian National Authority, Syria, Tunisia, Turkey, and Cyprus, with observer status for Libya and Mauritania). The free trade area planned for 2010, energy supply security issues, and environmental projects in the energy industry are at the forefront here. Finally, since March 2002, the European Commission has been submitting proposals for setting up a regional electricity and natural gas single market

in southeast Europe, and it has signed corresponding agreements (Albania, Bosnia and Herzegovina, the Federal Republic of Yugoslavia, Bulgaria, the former Yugoslav Republic of Macedonia, Greece, Turkey, Croatia, Romania, Kosovo, with observer status for the Republic of Moldavia, Slovenia, Austria, Hungary, and Italy). Cooperation and planning work has been undertaken by the different state and industrial institutions in the so-called Athens Forum.

In addition to these comparatively strongly institutionalized external relations, the European Union has defined three areas of particular interest for EU energy policy. These include, first, the so-called Northern Dimension, which aims to increase cooperation between Baltic states and Russia (especially in the field of nuclear security and waste disposal). Second, the European Union has defined particular interests in relation to the oil and gas reserves in the Caspian Basin. Facilitating the transportation of Caspian resources to Europe is a key issue here. For this purpose, the development of transport infrastructure across Turkey and Iran is under consideration; great importance is also attached to cooperation with Azerbaijan and Kazakhstan, however. Third, Ukraine plays a particular role in EU energy policy; it is the most important country in the transit of Russian natural gas to Western Europe. However, development of a cooperative relationship is at a very early stage here.

Particularly with Russia and the other states of the former Soviet Union in mind, the Energy Charter may become very significant, increasing investment security and access to network infrastructures. The extent to which this instrument may affect multi-lateral cooperation between the energy industry and energy policy will mainly be decided by Russia's attitude toward this agreement, and this is somewhat unpredictable.

SEE ALSO THE FOLLOWING ARTICLES

Equity and Distribution in Energy Policy • Fuel Economy Initiatives: International Comparisons • Geopolitics of Energy • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • National Energy Policy: Japan • National Energy Policy: United States • Nationalism and Oil • Research and Development Trends for Energy

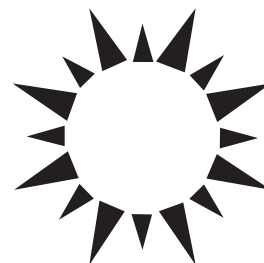
Further Reading

- Cames, M., Matthes, F. C., *et al.* (2000). "Instruments and Options for Environmental Policy during the Accession Process of EU Associated Countries in the Area of Environment and Energy." UBA Texte 42-01. Federal Environmental Agency, Berlin.
- Commission of the European Communities (CEC). (2000). Action plan to improve energy efficiency in the European Community. "Communication from the Commission to the Council, the European Parliament, the Economic and Social Committee and the Committee of the Regions." COM (2000) 247 final, 26.04.2000. CEC, Brussels.
- Commission of the European Communities (CEC). (2003). "Communication from the Commission to the Council and the European Parliament on the Development of Energy Policy for the Enlarged European Union, Its Neighbours and Partner Countries." COM (2003) 262 final, 13.05.2003. CEC, Brussels.
- European Commission (EC). (2001). Green Paper. "Towards a European Strategy for the Security of Energy Supply." EC, Luxembourg.
- European Environment Agency (EEA). (2002). "Energy and Environment in the European Union." Environmental Issue Report No. 31. EEA, Luxembourg.
- European Parliament (EP). (2001). "The European Parliament and the Euratom Treaty: Past, Present and Future." Energy and Research Series, ENER 114 EN. EP, Luxembourg.
- Matthes, F. C., and Timpe, C. (2000). "Sustainability and the Future of European Electricity Policy." Heinrich Boell Foundation Documentations, Papers and Reports No. 4, Berlin.

Evolutionary Economics and Energy

QUENTIN DUROY and JOHN GOWDY

Rensselaer Polytechnic Institute
Troy, New York, United States



-
1. Energy and the Evolution of Human Society
 2. Prices, Energy, and Scarcity
 3. Energy and Economic Growth
 4. Energy and Evolutionary Economics
 5. Conclusion

Glossary

energetics The study of the sources of energy, direct and indirect, available for mechanical work.

entropy In a closed system concentrations tend to disperse, structure tends to disappear, and order becomes disorder.

exergy The amount of energy converted into useful work.

exosomatic evolution The evolution by humans of outside-the-body tools and technology to supplement the limitations of our physically evolved bodies.

neoclassical economics The dominant school of economics whose core is the notion of general equilibrium and Pareto efficiency.

Pareto efficiency Exists when no further trading can make one party better off without making another worse off.

punctuated equilibrium The pattern of evolution tends to be one of long periods of stasis punctuated by periods of rapid change.

self-organization The tendency of an organic system, such as an ecosystem or the human economy, to actively shape its evolutionary environment.

Economic evolution has not been smooth or continuous. The history of the human occupation of Earth is characterized by long periods of stasis interrupted by periods of rapid movement to another system of socioeconomic organization. Major punctuations in human history include the use of fire, agriculture as the dominant mode of production, and the industrial revolution. All these changes involved the intensification of energy use and all of them signaled dramatic

changes in social and economic organization. The past 100 years or so have seen the most rapid and dramatic increase in the human population and in the planetary impact of our species in the 100,000-year history of *Homo sapiens*. The 5-2 fold increase in population and the approximately 40-fold increase in the value of economic output since 1850 were made possible by the era of petroleum. Harnessing fossil fuels has led to unimaginable material wealth for a large percentage of the world's population. However, the orgy of economic growth has left billions in poverty and has also brought humans to the limits of Earth's life-support systems. The tools of mainstream welfare economics have proved to be of limited value in understanding the role of energy in the evolution of the economy. Standard neoclassical economics is a theory of the allocation of scarce resources among alternative ends. It has a lot to tell us about the nature of market exchange, but it does not contain a useful theory of economic evolution or a notion of humans as a biological species inhabiting a finite planet. Fortunately, other tools are available to economists in addition to the general equilibrium model of constrained optimization. An evolutionary approach to energy use can give us an appreciation of the uniqueness of energy as a primary input to economic activity and of the importance of the fossil fuel age in reshaping the planet we inhabit.

1. ENERGY AND THE EVOLUTION OF HUMAN SOCIETY

Two major energy-induced milestones in human cultural and technological evolution were the use of fire and the use of fossil fuels. The earliest evidence of the use of fire is associated with a human ancestor known as *Homo erectus*, who first appeared

approximately 1.9 million years ago (there is controversy here, and some classify the earliest fire user as another species called *Homo ergaster*). Fire made possible the expansion of human occupation of the planet from Africa to the cooler areas of Europe and Asia. *Homo erectus* is arguably the most successful evolutionary human experiment in terms of having a sustainable technology (called the *Acheulean*) that lasted, essentially unchanged, for approximately 1 million years. Later humans used fire to dramatically change local ecosystems and landscapes. The Aborigines of Australia and Native Americans used fire to create habitats for large mammals that could be hunted. Much of the landscape encountered by the first Europeans in Australia and North America was the product of human alteration by fire.

Although coal has been used for heating and cooking for millennia, the age of fossil fuels was ushered in by the invention and refinement of the steam engine by Thomas Savery (1698), Thomas Newcome (1712), and James Watt (1769). In the 19th century, coal began to be heavily used in the booming industrialization of western Europe and the United States, especially in iron smelting. This development led to a dramatic increase in coal mining, forcing mining companies to dig deeper into the ground. Eventually, extraction activities became severely constrained by groundwater flooding. To remedy this problem, the mining industry relied on steam pumps. The use of these pumps was the first industrial application of the steam engine. The subsequent development and domination of steam and heat engines in the late 19th and early 20th centuries can be attributed, in retrospect, to an evolutionary accident that led to a technological lock-in and the current dependence of most nations on fossil fuels. What future historians will call the oil era will last only approximately 250 years, and we are now closer to the end of this era than to the beginning. The use of fossil fuel to power the industrial revolution not only had a dramatic effect on the level of economic output but also changed human society and the relationship between humans and the rest of the planet.

In the 19th century, Lewis Henry Morgan and Edward Tylor were among the first social scientists to apply an evolutionary approach to the study of traditional societies. However, it was the work of Leslie White and Fred Cottrell in the 1940s and 1950s that highlighted the central role of energy in the evolution of social systems. White argued that the evolution of human culture is conditioned by the ability of humans to harness and control energy:

“Culture evolves as the amount of energy per capita per year is increased, and/or as the efficiency of the instrumental means of putting the energy to work is increased.” He proposed a systemic analysis of cultural evolution. Human society, in his view, is organized around three systems: technological, sociological, and ideological. Via exosomatic evolution (the emergence of technology and material artifacts), humans managed to successfully harness and control different forms of energy. An example of exosomatic evolution is one of the most important “energy revolutions”—the development of the spear thrower, or atlatl, approximately 30,000 years ago. The additional power that this invention imparted to the human body revolutionized the hunting of big game animals in the Upper Pleistocene. According to White, the evolution of technology inevitably led to the creation of social institutions that, in turn, engendered a specific worldview or ideology (these three systems coevolved). White argued that cultural evolution was marked by two revolutions: the agricultural (harnessing and controlling cultivation and domestication) and the industrial (harnessing and controlling fossil fuels). Each revolution was subsequently characterized by an increase in technology that eventually reached a plateau due to inner systemic limits. A given technological system reaches its limit in terms of innovation, given the type of energy used, and the social system is no longer adequate to deal with the rapid growth of social units. Further cultural advance requires harnessing and controlling different energy forms, thus leading to a new revolution.

Much like White, Cottrell focused his analysis on the level of energy use as one of the key determinants in the emergence of complex cultural forms. Cottrell’s historical account of energy use and social evolution and his methodology are similar to those of Julian Steward and subsequent ecological anthropologists. Steward criticized the unilinear aspect of White’s evolutionary theory for failing to capture the complexity and diversity of human cultures. For Steward and Cottrell, there was no systemic reason for every culture to follow a given universal developmental path. Cottrell saw institutions and ideology as limiting factors toward the creation of a single world culture. Steward believed that regularities existed in cultural change but that local variations prevented the occurrence of universal stages of development. Nevertheless, beyond their theoretical and methodological differences, White, Cottrell, and Steward understood the importance of the laws of thermodynamics and the biophysical

constraints these laws impose on the availability and use of energy. More generally, their work emphasized the need for interdisciplinary approaches to the study of the role of energy in sociocultural evolution.

As documented by Cleveland, Daly, and Martinez-Alier, the development of energetics—the study of the role of energy in economic production—was an important contribution of the social sciences in the 19th and early 20th centuries. However, it was the growing concern over increasing environmental degradation and energy shortages in the 1970s that made the work of Patrick Geddes, Frederick Soddy, and others all the more relevant. The environmental movement led scholars, across disciplines, to focus on the role of energy in social and economic systems. In the 1970s, Adams, Ayres, and Georgescu-Roegen, among others, wrote extensively about the significance of the laws of thermodynamics to social and economic evolution. These authors taught us that socioeconomic processes are subject to the first and second laws of thermodynamics. The first law states that the total matter and energy for any system and its surroundings does not change over time. Matter and energy are conserved. The second law states that entropy always increases in an isolated system. Low-entropy (concentrated and useful) energy is irreversibly transformed in a closed system into high-entropy energy (dispersed and unusable). For Adams, exosomatic evolution provides human beings with the capacity to capture and degrade energy more rapidly and more effectively. Consequently, human activities speed up the march of entropy.

For Georgescu-Roegen, the unfolding of the entropy law is what the economic process is fundamentally about. Although there is little doubt that social and economic evolution is subject to the laws of thermodynamics, many issues are still unresolved. Georgescu-Roegen pushed the theoretical boundaries of the relevance of thermodynamics to economics by proposing a fourth law, which states that complete recycling is impossible (i.e., perpetual motion of the third kind is not possible). Ayres contested this proposition, arguing that although the creation of a “wastebasket” is inevitable due to the second law, this wastebasket, over a long period of time, can be a source of usable energy and matter. As Kozo Mayumi pointed out, the fourth law, as a consequence, needs to be qualified as follows: On an economic timescale, given the state of current technology and the limits to energy inputs, complete recycling is impossible.

With the growing literature and debates over the relevance of thermodynamics to social sciences during the past three decades, a consensus has developed

about the issues of concern to environmental, energy, and ecological economists. First, economic production and consumption processes are constrained by biophysical limits and therefore are subject to the laws of thermodynamics. Second, increased human activity (economic growth) speeds up the entropic process (increases in heat and material wastes) and puts more pressure on the natural environment. Third, if unchecked, this will eventually limit the potential for economic and social evolution.

2. PRICES, ENERGY, AND SCARCITY

The relationship between energy and economic value has intrigued economists since the inception of economics as an academic discipline. Members of the first recognized school of economics, the 18th-century French physiocrats, saw solar energy as being the ultimate source of wealth in agricultural production. Later, Sadi Cournot experimented with the steam engine to show how much useful work could be obtained from a given amount of energy. William Jevons recognized the importance of coal to the British economy and warned of the consequences of its increasing scarcity. In the late 19th and early 20th centuries, a number of physical scientists, including Geddes and Soddy, wrote extensively about economic issues and warned of the consequences of ignoring the importance of energy to economic value. Geddes is particularly interesting because he is perhaps the first to criticize—in a letter to Leon Walras in 1883—the general equilibrium approach of mathematical economics for ignoring the physical nature of production and the biological nature of the human species. Geddes did not advocate an energy theory of value, but he did point out that there is more to economic value than mere “feelings.” Coal, for example, was not just another commodity but rather an essential input for producing food and other necessities for human existence. Geddes and Soddy pioneered the field of energetics—that is, calculating the amount of energy required to produce various economic goods.

In the 1960s and 1970s, the work of Kenneth Boulding and Nicholas Georgescu-Roegen was tremendously influential in turning the attention of economists to biophysical constraints on economic activity. Georgescu-Roegen’s monumental work, “The Entropy Law and the Economic Process,” inspired a generation of economists to consider the environmental consequences of economic growth. Georgescu-Roegen was emphatic in rejecting market

prices as a means of allocating irreplaceable natural resources or irreducible pollution. Since future generations cannot bid in energy markets, the allocation of these resources between generations should be a matter of public policy, not isolated individuals bidding at a particular point in time in ill-informed markets.

Kenneth Boulding's classic article, "The Economics of the Coming Spaceship Earth," captured the spirit of the new environmental movement by emphasizing that pollution and resource exhaustion were part of any economic process. Except for a few isolated voices, however, most economists were silent on the energy issue until the oil crisis in 1973. Even after energy prices quadrupled in a few months, mainstream economists such as Robert Solow downplayed the role of energy, stressing the ability of the economy to substitute other inputs for energy in the face of relative price changes. The energy economy debate frequently led to ludicrous statements on both sides. The economist Julian Simon stated that there is enough energy to fuel continual economic growth for the next 7 billion years. Although not in the same league as Simon in terms of fantasy, many environmentalists made overly pessimistic predictions, such as those of the Club of Rome warning of "a rather sudden and uncontrollable decline in both population and industrial capacity" because of resource shortages.

The relationship between energy prices and resource scarcity has been the subject of intense debate among economists. In the 1930s, Harold Hotelling described the ideal exhaustion path of an absolutely scarce resource. Hotelling assumed, among other things, that economic actors are aware of the total amount of the scarce resource, of present and future extraction costs, and, implicitly, of future extraction technology. Given these assumptions, rising prices will reflect the increasing scarcity of this resource. Numerous writers (e.g., Barnett and Morse) have incorrectly used Hotelling's result to suggest that since real energy prices have not significantly increased during the past 100 years, energy is not a scarce resource. Norgaard pointed out that this argument is logically flawed since it goes from Hotelling's conclusion (economic indicators will reflect scarcity) to his premise (if resources are scarce) while ignoring his caveat that resource allocators must be fully informed of future scarcity.

After more than 30 years of debate, punctuated by real-world energy crises, it is clear that changes in energy availability can have drastic consequences for the economy. It is also clear that market economies

have a remarkable ability to adjust to energy shocks and, at least in the short term, to substitute technology for increasingly scarce resources. However, the role of essential inputs in production is one of the major unresolved issues in economics. We know that fossil fuel energy is an essential factor of production and we know that the supply of fossil fuels is finite. We also know that higher prices will call forth substitutes and more efficient technologies. So far at least, neoclassical theory has not provided much guidance to understanding the relationship between energy and economic growth. Crude estimates of the marginal rates of substitution between "energy" and "nonenergy" are methodologically flawed and do not tell us much about the ability of the economy to adjust to resource exhaustion. Evolutionary economics can give us guidance about the role of energy in the economy and suggest some policies to perhaps ensure a smoother transition to non-fossil fuel energy sources.

3. ENERGY AND ECONOMIC GROWTH

Among world leaders, economic growth is seen as the unique source of present and future well-being. Since energy is crucial to economic growth, one would think that its role should be an integral part of all growth theories. However, energy fails to appear as a variable in most mainstream models of economic growth. The traditional Solow model explains economic growth through the changes in two factors of production: capital (K) and labor (L). According to the model, output (Q) is generated through a production mix of K and L , whose physical properties are described by a Cobb–Douglas production function:

$$Q = AK^\alpha L^\beta, \quad (1)$$

where α and β are the partial output elasticities (and output shares) of K and L , respectively. In this model, one factor of production can potentially be completely replaced by another. Empirical estimations of the Solow model failed to explain most of the variation in economic growth. The part of the growth unaccounted for by capital and labor (approximately 87% of the total) was attributed by Solow to technical progress: The "Solow residual" (A in Eq. 1) represents a "catch-all" category dubbed technological change. Under the pressure of criticism from within and from outside the field, economists have developed more sophisticated growth models.

Recent models are based on endogenous growth and consider a broader array of relevant variables, including the Solow residual (hence the endogenous character of the theory), a research and development component, human capital, and natural resources. Even though the models possess a greater explanatory power, they still assume potential perfect substitution between all inputs. Moreover, in empirical estimations, a large part of the variation in economic growth remains unexplained.

To resolve some of the problems encountered in more conventional models, and to anchor them in biophysical reality, Charles Hall and colleagues developed economic growth models including energy as a factor of production. Hall's approach is to include energy in the production function as a full-fledged factor of production in a traditional Solow model. However, he explicitly rejects the assumption of perfect substitution because the unlimited substitution of energy by another form of capital would clearly violate the laws of thermodynamics. In a biophysical sense, energy cannot disappear during the production process (as the first law states) and economic production cannot take place without systematic inputs of low-entropy energy and subsequent dissipation of high-entropy energy (as the second law states). Hall's model successfully reproduced economic growth over a 30-year period for three industrialized countries—Japan, the United States, and West Germany. For Japan and West Germany, the model leaves unexplained less than 10% of the variation in output, even though the two countries underwent significant structural changes as a consequence of the two oil crises in the 1970s. For the United States, 98.3% of the variability in production is explained by the variables for the 1960–1977 period. The results are less conclusive for the 1978–1993 period. However, in this model, technological change is accounted for by a simple variable (creativity) that fails to capture the complex structural changes that resulted from the rapid development of energy-conservation technologies beginning in the 1970s. Hall's results clearly illustrate the prominent role played by energy in economic production.

A model by Ayres and Warr focuses on the concept of exergy, a measure of the conversion of energy into useful work. Exergy measures how much physical work can potentially be done by a system. Unlike energy, exergy is not conserved: It is used up by the system. Ayres and Warr introduce a measure of "useful work" (U) in the Solow model and they find that, since 1900, improvements in energy

conversion into physical work explain most of the variation in technical progress; the unexplained residual is only 12%.

These findings emphasize the importance of the role of energy in economic growth. To a large extent, they constitute empirical evidence of the theoretical work of White, Cottrell, and other earlier writers on the role of energy in economic and cultural evolution. They also reflect an intuitive sense that increased energy use and efficient energetic conversion lead to increased growth and development. Figure 1 shows a clear correlation between gross domestic product (GDP) and energy consumption in the United States. The problem of causality remains, however, since the figure does not indicate whether increased GDP leads to increased energy use or vice versa. Cutler Cleveland and colleagues demonstrated, taking into account energy quality, that energy use is indeed the independent variable and economic growth the dependent one. Growing energy use and improving energy efficiency fueled the growth in GDP and provoked dramatic changes in cultural evolution.

Figure 2 illustrates the significant differences between developing and advanced nations in terms of energy use per capita. Industrialized countries have undergone major technological developments and a major improvement in their standard of living during the past 150 years. They are also the countries with very high levels of energy use. Of course, a contributing factor in the energy/GNP relationship is the ability of high-income countries to keep the energy flowing to them. Figures 1 and 2 suggest an essential relationship between energy use and economic activity—a relationship difficult to determine using traditional economic models.

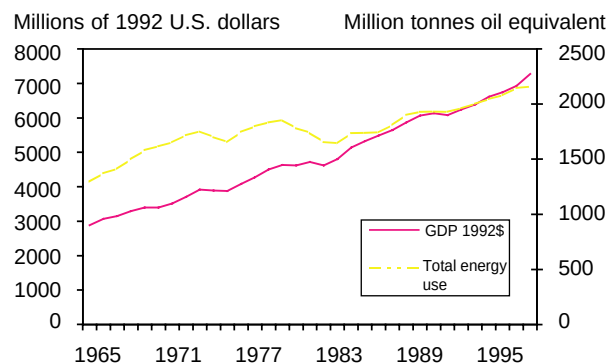


FIGURE 1 Evolution of real GDP and total primary energy use in the United States, 1965–1997. Data from the U.S. Department of Commerce and "BP Statistical Review of World Energy."

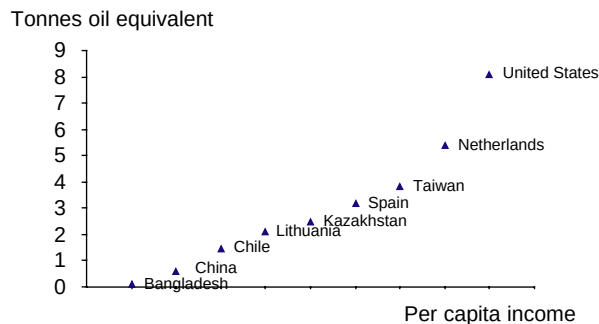


FIGURE 2 Energy use per capita in tonnes oil equivalent for selected countries, 2000. Data from the Energy Information Administration and “BP Statistical Review of World Energy.”

4. ENERGY AND EVOLUTIONARY ECONOMICS

Evolutionary economics is not a well-defined field of economic theory. Economists who call themselves “evolutionary” hold a wide variety of views about the pace of evolutionary change, the nature of economic equilibrium, and the role of prices in economic theory and policy. Nevertheless, a number of useful concepts from evolutionary economics can be applied to the study of the role of energy in the human economy. Most of the debate within economics about resources and the economy has taken place within the framework of neoclassical growth theory, which merely extends the optimal factor allocation framework to the allocation of factor growth rates. In contrast, an evolutionary framework can examine discontinuous change, exogenous as well as endogenous factors, and biophysical limits. Three useful concepts from evolutionary economics are the notion of punctuated equilibrium, the existence of multiple equilibria, and the phenomenon of self-organization.

Punctuated equilibrium as a description of the fossil record was first proposed by Niles Eldredge and Stephen Jay Gould in 1972. They argued that evolution proceeds not only by gradual changes at the margin but also by quantum leaps from one form to another. Therefore, the evolutionary history of a particular species is characterized by periods in which it remains virtually unchanged, punctuated by abrupt events leading to a rapid transformation into another species. As discussed previously, the human epoch can be seen as a series of fairly stable regimes interrupted by periods of rapid change. The major transitions in human physical evolution 2.4 million years ago (*Homo habilis*), 1.9 million years ago (*H. erectus*), and 100,000 years ago (*H. sapiens*) were abrupt in terms of the fossil record. Likewise, human cultural

evolution is also characterized by abrupt changes from one technology to another. The transition to agriculture was essentially complete within a few thousand years and the transition to the industrial age was even more rapid. Georgescu-Roegen identified three “Promethean” technologies as fundamental qualitative changes in human history: fire, agriculture, and the steam engine. These were fundamental bioeconomic mutations that redefined the human presence on the planet.

What caused these punctuations? Increasing evidence suggests that evolutionary changes, including human evolution, have been driven by abrupt climate events. Good evidence exists linking the origin of the genus *Homo* to the onset of glaciation. The appearance of *H. erectus* coincided with a dramatic change to a cooler and drier climate in Africa. The transition to agriculture that occurred throughout the world after the end of the last ice age ushered in a warmer and uniquely stable climate. The industrial age coincided with an unusually warm period in Europe. Another punctuation will likely occur in the near future as the temperature of the planet adjusts to the massive pumping of carbon into the atmosphere by the burning of fossil fuels.

Evolution involves two critical features: variation and selection. It is increasingly accepted that Darwinian selection can occur at several levels, from the gene to the organism and the group. In contrast to the evolution of most other species, human traits may be culturally selected through various systems of rewards and punishments. This accounts for the incredible variety of human cultures. The existence of cultural “filters” for human traits means that institutions and institutional change are critical to economic evolution. The availability and use of energy sources depend on consumer attitudes, history and historical lock-in, as well as political intrigue. The experience of the energy price shocks of the 1970s holds two particularly valuable lessons. First, these shocks demonstrated the vulnerability of the global economy to changes in energy availability. Second, they showed the ability of markets to substitute energy efficiency and other sources of energy for petroleum.

The ability of the economy to adjust to energy shortages should not be overestimated. The short-term energy disruptions of the 1970s and early 1980s caused much economic damage and, being politically driven, they did not signal impending fossil fuel exhaustion. Another lesson is that the ability of policy intervention to change patterns of energy consumption should not be underestimated. Policies

promoting energy efficiency played a major role in softening the impact of the energy shocks, especially in Japan and Europe. The energy crises demonstrated, as careful commentators have been noting for more than a century, that energy is not just another commodity. It is essential to the production process. Cheap energy has been the ultimate source of the unprecedented economic growth of the past 150 years and the transition to a non-fossil fuel economy will be another punctuation in human history. The most important by-product of energy use, increased carbon in the atmosphere, will most likely cause another punctuation as the climate continues to respond to the perturbation caused by the fossil fuel age.

Multiple equilibria also characterizes the evolution of the human economy. Ecologists tell us that there is nothing inevitable about the evolution of a particular ecosystem. Small changes can cause an ecosystem to change to another equilibrium that might look quite different than the original one. Ecologists no longer accept the notion of a “climax forest,” an inevitable and unique final form that would automatically evolve in a given area. Natural systems show that many equilibrium states are possible with any given set of initial conditions. Likewise, the market economy—contrary to the model of welfare economics that implicitly assumes a single, unique, global equilibrium—can have a large variety of stable equilibria. Human social systems have even more possibilities than natural ecosystems since cultural systems of rewards and punishment can create any conceivable social equilibrium. Also, there is no reason to say that one system is better than another.

The implications of multiple equilibria for energy use are both positive and negative. On the one hand, change can be abrupt and unpredictable. On the other hand, human institutions, including the market, can adjust rapidly to suddenly changing parameters. The global economy can most likely adjust to new energy shocks and probably even to the end of the fossil fuel age, but what the next economic system will look like is anyone’s guess. Likewise, many surprising outcomes are possible as a result of the current episode of global climate change. The planet could continue to warm gradually or suddenly change to another state (much colder or much warmer).

In terms of policy choices, the existence of multiple equilibria implies that we need not be afraid of moving away from the one true optimum. The notion of a unique and stable Pareto optimum toward which the economy naturally moves is an

illusion. Likewise, Pareto efficiency in allocation is a goal chosen by politicians and abetted by neoclassical economists, but there is nothing scientific or sacred about it. As Daniel Bromley argues, efficiency as the most important policy goal is an ideology, not a science-based conclusion. Rather than searching for the one true optimum, a more fruitful research project would be determining what categorizes different kinds of equilibria. What keeps a system in equilibrium in the face of fairly severe changes in the environment? Can we anticipate when a system is about to change to another state?

Self-organization refers to the process of the emergence of order in organic systems. It is a development process that is cumulative, discontinuous, irreversible, and creative. It incorporates processes at one level and creates new structures at higher levels. According to Stuart Kaufman, much of the order we see in the biological world may not be the result of natural selection at all but rather the spontaneous order of self-organized systems. Likewise, the (usual) order we see in economic systems may not be the result of the invisible hand of market selection but rather a spontaneous order of self-organized systems having little to do with prices or efficiency. The relationship between self-organization and thermodynamics is obvious. Self-organizing systems prosper by drawing down energy and information from surrounding systems. The global market economy expands by drawing down stocks of low-entropy fossil fuels, stocks of renewable resources such as forest and ocean systems, and, arguably, human cultural diversity.

The fossil record shows that the evolution of species is driven by interplay between internal and external forces. On the one hand, species compete according to the laws of Darwinian natural selection in self-organizing and adaptive systems. On the other hand, evolutionary outcomes are sometimes dominated by rapid and unpredictable external changes. Evolutionary systems cannot be understood without an appreciation of the interplay between external and internal, micro and macro, order and disorder. Like biological systems, the economy is an evolving complex system buffeted by constant changes in boundary conditions.

The current self-organization of the world economy is driven by the world’s dowry of fossil fuels. Increases in economic activity mean increases in energy use, which also increase the dissipation of high-entropy energy, putting irreversibly more pressure on the natural environment and speeding up the emergence of scarcity in both renewable (e.g., via

pollution) and nonrenewable resources (via depletion and deterioration). Therefore, the increased dependence on fossil fuels may eventually limit the potential for diversity. As fewer resources are available for human beings to harness, control, and transform, fewer potential developmental paths are available. With fewer developmental paths to choose from, evolution becomes limited. To a large extent, evolution depends on diversity and variation, and if diversity is low, sorting mechanisms such as natural selection have less raw material with which to work.

5. CONCLUSION

The dependence of the market economy on fossil fuel energy, together with the fact that fossil fuels are being rapidly exhausted, paints a disturbing picture of the human prospect. Added to this is the fact that the major punctuations in human physical and cultural evolution have apparently been driven by climate change—occurring today because of carbon emissions from fossil fuels. On the other hand, our evolutionary history shows that we are an adaptable species, capable of making drastic changes in lifestyles and technology to meet new challenges. From an evolutionary perspective, market economies are not as stationary or as fragile as the standard economics model shows them to be. The global economy withstood a number of monumental shocks and it eventually adjusted to the new conditions. Also, there is no reason to believe that a “free” market equilibrium is superior. A number of stable states are possible, and a number of policy tools are available to make the adjustment to the end of the fossil fuel age smoother. We should not hesitate to use them out of fear of disturbing a “natural” free market system. The end of fossil fuels does not mean the end of civilization.

The technology already exists to enable us to move to a non-fossil fuel economy. With a little help from sound energy policy initiatives, the prospects for changing to another Promethean technology are good. Technological adjustment need not mean using

ever more sophisticated technologies to hang on to the bitter end to an unsustainable system.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Cultural Evolution and Energy • Economic Growth and Energy • Economic Thought, History of Energy in • Entropy and the Economic Process • Exergoeconomics • Exergy • Prices of Energy, History of • Transitions in Energy Use

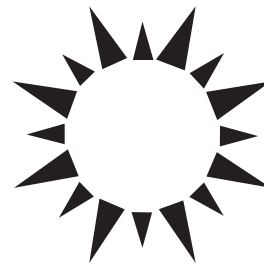
Further Reading

- Adams, R. (1981). Natural selection, energetics, and cultural materialism. *Curr. Anthropol.* **22**, 603–608.
- Ayres, R., and Warr, B. (2004). Accounting for growth: The role of physical work. *Struct. Change Econ. Dynamics*, in press.
- Ayres, R., d'Arge, R., and Kneese, A. (1970). “Aspects of Environmental Economics: A Material Balance, General Equilibrium Approach.” Johns Hopkins Univ. Press, Baltimore.
- Bergh, J. C., and van den, J. M. (2002). Evolutionary analysis of economic growth, environment and resources. In “Scarcity and Growth in the New Millennium” (R. Ayres, D. Simpson, and M. Toman, Eds.). Resources for the Future, Washington, DC.
- Bromley, D. (1990). The ideology of efficiency: Searching for a theory of policy analysis. *J. Environ. Econ. Management* **19**, 86–107.
- Calvin, W. (2002). “A Brain for all Seasons: Human Evolution and Abrupt Climate Change.” Univ. of Chicago Press, Chicago.
- Cleveland, C. (1999). Biophysical economics: From physiocracy to ecological economics and industrial ecology. In “Bioeconomics and Entropy: Essays in Honor of Nicholas Georgescu-Roegen” (K. Mayumi and J. Gowdy, Eds.). Elgar, Cheltenham, UK.
- Daly, H. (1977). “Steady State Economics.” Freeman, San Francisco.
- Georgescu-Roegen, N. (1971). “The Entropy Law and the Economic Process.” Harvard Univ. Press, Cambridge, MA.
- Hodgson, G. (1993). “Economics and Evolution: Putting Life Back into Economics.” Univ. of Michigan Press, Ann Arbor.
- Martinez-Alier, J. (1987). “Ecological Economics: Energy, Environment and Society.” Blackwell, Oxford, UK.
- Potts, J. (2000). “The New Evolutionary Microeconomics.” Elgar, Cheltenham, UK.
- Stern, D. (2000). A multivariate cointegration analysis of the role of energy in the U.S. macroeconomy. *Energy Econ.* **22**, 267–283.

Exergoeconomics

ENRICO SCIUBBA

University of Roma 1 “La Sapienza”
Rome, Italy



-
1. Introduction
 2. The Concept of Exergy
 3. The Economics of a Commodity-Producing Process
 4. Second vs First Law Optimization
 5. Structural Representation of Energy Conversion Systems
 6. Thermoeconomics
 7. Exergoeconomics and Extended Exergy Accounting
 8. Conclusions

Glossary

commodity In economics, a physical object produced by a man-made process. In this article, commodity also includes dematerialized services.

objective function In optimization procedures, a mathematical expression depending on process parameters, whose extremum (minimum or maximum) corresponds to the “most convenient” combination of production factors for the assigned output.

Exergoeconomics (ee) is the branch of applied thermodynamics that deals with second law-based process synthesis and optimization methods. Per se, EE does not include optimization techniques: It provides the objective function, but the actual optimization of process performance (i.e., the achievement of the specified output with the most convenient combination of production factors) is carried out with one of the available mathematical techniques. Therefore, this article does not deal with the numerical procedures that lead to the optimized system but rather describes and discusses the engineering background of the methods that lead to the formulation of the objective function.

1. INTRODUCTION

Exergoeconomics (EE) is a relatively new technique. It originated in the 1960s, first in Europe, where

Wolfgang Fratscher (Germany), Jan Szargut (Poland), and Valeriy Brodyanskii (USSR) published papers on the “economic value of the concept of exergy,” and then in the United States, where Myron Tribus, Yehia El-Sayed, and Robert Evans published a series of papers on a mathematical cost-optimization procedure based on exergy (which they called availability). The term thermoeconomics (TE) was first used in Tribus’s MIT course notes in 1960 and in Evans’ doctoral thesis in 1961. The concept was developed both in breadth and in depth in academic circles, but it did not impact process engineers, industrial managers, and energy planners until approximately 20 years later, mostly due to the efforts of El-Sayed, Richard Gaggioli, Tadeusz Kotas, and Michael Moran in the early 1980s and Antonio Valero and George Tsatsaronis in the late 1980s. In a series of papers from 1986 to 1989, Valero derived a general and completely formalized costing theory to calculate the exergetic cost of a product from the exergetic input into the process and the structure of the production process. In the 1990s, TE was extended (mostly by Christos Frangopoulos and Michael von Spakowsky) by adapting it to off-design conditions and time-dependent problems and by including environmental effects. The literature on TE is abundant, and a comprehensive review has not been attempted. Regarding EE, there is a qualitatively very interesting debate, expressed by the works of Valero, Göran Wall, Szargut, and Enrico Sciubba, that seems to lead to an exergetic theory of value on which to base societal planning (especially with respect to the much-debated issues of sustainable-development and limited-growth models).

In this article, the concept of exergy and cost optimization is briefly discussed, followed by separate treatments of TE and EE. A caveat is necessary: Some authors, most notably Tsatsaronis and von Spakowsky, use TE and EE as synonyms and consider irrelevant whether the optimization is carried using money or exergy as a quantifier. Here,

TE is considered as a special case of EE, and treatment of these is developed with this guideline.

2. THE CONCEPT OF EXERGY

Energy is considered an a priori concept in thermodynamics: We are taught that “energy is an extensive property of a system such that its change in value between two states is equal to the adiabatic work between these states.” However, energy is known to manifest itself in several forms, each endowed with a different quality. The concept of quality can be fully explained only by referring to the second law, but in general the quality of an energy flux is a measure of its capacity to cause change. We can identify “ordered” or “high-quality” energy forms (potential, kinetic, mechanical, and electrical) that can be ideally converted into each other with 100% efficiency so that even after a number of transformations, as shown in Fig. 1A, the output flux of useful energy, En_{out} , is quantitatively equal to the flux of useful input, En_{in} . However, other forms (internal energy, chemical energy, thermal radiation, and turbulent kinetic energy) cannot be converted into each other, much less into high-quality energy, without an intrinsic efficiency loss, so the useful output En_{out} is now always lower than the useful input En_{in} (Fig. 1B). Since energy is conserved (this is a basic postulate in all physical sciences), the

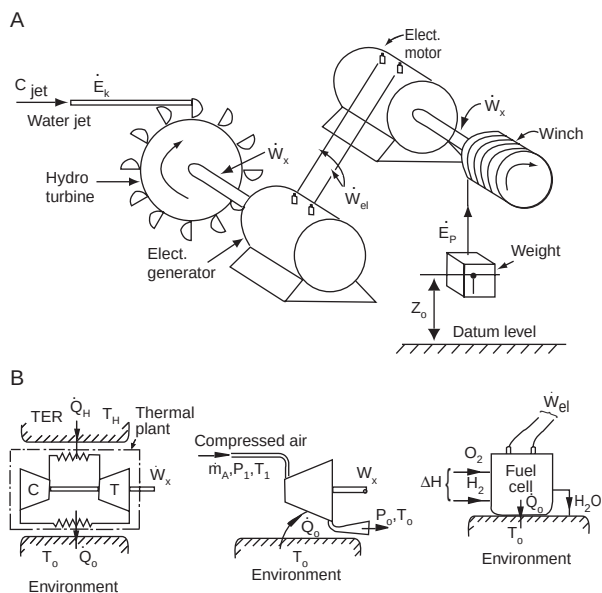


FIGURE 1 (A) A chain of ideal energy conversions of high-quality energy forms into each other. (B) Examples of conversion of low-quality into high-quality energy forms.

difference $En_{in} - En_{out}$ must be accounted for: Indeed, we invariably find that this difference has been dissipated into a low-temperature flux absorbed or provided by the environment. The “environment” is meant here in its usual layman sense—that is, as a system so large that its properties (pressure, temperature, chemical composition, kinetic or potential energy, etc.) are not affected by the interaction with the man-made system. These considerations, though formally incomplete from a scientific standpoint, convey the correct information in an easily understandable form.

The concept of exergy has been developed to provide a congruent and coherent quantification to the quality of an energy form. Imagine that at a certain time t_0 we place a system S identified by a certain set of thermodynamic properties (V, z_B, p, T, c, m, μ , etc.) in contact with the environment E and arrange things in such a way that S may exchange mass and energy only with E (Fig. 2). Our intuition tells us and experiments confirm that after some time t_{relax} , which depends on the extension of S and on the difference between its initial properties and those of E , S will come to equilibrium with E ; that is, even possible dynamic interactions will occur at a macroscopically stationary state and will result in no net change in the mass and energy contents of either system. Exergy is defined as the maximum work developed in this ideal process. Since the interaction is between E and S , and we assume that all processes can be described by a succession of quasiequilibrium states (so that the time interval t_{relax} drops out from the problem formulation), exergy will be a function only of the initial and the final state of S (the final state being equal to that of the environment) so that it is an attribute of the pair (S, E) and not of S alone. S and E exchange energy fluxes under different forms: kinetic (all parts of S in relative motion with respect to E come to rest), potential (the barycentric position of S reaches a fixed elevation in the gravity field of E),

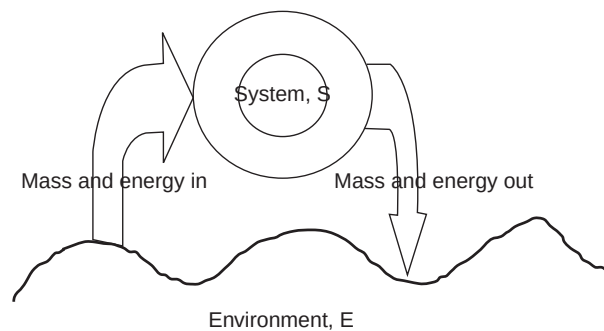


FIGURE 2 A system S and its environment E .

thermal (heat flows from S to E or vice versa depending on the initial temperature difference $T_S - T_E$), work interaction (E performs or receives work from its boundary interactions with S), and mass exchange (mass fluxes from S to E and vice versa carry different energies). Let us consider here only stationary interactions between S and E (this is the so-called control volume approach) in which flows of energy and matter from S to E are continuous and (on average) steady in time. By applying energy and entropy balances to the complex system composed of S and E, we obtain

$$\sum_j \left(1 - \frac{T_E}{T_j}\right) Q_j - W + \sum_{\text{in}} m_i e_i - \sum_{\text{out}} m_{\text{out}} e_{\text{out}} = E_\lambda, \quad (1)$$

where the first term represents the thermal exergy flow, the third and fourth the exergy in- and outflows due to material exchanges, and E_λ is the exergetic destruction. In Eq. (1), we make use of the definition of specific exergy:

$$e = h - h_E - T_E(S - S_E) + \sum_k (c_k \mu_k - c_E \mu_{k,E}) + 0.5V^2 + gz. \quad (2)$$

From Eqs. (1) and (2), we can draw several conclusions:

1. The initial kinetic and potential exergies of S are entirely “recovered” into useful work; in fact, they might be included in the enthalpic term by considering a baroclinic total enthalpy, $h_{\text{tot}} = h + 0.5V^2 + gz$.

2. Work exchanges between S and E are entirely recovered in the exergy balance. This derives from the fact that work is exergy and, as a consequence, the kinetic and potential energy contents are entirely equivalent to exergy. If we define a quality or exergetic factor as the ratio between the energy and the exergy content of an energy flux, and since electrical energy is equivalent to work, all these high-quality forms are seen to have an exergetic factor equal to 1.

3. Thermal energy has an exergetic factor equal to its associated Carnot efficiency: This reflects the fact that, under the given assumptions, the maximum work that can be extracted from a quantity of heat Q available at a certain temperature T is equal to $W_Q = (1 - T_E/T) Q$.

4. If S is initially at a lower temperature than E, thermal energy will flow from E to S, with a corresponding exergy flow equal to $E_Q = (1 - T/T_E) Q$. This amount is always positive.

5. For a simple substance whose states are defined by temperature and pressure (e.g., most gases for most practical engineering purposes), exergy is always positive, except for particular combinations of p_S/p_E and T_S/T_E .

6. The chemical potential of the elements in S cannot be entirely converted to work: Since S must come to equilibrium with E, the most we can recover for the generic k th component is the difference between the values of its Gibbs function in S and in E, each weighted by the respective concentration. In other words, we can ideally transform only a portion of the initial ΔG_S into useful work because the products of reaction must be at their respective environmental concentrations and chemical potentials at T_E and p_E .

7. Exergy is not conserved: In every real process there is an exergy destruction caused by irreversibility effects. In the simplest case (work $\Rightarrow$ work transformation), this exergy destruction is equal to the irreversible entropy generation due to frictional effects. This can be seen by considering that S exchanges only adiabatic work with E, and the loss is equivalent to frictional heat. If there is only one medium participating in the exchange, from Eqs. (1) and (2) we obtain

$$\begin{aligned} E_\lambda &= -W + m(e_S - e_E) \\ &= -W + m[h_S - h_E - T_E(s_S - s_E)] \\ &= -T_E(s_S - s_E)[W]. \end{aligned}$$

Since $s_S - s_E$ must be negative (S may spontaneously evolve only toward higher entropies), the destroyed exergy flux E_λ is positive and equal to the “lost work” defined by Gouy and Stodola.

8. From Eqs. (1) and (2), we obtain the general exergy equation for flow systems:

$$e_S = e_{\text{ph}} + e_{\text{ch}} + e_k + e_p + e_j [\text{J/kg}], \quad (3)$$

where the suffixes indicate the usual denomination of the various terms: physical, chemical, kinetic, and potential exergy, respectively. If other forms of energy fluxes are involved (magnetic, nuclear, etc.), one needs only to identify the suffix j and apply the corresponding definition.

9. Exergy is an extensive property so that $E_{(S_1 + S_2)} = E_{S_1} + E_{S_2}$.

10. If two systems S_a and S_b interact, the maximum work obtainable from their exclusive interaction is $E_a - E_b$. Furthermore, if a stream a undergoes a series of transformations $i, j, k, \dots, z$ in which it receives the exergies $E_i \dots E_z$, its final exergy content is the sum of $E_a + E_i + E_j + E_k + \dots E_z$. Exergy is thus an additive property. For engineering

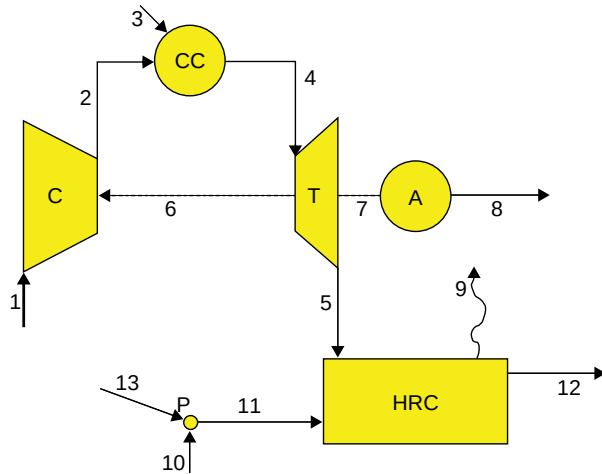


FIGURE 3 Schematic representation of a cogeneration plant.

(and EE) calculations, it is convenient to define a standard environment to which all exergy values can be referred, but this is by no means necessary, as Gaggioli and coworkers have shown.

Practical applications of the exergy concept are important: It can be shown that a second law efficiency defined on the basis of exergy fluxes provides more information than the first law efficiency based on energy fluxes. For example, consider the cogeneration plant shown in Fig. 3. From the design data given in Table I, the efficiencies shown in Table II can be calculated and the following considerations apply:

1. First law efficiency attains a value of 0.655: This means that 65% of the fuel chemical energy is converted into useful energy output. This is misleading, however, because it does not account for the temperature-dependent quality of the heat flux Q_{cog} . An exergetic analysis, in which all inputs and outputs are expressed in exergetic terms, provides an η_{II} of 0.347, correctly indicating that a substantial portion of the chemical exergy of the fuel has been “degraded” into heat (i.e., into a low-quality energy form).

2. Component exergetic efficiencies (η_{II}) also provide interesting information, especially if compared with their energetic counterparts (η_{I}). For example, whereas the quasi-adiabatic compressor and turbine display similar values for η_{I} and η_{II} , both the heat-recovery boiler and the condenser have $\eta_{\text{II}} < \eta_{\text{I}}$. This indicates that it is useless to invest excessive resources to improve boiler performance because this component is very efficient in transmitting the heat flux it is required to transmit

TABLE I

Design Data for the Cogeneration Plant Shown in Fig. 3

$T_{\text{env}} = 300 \text{ K}$	$p_{\text{env}} = 0.1 \text{ MPa}$	$\Delta p_{2-4} = 0.02 p_2$	$\eta_{p,C} = 0.85$
$T_1 = 300 \text{ K}$	$p_1 = 0.1 \text{ MPa}$	$\Delta p_{5-9} = 0.08 p_9$	$\eta_{p,T} = 0.9$
$T_4 = 1300 \text{ K}$	$p_2 = 0.8 \text{ MPa}$	$\Delta p_{11-12} = 0.1 \text{ MPa}$	$\eta_{\text{el}} = 0.98$
$T_9 = 353 \text{ K}$	$p_9 = 0.1 \text{ MPa}$		$\eta_{\text{mec}} = 0.95$
$T_{10} = 673 \text{ K}$	$p_{10} = 0.1 \text{ MPa}$	$c_p = 1.1 \text{ J/(kgK)}$	$\eta_{\text{CC}} = 0.98$
	$P_{11} = 1 \text{ MPa}$	$R_{\text{air}} = 287 \text{ J/(kgK)}$	$\eta_P = 0.8$
		$\kappa = 1.4$	$\eta_{\text{HRB}} = 0.9$

TABLE II

Calculated Process Parameters for the Cogeneration Plant Shown in Fig. 3

$T_2 = 603 \text{ K}$	$p_4 = 0.798 \text{ MPa}$	$\eta_{1,C} = 0.85$	$\eta_{2,C} = 0.76$
$T_5 = 777 \text{ K}$	$p_5 = 0.108 \text{ MPa}$	$\eta_{1,T} = 0.9$	$\eta_{2,T} = 0.87$
$T_{12} = 403 \text{ K}$		$\eta_{1,\text{CC}} = 0.94$	$\eta_{2,\text{CC}} = 0.59$
		$\eta_{1,\text{HRB}} = 0.81$	$\eta_{2,\text{HRB}} = 0.68$
$\eta_{\text{plant,el}} = 0.2$			
$\eta_{1,\text{plant}} = 0.658 \quad \eta_{2,\text{plant}} = 0.34785$			

($\eta_{1,B} = 0.81$). The boiler low exergetic efficiency is intrinsic and not due to design errors: It degrades a high-temperature, high-exergy heat flux into a low-temperature, low-exergy one. Similar considerations apply to the combustor: It is the degradation of the relatively high-quality chemical energy into heat, and not its energy-transfer capacity, that lowers its exergetic efficiency: Energywise, the performance of a combustor is more than satisfactory (≈ 0.94). This is the rationale for the use of fuel cells, which exploit the ΔG of the fuel by means of chemical reactions, without combustion.

3. THE ECONOMICS OF A COMMODITY-PRODUCING PROCESS

Consider a process that produces a certain commodity (a good or a service) (Fig. 4), and assume that both its internal structure and its operating parameters are known so that all of its outputs $\mathbf{O}$ can be calculated if all of the inputs $\mathbf{I}$ are known (the transfer function Π of process P is known, where $\mathbf{O} = \Pi \times \mathbf{I}$). To produce a cumulative amount O_1 of the commodity over a period of time T , a certain amount of inputs $I_1, I_2, \dots, I_N$ are used together with a certain amount of process exergy E_{in} . In general,

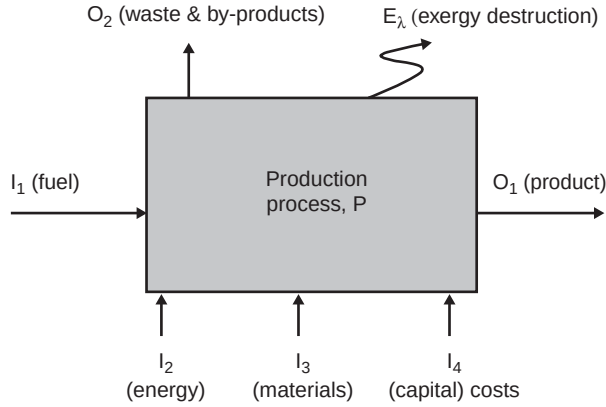


FIGURE 4 A commodity-generating process.

some waste O_3 and some by-product O_2 are also produced. Two additional production factors, namely labor and capital, must also be accounted for. From the process mass and energy balances we can compute the amounts of $m_{I1}, m_{I2}, \dots, m_{IN}$ and the corresponding exergies $e_{I1}, e_{I2}, \dots, e_{IN}$ necessary to produce one unit of the desired commodity O_1 (1 kg, one piece, or 1 kJ, depending on the type of product). We can also compute the unit expense for labor (man-hours for unit output) and for capital (€ of rate of return on investment in the equipment per unit output). Over the time period t_p (usually 1 year), we can thus compute O_1 production cost as the ratio of the sum of all expenses (materials, exergy, labor, and capital) to the cumulative amount of production of O_1 , both computed in t_p :

$$c_{O_1} = \frac{\text{Cumulative expenses}}{\text{Cumulative production}} = \int_{t_p} \left(\frac{C_I + C_E + C_L + C_C}{O_1} \right) dt \quad (\text{€/kg, €/unit, or €/kWh}) \quad (4)$$

where O_1 is measured in proper units (kg/s, units/s, or kW). Such a production cost is dependent on both the structure of the production line and its efficiency. If we improve the efficiency, we can obtain a larger amount of O_1 with the same amount of inputs or, conversely, the same O_1 with a reduced input. If we decrease the number of components in the plant, we can lower the capital expenditures for the same amount of output. In practice, the functional relation between efficiency and capital cost is such that higher efficiency components have a higher initial cost, and thus lead to higher capital expenditures. However, the relationship is not linear, and this indicates the possibility of performing process optimization (i.e., searching for process structures and/or operating

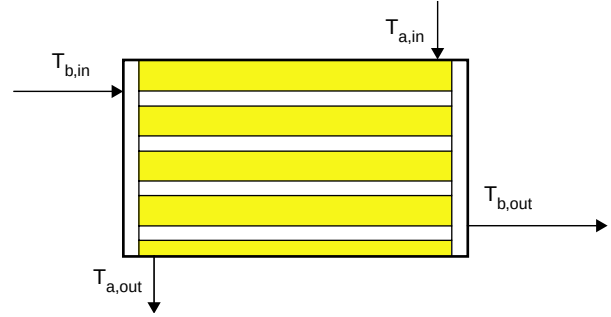


FIGURE 5 Schematic representation of a heat exchanger.

parameters that lead to a lower production cost C_{O_1}). A whole branch of engineering, called engineering economics, deals with these issues: We can explore some of its implications by examining a simple example consisting of an elementary energy conversion device, a heat exchanger.

The function of a heat exchanger is to transfer a quantity of energy Q to a medium at lower temperature. In Fig. 5, fluid a (the heat donor) transfers a portion of its sensible heat, $m_a c_{pa} (T_{a,in} - T_{a,out})$ to fluid b (heat acceptor), which increases its temperature from $T_{b,in}$ to $T_{b,out}$. The process can be described by the following equations:

$$\text{Energy balance : } m_a c_{pa} (T_{a,in} - T_{a,out}) \eta_{\text{exchanger}} = m_b c_{pb} (T_{b,out} - T_{b,int}) \quad (5)$$

$$P_{\text{pump,tot}} = \frac{m_a \Delta p_a}{\rho_a \eta_{\text{pump,a}}} + \frac{m_b \Delta p_b}{\rho_b \eta_{\text{pump,b}}} \quad (6)$$

$$\text{Heat transfer at the interface : } Q_{a \rightarrow b} = m_b c_{pb} (T_{b,out} - T_{b,int}) = UA_{\text{HE}} \text{LMTD} \quad (7)$$

In Eq. (7),

$$\text{LMTD} = \frac{(T_{a,in} - T_{a,out}) - (T_{b,in} - T_{b,out})}{\ln \left(\frac{T_{a,in} - T_{b,out}}{T_{a,out} - T_{b,in}} \right)}$$

is the log-mean temperature difference calculated under the following assumption: For fixed mass flow rates, the amount of heat transferred from a to b increases with LMTD. Since $T_{a,in}$ and $T_{b,out}$ are specified, this means that a smaller terminal temperature difference, $\text{TTD} = T_{a,in} - T_{b,out}$, is desirable. However, to lower the TTD more, heat exchange (HE) area must be used, and this leads to an increase in the pressure losses Δp_a and Δp_b in the heat exchanger passages and to a higher pump work. Here, Reynolds effects on the overall heat-exchange coefficient U and on the friction factor f are neglected, both of which are assumed to remain

fixed throughout. The design problem can be rephrased as follows:

1. The effectiveness of the HE increases with an increase in the active HE surface so that more output $[m_b \times c_{p,b} \times (T_{b,out} - T_{b,in})]$ is generated by the same input $[m_a \times c_{p,a} \times (T_{a,in} - T_{a,out})]$.
2. The parasitic losses that the pumps and fans must compensate for also increase with the active HE surface so that a proportionally larger amount of electrical energy must be purchased.

Let us formulate a costing problem on the following bases: The cost of the unit of heat acquired by stream b is given by the rate of cash outflow c_{HE} corresponding to the installed cost C_{HE} of the exchanger (for simplicity, only the active area A_{HE} is considered here) plus the cost of stream a c_a plus the cost of the electricity c_{el} needed to power the pumps. The following expressions may be used:

$$c_{HE} = CRF C_{HE} \quad (8)$$

$$CRF = \frac{i(1+i)^N}{(1+i)^N - 1}, \quad (9)$$

where $i = 0.05$ is the yearly interest rate and $N = 15$ years is the assumed technical life of the HE:

$$C_{HE} = K_{HE} \left(\frac{Q_{a \rightarrow b}}{U \times LMTD} \right)^x = K_{HE} A^x, \quad (10)$$

with $K_{HE} = 2000 \text{ €/m}^2$ and $x = 0.6$.

where $K_{HE} = 2000 \text{ €/m}^2$ and $x = 0.6$.

$$c_a = K_a \times m_a \times c_{p,a} (T_{a,in} - T_E) \quad (11)$$

$$c_{el} = 7.7 \text{ ¢€/kWh} \quad (12)$$

$$\Delta p_b = f \times m_b^2, \text{ with } f = 0.01 \text{ 1/(m} \times \text{kg)} \quad (13)$$

where $f = 0.01[1/(m \times \text{kg})]$. For a fixed c_a and assuming a zero cost $c_{b,in}$ for stream b at the entrance of the heat exchanger, the specific cost of the heating of stream b is given by

$$c_{b,out} = \frac{c_{HE} + c_{el}}{m_b} = CRF \frac{K_{HE} A_{HE}^x}{m_b} + \frac{c_{el} P_{\text{pump,tot}}}{m_b}. \quad (14)$$

Since m_b increases almost linearly with A (Eq. 7), whereas P is proportional to the square of m_b (Eq. 13), the area-related cost term in Eq. (14) will decrease with an increasing heat exchange area and the pressure-loss term will increase linearly (because f is constant). Figure 6 depicts the situation: The total cost $c_{b,out}$ exhibits a minimum that can be found by optimizing, under the given assumptions and constraints, Eq. (14) with respect to A_{HE} . By manipulating the costing equations, the following

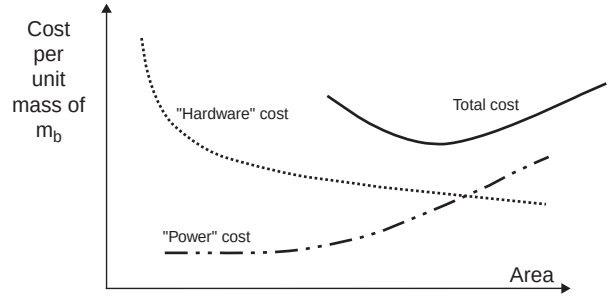


FIGURE 6 Determination of the optimal cost of the heated stream.

(approximate) expression is obtained:

$$c_{b,out} \approx C_{\text{area}} \times A_{HE}^{-0.4} \times C_{\text{power},1} \times A_{HE}^{-1} + C_{\text{power},2} \times A_{HE}^2 \quad (15)$$

which derived w.r.t. A_{HE} and equated to 0 provides an expression for the optimal heat exchange area A_{opt} :

$$-0.4 \times C_{\text{area}} \times A_{\text{opt}}^{1.4} - C_{\text{power},1} \times A_{\text{opt}}^{-2} + 2 \times C_{\text{power},2} \times A_{\text{opt}} = 0 \quad (16)$$

where

$$C_{\text{area}} = \frac{CRF \times K_{HE} \times c_{p,b} \times (T_{b,out} - T_{b,in})}{U \times LMTD}$$

$$C_{\text{power},1} = \frac{c_{el} \times m_a \Delta p_a \times c_{p,b} \times (T_{b,out} - T_{b,in})}{\rho_a \times \eta_{\text{pump},a} \times U \times LMTD} \text{ and}$$

$$C_{\text{power},2} = \frac{c_{el} \times f}{\rho_b \eta_{\text{pump},b}} \left[\frac{U \times LMTD}{c_{p,b} (T_{b,out} - T_{b,in})} \right]^2.$$

The approximation here is due to the fact that the constants depend on LMTD and $(T_{b,out} - T_{b,in})$, which in turn vary with A .

The method of engineering economics is employed in process and design engineering to assess the solution to a design problem not only from a thermodynamic standpoint but also from an economic one. As demonstrated in the previous example, this method is based on the following assumptions:

1. Mass and energy balances exist for the process under consideration.
2. A suitable cost equation can be derived for all products of interest.
3. The coefficients in this cost equation may be expressed in terms of process parameters.
4. The specific cost of all incoming streams is known.

Of course, in real engineering practice problems are not so crisply posed as purported here. The

following are the most common difficulties encountered in industrial cost analysis:

1. Only approximate or incomplete mass and energy balances exist. In this case, good technical common sense must be used to approximately close all balances.

2. Cost equations are not known for all production factors (material, energy, labor, and capital), or they are known only for plants of a different scale. The missing data must be judiciously extrapolated from the existing ones or prudently estimated on the basis of known data in related fields.

3. The cost of incoming streams is not known since they originate from the environment or are cogenerated in the same process. For the flows of raw materials of unknown origin, usually transportation costs may be used as a good approximation. For water or air at ambient temperature, the cost is usually assumed to be equal to zero. In this case, however, a better choice is to charge the incoming stream with the (real or virtual) effluent treatment costs necessary to restore its original environmental quality.

4. Labor costs are site dependent so that the same process may have different economic optima in different locations. The location becomes a design parameter.

5. Capital costs (interest rate, social costs, taxation level, etc.) are site dependent. Furthermore, they are also time dependent. Economic optimization must be performed independently for each site. Regarding the time dependency, levelization methods exist that allow the engineer to average out both market fluctuations and changing financial conditions by referring the calculations to a fixed economic state of reference (current money and constant money formulations).

6. There are more than one useful outputs, and their production lines share some of the components, materials, labor, and energy inputs. Allocation decisions must be made about what share of capital, material, labor, and energy costs must be allocated to which product. The optimal point is substantially affected by the allocation criteria.

Engineering economics encompasses more complex topics than those described here: the treatment of cash flows, the allocation of equipment depreciation, the inclusion of inflation and taxes in the economic calculations, etc. Interested readers may consult the book by Bejan, Tsatsaronis, and Moran.

The limits of engineering economics, in the formulation presented here, depend on the fact that

it entirely neglects second law considerations. The efficiencies that appear in its formulation are first law efficiencies and as such are incapable of assessing the thermodynamic quality of participating streams. This is particularly inconvenient when the inputs and/or outputs of the process under consideration have different qualities, as in the frequent case of combined or cogenerating power plants. In these cases, engineering economics provides the wrong answer: Both the optimum it suggests and the relative cost ratio between different products are incorrect. It is from a critical analysis of these weaknesses that TE was first formulated.

4. SECOND VS FIRST LAW OPTIMIZATION

Suppose we wish to find the optimal pressure ratio β for the turbogas of the cogeneration plant shown in Fig. 3. This is indeed a credible design task. An increase in β leads to higher efficiency, but also to a lower air mass flow rate, and for a fixed turbine inlet temperature T_4 , it leads to a decrease in T_5 so that for a fixed water return temperature T_{10} , the amount of generated steam decreases as well. A decrease in β , on the other hand, reduces the efficiency of the gas turbine cycle and, for a fixed required power output, increases the mass flow rate at compressor inlet, produces a higher T_5 , and produces a higher steam mass flow rate. Figure 7 displays the sensitivity of the plant outputs (P , Q_{12} , and E_{Q12}) to the variation of β . First and second law efficiencies are shown as well. It is clear that η_I and η_{II} peak for different values of β . Which objective function is correct—the one based on η_I or the one based on η_{II} ? From a purely thermodynamic standpoint, η_{II} provides the correct solution, and the reason is apparent if we look at the energy and exergy flow diagrams of the process, which are called the Grassmann (for exergy) and

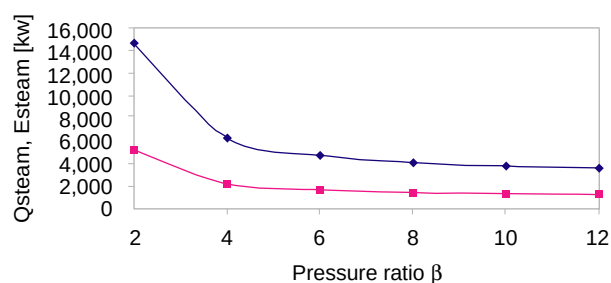


FIGURE 7 Cogeneration plant: sensitivity of the global performance indicators to β .

Sankey (for energy) diagrams (Fig. 8). These diagrams constitute a very effective pictorial representation of the energy and exergy flow within the process: They consist of boxes interconnected by arrows, with each arrow representing a stream and its width being proportional to the relative energy or exergy flux associated with it (usually normalized with respect to the main fuel input). Each box represents a component. For the plant under consideration, the energy diagram indicates that the losses, measured as percentage of the total input, are approximately the same for all components: 20% for the pump, 15% for the compressor and the boiler, 10% for the turbine, and 5 and 2% for the alternator and combustor, respectively. On the basis of this information, if we wanted to improve the design, we would concentrate on the pump first, then on the compressor and boiler, and so on. The diagram provides a ranking of the relative importance of each component in the energy dissipation during the process. The exergy diagram tells a different story, however: The highest losses (approximately 40%) are in the combustor and the boiler. This reflects more accurately the physics of the phenomena: In these two components, we enforce strongly irreversible transformations of high-quality into low-quality energy, irrevocably destroying exergy. A more

accurate analysis (not performed here) would show that these low exergetic efficiencies are almost independent of the process parameters: They cannot be cured by improving, for instance, the thermal insulation of the combustor or by increasing the global heat transfer coefficient in the boiler. The irreversibility is introduced by the type of transformation these two components enact, and to remove the corresponding exergy destruction we would have to select a different device (e.g., a fuel cell instead of the combustor and a back-pressure steam turbine instead of the boiler). Thus, the proper objective function to use in the thermodynamic optimization is the one based on second law concepts.

Caution should be exercised here because energy conversion systems display a relatively high degree of interdependency: Components interact in complicated ways with each other, and an improvement in the efficiency of one component may lead to increased exergy destruction in other units, resulting in no net change or even a decrease in η_{II} . Furthermore, the installation cost of a component depends in a nonlinear way on its efficiency, and it may also affect the magnitude of other production factors, such as labor or material input. Therefore, an exergy analysis by itself does not tell the entire story. Methods are needed that combine exergy

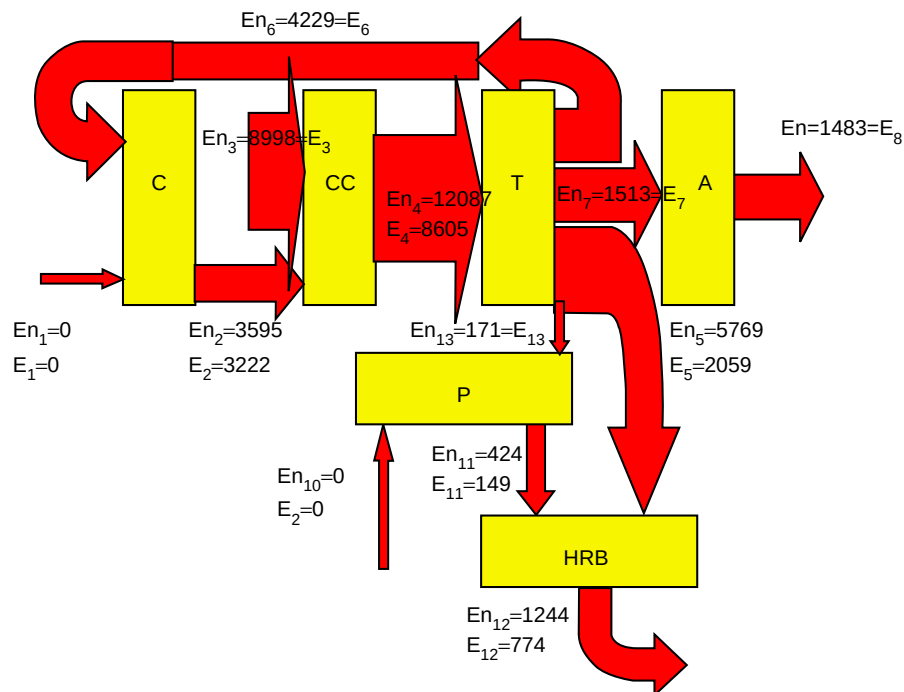


FIGURE 8 Energy (Sankey) and exergy (Grassmann) flow diagrams for the cogeneration plant shown in Fig. 3. En , energy flux; E , exergy flux (both in kW).

analysis and costing, as engineering economics does with energy analysis and costing. The simplest of these methods is TE.

An entirely different type of optimization problem encountered in practice is the so-called design synthesis problem, in which the goal is to find the process structure (interconnectivity) that yields a maximum of the objective function. This is a much more difficult task, and some of the possible solution methods are discussed by Frangopoulos. In general, second law analysis provides more insight than its first law counterpart. Especially for synthesis problems, it is important to correctly assess the contribution to the overall process performance of each component, and the exergy flow diagrams offer an immediate glimpse into the real causes of the irreversibilities occurring in the process.

5. STRUCTURAL REPRESENTATION OF ENERGY CONVERSION SYSTEMS

One of the major difficulties in the assessment of the optimal performance of an energy conversion system regards the complex interconnection of its components. As a result, the response of the system to variation in the performance of one of its subunits is not easily predictable: It may well happen that an increase in the efficiency of one component leads to an overall performance deterioration. Several methods have been proposed to represent these cross-effects: The symbolic exergoeconomic method, first formalized by Valero and coworkers in 1988 and 1989, is discussed in detail here, complemented by a formalism adapted from the book by Kotas.

Consider again the cogeneration process of Fig. 3, which we refer to as **P** in this section, and define a matrix **CM** (connectivity matrix) constructed according to the following algorithm:

1. The j th column ($j = 1, \dots, N$) of **CM** represents the j th process streams.
2. The i th row of **CM** ($i = 1, \dots, M$) represents the i th component.
3. The generic element CM_{ij} is 0 if flux j does not interact with component i , 1 if flux j enters component i , and -1 if flux j exits component i .

For $M=6$ and $N=13$, Fig. 9 represents the **CM**(6,13) for the plant shown in Fig. 3. This matricial representation of a process can also be derived from an application of graph theory (the

Components	Fluxes												
	1	2	3	4	5	6	7	8	9	10	11	12	13
C	1	-1	0	0	0	1	0	0	0	0	0	0	0
CC	0	1	1	-1	0	0	0	0	0	0	0	0	0
T	0	0	0	1	-1	-1	-1	-1	0	0	0	0	0
A	0	0	0	0	0	0	0	0	0	0	0	0	0
HRB	0	0	0	0	1	0	0	0	0	-1	0	1	-1
P	0	0	0	0	0	0	0	0	1	0	1	-1	0

FIGURE 9 The connectivity matrix associated with the cogeneration plant shown in Fig. 3.

Leontief input–output model) and is the basis for a compact and elegant computational procedure. By examining the streams in the cogeneration plant, we see that some of them transport mass and exergy (1–5 and 9–12) and some only exergy (6–8 and 13). We can correspondingly define a mass vector **vm**(13), every element of which represents the mass flow associated with the corresponding stream, and two energy and exergy vectors **vh**(13) and **ve**(13), which contain the energy and exergy values of each stream, respectively. As a further convention, **vm** contains a 0 in all positions corresponding to a “massless exergy carrier” (a flux of mechanical or electrical energy). It turns out that the mass and energy balances as well as the exergy accounting of the process are described by the scalar product of the matrix **CM** with the three vectors **vm**, **vh**, and **ve**, respectively.

ve is a vector that contains the exergy destructions in each component. On the basis of this formalism, Valero developed a procedure to symbolically compute the general transfer function of the process—that is, the matricial function **Π** that premultiplied by any property vector of the input streams (the fuels in Valero’s terminology) provides the resulting property vector for the output streams:

$$\begin{aligned}\Pi \times \mathbf{mv}_I &= \mathbf{mvo} \\ \Pi \times \mathbf{hv}_I &= \mathbf{hvo} \\ \Pi \times \mathbf{ev}_I &= \mathbf{evo}\end{aligned}\quad (17)$$

There is a direct relationship between **CM** and **Π**, the second being a sort of augmented form of the former, but with the 0 and 1 substituted by influence coefficients. Note that **Π** is necessarily square, whereas **CM** is not. This transfer function is of course a function of the internal characteristics of each component, particularly the efficiencies. Once a closed-form expression for **Π** is known, we can study the sensitivity of the process performance to that of each one of its components. It is convenient here to adopt the formalism of the influence coefficients, first

proposed independently by Beyer and Brodyanskii in the 1970s and later systematized by Kotas. For each component, we thus define a coefficient of structural bonds σ_{ij} and a coefficient of influence π_{ij} :

$$\sigma_{k,i} = \frac{(\partial E_{\lambda,P}/\partial x_i)}{(\partial E_{\lambda,k}/\partial x_i)} \quad (18)$$

$$\pi_{j,i} = \left(\frac{\partial E_{in}}{\partial E_j} \right)_{x_i=\text{variable}} \quad (19)$$

In Eq. (18), $E_{\lambda,P}$ and $E_{\lambda,k}$ are the exergy destructions of the entire process and of the k th component, respectively, linked to the efficiency by

$$\eta = 1 - \frac{E_{\lambda}}{E_{\text{input}}} \quad (20)$$

If the i th thermodynamic variable x_i is varied, it produces in general a series of modifications in the efficiencies of some of the components: The sensitivity of each component is measured by $\partial E_{\lambda,k}/\partial x_i$. At a higher aggregation level, the coefficient of structural bonds $\sigma_{k,i}$ measures the sensitivity of the process to a variation of the efficiency of the k th component. However, the variation of the efficiency of components that operate on quantitatively small exergy flows is not as important for the process as that of a major components that elaborate the main fluxes of exergy: This aspect is quantified by the influence coefficient π_{ij} . If a variation of the i th thermodynamic variable x_i causes a variation in the magnitude of a certain exergy flux E_j , π_{ji} measures the influence on the overall input of the process.

Once we know σ and π , we can formally analyze (i.e., without to a numerical process simulation) the influence of a variation in a design or operating parameter of the plant. The problem, of course, is that to obtain these coefficients we must perform either an extensive experimental campaign or numerical process simulations.

6. THERMOECONOMICS

Thermoeconomics, in Tsatsaronis' definition, is exergy-aided cost minimization. It is a relatively new branch of engineering sciences that combines engineering economics and exergy analysis to answer the question of the "optimal" process structure and design. The major assumption in TE is that exergy is the only rational basis for constructing a costing function. The difference between TE and EE lies in the treatment of the monetary production factors: TE considers a monetary product cost optimized in

terms of exergy considerations, whereas EE transforms monetary expenses into equivalent exergy fluxes and performs an optimization operating solely on exergy fluxes.

In TE, a cost is associated with each stream according to its exergy content. For the generic component shown in Fig. 10, a cost balance shows that the sum of the cost rates associated with the output flows is equal to that of the cost rates associated with the input flows plus the cost rates due to the nonenergetic production factors (capital, labor, and environmental costs). The latter are computed exactly as they would be in engineering economics, taking into account interest rates, inflation, taxes, plant depreciation, etc., and they are denoted here by Z (€/s). The cost balance equation for the component of Fig. 10 (a boiler) is

$$\begin{aligned} \sum_{\text{in},j} (c_{\text{in}} m_{\text{in}} e_{\text{in}})_j + c_q E_q + c_w E_w + Z \\ = \sum_{\text{out},i} (c_{\text{out}} m_{\text{out}} e_{\text{out}})_i, \end{aligned} \quad (21)$$

where the mass flow rates m_j , the corresponding exergy values e_j , and the heat (E_q) and work (E_w) exergy flows are known from previous process calculations. If there are N inputs and M outputs, there will be $N + M$ exergetic cost coefficients c_k . For the analysis to be feasible, the N coefficients relative to the input flows must be known, either from the design specifications or from the analysis of upstream components. We are left with M unknowns, and we need $M - 1$ auxiliary equations to supplement Eq. (22). These auxiliary equations are derived on the basis of thermodynamic cost-allocation considerations. In essence, these auxiliary relations express the designer idea of the purpose of that particular component. Note that they are needed to make the transfer function matrix Π square. In our case, the reasoning is as follows:

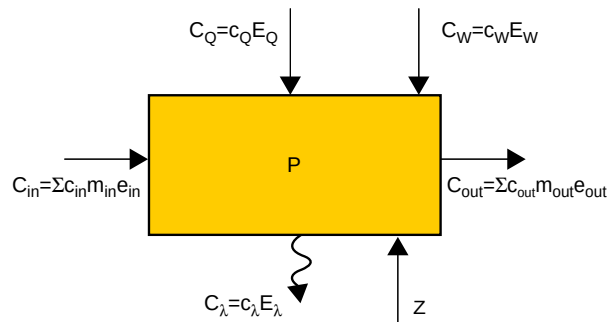


FIGURE 10 Thermoeconomic representation of a process P . All flows in €/s.

1. Since the flue gas must pass through a gas-treatment unit before being discharged into the environment, the cost rate associated with the installation and operation of this unit must be charged to this stream:

$$c_3 = -Z_{\text{Desox}}/m_3e_3. \quad (22)$$

2. Similarly, the ash must be charged with the costs of the ash-handling and -disposing equipment:

$$c_4 = -Z_{\text{inert}}/m_4e_4. \quad (23)$$

3. Regarding the boiling and the superheating sections of the tubes, we can consider that the boiler must provide both streams with the respective exergy increase at the same time; therefore, for both streams the cost increase is proportional to the exergy increase with the same constant of proportionality:

$$\frac{m_6c_6 - m_5c_5}{m_6e_6 - m_5e_5} = \frac{m_8c_8 - m_7c_7}{m_8e_8 - m_7e_7}. \quad (24)$$

Bejan, Tsatsaronis, and Moran provide examples of commonly used auxiliary equations for various components.

Once we have M equations in M unknowns, we can solve the TE problem and compute the cost of each stream, particularly that of the final products. If we want to execute an optimization, however, we need additional functional relationships to link the exergetic efficiency of each component to its installation cost (which is reflected in the capital flow factor Z). In the past decade, data have been published by several authors: we recommend the 1998 compilation by El-Sayed.

TE is the most advanced and powerful technique for the cost optimization of complex production systems. It requires an understanding of the exergy method of system analysis and a working knowledge of engineering economics: Both topics are included in most engineering curricula. In every monthly issue of every energy-related technical journal, one may find original applications of TE or a revisitation of an old application. Its impact on the engineering community has been profound because it has promoted the shift, advocated by exergy analysis methods, toward the consideration of irreversibilities and of their real causes in industrial processes. However, it has two weak points. First, being in essence a revised version of engineering economics, it suffers from the adoption of two separate quantifiers: one (exergy) to assess the technical worth of a process, and the other (money) to assess

the profitability of the product. There is something inherently awkward in the “pricing” of a physical stream that makes one consider whether it would not be more convenient to base decision tools on purely physical parameters. Second, as a consequence of the first weakness, a correct assessment of environmental issues is problematic in TE. Despite some attempts, for instance, by Frangopoulos and von Spakowsky and independently by Szargut, the problem of attributing a cost to environmental damage remains an open and controversial issue. The current solution is to use the results of large-scale medical studies to attribute a “social cost” to the unit of pollutant discharged into the environment and then charge this cost to the product(s) of the plant. Since pollution effects are inherently nonlinear, and the translation of “a tolerable amount of health risk” in monetary terms is somewhat perplexing, this issue has been intensely debated, resulting in the formulation of a method for process costing based solely on exergy. This method should be called exergoeconomics, and it is described as such here. However, since many engineers use TE and EE as interchangeable terms with the same meaning, it has been published under the name of extended exergy accounting (EEA).

7. EXERGoeconomics AND EXTENDED EXERGY ACCOUNTING

If we want to unify the TE costing method and express all expenses by means of a single quantifier, how can we convert monetary expenses into exergy fluxes? The answer has been provided by a controversial energy-accounting theory, *emergy* (energy memory) analysis, formulated in the 1970s by Howard Odum and mainly used to study the energy cascade of biological systems. Odum maintains that the monetary capital that circulates in a society (a town, county, or nation) is entirely supported by, and therefore equivalent to, the amount of solar energy that maintains life in the environmental niche in which that society thrives. This idea, not developed to its full potential in *emergy* theories, has been echoed in Herman Daly’s critical assessment of the economic theories of growth: An economist by trade, Daly (echoing and developing concepts formulated by others, such as Nicholas Georgescu-Roegen, Lotka, and Robert Costanza) distinguishes between natural capital and monetary capital and demonstrates that the two are not completely substitutable for each other but rather complementary. The

conclusion is that it is the amount of resources we avail ourselves of, and not our capital intensity, that is the key to our survival as an industrial society on Earth. From a completely different point of view, Szargut reached similar conclusions. He devised a procedure to track all amounts of exergy used in the production of a commodity from the mining of the raw materials to the final distribution and disposal, and he proposed to associate a cumulative exergy content to each good or service to express its value in purely exergetic terms. However, both Szargut and Daly left unanswered the problem of how to convert nonenergetic expenditures into resource consumption indices. Extended exergy accounting is the first theory to provide a unified treatment of these issues, along the guidelines discussed here.

EEA adopts the standard exergy accounting method of Szargut to embody into a product all of the exergetic expenditures incurred during its production: Extraction, refining, transportation, preprocessing, final processing, distribution and disposal activities are computed in terms of exergy consumption. Exergy, unlike energy, is not conserved, and in each step of the extended production line a portion of it is irrevocably destroyed in irreversible entropy generation. Thus, the exergetic production factors can be accounted for in full. Regarding the three nonenergetic externalities (labor, capital, and environmental remediation costs), EEA proposes the treatment discussed in the following sections.

7.1 Capital

With regard to including capital and general financial costs in the balance sheet of an energy conversion system, there seems to be a general agreement: We employ money and its time value as the sole quantifier for goods, services, labor, resources, etc. This paradigm of money as a basic metric for energy-related activities is so deeply rooted in our culture that most industrial energy balances are expressed in monetary price per unit of product rather than per energetic (not to mention exergetic) content. There are two reasons to support the necessity of maintaining a purely monetary criterion:

1. The society within which energy conversion systems operate consists of a complex network of related systems, and all exchanges among them are already assessed on a monetary basis.
2. On an average world basis, the contribution of energy-related costs to the global gross national product varies between 2 and 9%; therefore,

money (capital), and not energy, is the main quantifier of human activities.

It is perfectly logical to reverse this point of view by considering that economic systems are ecosystems that function only because of the energy and material fluxes that sustain human activities. All agricultural, industrial, and economic activities can only exist as long as they exploit (use) biophysical resources taken from a reservoir of noninfinite mass capacity but of practically infinite exergy capacity. From this point of view, it clearly appears that exergetic content, and not capital, is the correct measure for the value of a commodity or a service, and that the monetary price should reflect this new scale of values.

Naturally, to support the cancellation of monetary prices would be unrealistic and destabilizing: The very structure in which we live and function demands the conservation of the “price-tag” concept in everyday economic activities and on a worldwide scale. EEA advocates that this tag be calculated on the basis of the extended exergetic content (EEC) of a good or service. The numerical relation between the EEC and the price of a product is also immediately clear: The conversion factor is the ratio of some measure of the monetary circulation to the global exergetic input; just as the monetary flux is the driving agent of societal development according to the current paradigm, the exergy flow is its physical driving force. The choice of a proper indicator for the monetary flux, of course, is somewhat arbitrary: EEA adopts the absolute measure of the global monetary circulation ($M2$), which for each country is computed and published by the central bank (or other organization):

$$e_{\text{Cap, Sector}} = \frac{E_{\text{in, Sector}}}{M2} [DJ/\epsilon]. \quad (25)$$

When processes are analyzed, it is the exergetic equivalent of the financial expenditures that must be inserted in the balance as an input, and the same applies to the revenues. It is noteworthy that for most state-of-the-art processes the substitution of the equivalent exergetic value for the monetary price causes a significant discrepancy in the exergy balance: the reason is clearly the over- or underestimation of the real extended exergetic content of materials, feed stocks, labor, and energy flows. For the EEA method to be optimally applicable, it is necessary that the economic and exergetic values become locally consistent in the long term, allowing the two value scales to reach a sort of local fixed parity. The same definition of the exergy equivalent implies that different countries may have different

e_{Cap} due to their different productive and economic structures and lifestyles, and that for a country e_{Cap} may vary over time due to an evolving social structure.

7.2 Labor

The capability of attaching a properly computed exergetic value to a labor input (taken here to include all service-related, blue- and white-collar human activities) is perhaps the most relevant novelty of EEA. In all practical applications of engineering cost accounting, including TE, labor is accounted for on a purely monetary basis, but very reasonable alternative approaches have been formulated. Odum's previously mentioned emergetic approach, for instance, computes a value for labor on the basis of the total time-averaged embodied energy content of the human life in the particular area where a worker lives and works. The problem is that the attribution of a purely monetary value to labor is unsatisfactory because it assigns a higher weight to market conditions and financial considerations than to social, technical, and environmental issues that although properly valued may so distort the objective function that new solutions emerge.

EEA proposes to assign labor, and human services in general, in each portion of a society an exergetic value computed as the total (yearly averaged) exergetic resource input into that portion divided by the number of working hours generated therein:

$$e_{\text{L,Sector}} = \frac{E_{\text{in,Sector}}}{n_{\text{work hours,Sector}}} \quad [\text{J}/\text{man-hour}]. \quad (26)$$

7.3 Environmental Impact

All methods for assigning a monetary cost to environmental damage in essence suggest that once a substance is acknowledged to be (in some sense) harmful, it become regulated (i.e., legal upper limits are set for its free release into the environment). This is equivalent to setting, for a given technology, an upper price for the cleanup costs for that particular effluent. For many pollutants, however, the environmental situation has deteriorated to a point at which the overall balance of the biosphere has already been affected, and even a small amount of each one of the pollutants is, strictly speaking, intolerable. Therefore, to make a nonzero limit acceptable, the risks to humans are assessed in terms of monetary health- and life-expectancy parameters, and an upper bound is set a priori to the expenditures in such a way so as

to remain below a certain statistical probability of incurring that risk. It is claimed that the consequent economical burden is transferred both ways in the economy: downwards, affecting the price of resources that are perceived to be more or less clean, and upwards, increasing the price of products seen as less environmentally friendly. The fallacy of this assertion may be exposed by considering that willingness to pay and the attitude toward a sustainable resource exploitation are different in different countries and may well vary in time. Therefore, in reality, the disturbances caused by a local modification in the environmental regulations affect regions possibly far away from the point of origin, and the method actually promotes an unfair transfer of not only the pollution but also the health risks from a region to another. A solution to this situation has been sought by linking the monetary structure of the environmental levies to energetic considerations: This is the rationale behind pollution commodity trading and the exergy tax. These are remedial measures, aimed at a fairer redistribution of the environmental pressure on a global scale, and they do not address the issue of how high the actual environmental cost is (all methods use the currently regulated values as a basis for their calculations).

EEA advocates a substantially different approach. Consider a process P (Fig. 11A), and assume that its only effluent is a stream that contains hot chemicals, some of which are not at their standard environmental concentration. To achieve a zero environmental impact, these chemicals would have to be brought to both thermal and chemical equilibrium with the surroundings. Thus, the real exergetic cost of the zero impact is proportional to the physical exergy of the effluent because it corresponds to the extended exergy (the sum of the physical exergy spent in the cleanup process plus the invested exergy—labor and capital—required by the installation and operation of the effluent cleanup devices) ideally required to cool the effluent to T_0 and break it up into its constituents such that each one is in equilibrium conditions with the surroundings. A possible representation of such an effluent treating process is shown in Fig. 11B: The additional process P_t requires an energetic input—possibly some auxiliary materials, labor, and invested exergy—but its output will have a zero physical exergy. The additional exergetic expenditures required by P_t must be charged to the effluent O_2 , whose extended exergy will now be higher than its original one. Therefore, the overall conversion efficiency of the joint process $(P + P_t)$ is decreased. There may be

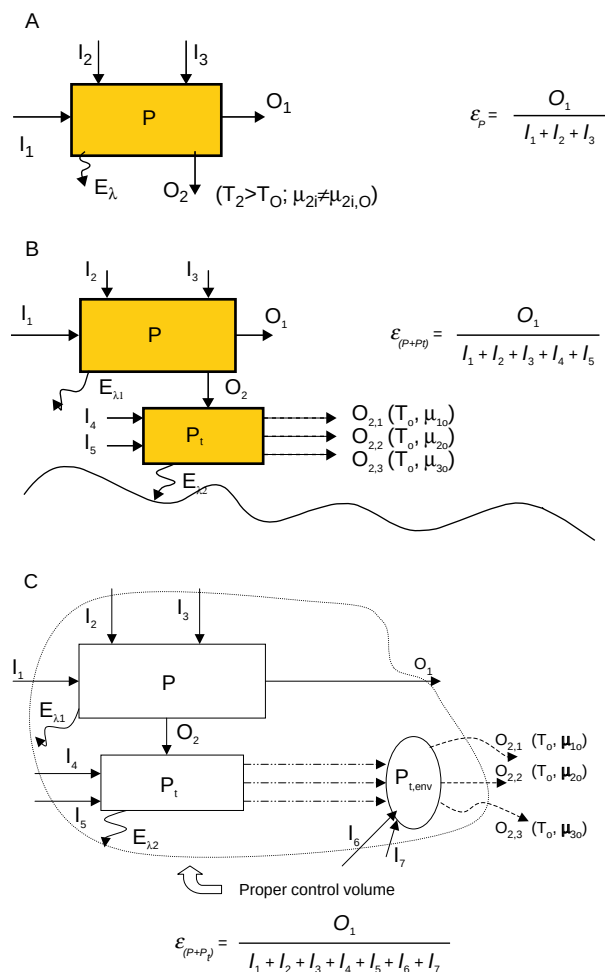


FIGURE 11 Environmental cost assessment in EEA. (A) The effluent O_2 is not at reference conditions. (B) Treatment of the effluent O_2 . Each final effluent is at its reference conditions. (C) Only a portion of the cleanup is performed by man-made treatment processes; the remainder occurs spontaneously by biodegradation in the immediate surroundings.

effluents for which some of the chemical decomposition reactions take place spontaneously, in a short time and in the immediate surroundings of the emitting source. In such cases (Fig. 11C), the reactions must draw on some exergy source within the environment (a certain particular chemical catalyst, oxygen, water, solar radiation, or even a biological system), and this exergy flow must be accounted for as well.

EEA thus allows for a consistent incorporation of the effects of effluent treatment in the extended exergetic balance of a process, and it provides an absolute order-of-magnitude estimate of the minimum exergy consumption necessary to achieve zero impact. If an acceptable level of pollutant is specified,

then the minimum exergetic expenditure will be proportional to the difference between the values of the physical exergies of the effluent stream between the point of its release and the regulated state point. Here, we use one of the desirable features of environmental cost estimates and at the same time effectively avoid the considerable effort required to determine what the “tolerable environmental impact limit” for a certain pollutant would be.

Once all fluxes (which can be equally regarded as production factors) have been computed in terms of exergy equivalents, EEA makes use of the same structured representation of TE to compute the costs of the final products and to study their dependence on a variation of process parameters. With all terms now expressed in terms of a uniform quantifier, the procedure is somewhat simpler and, more important, absolutely independent of external factors such as market fluctuations or time-varying currency exchange levels.

8. CONCLUSIONS

Current engineering practice uses a mixed energy-monetary formulation to study the cost-formation process of industrial products. TE represents a substantial improvement because its exergy-monetary procedures allow for an exact treatment of irreversibilities and for a better (more precise) accounting of the real causes of exergetic waste, which is equivalent to production cost (if we could find abundant amounts of a certain finished product in nature, its cost would be zero). EE, or EEA, broadens the objective and bases its treatment on a purely exergetic paradigm. TE is becoming increasingly popular, and modern production lines are designed according to its dictates. It is hoped that in the near future, EEA will constitute the basis for our cost-accounting procedure. When (and if) this occurs, we will be able to assess our industrial processes on performance indicators that correctly reflect the resource to final use (including disposal, in a omnicomprehensive cradle-to-grave approach) of our exergy resources.

SEE ALSO THE FOLLOWING ARTICLES

Economic Thought, History of Energy in • Entropy and the Economic Process • Evolutionary

Economics and Energy • Exergy • Thermodynamics and Economics, Overview • Thermodynamics, Laws of

Further Reading

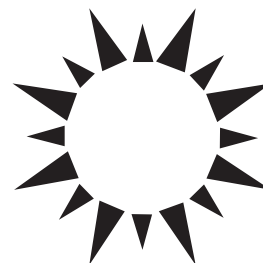
- Ahern, J. E. (1980). "The Exergy Method of Energy Systems Analysis." Wiley, New York.
- Baehr, H. D. (1962). "Thermodynamik. Eine Einführung in die Grundlage und ihre technischen Anwendungen." Springer-Verlag, Berlin.
- Bejan, A., Tsatsaronis, G., and Moran, M. J. (1996). "Thermal Design and Optimisation." Wiley, New York.
- Bosnjakovic, F. (ed.). (1961). "Fachheft Exergie, Die praktische Bedeutung der Exergie," Vol. 13, No. 11.
- Daly, H. (1996). "Beyond Growth: The Economics of Sustainable Development." Beacon, Boston.
- El-Sayed, Y. (1998). A short course in thermoeconomics, Notes for summer school, Ovidius University, Constantza, Romania.
- Georgescu-Roegen, N. (1971). "The Entropy Law and the Economic Process." Harvard Univ. Press, Cambridge, MA.
- Moran, M. J. (1982). "Availability Analysis: A Guide to Efficient Energy Use." Prentice-Hall, Englewood Cliffs, NJ.
- Odum, H. T. (1971). "Environment, Power and Society." Wiley, New York.
- Sciubba, E., and Wall, G. (2003). Exergy—A brief commented history. *Int. J. Exergy*.

Exergy

GÖRAN WALL

Exergy Consultant

Mölndal, Sweden



1. Thermodynamics and Exergy

2. Common Forms of Exergy

3. Exergy Tools

Glossary

energy Motion or ability of motion (e.g., the disordered motions of the hot molecules in a cup of coffee); measured in joules (J).

entropy A physical concept of disorder; measured in joules per kelvin (J/K).

exergy Work (i.e., ordered motion) or ability of work (e.g., the motion of a falling object); measured in joules.

exergy analysis Calculation of the total input and output of exergy of a product or service.

exergy destruction Exergy that vanishes in an irreversible process; also called irreversibility.

information Physical capacity of carrying a message; measured in binary units (bits).

irreversible Real processes are always irreversible (i.e., exergy is destroyed) in order to occur.

The exergy concept has its roots in the early work of what would later become thermodynamics. In 1824, N. L. Sadi Carnot stated that the extractable work of a heat engine is proportional to the temperature difference, which led to the definition of the second law of thermodynamics. Willard Gibbs introduced the notion of available work, including the diffusion term, in 1873. He simply stated the following: We will first observe that an expression of the form

$$-\varepsilon + T\eta - Pv + M_1m_1 + M_2m_2 + \dots + M_nm_n \quad (1)$$

denotes the work obtainable by the formation (by a reversible process) of a body of which ε , η , v , m_1 , $m_2, \dots, m_n$ are the energy, entropy, volume, and the quantities of the components, within a medium having the pressure P , the temperature T , and the potentials M_1 , M_2 , ..., M_n . (The medium is taken to be so large that its properties are not sensibly altered in any part by the formation of the body.)

Equation (1), originally referred to as Eq. (54) in Gibbs' work, is in exact correspondence with the present definition of exergy. Exergy is a thermodynamic potential; it is a general measure of "difference" or contrast. Thus, two states must be defined—the actual state and an equilibrium or a reference state, often defined by the environment. In 1953, the term exergy was first expressed by Z. Rant to denote "technical working capacity." In 1961, Tribus linked classical thermodynamics with information theory through the entropy concept. In 1977, a concise theory of exergy was offered and exergy was introduced as a useful concept not only for engineering but also for resource accounting in order to reduce environmental destruction. The term exergy is gaining in popularity and gradually replacing terms such as availability, available energy, essergy, utilizable energy, work potential, available work, and convertible energy.

1. THERMODYNAMICS AND EXERGY

Thermodynamics provides the concepts of temperature T , pressure P , heat Q , work W , internal energy U , entropy S , and four laws of thermodynamics. First, the zeroth law establishes the concept of temperature. Second, the first law defines energy as a conserved quantity (i.e., added heat ΔQ is equal to the increase of internal energy ΔU and the extracted work ΔW):

$$\Delta Q = \Delta U + \Delta W. \quad (2)$$

The extracted work is force times distance, or pressure times volume change (i.e., PdV). Pressure is an intensive variable (i.e., independent of the amount of matter). Volume is an extensive quantity that obviously depends on size or amount of matter:

$$dW = PdV, \quad (3)$$

TABLE I

Energy versus Exergy

Energy	Exergy
The first law of thermodynamics	The second law of thermodynamics
Nothing disappears.	Everything disperses.
Energy is motion or ability to produce motion.	Exergy is work (i.e., ordered motion) or ability to produce work.
$\Delta Q = \Delta U + \Delta W$ (2)	$E = T_0 \Delta S^{\text{tot}}$ (6)
where ΔQ is the total heat supplied to the system, ΔU is the total increase in the internal energy U of the system, and ΔW is the total increase in the external energy of the system or the total work done by the system.	where E is exergy, T_0 is the temperature of the environment, and ΔS^{tot} is the change of the entropy of the total system S^{tot} (i.e., the system and the environment).
$U = mc^2$ (7)	$E = k \ln 2T_0 I$ (8)
Energy U and matter m are “the same thing.”	Exergy and information I is “the same thing.”
Everything is energy.	Contrast is exergy.
Energy is always conserved (i.e., in balance); it can neither be produced nor consumed.	Exergy is only conserved in a reversible process but partly consumed in an irreversible process, (i.e., real processes). Thus, exergy is never in balance for real processes.
Energy is a measure of quantity.	Exergy is a measure of quantity and quality.

where V is the volume. Third, the second law defines the concept of entropy through heat:

$$dS \geq 0, dS_{\text{reversible}} = 0 \quad (4)$$

$$dQ = TdS, \quad (5)$$

analogous to Eq. (3). Finally, the third law defines the zero level for entropy.

Thermodynamics only treats reversible processes (i.e., nonreal processes) and systems in equilibrium states. Thus, dynamics is not treated and a more proper name is thermostatics. Even so, thermodynamics remains one of the most successful pieces of physics, particularly in engineering.

The first law (i.e., energy is conserved) generally fails to identify losses of work and potential improvements or the effective use of resources (e.g., in an adiabatic throttling process). The second law of thermodynamics shows that for some energy forms, only a part of the energy is convertible to work (i.e., the exergy). Thus, exergy is a very useful tool in engineering. Table I summarizes the main differences between energy and exergy.

1.1 Calculation of Exergy

The exergy of a system is expressed by the formula

$$E = U + P_0 V - T_0 S - \sum_i \mu_{i0} n_i, \quad (9)$$

where U , V , S , and n_i denote extensive parameters of the system (energy or internal energy, volume,

TABLE II

System A in a Much Larger Environment A_0

Intensive quantities	Extensive quantities
T_0 = temperature	U = internal energy
P_0 = pressure	V_0 = volume
μ_{i0} = chemical potential for substance i	S_0 = entropy
	n_{i0} = number of moles of substance i

entropy, and the number of moles of different chemical materials i), and P_0 , T_0 , and μ_{i0} are intensive parameters of the environment (pressure, temperature, and chemical potential). This expression is exactly what Gibbs stated in 1873 (Eq. 1). The chemical potential μ_{i0} may be generalized to also include exergy from electromagnetic forces, stress, nuclear forces, etc. This makes exergy the most general concept of work or ability of work. The situation is illustrated in Table II. System A is surrounded by the environment A_0 that is assumed to be much larger than the system (i.e., almost infinite in size). These two systems define two unique thermodynamic states—a difference or a contrast (Table II).

System A_0 is assumed to be unchangeable in order to initially also define the final state (i.e., the equilibrium state), when both systems have interacted in such a way that the contrast is exhausted. In

the real world, the situation is often like this. The ambient is usually so immense that it may be regarded as unchangeable. Each system is described by intensive and extensive quantities. The intensive quantities are actual measures of the contrast, such as the temperature T , the pressure P , and the chemical potential μ for every substance i . These quantities are not affected by the size of the system but only by the level of contrast. The extensive quantities, on the other hand, are directly proportional to the size of the system. However, it has been determined that human action (e.g., from combustion of fossil fuels) has a destructive impact on the environment in terms of changing its intensive quantities, such as temperature and composition, through the so-called greenhouse effect. The unpredictable dynamics of the environment makes this a dangerous activity.

The external energy P_0V , energy as “ambient heat” T_0S , and energy as “ambient substances” $\mu_{i0}n_i$ are the product of an intensive quantity of the environment and an extensive quantity of the system. This is the expression of the available work from system A when it is brought to equilibrium with the environment, system A_0 —that is, when all contrasts have been reversibly transformed into work. This expression is composed of four different terms. First, U is the energy carried within the system, which is also called internal energy. However, part of this energy is useless as work (i.e., exergy). The value of this useless part is handled by the last two terms. P_0V is the work as a volume V that occupies a space of the environment of pressure P_0 . This is pure work. Thus, by reducing the volume of the system A work can be extracted. These two terms are positive contributions to the work. The energy or internal energy unfortunately also carries two parts of useless energy or disorder, which must be deducted. T_0S is the part of the energy U that is useless due to lack of order, disorder, or entropy S (i.e., heat at ambient temperature T_0). Similarly, the last term is another useless part of the energy U ; in this case, as substances completely identical to substances in the environment (i.e., substances at ambient states).

Obviously, these two terms are not in contrast to the environment; thus, they cannot generate work. The energy that is carried by substances can only be used “down to” the level that is given by the environment (cf. with the available potential energy of a waterfall or the carrying capacity of a ship, which is the total capacity minus the ballast).

In summary, exergy expresses the maximum available work, in this case with reference to a specified environment (i.e., to what extent the contrast that is carried by the system can be used as work). Work and heat are concepts related to a process (i.e., that something is happening). They are not state variables. They are phenomena of time; that is, the existence of time is necessary in order to experience these phenomena. When it is said that work and heat are stored in a body, this means that the ability to create work and heat is stored. What is stored is exergy as some form of time-independent contrast.

By using the so-called Gibbs relation,

$$U = ST - VP + \sum_i n_i \mu_i, \quad (10)$$

the following expression is derived from Eq. (9):

$$E = S(T - T_0) - V(P - P_0) + \sum_i n_i(\mu_i - \mu_{i0}), \quad (11)$$

from which it is seen that E vanishes at equilibrium with the environment; that is, when $T = T_0$, $P = P_0$, and $\mu_i = \mu_{i0}$. This is the same as saying that with no contrast there is no exergy and no potential of action.

Analogously, the exergy of a flow can be written as

$$E = H - T_0S - \sum \mu_{i0}n_i, \quad (12)$$

where H is the enthalpy.

Exergy is the most general expression of work. Other, less general but more common expressions of thermodynamic potentials (e.g., Gibbs and Helmholtz free energies and enthalpy) are special cases of exergy as described in Table III. From Eq. (9) and the limitations specified in Table III, the equivalences

TABLE III
Relations between Differences in Exergy and in Other Thermodynamic Potentials

Limitations	$\Delta E = \Delta X$	Definition of X	Usually named
$\Delta n_i = 0, P = P_0, T = T_0$	$\Delta E = \Delta G$	$G = U + PV - TS$	Gibbs free energy
$\Delta n_i = 0, \Delta V = 0, T = T_0$	$\Delta E = \Delta F$	$F = U - TS$	Helmholtz free energy
$\Delta n_i = 0, \Delta S = 0, P = P_0$	$\Delta E = \Delta H$	$H = U + PV$	Enthalpy

between the difference in exergy and the difference in these other potentials can be derived.

1.2 Reference States

The major conceptual issue is the choice of reference state to which the chemical reactant under consideration is to be compared. Usually, the standard state of each element is defined as its pure elementary state, such as C for carbon, O₂ for oxygen, N₂ for nitrogen, and so on. The pure elements are normally assigned zero values in Gibbs free energy.

The standard exergies of chemical substances are often calculated on the assumption that the environment consists of a number of reference substances, one of each chemical element, with standard concentrations based on the average concentration in the natural environment. The reference substances selected are classified into three groups: gaseous compounds of the atmosphere, solid substances from the lithosphere, and both ionic and nonionic substances from the oceans. These groups are not in thermodynamic equilibrium, neither internally nor externally. Minor deviations of the actual environment from the standard environment only occasionally introduce significant errors to the exergy values.

It should also be noted that exergy may well be defined with reference to any other state (i.e., not necessarily a reference state). Thus, a locally defined environmental state is sometimes preferred. However, such a definition will imply a less general applicability of its use.

2. COMMON FORMS OF EXERGY

2.1 Work

Since exergy is defined as the maximum work potential, they are equivalent in every respect:

$$W = E - T_0 \Delta S. \quad (13)$$

It is stated in the second law that ΔS is always positive. Thus, from Eq. (13) it is obvious that exergy gives an upper limit to the work that is extractable from the process. The maximal amount of work is achieved when ΔS is minimal.

2.2 Pressure

Exergy is work, which is force F times distance S (FS) or pressure P times volume V (PV). This is illustrated by the movement of a piston in a cylinder in Fig. 1.

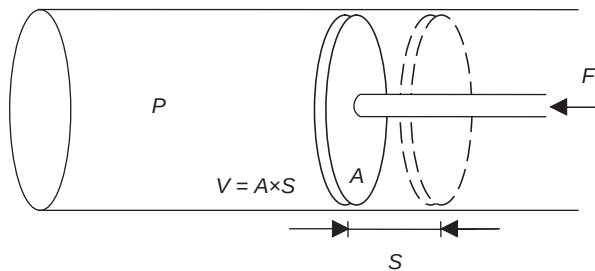


FIGURE 1 Pressure P is acting on a piston with area A that is moved the distance S by applying the force F .

2.3 Heat and Cold

An expression can be derived for the exergy of systems at different temperatures than that of the environment, both higher and lower. This will generate heat transfer when they are allowed to interact with each other or with the environment.

In Fig. 2, heat Q is transferred between two reservoirs at temperatures T and T_0 , and $T > T_0$. A reversible heat engine (i.e., a Carnot engine) working between the temperatures T and T_0 gives the maximal available work. By assuming that the heat is completely lost (i.e., maximum entropy production), the exergy being lost (i.e., the available exergy that is able to be utilized) can be calculated. The ambient temperature is T_0 (Fig. 2).

When $T > T_0$, the following familiar relation holds:

$$E = T_0 S^{\text{tot}} = T_0 \left(\frac{Q}{T_0} - \frac{Q}{T} \right) = Q \left(1 - \frac{T_0}{T} \right). \quad (14)$$

If the temperature $T < T_0$, this only changes the direction of the heat flow Q , which then goes from T_0 to T ; the exergy becomes

$$E = T_0 \left(\frac{Q}{T} - \frac{Q}{T_0} \right) = Q \left(\frac{T_0}{T} - 1 \right). \quad (15)$$

The ratio between the exergy E and the energy of the transferred heat Q is the exergy factor E/Q . Thus, exergy is obtained by multiplying the energy by the exergy factor. The exergy factor of heat from a heat reservoir (i.e., at a constant temperature T that differs from the ambient temperature T_0) is

$$\frac{E}{Q} = \left| \frac{T - T_0}{T} \right|. \quad (16)$$

This is a generalization of Eqs. (14) and (15) that holds at both $T > T_0$ and $T < T_0$. This expression is

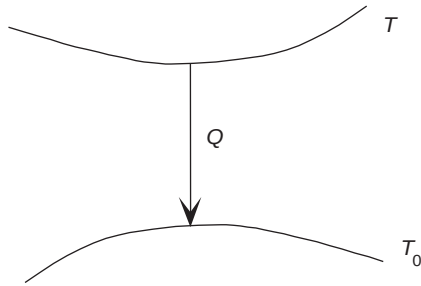


FIGURE 2 Heat Q is transferred between two reservoirs at temperatures T and T_0 .

valid for heat from phase changes, i.e.,

$$E = Q \left| \frac{T_p - T_0}{T_p} \right| = L \left| \frac{T_p - T_0}{T_p} \right|, \quad (17)$$

where T_p is the temperature during the phase change and L is latent heat, which depends on the substance and type of phase change.

Figure 3 shows how E/Q varies with temperature T according to Eq. (16). When $T < T_0$, energy is often regarded as negative, but the exergy is still positive. Figure 3 shows that E/Q strongly increases with low temperature. When T tends to absolute zero temperature (i.e., 0 K or -273.15°C), E/Q tends to infinity. This explains the difficulty in reaching absolute zero temperature. At high temperature, the E/Q comes closer to 1, which means that the exergy (E) and energy (Q) of the heat become almost the same. Higher temperatures (i.e., larger temperature differences) generate more work in a heat engine, which was stated by Carnot in 1824. This equation is useful to describe the exergy needed to maintain the temperature in hot spaces (e.g., furnaces, stoves, ovens, or houses) as well as in a cold storage (e.g., a refrigerator or a freezer) or to keep a pleasant indoor climate by air-conditioning on a hot summer day. Exergy reflects better than energy that heat or cold are more “expensive” when they are needed most. Thus, it is easier and less costly to generate cold in winter and heat in summer, and the storage of ice from winter to summer was once a profitable business. Also, waste heat to the environment carries a higher exergy together with a higher environmental impact in winter than in summer.

The exergy of a body with temperature T (i.e., exergy as sensible heat) is slightly more complicated to calculate since the temperature of the body as well as the delivered heat decrease when heat is transferred. Assume that the heat capacity of the body is C

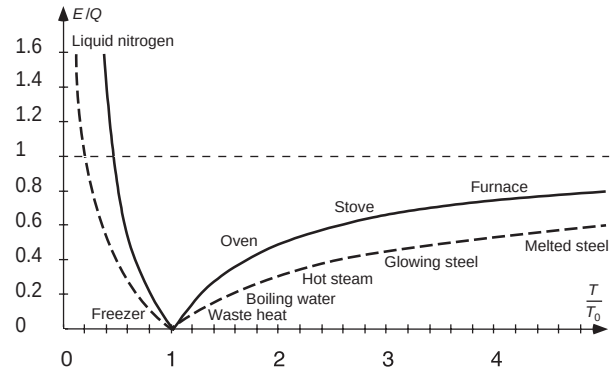


FIGURE 3 E/Q of heat as a function of its temperature; the solid line relates to Eq. (16), and the dashed line relates to Eq. (18).

(i.e., constant); the exergy becomes

$$E = C(T - T_0) \left(1 - \frac{T_0}{T - T_0} \ln \frac{T}{T_0} \right) \\ = C \left(T - T_0 - T_0 \ln \frac{T}{T_0} \right). \quad (18)$$

This relation is plotted in Fig. 3, and a number of examples are given, such as waste heat and melted steel. In many cases, such as for melted steel, the total exergy at different temperatures is a combination of latent heat due to phase changes and sensible heat (Eqs. 16 and 18, respectively). Assume that a power station produces 1000 MW of electricity and 2000 MW of waste heat as energy power at a temperature of 310 K by using cooling water at a local ambient temperature of $T_0 = 300$ K. From Eq. (18), the exergy of the waste heat in the cooling water is

$$E = Q \left(1 - \frac{T_0}{T - T_0} \ln \frac{T}{T_0} \right) \\ = 2000 \left(1 - \frac{300}{310 - 300} \ln \frac{310}{300} \right) \\ \approx 2000 \times 0.0165 = 33 \text{ MW}.$$

The discharge of 33 MW waste heat as slightly warmed water causes an increase in the local environmental temperature, which gradually changes the local ecosystem. Most likely, the more exergy that is discharged into the environment, the more damage will be done to the local environment. Thus, exergy is also a suitable environmental indicator.

2.4 Light

The exergy per unit area of light (i.e., black body radiation, which is radiating from a body of temperature T and received by a body of temperature

T_0) can be derived as

$$\dot{e} = \dot{u} \left[1 + \frac{1}{3} \left(\frac{T_0}{T} \right)^4 - \frac{4}{3} \frac{T_0}{T} \right], \quad (19)$$

where $\dot{u}$ is energy emission per unit area and $\dot{u} = \sigma T^4$, where $\sigma = 5.67 \times 10^8 \text{ W K}^{-4} \text{ m}^{-2}$; this is the Stefan-Boltzmann constant.

Sunlight supports life on Earth. The temperature of the surface of the sun is estimated to be approximately $T_{\text{sun}} = 6000 \text{ K}$ and that of the earth approximately $T_{\text{earth}} = T_0 = 300 \text{ K}$. The exergy factor of the sunlight on the earth (i.e., the ratio of the exergy flux to the total energy flux) is

$$\left(\frac{\dot{e}}{\dot{u}} \right)_{\text{sunlight}} = 1 + \frac{1}{3} \left(\frac{300}{6000} \right)^4 - \frac{4}{3} \times \frac{300}{6000} \approx 0.933.$$

Similarly, by considering the temperature of space that is assumed to be 3 K , the exergy factor of the earth's radiation to space is approximately 0.987 . Strangely, more exergy leaves the earth with reference to space than is received on the earth as sunlight. However, to our knowledge, this radiated exergy is not part of any life-support system. Transmission through the atmosphere changes the spectral distribution of solar radiation due to the selective absorption of radiation. It also varies with the composition of the atmosphere and the length of the path through the atmosphere, which is determined by the position of the sun. The total solar energy and exergy power to the earth are approximately $1.7 \times 10^{17} \text{ W}$ and $1.6 \times 10^{17} \text{ W}$, respectively; that is, the solar energy and exergy power at right angles to the inflow are approximately 1367 and 1275 W/m^2 , respectively, before entering the atmosphere.

If the ambient temperature on the earth increases due to the greenhouse effect (i.e., from the accumulation of gases such as carbon dioxide in the atmosphere), this implies a loss of solar exergy. For every higher degree in ambient earth temperatures the loss is approximately 1.5 times the total exergy use in society or approximately $2.5 \times 10^{13} \text{ W}$. This exergy will now, instead, add to the exergy of the heat radiation from the earth into space.

2.5 Material Substances

Equation (11) gives the following relation if the pressure P and the temperature T are constant (i.e., $P = P_0$ and $T = T_0$):

$$E = \sum_i n_i (\mu_i - \mu_{i,0}). \quad (20)$$

Furthermore, assume that the substance only departs from the environment in terms of chemical potential μ and concentration c . Also, assume that the chemical potential can be written as

$$\mu_i = \mu_i^0 + RT_0 \ln c_i, \quad (21)$$

where μ_i^0 is the chemical potential for the material i in relation to its standard state, obtained from chemical tables, usually the pure element in solid, liquid, or gas phase. Then the exergy becomes

$$E = \sum n_i (\mu_i^0 - \mu_{i0}^0) + RT_0 \sum n_i \ln \frac{c_i}{c_{i0}}, \quad (22)$$

where μ_{i0}^0 is the chemical potential for the material in the environment in relation to its standard state. Data on the chemical potential of different substances are found in the chemical literature. For only one material, the following relation holds:

$$E = n \left[\mu^0 - \mu_0^0 + RT_0 \ln \frac{c}{c_0} \right]. \quad (23)$$

Chemical reactions occurring in combustion processes often result in extensive exergy destruction. Evidently, no energy losses occur in the reaction. Only exergy calculations show that the production of entropy can cause a large percentage of potential work to be destroyed in the reaction. This indicates that an electrochemical membrane reactor, such as a fuel cell, could reduce exergy losses dramatically.

For pure reference components, which exist in the environment, exergy consists of the exergy that can be obtained by diffusing the components to their reference concentration. For gases and when the ideal gas law is applicable, this can be written as

$$E = nRT_0 \ln \frac{P_i}{P_{i0}}, \quad (24)$$

where P_i and P_{i0} are the partial pressures of the gas in the emission and in the environment, respectively.

Assume a power plant is fueled by pure carbon, C, with the exergy content of 410 kJ mol^{-1} . Thus, at complete combustion and with no exergy losses besides waste heat, approximately 7.3 kmol s^{-1} of CO_2 is produced as exhaust gases. From Eq. (24), this emission of CO_2 carries an exergy into the environment as waste of

$$E = nRT_0 \ln \frac{P_i}{P_{i0}} \approx 7300 \times 8.31 \times 300 \times \ln \frac{1}{300 \times 10^{-6}} \approx 148 \text{ MW},$$

where the concentration of CO_2 in the atmosphere is assumed to be 300 ppm . Thus, this is a considerable part of the exergy use that also reduces the efficiency

of the process. This emission will cause environmental effects such as global warming by influencing the heat radiation from the earth. When a substance does not exist in the reference environment, it often reacts to reference substances in order to get in equilibrium with the environment.

2.6 Nuclear Fuel

When the nuclear particles (i.e., nucleons) regroup into lower energy levels, exergy is released. In fission, this occurs by splitting heavy atoms, and in fusion it occurs by combining light atoms to form heavy atoms. Since a part of the transformed exergy in most cases appears as neutrinos, which can be regarded as invisible and noninteractive (i.e., not able to convert to work), nuclear exergy may be written as

$$E = \Delta mc^2 - E_{\text{neutrino}}, \quad (25)$$

where c is the speed of light in a vacuum (approximately $3 \times 10^8 \text{ m s}^{-1}$). The exergy of the neutrinos is often approximately 5% of the totally transformed exergy; thus, the exergy factor of nuclear energy is approximately 0.95.

2.7 Information

Exergy appears in daily life basically as three different phenomena: energy, matter, and information. Of these, energy and matter dominate in terms of amount of exergy, whereas information is exergy of a different magnitude, often several magnitudes lower than that of the energy. The exergy of a sheet of paper that carries a written message consists of the exergy of the paper and the exergy of the message (i.e., the information that it represents). The exergy of the paper (i.e., a material substance that may be combusted) amounts to approximately 50 kJ for an A4 sheet weighing 3 g (i.e., 17 MJ/kg). If this sheet carries a message of 5000 alphabetic characters (e.g., an English text), the exergy of this information is approximately $1.5 \times 10^{-20} \text{ kJ}$ (Eq. 8) since each character in reality only carries approximately 1 bit of information due to redundancy, which will be explained later. Exergy as material substances amounts to more than 10^{21} times the exergy as information on this sheet of paper. Thus, if it were possible to give an exact figure for the total exergy of this printed paper, the exergy of the information would not appear until in the 21st digit. Thus does not account for the exergy of the ink, which will also exceed that of the information by several magni-

tudes. This indicates that exergy of information should be treated differently than exergy of energy and matter. Therefore, special attention is given to the exergy of information.

Thermodynamics deals with the physical laws of systems containing many identical particles. From a macroscopic standpoint, the systems may be as small as 10^{-21} m^3 , but from a microscopic standpoint they are still large. 10^{-21} m^3 contains approximately 10^9 atoms. Thus, it is impossible to have a detailed knowledge of the motion of each individual particle. In statistical mechanics, which provides the theoretical basis for the understanding of thermodynamics, advantage is taken of the large number of particles and statistical methods are used to describe macroscopic properties as average properties resulting from the motion and interaction of many particles.

Information theory deals with systems of incomplete information. The general problem of how to take into account, in an unambiguous way, the available and incomplete information was successfully dealt with by Jaynes in 1957. As Jaynes showed, statistical mechanics is a special case of information theory. Conversely, powerful concepts and methods, developed in statistical mechanics, may be extended into other fields through information theory. In 1961, Tribus bridged the gap to thermodynamics in his classic book, *Thermostatics and Thermodynamics*. Thus, the theory of thermodynamics survived the breakthrough of quantum physics and was incorporated as a special branch of information theory through statistical mechanics. Reality will now be approached from this microscopic view (i.e., where everything is built up from discrete energy quanta). Thus, a system can only appear in a limited number of discrete states, and if energy is added to the system this will occur step by step with not less than one energy quantum at the time.

Assume that the total number of possible states of the system is Ω . The probability that the system is in the j th state is indicated by P_j and the sum of all probabilities is normalized to 1:

$$\sum_{j=1}^{\Omega} P_j = 1. \quad (26)$$

In statistical mechanics, the entropy S is defined as

$$S = -k \sum_{j=1}^{\Omega} P_j \ln P_j, \quad (27)$$

where k is the Boltzmann constant. This gives a completely new insight into the concept of entropy.

From classical thermodynamics, entropy is just a state variable with no physical interpretation, which for generations of students made entropy a mysterious concept. Now it may be regarded as disorder and related to a lack of information or knowledge.

If the system is known to be in a specific state, the probability for this state is equal to 1. Then, the probability for all other states must be equal to 0 since the sum of all probabilities is set at 1 according to Eq. (26). The knowledge of the system is complete and from Eq. (27) the entropy becomes 0: $S = 0$ [$0 \times \ln(0) = 0$ and $1 \times \ln(1) = 0$]. The state of the system is known. On the other hand, suppose nothing is known about the possible states in which the system is in—the knowledge is nil. All possible states are then equally probable. Since the number of possible states is Ω , then the probability of each state must be $1/\Omega$ in order to sum up to 1, according to Eq. (26). The entropy of the system now reaches its maximum value (i.e., $S = k \ln \Omega$). From our knowledge, the system is in complete disorder, like a library in which all the books are placed randomly. In 1896, Ludvig Boltzmann realized this expression of entropy for a thermodynamic system in equilibrium. Tragically, he never gained recognition for this fundamental relation during his lifetime.

The probabilities at equilibrium P_j^0 are those that maximize entropy S , under whatever constraints on the system that may be given

$$S_{\text{eq}} = S_{\text{max}} = -k \sum_{j=1}^{\Omega} P_j^0 \ln P_j^0. \quad (28)$$

The entropy difference between a given state and the equilibrium state of a system then becomes

$$S - S_{\text{eq}} = -k \left(\sum_{j=1}^{\Omega} P_j \ln P_j - \sum_{j=1}^{\Omega} P_j^0 \ln P_j^0 \right), \quad (29)$$

which then becomes a negative number since $S \leq S_{\text{eq}}$. Therefore, the concept of negentropy is sometimes defined as $S_{\text{eq}} - S$, which instead becomes positive and increases with order.

According to information theory, the information content or information capacity of a system is, in binary units (bits),

$$I = \frac{1}{\ln 2} \left(\sum_{j=1}^{\Omega} P_j \ln P_j - \sum_{j=1}^{\Omega} P_j^0 \ln P_j^0 \right). \quad (30)$$

From Eqs. (29) and (30), it is seen that entropy, or rather negentropy, is very closely linked to information:

$$S_{\text{eq}} - S = k'I, \quad (31)$$

where $k' = k \ln 2 \approx 1.0 \times 10^{-23}$ J/K. Thus, 1 bit of information is equivalent to 1.0×10^{-23} J/K of negentropy.

Assume a system of N different particles with two possible states each (e.g., 0 or 1). Then, $\Omega = 2^N$. If there are no other restrictions, then all P_j^0 must be 2^{-N} ; see Table IV, in which different particles are indicated by boxes in a row that can carry either 0 or 1.

Total information about the system (i.e., when one of P_j equals 1 and all the others equal 0) gives $I = N$, according to

$$\begin{aligned} I &= \frac{1}{\ln 2} \left[0 - \sum_{j=1}^{2^N} 2^{-N} \ln(2^{-N}) \right] \\ &= \frac{1}{\ln 2} \left[-2^N \frac{1}{2^N} (-N) \ln 2 \right] = N. \end{aligned}$$

For every particle, there is information corresponding to a “yes” or “no” to a specific question. Every such answer corresponds to 1 bit of information. Equation (6) (i.e., $E = T_0 \Delta S^{\text{tot}}$) gives

$$E = T_0 (S_{\text{eq}}^{\text{tot}} - S^{\text{tot}}), \quad (32)$$

where tot is the total system, which in this case is only the system. Equations (31) and (32) give the following relation between exergy and information:

$$E = k'T_0 I. \quad (33)$$

Thus, $k'T_0 \approx 2.9 \times 10^{-21}$ J is the amount of exergy connected to 1 bit of information at room temperature. It should be noted that information is a measure not necessarily of actual meaningful information but rather of information capacity. It can also be thought of as a measure of order or structure.

The net flow of information capacity to the earth from the sun may now be calculated from Eq. (33) and the fact that the net inflow of exergy as sunlight (i.e., after reduction for reflection) is approximately

TABLE IV

Number of Possible States and Probabilities of a System of N Different Particles with Two Possible States Each

N	Boxes	Possible states	$\Omega = 2^N$	P_j^0
1	□	0, 1	$2^1 = 2$	$\frac{1}{2}$
2	□□	00, 01, 10, 11	$2^2 = 4$	$\frac{1}{4}$
3	□□□	000, 001, 010, 100, 011, 101, 110, 111	$2^3 = 8$	$\frac{1}{8}$
etc.	...	...	...	...

1.1×10^{17} W. This gives an enormous amount of information, or

$$E = \frac{I}{k'T_0} = \frac{1.1 \times 10^{17}}{2.9 \times 10^{-21}} \approx 3.8 \times 10^{37} \text{ bits/s.}$$

Of this, man only uses a tiny fraction, approximately 5×10^{13} bits/s. Thus, the relation between used and available information capacity is approximately 10^{-24} . The same relation for energy conversion is approximately 10^{-5} . Therefore, most of the information flow from the sun (i.e., the exergy of sunlight) remains to be used.

3. EXERGY TOOLS

3.1 Exergy Losses

Energy is always in balance; however, for real processes exergy is never in balance. This unbalance is due to irreversibilities, which are usually called exergy destruction. The exergy output consists of the used output and the nonused output (i.e., exergy of waste output E_{waste}). It is very important to distinguish between exergy destruction caused by irreversibilities and exergy waste due to unused exergy (i.e., exergy flow to the environment). Both represent exergy losses, but irreversibilities have, by definition, no exergy and, thus, no environmental effects. However, due to high exergy losses from irreversibilities, more exergy is used to compensate for this loss, and this in turn may have an environmental impact (i.e., an indirect effect). In the literature, exergy destruction is commonly referred to as availability destruction, irreversibility, and lost work.

Exergy destruction is related to the entropy production by

$$E_{\text{in}}^{\text{tot}} - E_{\text{out}}^{\text{tot}} = T_0 \Delta S^{\text{tot}} = \sum_i (E_{\text{in}} - E_{\text{out}})_i, \quad (34)$$

where ΔS^{tot} is the total entropy increase, $E_{\text{in}}^{\text{tot}}$ is the total exergy input, $E_{\text{out}}^{\text{tot}}$ is the total exergy output, and $(E_{\text{in}} - E_{\text{out}})_i$ is the exergy destruction in process i .

By definition, an exergy balance only exists for reversible processes. Thus, for real processes [i.e., irreversible processes ($\Delta S^{\text{tot}} > 0$)], exergy is never in balance because the total exergy input always exceeds the total exergy output (i.e., $E_{\text{in}}^{\text{tot}} > E_{\text{out}}^{\text{tot}}$).

By calculating the exergy loss (i.e., destruction and waste), possible process improvements are visualized. In general, when the exergy loss is high, this part should be considered for improvement first.

However, this “tackle the biggest loss first” approach is not always appropriate because every part of the system depends on the others, so an improvement in one part may cause increased losses in other parts. As such, the total losses in the modified process may in fact be equal to or even larger than those in the original process configuration. Also, the use of renewable and nonrenewable resources must be considered. Therefore, the problem needs a more careful approach, which is described later.

3.2 Exergy Efficiencies

Exergy efficiency, which is also called second law efficiency, effectiveness, or rational efficiency, is usually defined as utilized exergy divided by used exergy. This must be a number between 0 and 1 since all real processes involve exergy destruction. This is in distinction to energy efficiency, which may well exceed 1. However, there are several ways to define the utilized exergy and used exergy, which add ambiguity to the application of this concept. Also worth noting is that exergy efficiency is often also defined as utilized exergy divided by the exergy that is theoretically possible to utilize due to, for example, time constraints. The faster a process operates the more exergy is lost due to increased irreversibility.

A simple definition of efficiency expresses all exergy input as used exergy and all exergy output as utilized exergy. So the exergy efficiency $\eta_{\text{ex},1}$ becomes

$$\eta_{\text{ex},1} = \frac{E_{\text{out}}}{E_{\text{in}}} = 1 - \frac{E_{\text{in}} - E_{\text{out}}}{E_{\text{in}}}, \quad (35)$$

where the definition of exergy destruction given previously is added. However, this efficiency does not always provide an adequate characterization of the thermodynamic efficiency of processes, such as heat transfer, separation, and expansion. Often, there exists a part of the output exergy that is unused (i.e., an exergy waste E_{waste} to the environment). Thus, the utilized exergy is given by $E_{\text{out}} - E_{\text{waste}}$, which is called the exergy product E_{pr} . The output consists of two parts:

$$E_{\text{out}} = E_{\text{pr}} + E_{\text{waste}}. \quad (36)$$

The exergy efficiency $\eta_{\text{ex},2}$ now instead becomes

$$\eta_{\text{ex},2} = \frac{E_{\text{out}} - E_{\text{waste}}}{E_{\text{in}}} = \frac{E_{\text{pr}}}{E_{\text{in}}} = \eta_{\text{ex},1} - \frac{E_{\text{waste}}}{E_{\text{in}}}. \quad (37)$$

Sometimes, part of the exergy going through the system is unaffected. This part has been called transit exergy E_{tr} (Fig. 4).

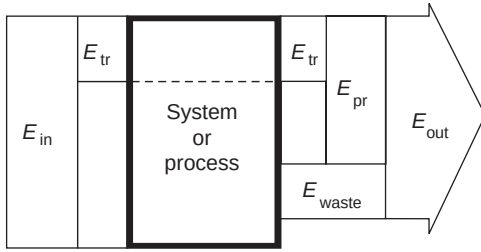


FIGURE 4 The input and output of exergies for a system.

If the transit exergy E_{tr} is deducted from both the input and the output exergy (or, rather, from the exergy product), the exergy efficiency $\eta_{ex,3}$ becomes

$$\eta_{ex,3} = \frac{E_{out} - E_{waste} - E_{tr}}{E_{in} - E_{tr}} = \frac{E_{pr} - E_{tr}}{E_{in} - E_{tr}}. \quad (38)$$

Due to the difficulties of calculating transit exergy and its lack of recognition, the exergy efficiency $\eta_{ex,2}$ is normally used:

$$\eta_{ex} = \eta_{ex,2} = \frac{E_{out} - E_{waste}}{E_{in}} = \frac{E_{pr}}{E_{in}}. \quad (39)$$

These latter definitions are compared by applying them to a system with two different processes A and B (Fig. 5). The exergy efficiencies for process A are $\eta_{ex,2} = 91\%$ and $\eta_{ex,3} = 10\%$, and those for process B are $\eta_{ex,2} = \eta_{ex,3} = 50\%$. Thus, determining which is the most efficient process is a matter of defining efficiency. However, from Fig. 5, it is seen that the exergy destruction of process A is larger than that of

process B (9 vs 5); thus, process A should really be improved first.

From this comparison it is seen that a better insight is offered by using exergy flow diagrams. From an exergy flow diagram, one can learn

- the exergy efficiencies of the various parts of a system
- the different exergy inputs and outputs
- where the various exergy flows come from and go to
- the amount of transit exergy
- how much exergy is destroyed in each processes

This often leads to a better insight and understanding of the system and its characteristics, which implies a better basis for improvement considerations.

3.3 Exergy Flow Diagrams

From the previous discussion, it is known that ambiguity reduces if an exergy flow diagram is used to demonstrate an exergy transfer instead of a ratio. In engineering, these diagrams are often used to describe the energy or exergy flows through a process (Fig. 6).

Figure 6 shows a typical power plant, its main components, and the energy and exergy flows of the plant. This diagram shows where the energy and exergy losses occur in the process and also whether exergy is destroyed from irreversibilities or whether it is emitted as waste to the environment. This information is very useful when considering the ecological effects of an activity. It also gives a hint

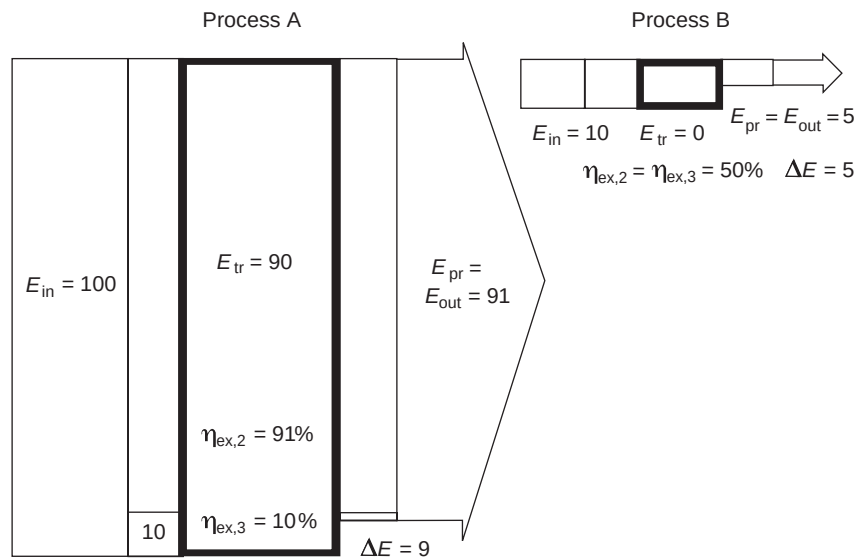


FIGURE 5 Comparison of efficiencies of exergy flows for two processes, A and B.

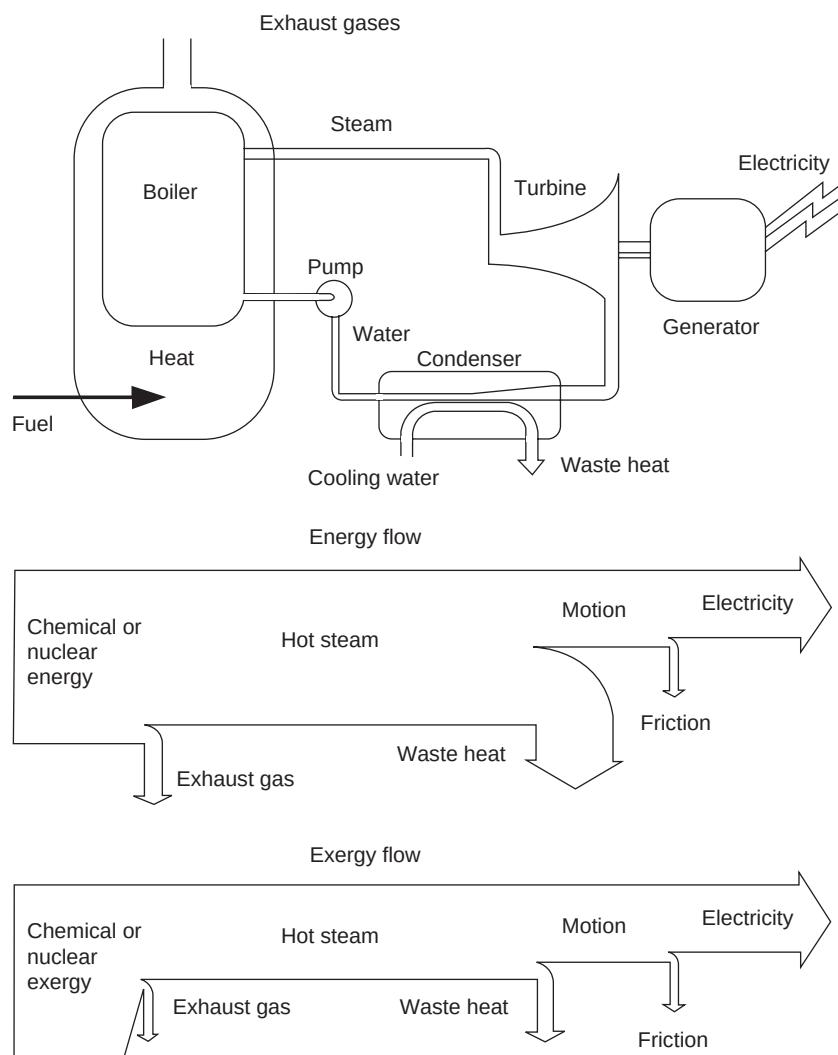


FIGURE 6 Energy and exergy flow diagrams of a condensing power plant.

about the possibilities of improving the process and where to direct efforts for improvement. In the energy flow diagram, energy is always conserved, as stated by the first law of thermodynamics. The waste heat carries the largest amount of energy into the environment, far more than is carried by the exhaust gases. However, in the exergy flow diagram the temperature of the waste heat is close to ambient so the exergy becomes much less. The exergy of the exhaust gas and that of the waste heat are comparable. Energy and exergy losses from friction have little direct impact on the environment. The ecological effects are more related to the exergy flows than to the energy flows, which makes exergy flows better ecological indicators.

The widths of the inflows and outflows in the energy and the exergy diagrams are almost equal

because both the inflows and the useful outflows are very high-quality energy forms, and for them energy and exergy are almost equal. For electricity, the energy and the exergy values are identical since electricity does not carry any entropy. The losses in the two diagrams, however, are quite different. First, there are losses in the furnace, where fuel is converted into heat. In an oil- or coal-fueled power plant the flame temperature is approximately 2000°C , but in a nuclear power plant the temperature is only approximately 300°C due to safety regulations. The heat is then transmitted through heat exchangers to boil water (i.e., produce steam in what is often called a steam generator). The pressure is high to make the water boil at high temperatures. Through the walls of the furnace, and through pipes, heat is lost to the environment. Heat is also lost with the exhaust gases,

usually in order to keep the chimneys free from condensed water. These losses represent only a minor part of the total energy flow. However, from the exergy diagram, we see that more than one-third of the exergy is lost at this stage. We also see that exergy disappears and only a small fraction leaves the plant. The exergy flow literally shrinks. Large quantities of entropy are created because the steam that leaves the boiler has a relatively low temperature and pressure due to limitations in the fatigue strength of the components in the process, principally the boiler and the turbine blades. In a nuclear power plant, more than half the converted exergy is lost during this stage of the process.

The energy and exergy of hot steam in Fig. 6 differ a great deal. In the energy diagram, major losses occur in the condenser as waste heat in the cooling water. The waste heat in this case, however, is heat at a temperature slightly above the ambient temperature and its exergetic value is therefore very low. This is obvious from the exergy diagram, in which the waste heat is a minor flow. Thus, for a conversion of waste heat into work by an ideal process (i.e., a Carnot engine), only the amount indicated in the

exergy diagram would become work, and the loss would be waste heat at ambient temperature (i.e., exergy equals zero). It is not possible to produce any more work or electrical energy from the steam, or from any other energy form, than what is determined by its exergy content.

When converting mechanical energy into electrical energy, both of which have an exergy factor of 1, small losses arise from friction. These losses are on the whole equally small in both diagrams. Part of the friction losses is due to mechanical fatigue (i.e., wearing out of shafts and bearings).

Figure 7 illustrates the energy and exergy flows of an oil furnace, an electric heater, an electric heat pump, and a combined power and heat plant (i.e., a cogeneration plant). In the oil furnace, the energy efficiency is assumed to be approximately 85%, with losses due mainly to the hot exhaust gases (Fig. 7). The exergy efficiency is very low (approximately 4%) because the temperature difference is not utilized when the temperature is decreased to a low of approximately 20°C, a comfortable indoor climate.

Electric heating by short-circuiting in electric resistors has an energy efficiency of 100% by

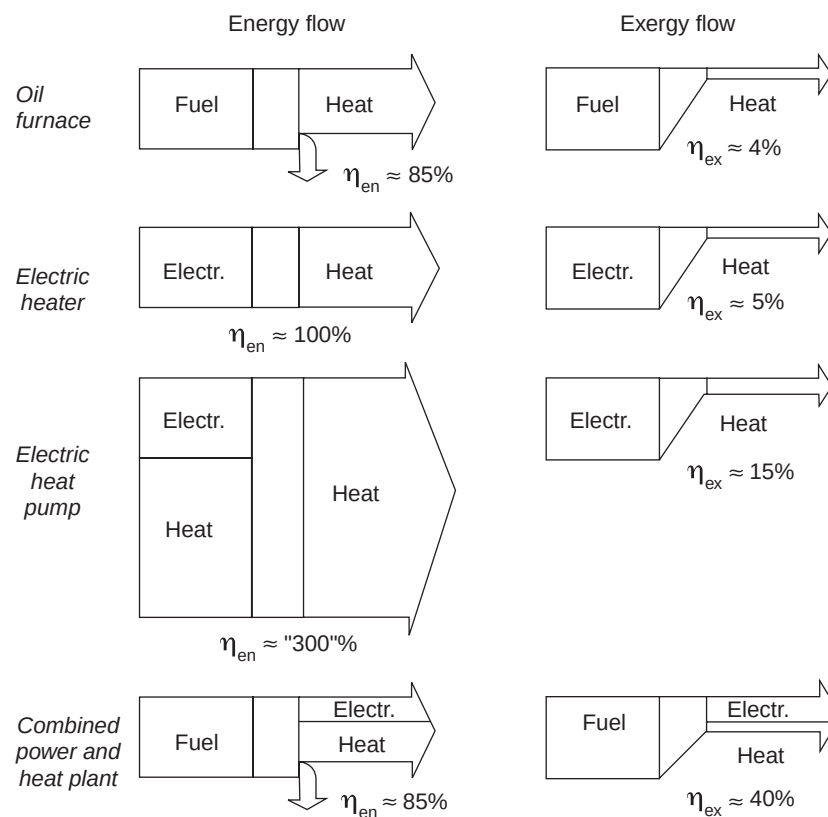


FIGURE 7 Energy and exergy flows through some typical energy systems.

definition of energy conservation. The energy efficiency of an electric heat pump is not limited to 100%. If the heat originating from the environment is ignored in the calculation of the efficiency, the conversion of electrical energy into indoor heat can be well over 100% (e.g., 300% as in Fig. 7). The exergy flow diagram of the heat pump looks quite different. The exergy efficiency for an electric heater is approximately 5% and for the heat pump 15%.

In Fig. 6, the energy and exergy efficiencies are the same because the inflow of fuels and the outflow of electricity both have an exergy factor of approximately 1. For a combined power and heat plant (Fig. 7), the exergy efficiency is about the same as that for a condensing power plant (Fig. 6). The main exergy loss occurs in the conversion of fuel into heat in the boiler. Since this conversion is practically the same in both the condensing and the combined power plants, the total exergy efficiency is the same

(i.e., approximately 40%). The maximum energy efficiency of an ideal conversion process may be more than 100%, depending on the definition of efficiency. The exergy efficiency, however, can never exceed 100%.

Exergy flow diagrams may also include the total use of physical resources (i.e., energy and other resources for a system). Figure 8 shows the main conversions of energy and materials in Swedish society in 1994. The flows of resources go from left to right in the diagram (i.e., from the resource base to the consumption sector). Thus, the diagram basically represents the resource supply sector in which resources, such as crops and minerals, are turned into consumer goods, such as food, transport, and thermal comfort. The width of the flows is defined by their exergy content and the unit of the flows is Joules. The accuracy of the flows varies a great deal between the different areas. For the electricity system, the accuracy is quite high, whereas for

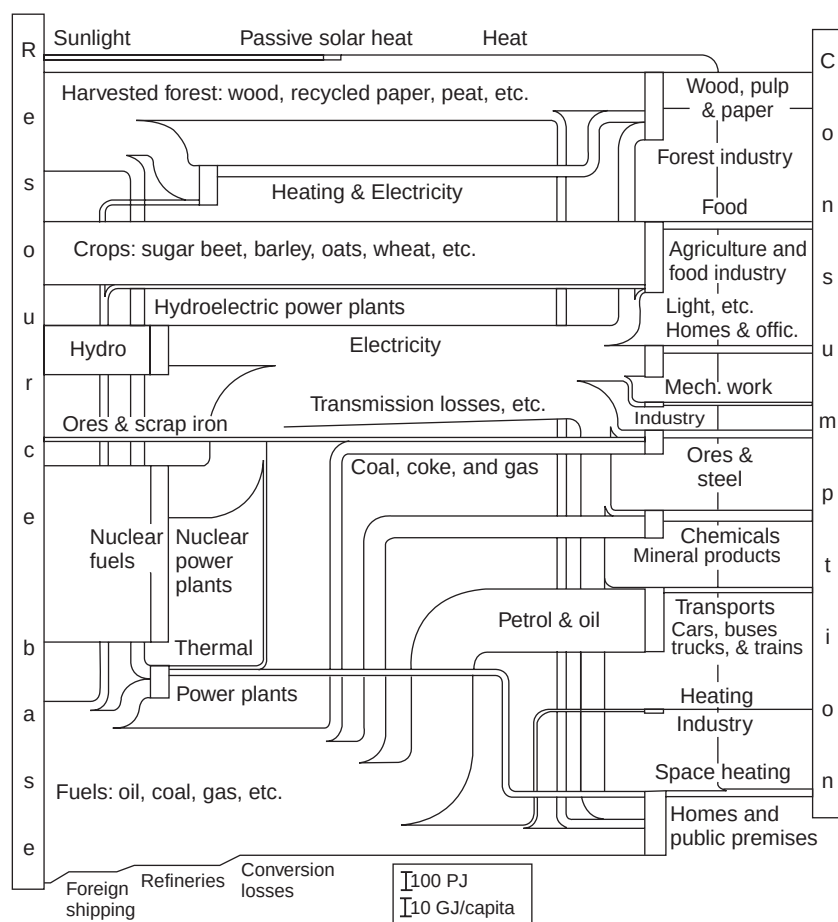


FIGURE 8 The exergy conversion system in the Swedish society in 1994. Total input was approximately 2720 PJ or 310 GJ per capita, and total output was approximately 380 PJ or 43 GJ per capita.

sectors related to agriculture and forestry it is not. In order to not make the diagram too complicated, only exergy flows exceeding 5 PJ are included. The inflows are ordered according to their origins. Sunlight is thus a renewable natural flow. Besides a minor use of wind power, far less than 5 PJ, this is the only direct use of a renewable natural flow. Harvested forests, agricultural crops, and hydropower are renewable exergy flows derived from funds, which of course are founded on the renewable natural flow of sunlight. Iron ore, nuclear fuels, and fossil fuels are non-renewable exergy flows from deposits, which are exhaustible and also carry toxic substances. The unfilled boxes represent exergy conversions, which in most cases represent a huge number of internal conversions and processes. The resources actually demanded by society appear as outflows on the right side of the diagram. The total inflow of resources during 1994 was approximately 2720 PJ or 310 GJ per capita, and the net output was 380 PJ or 40 GJ per capita. Thus, the overall efficiency of the supply sector can be estimated at less than 15%, which must be regarded as poor. As can be seen, some sectors are extremely inefficient. Some resource conversion systems have a ridiculously poor efficiency. For nuclear fuel to space heating through short-circuit heaters the utilization is less than 0.025% if the physically available nuclear exergy is considered.

Exergy is a concept to be used by society at large, e.g., in town and country planning. In addition to

having an important pedagogical value, it bridges gaps between different areas of science (e.g., thermodynamics, information theory, economics, and ecology). Thus, exergy will play a central role in disciplines that are dedicated to sustainable development.

SEE ALSO THE FOLLOWING ARTICLES

Entropy • *Entropy and the Economic Process* • *Exergoeconomics* • *Exergy Analysis of Energy Systems* • *Exergy Analysis of Waste Emissions* • *Exergy: Reference States and Balance Conditions* • *Thermodynamics and Economics, Overview* • *Thermodynamics, Laws of* • *Thermodynamic Sciences, History of*

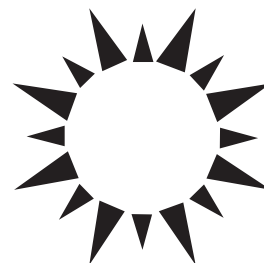
Further Reading

- Encyclopedia of Life Support Systems*, <http://www.eolss.com> and <http://www.eolss.net>.
Exergy, an International Journal, <http://www.exergyonline.com>.
 Tribus, M. (1961). *Thermostatics and Thermodynamics*. Van Nostrand, New York.
 Wall, G. (1986). Exergy—A useful concept, Ph.D. thesis. Chalmers University of Technology, Sweden; <http://www.exergy.se/goran/thesis/index.html>.
 Wall, G. (1997). *Exergetics*. <http://www.exergy.se/ftp/exergetics.pdf>.

Exergy Analysis of Energy Systems

MARC A. ROSEN

University of Ontario Institute of Technology
Oshawa, Ontario, Canada



-
1. Exergy
 2. Exergy Analysis
 3. Survey of Applications of Exergy Analysis to Energy Systems
 4. Examples of Applications of Exergy Analysis to Energy Systems
 5. Simple Device: Electrical Resistance Space Heater
 6. Simple System: Thermal Energy Storage System
 7. Complex System: Coal-Fired Electrical Generating Station
 8. Macrosystem: The Aggregate Energy Systems of a Country
 9. Conclusions

Glossary

chemical exergy The maximum work obtainable from a substance when it is brought from the environmental state to the dead state by means of processes involving interaction only with the environment.

dead state The state of a system when it is in thermal, mechanical, and chemical equilibrium with a conceptual reference environment, which is characterized by a fixed pressure, temperature, and chemical potential for each of the reference substances in their respective dead states.

environmental state The state of a system when it is in thermal and mechanical equilibrium with the reference environment—that is, at the pressure and temperature of the reference environment.

exergy (i) A general term for the maximum work potential of a system, stream of matter, or a heat interaction in relation to the reference environment as the datum state. Also known as available energy, availability, essergy, technical work capacity, usable energy, utilizable energy, work capability, work potential, and xergy. (ii) The unqualified term exergy or exergy flow is the maximum amount of shaft work obtainable when a

steady stream of matter is brought from its initial state to the dead state by means of processes involving interactions only with the reference environment.

exergy analysis An analysis technique, based on the second law of thermodynamics, in which process performance is assessed by examining exergy balances, losses, and efficiencies.

exergy consumption The exergy consumed or destroyed during a process due to irreversibilities within the system boundaries. Also known as dissipation, irreversibility, and lost work.

physical exergy The maximum amount of shaft work obtainable from a substance when it is brought from its initial state to the environmental state by means of physical processes involving interaction only with the environment. Also known as thermomechanical exergy.

reference environment An idealization of the natural environment that is characterized by a perfect state of equilibrium—that is, the absence of any gradients or differences involving pressure, temperature, chemical potential, kinetic energy, and potential energy. The environment constitutes a natural reference medium with respect to which the exergy of different systems is evaluated.

reference state A state with respect to which values of exergy are evaluated. Several reference states are used, including the environmental state, dead state, standard environmental state, and standard dead state. Also known as ground state.

thermal exergy The exergy associated with a heat interaction—that is, the maximum amount of shaft work obtainable from a given heat interaction using the environment as a thermal energy reservoir.

Energy analysis is based on the first law of thermodynamics, which embodies the principle of the conservation of energy. Exergy analysis is a thermodynamic analysis technique for systems and processes that is based on the second law of thermodynamics and has several advantages over energy analysis.

First, more meaningful efficiencies are evaluated with exergy analysis since exergy efficiencies are always a measure of the approach to the ideal. Second, the inefficiencies in a process are better pinpointed with exergy analysis in that the types, causes, and locations of the losses are identified and quantified. For these reasons and others, exergy analysis has been increasingly applied during the past several decades. In this article, the role of exergy analysis in the assessment and improvement of energy systems is discussed. First, exergy and its use as a tool through exergy analysis are briefly described. Then, the range of energy systems that have in the past been assessed using exergy analysis are surveyed. Finally, several examples of applications of exergy analysis are presented, ranging from simple energy devices and systems to complex energy systems and large macrosystems.

1. EXERGY

Exergy is defined as the maximum amount of work that can be produced by a stream or system as it is brought into equilibrium with a reference environment, and it can be thought of as a measure of the usefulness or quality of energy. Exergy is consumed during real processes due to irreversibilities and conserved during ideal processes.

Exergy quantities are evaluated with respect to a reference environment. The intensive properties of the reference environment in part determine the exergy of a stream or system. The exergy of the reference environment is zero. The reference environment is in stable equilibrium, with all parts at rest relative to one another. No chemical reactions can occur between the environmental components. The reference environment acts as an infinite system and is a sink and source for heat and materials. It experiences only internally reversible processes in which its intensive state remains unaltered (i.e., its temperature T_o , pressure P_o , and the chemical potentials μ_{ioo} for each of the i components present remain constant).

Energy and exergy balances can be written for a general process or system. An energy balance for a system may be written as

$$\text{Energy input} - \text{energy output} = \text{energy accumulation.} \quad (1)$$

Energy input and output refer respectively to energy entering and exiting through system boundaries.

Energy accumulation refers to buildup (either positive or negative) of the quantity within the system. Clearly, energy is subject to a conservation law.

By combining the conservation law for energy and nonconservation law for entropy (which states that entropy is created during a process due to irreversibilities), an exergy balance can be obtained:

$$\text{Exergy input} - \text{exergy output} - \text{exergy consumption} = \text{exergy accumulation.} \quad (2)$$

Exergy is consumed due to irreversibilities. Exergy consumption is proportional to entropy creation. Equations (1) and (2) demonstrate an important difference between energy and exergy: Energy is conserved, whereas exergy, a measure of energy quality or work potential, can be consumed.

2. EXERGY ANALYSIS

Exergy analysis involves the application of exergy concepts, balances, and efficiencies to evaluate and improve energy and other systems. Many engineers and scientists suggest that devices can be well evaluated and improved using exergy analysis in addition to or in place of energy analysis. Exergy analysis has been applied to a wide range of processes and systems. This article focuses on applications to energy systems.

2.1 Exergy Analysis and Limitations on Increased Efficiency

Increases in efficiency are subject to two constraints, which are often poorly understood: (i) theoretical limitations, which establish the maximum efficiency theoretically attainable for a process by virtue of the laws of thermodynamics, and (ii) practical limitations, which further limit increases in efficiency.

First, consider practical limitations on efficiency. In practice, when selecting energy sources and utilization processes the goal is not to achieve maximum efficiency but rather to achieve an optimal trade-off between efficiency and such factors as economics, sustainability, environmental impact, safety, and societal and political acceptability. This optimum is dependent on many factors controllable by society; furthermore, these factors can be altered to favor increased efficiency (e.g., governments can offer financial incentives that render high-efficiency technologies economically attractive and/or provide disincentives for low-efficiency alternatives through special taxes and regulations).

Now, consider theoretical limitations on efficiency. To assess the potential for increased efficiency, the theoretical limits must be clearly understood. In the past, lack of clarity on this issue has often led to confusion, in part because energy efficiencies generally are not measures of how close the performance of a process or device approaches the theoretical ideal. The consequences of such confusion can be very significant. For example, extensive resources have occasionally been directed toward increasing the energy efficiencies of devices that in reality were efficient and had little potential for improvement. Conversely, at other times devices have not been targeted for improved efficiency, even though the difference between the actual and maximum theoretical efficiencies, which represents the potential for improvement, has been large.

The difficulties inherent in energy analysis are also attributable to the fact that it considers only quantities of energy and ignores energy quality, which is continually degraded during real processes. Exergy analysis, as discussed earlier, overcomes many of the problems associated with energy analysis.

2.2 Applications of Exergy Analysis beyond Thermodynamics

Exergy concepts can be applied beyond thermodynamics. One area in which applications of exergy are increasing is that of environmental impact. Many suggest that the impact of energy resource utilization on the environment and the achievement of increased resource-utilization efficiency should be addressed by considering exergy, among other factors. Although the exergy of an energy form or a substance is a measure of its usefulness, exergy is also a measure of its potential to cause change. The latter point suggests that exergy may be, or provide the basis for, an effective measure of the potential of a substance or energy form to impact the environment.

Another area in which applications of exergy are increasing is that of economics. In the analysis and design of energy systems, techniques are often used that combine scientific disciplines (especially thermodynamics) with economics to achieve optimum designs. For energy systems, costs are conventionally based on energy. Many researchers, however, have recommended that costs are better distributed among outputs, based on exergy. Methods of performing exergy-based economic analyses have evolved and are referred to as thermoeconomics, second law costing, and exergoeconomics. These analysis tech-

niques recognize that exergy, more than energy, is the commodity of value in a system, and they assign costs and/or prices to exergy-related variables. These techniques usually help determine the appropriate allocation of economic resources so as to optimize the design and operation of a system and/or the economic feasibility and profitability of a system (by obtaining actual costs of products and their appropriate prices).

2.3 Steps for Energy and Exergy Analyses

A simple procedure for performing energy and exergy analyses involves the following steps:

- Subdivide the process under consideration into as many sections as desired, depending on the depth of detail and understanding desired from the analysis.
- Perform conventional mass and energy balances on the process, and determine all basic quantities (e.g., work and heat) and properties (e.g., temperature and pressure).
- Based on the nature of the process, the acceptable degree of analysis complexity and accuracy, and the questions for which answers are sought, select a reference environment model.
- Evaluate energy and exergy values, relative to the selected reference environment model.
- Perform exergy balances, including the determination of exergy consumptions.
- Select efficiency definitions, depending on the measures of merit desired, and evaluate the efficiencies.
- Interpret the results, and draw appropriate conclusions and recommendations relating to such issues as design changes, retrofit plant modifications, etc.

3. SURVEY OF APPLICATIONS OF EXERGY ANALYSIS TO ENERGY SYSTEMS

Many applications of exergy analysis to energy systems have been reported in the literature, particularly during the past several decades. Some extensive bibliographies have been compiled, including one by Goran Wall (<http://exergy.se>). In addition, a journal devoted to exergy matters, *Exergy: An International Journal*, has been established.

The types of applications of exergy methods to energy systems that have been reported in the past are extremely varied and include the following:

- Electricity generation using both conventional devices, such as fossil and nuclear power plants, and alternative devices, such as fuel cells and solar energy systems
- Engines of various types
- Transportation systems for land, air, and water transport
- Heating and cooling systems for building systems and industrial applications
- Cogeneration systems for producing heating and electrical needs simultaneously
- Petrochemical processing and synthetic fuels production
- Chemical processes, such as sulfuric acid production, distillation, and water desalination
- Metallurgical processes, such as lead smelting
- Combustion technologies and systems
- Energy storage systems, such as batteries, pumped storages, and thermal energy storages

4. EXAMPLES OF APPLICATIONS OF EXERGY ANALYSIS TO ENERGY SYSTEMS

Examples of applications of exergy analysis are presented in the following four sections. Because the range of energy systems to which exergy has been applied is great, the four examples described here were selected to cover the breadth of complexity. Consequently, we start with a simple device (an electrical resistance space heater) and then consider a simple system (a thermal energy storage system), where simple is taken here to imply that the system has a very limited number of components. Next, we examine a more complex system (a coal-fired electrical generating station). Finally, we investigate the application of exergy analysis to a “macrosystem”—that is, a large conglomerate of energy systems that cover a region such as a city, province, or country.

5. SIMPLE DEVICE: ELECTRICAL RESISTANCE SPACE HEATER

An electrical resistance space heater converts electricity to heat at a temperature suitable for keeping a room at a comfortable temperature. It is illustrated in Fig. 1.

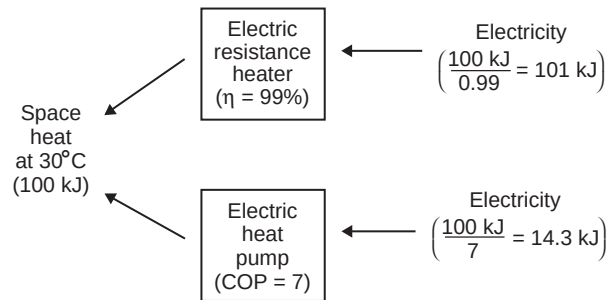


FIGURE 1 Comparison of the quantity of electricity required to provide a fixed amount of space heat (100 kJ) using two different heating devices: an electric resistance heater and an electric heat pump. η , energy efficiency; COP, coefficient of performance.

The energy efficiency of electric resistance space heating is often said to exceed 99%. The implication is that the maximum possible energy efficiency for electric resistance heating is 100%, corresponding to the most efficient device possible.

This understanding is erroneous, however, because energy analysis ignores the fact that in this process high-quality energy (electricity) is used to produce a relatively low-quality product (warm air). Exergy analysis recognizes this difference in energy qualities and indicates that the exergy of the heat delivered to the room is approximately 5% of the exergy entering the heater. Thus, based on exergy, the efficiency of electric resistance space heating is approximately 5%.

We therefore obtain useful information from the exergy results. Since thermodynamically ideal space heating has an exergy efficiency of 100%, the same space heating can theoretically be achieved using as little as 5% of the electricity used in conventional electric resistance space heating. In practical terms, one can achieve space heating with a greatly reduced electricity input using a high-efficiency electric heat pump (Fig. 1), using 15% of the electricity that electric resistance heating would require for a heat pump with a coefficient of performance of 7.

6. SIMPLE SYSTEM: THERMAL ENERGY STORAGE SYSTEM

A thermal energy storage (TES) system receives thermal energy and holds the energy until it is required. TESs can store energy at temperatures above or below the environment temperature, and there are many types of systems (e.g., tanks, aquifers, ponds, and caverns).

The evaluation of a TES system requires a measure of performance that is rational, meaningful, and practical. The conventional energy storage efficiency is an inadequate measure. A more perceptive basis for comparison is needed if the true usefulness of thermal storages is to be assessed and thus permit maximization of their economic benefit. Efficiencies based on ratios of exergy provide rational measures of performance since they can measure the approach of the performance of a system to the ideal.

That the energy efficiency is an inappropriate measure of thermal storage performance can best be appreciated by a simple example. Consider a perfectly insulated thermal storage containing 1000 kg of water, initially at 40°C. The ambient temperature is 20°C, and the specific heat of water is taken to be constant at 4.2 kJ/kg K. A quantity of 4200 kJ of heat is transferred to the storage through a heat exchanger from an external body of 100 kg of water cooling from 100 to 90°C. This heat addition raises the storage temperature 1.0°C to 41°C. After a period of storage, 4200 kJ of heat is recovered from the storage through a heat exchanger that delivers it to an external body of 100 kg of water, raising the temperature of that water from 20 to 30°C. The storage is returned to its initial state at 40°C.

For this storage cycle the energy efficiency, which is the ratio of the heat recovered from the storage to the heat injected, is $4200/4200 \text{ kJ} = 1$, or 100%. However, the recovered heat is at only 30°C and of little use, having been degraded even though the storage energy efficiency was 100%. The exergy recovered in this example is evaluated as 70 kJ and the exergy supplied as 856 kJ. Thus, the exergy efficiency, which is the ratio of the thermal exergy recovered from storage to that injected, is $70/856 = 0.082$, or 8.2%—a much more meaningful expression of the achieved performance of the TES. Consequently, a device that appears to be ideal on an energy basis is correctly shown to be far from ideal on an exergy basis, clearly demonstrating the benefits of using exergy analysis for evaluating TESs.

7. COMPLEX SYSTEM: COAL-FIRED ELECTRICAL GENERATING STATION

Energy and exergy analyses are applied to a coal-fired electrical generating station. The coal-fired Nanticoke Generating Station is selected as the representative station for the analysis. The station

has a net unit electrical output of approximately 500 MWe and is operated by the provincial electrical utility, Ontario Power Generation (formerly Ontario Hydro), in Ontario, Canada. This example clearly illustrates how exergy analysis allows process inefficiencies to be better determined than does an energy analysis and efficiencies to be more rationally evaluated. Consequently, the example identifies areas in which the potential for performance improvement is high and trends that may aid in the design of future stations.

7.1 Process Description

A detailed flow diagram for a single unit of the station is shown in Fig. 2. The symbols identifying the streams are described in Table I, in which data are provided for the streams. The diagram in Fig. 2 is divided into four main sections:

- *Steam generation:* Heat is produced and used to generate and reheat steam. In the station, eight pulverized coal-fired natural circulation steam generators each produce 453.6 kg/s steam at 16.89 MPa and 538°C and 411.3 kg/s of reheat steam at 4.00 MPa and 538°C. Air is supplied to the furnace by two 1080-kW, 600-rpm motor-driven forced-draft fans. Regenerative air preheaters are used. The flue gas passes through an electrostatic precipitator rated at 99.5% collection efficiency and exits the plant through two multiflued, 198-m-high chimneys.

- *Power production:* The steam produced in the steam generation section is passed through a series of turbine generators that are attached to a transformer. Extraction steam from several points on the turbines preheats feedwater in several low- and high-pressure heat exchangers and one spray-type open deaerating heat exchanger. The low-pressure turbines exhaust to the condenser at 5 kPa. Each station unit has a 3600-rpm, tandem-compound, impulse-reaction turbine generator containing one single-flow high-pressure cylinder, one double-flow intermediate-pressure cylinder, and two double-flow low-pressure cylinders. Steam exhausted from the high-pressure cylinder is reheated in the combustor.

- *Condensation:* Cooling water from Lake Erie condenses the steam exhausted from the turbines. The flow rate of cooling water is adjusted so that a specified temperature increase in the cooling water is achieved across the condenser.

- *Preheating:* The temperature and pressure of the feedwater are increased in a series of pumps and feedwater-heater heat exchangers.

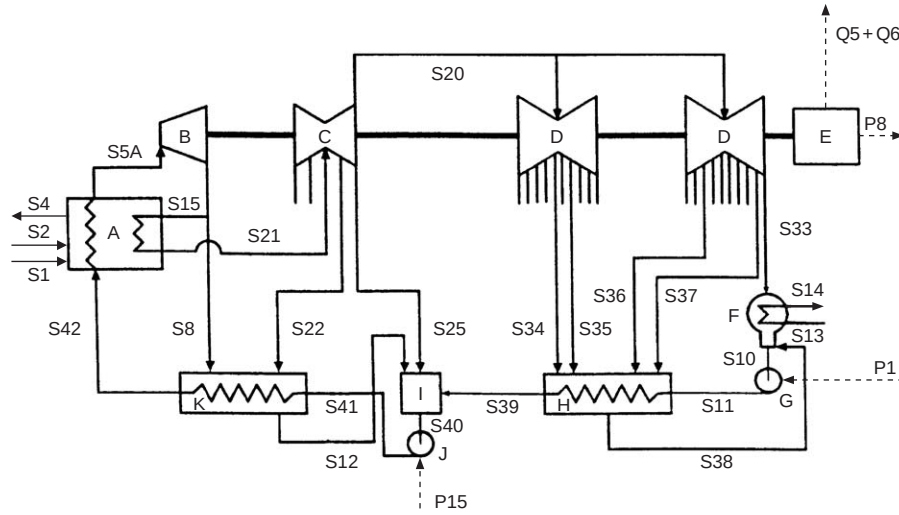


FIGURE 2 A unit of the coal-fired electrical generating station. Lines exiting the turbines represent extraction steam. The station has four main sections: Steam generation (A), power production (B–E), condensation (F), and preheating (G–K). The external inputs for device A are coal and air, and the output is stack gas and solid waste. The external outputs for device E are electricity and waste heat. Electricity is input to devices G and J, and cooling water enters and exits device F. A, steam generator and reheater; B, high-pressure turbine; C, intermediate-pressure turbine; D, low-pressure turbines; E, generator and transformer; F, condenser; G, hot well pump; H, low-pressure heat exchangers; I, open deaerating heat exchanger; J, boiler feed pump; K, high-pressure heat exchangers.

7.2 Analysis Assumptions and Simplifications

The reference environment model used here has a temperature of 15°C (the approximate mean temperature of the lake cooling water), a pressure of 1 atm, a chemical composition consisting of air saturated with water vapor, and the following condensed phases at 15°C and 1 atm: water (H₂O), gypsum (CaSO₄·2H₂O), and limestone (CaCO₃). For simplicity, heat losses from external surfaces are assumed to occur at the reference environment temperature of 15°C.

7.3 Results and Discussion

Energy and exergy values for the streams identified in Fig. 2 are summarized in Table I. Exergy consumption values for the devices are listed, according to process diagram sections, in Table II. Figures 3 and 4 illustrate the net energy and exergy flows and exergy consumptions for the four main process diagram sections described in the legend to Fig. 2.

Overall energy and exergy efficiencies are evaluated as

$$\text{Energy efficiency} = \frac{(\text{net energy output with electricity})}{(\text{energy input})} \quad (3)$$

and

$$\text{Exergy efficiency} = \frac{(\text{net exergy output with electricity})}{(\text{exergy input})} \quad (4)$$

Coal is the only input source of energy or exergy for the station, and the energy and exergy efficiencies are 37 and 36%, respectively. The small difference in the efficiencies is due to the fact that the specific chemical exergy of coal is slightly greater than its energy. Although the energy and exergy efficiencies for the station are similar, these efficiencies differ markedly for many station sections.

In the steam generation section, exergy consumptions are substantial, accounting for 659 MW (72%) of the 916 MW total exergy losses for the station. Of the 659 MW of exergy consumed in this section, 444 MW is due to combustion and 215 MW to heat transfer. The energy and exergy efficiencies for the steam generation section, considering the increase in energy or exergy of the water as the product, are 95 and 49%, respectively. The steam generation section thus appears significantly more efficient on an energy basis than on an exergy basis. Physically, this discrepancy implies that although 95% of the input energy is transferred to the preheated water, the energy is degraded as it is transferred. Exergy analysis highlights this degradation.

TABLE I
Flow Data for a Unit of the Coal-Fired Electrical Generating Station

Stream	Mass flow rate (kg/s) ^a	Temperature (°C)	Pressure (N/m ²)	Vapor fraction ^b	Energy flow rate (MW)	Exergy flow rate (MW)
S1	41.74	15.00	1.01×10^5	Solid	1367.58	1426.73
S2	668.41	15.00	1.01×10^5	1.0	0.00	0.00
S3 ^c	710.15	1673.59	1.01×10^5	1.0	1368.00	982.85
S4	710.15	119.44	1.01×10^5	1.0	74.39	62.27
S5A	453.59	538.00	1.62×10^7	1.0	1585.28	718.74
S8	42.84	323.36	3.65×10^6	1.0	135.44	51.81
S10	367.85	35.63	4.50×10^3	0.0	36.52	1.20
S11	367.85	35.73	1.00×10^6	0.0	37.09	1.70
S12	58.82	188.33	1.21×10^6	0.0	50.28	11.11
S13	18,636.00	15.00	1.01×10^5	0.0	0.00	0.00
S14	18,636.00	23.30	1.01×10^5	0.0	745.95	10.54
S15	410.75	323.36	3.65×10^6	1.0	1298.59	496.81
S20	367.85	360.50	1.03×10^6	1.0	1211.05	411.16
S21	410.75	538.00	4.00×10^6	1.0	1494.16	616.42
S22	15.98	423.23	1.72×10^6	1.0	54.54	20.02
S25	26.92	360.50	1.03×10^6	1.0	88.64	30.09
S33	309.62	35.63	4.50×10^3	0.93	774.70	54.07
S34	10.47	253.22	3.79×10^5	1.0	32.31	9.24
S35	23.88	209.93	2.41×10^5	1.0	71.73	18.82
S36	12.72	108.32	6.89×10^4	1.0	35.77	7.12
S37	11.16	60.47	3.45×10^4	1.0	30.40	5.03
S38	58.23	55.56	1.33×10^4	0.0	11.37	0.73
S39	367.85	124.86	1.00×10^6	0.0	195.94	30.41
S40	453.59	165.86	1.00×10^6	0.0	334.86	66.52
S41	453.59	169.28	1.62×10^7	0.0	347.05	77.57
S42	453.59	228.24	1.62×10^7	0.0	486.75	131.93
Q5					5.34	0.00
Q6					5.29	0.00
P1					0.57	0.57
P8					523.68	523.68
P15					12.19	12.19

^aThe composition of all streams is 100% H₂O, except that on a volume basis the composition of S1 is 100% carbon, that of S2 is 79% N₂ and 21% O₂, and that of both S3 and S4 is 79% N₂, 6% O₂, and 15% CO₂.

^bVapor fraction is listed as 0.0 for liquids and 1.0 for superheated vapors.

^cStream S3 (not shown in Fig. 2) represents the hot product gases for adiabatic combustion.

A large quantity of energy (775 MW for each unit) enters the condensers, of which nearly 100% is rejected, and a small quantity of exergy (54 MW for each unit) enters, of which approximately 25% is rejected and 75% is internally consumed. Thus, energy analysis results lead to the erroneous conclusion that almost all losses in electricity-generation potential for the station are associated with the heat rejected by the condensers, whereas exergy analysis demonstrates quantita-

tively and directly that the condensers are responsible for little of these losses (Fig. 4). This discrepancy arises because heat is rejected by the condensers at a temperature very near that of the environment.

In the power production and preheating sections, energy losses were found to be very small (<10 MW total), and exergy losses were moderately small (118 MW in the power production section and 23 MW in the preheating section). The exergy losses

TABLE II

Breakdown by Section and Device of Exergy Consumption Rates for a Unit of the Coal-Fired Electrical Generating Station

Section/device	Exergy consumption rate (MW)
Steam generation section	
Steam generator (including combustor)	659.0
Power production section	
High-pressure turbine	26.4
Intermediate-pressure turbine	22.3
Low-pressure turbines	59.2
Generator	5.3
Transformer	5.3
Total	118.5
Condensation section	
Condenser	43.1
Preheat section	
Low-pressure heat exchangers	10.7
Deaerating heat exchanger	5.1
High-pressure heat exchangers	6.4
Hot well pumps	0.1
Boiler feed pumps	1.1
Total	23.4
Total consumption rate	844.0

are almost completely associated with internal consumptions.

In assessing the thermodynamic characteristics of a coal-fired electrical generating station, several insights into the performance of such stations have been obtained. First, although energy and exergy efficiencies are similar for the station, energy analyses do not systematically identify the location and cause of process inefficiencies, whereas exergy analyses do. Specifically, energy losses are associated with emissions (mainly heat rejected by condensers) and exergy losses primarily with consumptions (mainly in the combustors). Second, since devices with the largest thermodynamic losses have the largest margins for efficiency improvement, efforts to increase the efficiencies of coal-fired electrical generating stations should focus on the combustors. For instance, technologies capable of producing electricity without combustion (e.g., fuel cells) or utilizing heat at high temperatures could increase efficiencies significantly. This suggestion, of course, is overly simplistic because such decisions require consideration of other technical and economic factors in addition to thermodynamic efficiency.

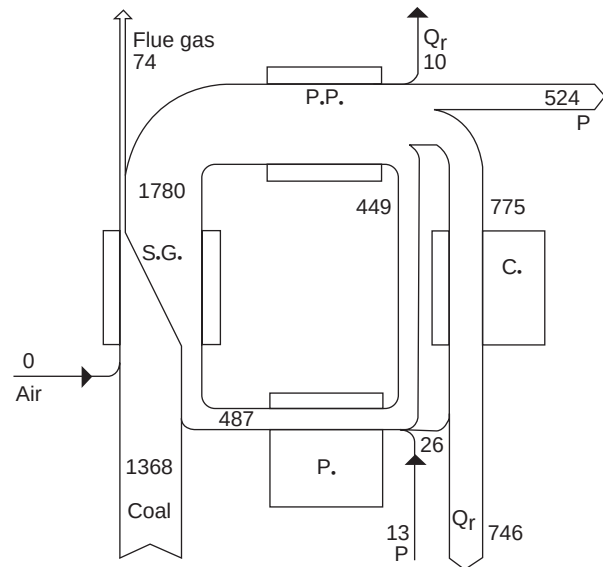


FIGURE 3 Simplified energy diagram for a unit of the coal-fired electrical generating station, indicating net energy flow rates (MW) for streams. Stream widths are proportional to energy flow rates. Station sections shown are steam generation (S.G.), power production (P.P.), condensation (C.), and preheating (P.). Streams shown are electrical power (P), heat input (Q), and heat rejected (Q_r).

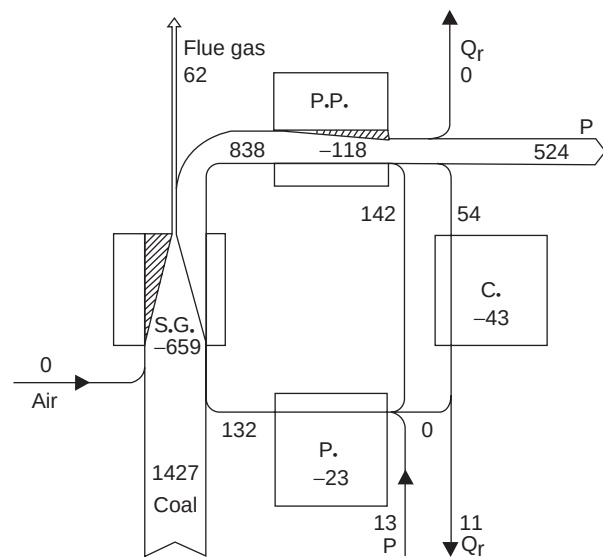


FIGURE 4 Simplified exergy diagram for a unit of the coal-fired electrical generating station, indicating net exergy flow rates for streams and consumption rates (negative values) for devices (MW). Stream widths are proportional to exergy flow rates, and shaded regions are proportional to exergy consumption rates. Other details are as in the legend to Fig. 3.

Third, the use of heat rejected by condensers only increases the exergy efficiencies by a few percent.

8. MACROSYSTEM: THE AGGREGATE ENERGY SYSTEMS OF A COUNTRY

The energy-utilization efficiency of a macrosystem, such as a region or country, can be assessed beneficially using exergy analysis. This section illustrates how the efficiency of energy utilization in a particular macrosystem, Canada, can be examined using both energy and exergy analyses. This example is based on a 1992 publication by the author titled "Evaluation of Energy Utilization Efficiency in Canada Using Energy and Exergy Analyses," which used Canadian energy data for 1986.

The relation between energy resources and sectors of a macrosystem can be modeled as in Fig. 5. The energy resources considered here are coal, petroleum, natural gas, nuclear energy, and hydraulic energy. Renewable energy resources (e.g., wind, solar, wave, and tidal) are neglected since the quantity used in Canada is minor relative to other resources. Four sectors are considered. The residential-commercial sector, which is driven mainly by fossil fuels and electricity, includes residential, public administration, commercial, and other institutional buildings and involves as principal devices space heaters, water heaters, cooking appliances, and clothes dryers. The transportation sector is primarily driven by fossil fuels and includes rail, marine, air, road and urban transport, pipelines, and retail pump sales. The industrial sector has as its main tasks work production and heating and includes mines, steam generation, pulp and paper mills, cement, iron and steel mills, petroleum refining, smelting and refining, other manufacturing, agriculture, chemicals, and non-energy uses. The utility sector includes processes for electricity generation from coal, petroleum, natural gas, and nuclear and hydraulic energy sources. The same reference environment model as used in the

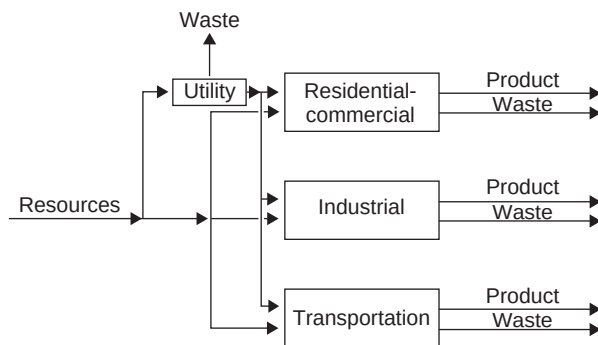


FIGURE 5 Illustration of the energy flows in a macrosystem.

previous section is used here but with a reference environment temperature of 25°C (except where noted).

8.1 Principal Types of Processes and Their Efficiencies

Energy and exergy efficiencies are based on the following general definitions:

$$\text{Energy efficiency} = \frac{(\text{energy in products})}{(\text{total energy input})} \quad (5)$$

and

$$\text{Exergy efficiency} = \frac{(\text{exergy in products})}{(\text{total exergy input})} \quad (6)$$

The main types of processes encountered, for which energy and exergy efficiencies are evaluated, are as follows:

- Heating: Electric and fossil fuel heating processes are taken to generate product heat at a constant temperature.
- Work production: Electric and fossil fuel work production processes produce shaft work.
- Electricity generation: Electricity is generated from fossil fuel, nuclear, and hydraulic sources. The hydraulic source is taken to be the change in kinetic and potential energy of flowing water (even though coal equivalents are often used to represent hydraulic energy inputs). The nuclear source is taken to be nuclear-derived heat at a temperature so high that the energy and exergy of the nuclear-derived heat are equivalent. The efficiency definitions for nuclear plants used here follow nuclear industry conventions (although these efficiency definitions are inadequate because they are based on the heat released from the nuclear fuel rather than its energy or exergy content).
- Kinetic energy production: Fossil fuel-driven kinetic energy production processes occur in some transportation-sector devices (e.g., turbojet engines) and increase the kinetic energy of matter.

8.2 Sector Analyses

8.2.1 Transportation Sector

In the first four columns of Table III, a breakdown is presented by mode of transport of sector energy use, whereas the last two columns show corresponding energy efficiencies. Since vehicles generally are not operated at full load, a distinction is made between

TABLE III
Transportation Sector Data

Mode of transport	Main fuel types	Energy used in the sector, by mode of transport		Energy efficiencies (%)	
		PJ	%	Rated load	Estimated operating
Rail	Diesel	75.39	4.6	35	28
Air	Jet fuel	131.2	8.1	35	28
Marine	Diesel, gasoline	74.82	4.6	—	15
Road and urban transit	Gasoline, diesel, propane, electricity	164.3	10.1	28	22
Retail pump sales	Gasoline, diesel, propane	1093	67.2	25	15
Pipelines	Natural gas	86.21	5.3	—	29

TABLE IV
Electricity-Generation Data for the Utility and Industrial Sectors

Mode of generation	Electrical energy generated (PJ) ^a	Energy input (PJ)	Energy efficiency (%)
Hydraulic ^b	1000	1176	85
Nuclear	230.1	767.0	30
Fossil	293.0	915.6	32

^aHydraulic and nuclear generation is entirely within the utility sector, whereas 262.3 PJ of fossil generation is within the utility sector and the remainder is within the industrial sector. The contribution by different fossil fuels is 91.2% coal, 6.7% petroleum, and 2.1% natural gas.

^bThe hydraulic source is taken to be the change in kinetic and potential energy of flowing water and not a fossil fuel equivalent.

rated load (full load) and operating (part load) efficiencies. The energy efficiencies in Table III are equal to the corresponding exergy efficiencies. Weighted means are obtained for the transportation-mode energy efficiencies in Table III, where the weighting factor is the fraction of the total sector energy input that supplies each transportation mode (fourth column in Table III). Thus, the weighted mean overall efficiency for the transportation sector based on energy, which is equal to that based on exergy, is 18%.

8.2.2 Utility Sector

For the different types of electricity generation, Table IV lists energy efficiencies and the quantities of electricity generated and energy inputs. The energy efficiencies in Table IV are equal to the corresponding exergy efficiencies. Using data from this table, the

TABLE V
Energy Consumption Data for the Residential–Commercial Sector

Devices	Energy used in the sector, by type (%)		
	Electrical	Fuel	Total
Space heaters	15.0	79.5	66.0
Water heaters	14.0	15.0	10.0
Cooking appliances	5.0	4.5	1.7
Clothes dryers	5.0	1.0	1.3
Others ^a	61.0	0.0	21.0

^aThis category includes fans, pumps, air conditioners, refrigerators, freezers, washers, lights, etc.

mean overall energy and exergy efficiencies for the utility sector are found to be equal at 53%.

8.2.3 Residential–Commercial Sector

Energy end-use quantities for this sector are listed in Table V. Operating data for the principal devices in the sector are listed in Table VI and are used to determine the device exergy efficiencies. Weighted mean energy and exergy efficiencies are calculated for the overall residential–commercial sector using a three-step process. First, weighted means are obtained for the electrical energy and exergy efficiencies for the devices in Table VI, where the weighting factor in both cases is the ratio of the electrical energy input to the device to the total electrical energy input to all of the devices (as given in Table V). These energy and exergy efficiencies are 89 and 20%, respectively. Second, weighted mean efficien-

TABLE VI

Device Data for the Residential–Commercial Sector^a

Devices	Product-heat temperature (K)		Heating efficiency (%)			
			Electrical		Fuel	
	Electrical	Fuel	Energy	Exergy	Energy	Exergy
Space heaters	328	328	100.0	17.1	65.0	11.1
Water heaters	350	374	93.0	25.4	62.0	14.0
Cooking appliances	394	—	80.0	22.5	—	—
Clothes dryers	350	350	50.0	9.6	50.0	10.3

^aCorresponding reference environment temperatures are 272 K for space heating and 283 K for all other devices.

cies for the fossil fuel-driven devices in Table VI are similarly determined as 64 and 12%, respectively. Third, overall weighted means are obtained for the energy and exergy efficiencies for the electrical and fossil fuel processes together, where the weighting factor is the ratio of the total energy (either fossil fuel or electrical) input to the residential–commercial sector to the total energy input to the sector. Since 34% of the 2129 PJ input to the residential–commercial sector is electrical energy and the remaining 66% is fuel energy, the weighted mean overall energy and exergy efficiencies for the sector are 73 and 14%, respectively. These means are representative of the sector since the devices in Table VI account for more than half of the total sector energy consumption.

8.2.4 Industrial Sector

The industrial sector is composed of many industries, each of which encompasses a wide variety of complex plants. The five most significant industries were identified, and a single representative commercial process for each was examined. The identified industries with the selected commercial processes are as follows:

- Pulp and paper (integrated paper mill plant)
- Iron and steel (blast furnace and steel mill plant)
- Cement (hydraulic cement factory)
- Petroleum (petroleum refining plant)
- Chemical [NEC (salt) chemical preparation plant]

These five industries are taken to be representative of the sector since they account for approximately half of the sector energy use.

Only heating and mechanical processes are considered since these processes account for 81% of the industrial-sector energy consumption (66% for heating and 15% for mechanical drives). Following

TABLE VII

Process Heating Temperature Range Categories and Energy Efficiencies for the Industrial Sector

Product heat temperature category	Product heat temperature range (°C) ^a	Heating energy efficiency (%)	
		Electrical	Fuel
Low	< 121	100	65
Medium	121–399	90	60
High	> 399	70	50

^aTemperature ranges were determined subjectively after examining many commercial processes.

the temperature categories in Table VII, process heat data for the five selected commercial processes are listed in Table VIII, with the percentages of the heat in each range supplied by fossil fuels and electricity. Canadian energy consumption in the form of electricity and fossil fuels for each of the five main industries is listed in Table IX, with a breakdown of the total energy demand. The values are obtained by averaging the percentage of electrical and fuel energy demands for each of the industries shown. Subsequently, steam (from utilities) is treated as an industrial-sector product and thus passes through the sector unchanged, and electricity generation in the industrial sector is included in the utility sector.

Three steps are used in evaluating industry efficiencies. First, energy and exergy efficiencies are obtained for process heating for each of the temperature range categories in Table VIII using data from Tables VII and VIII. Second, mean heating energy and exergy efficiencies for each of the five industries (see the fifth and sixth columns of Table IX) are evaluated using a two-step procedure: (i)

TABLE VIII

Process Heating Data for All Categories of Product Heat Temperature for the Industrial Sector

Industry	Product heat temperature category	Mean product heat temperature (°C)	Energy used in each product heat temperature category, by type (%)	
			Electricity	Fuel
Pulp and paper	Low	67	100.0	0.0
	Medium	160	0.0	83.0
	High	732	0.0	17.0
Iron and steel	Low	45	4.2	0.0
	Medium	—	0.0	0.0
	High	983	95.8	100.0
Cement	Low	31	91.7	0.9
	Medium	163	0.0	9.0
	High	1482	8.3	90.1
Petroleum	Low	57	10.1	13.8
	Medium	227	9.4	22.6
	High	494	80.4	63.6
Chemical	Low	42	62.5	0.0
	Medium	141	37.5	100.0
	High	—	0.0	0.0

TABLE IX

Process Data for the Industrial Sector

Industry	Energy used in sector, by type			Energy and exergy efficiencies (%)			
				Heating		Total	
	Electrical (PJ)	Fuel (PJ)	Total (%)	Energy	Exergy	Energy	Exergy
Pulp and paper	181.4	169.6	16.8	80.0	20.4	82.0	33.0
Iron and steel ^a	155.7	284.5	17.2	58.0	43.4	64.0	52.0
Cement	5.38	52.29	1.4	56.0	37.5	62.3	47.0
Petroleum	16.45	87.81	2.9	65.0	29.4	70.0	41.0
Chemical	63.00	140.8	7.4	71.0	18.5	74.5	32.0

^aThis category includes smelting and refining industrial operations.

Weighted mean efficiencies for electrical heating and for fuel heating are evaluated for each industry in Table VIII using as weighting factors the values listed in the last two columns of Table VIII, and (ii) weighted mean efficiencies for all heating processes in each industry are evaluated with these values using as weighting factors the ratio of the industry energy consumption (electrical or fuel) to the total consumption of both electrical and fuel energy (Table IX). Third, weighted mean overall (i.e., heating and mechanical drive) efficiencies for each industry are evaluated using the values in Table IX, assuming

mechanical drive energy efficiencies of 90% and using as weighting factors the fractions of the total sector energy input for both heating and mechanical drives (as noted earlier).

Weighted means for the weighted mean overall energy and exergy efficiencies (listed in the last two columns of Table IX) for the five principal industries in the industrial sector are obtained, using as weighting factors the fraction of the total industrial energy demand supplied to each industry (given in the fourth column of Table IX), as 73 and 41%, respectively.

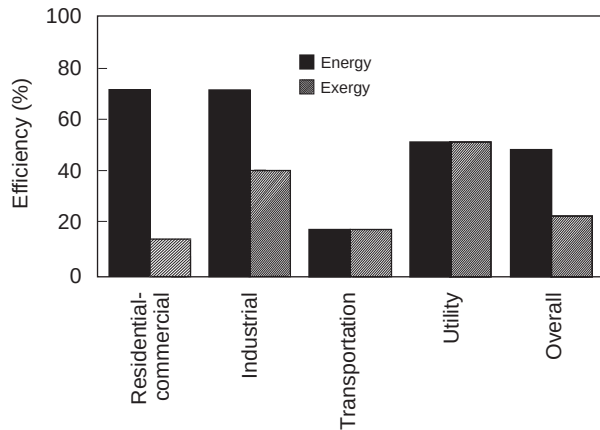


FIGURE 6 Energy and exergy efficiencies for the four main sectors of the Canadian economy and for the overall economy.

8.2.5 Sector Comparison

The efficiencies for the four sectors and for the overall Canadian economy are compared in Fig. 6. The energy efficiencies range from 18 to 73% and the exergy efficiencies from 14 to 53%. The most efficient sectors on energy and exergy bases are different. The residential-commercial sector exhibits the greatest differences for energy and exergy evaluations. The most significant differences between energy and exergy efficiencies are attributable to heating processes. For example, the overall exergy efficiency of the residential-commercial sector (14%) is much lower than the overall energy efficiency (73%) because

high-quality energy (i.e., high exergy) sources are used for low-quality energy (i.e., low exergy) demands. The higher exergy efficiency of the industrial sector is mainly attributable to the high temperatures required for most industrial heating ($\approx 400^\circ\text{C}$) compared to residential-commercial heating ($\approx 25^\circ\text{C}$). The industrial sector thus utilizes relatively more of the quality or work potential of fuels.

8.3 Analyses for the Overall Macrosystem

Figures 7 and 8 illustrate, respectively, the flows of energy and exergy in Canada for the year considered. The left sides of the figures consist of the total energy or exergy inputs to all sectors by fuel type. These quantities are then branched according to sector demand. Fuels such as coal, gas, and petroleum are routed to “fuel for electricity” and include fuels used to produce electricity by utilities as well as by industry. Similarly, “fuel for end uses” represents the net fuel requirement for all sectors and is routed together with electricity and steam flows to the sectors. The final branching into useful and waste energy is achieved by multiplying the energy for end use (fuel, steam, and electricity) in each sector by the corresponding efficiency (see the values summarized in Fig. 6).

The energy flow diagram indicates that nearly 50% of the total energy consumed in Canada is

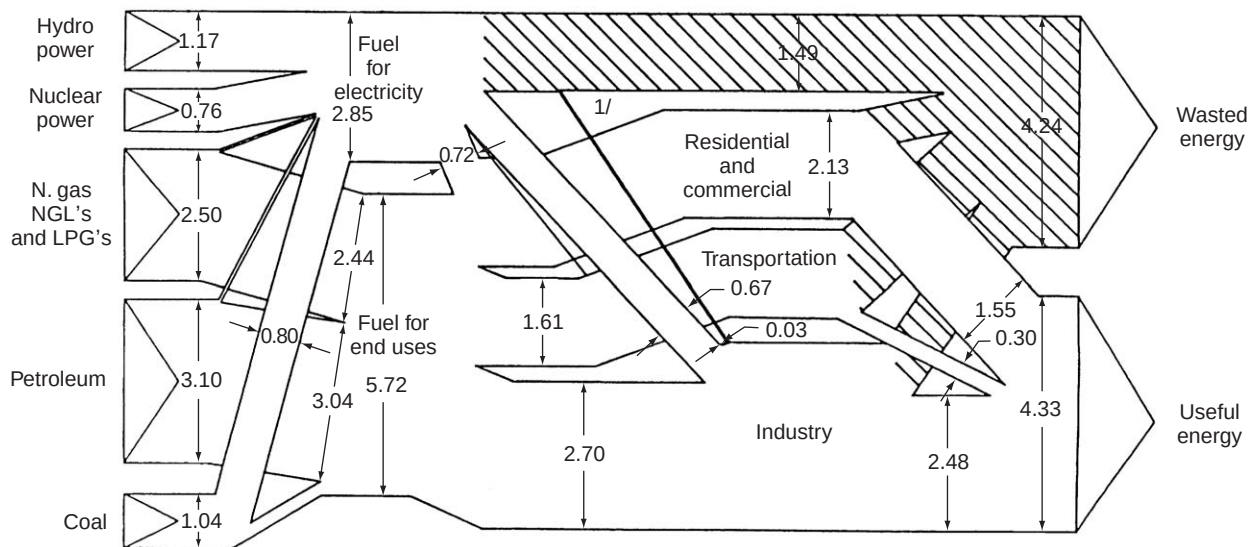


FIGURE 7 Flows of energy (EJ) in Canada. Widths of lines are proportional to the magnitude of the quantity represented. Hydraulic energy is reported in kinetic energy equivalent. The portions of coal, petroleum, and natural gas used for industrial-sector electricity generation are routed to the utility sector and then back to the industrial sector in the form of electricity. The flow associated with the symbol 1/ represents steam produced in the utility sector for the end-use sectors.

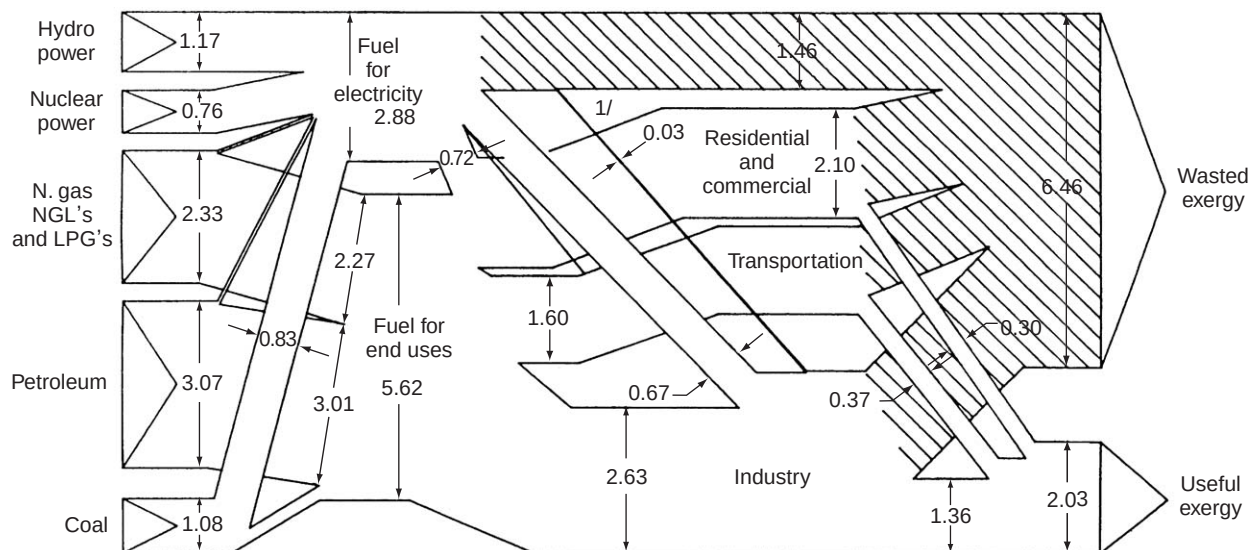


FIGURE 8 Flows of exergy (EJ) in Canada. Details are as given in the legend to Fig. 7.

converted to useful energy for end uses, and that the transportation and utility sectors produce the most waste. The exergy flow diagram differs markedly from the energy diagram in that only 24% of Canadian exergy consumption is converted to useful exergy for end uses, with the residential-commercial and industrial sectors producing the greatest waste. Thus, energy analysis indicates a more limited margin for efficiency improvement than exergy analysis. However, only exergy analysis indicates the margin for improvement that actually exists. Therefore, by using exergy rather than energy analysis, energy utilization efficiency in the macro-system is more clearly illuminated, and more rational assessments are obtained for allocations of energy-related research and development (R&D).

9. CONCLUSIONS

Exergy analyses of energy systems provide useful information that can have direct implications on designs, improvements, and application decisions. Furthermore, exergy analyses, more than energy analyses, provide insights into better directions for R&D. Here, "better" is loosely taken to mean "more promising for significant efficiency gains." There are two main reasons for this statement:

- Exergy losses represent true losses of the potential that exists to generate the desired product from the given driving input. This is not true in general for energy losses. Thus, if the objective is to

increase efficiency while accounting for energy degradation, focusing on exergy losses permits R&D to focus on reducing losses that will affect the objective.

- Exergy efficiencies always provide a measure of how closely the operation of a system approaches the ideal, or theoretical upper limit. This is not generally true for energy efficiencies. By focusing R&D on those plant sections or processes with the lowest exergy efficiencies, effort is directed to those areas that inherently have the largest margins for efficiency improvement. By focusing on energy efficiencies, on the other hand, one can inadvertently expend R&D effort on areas for which little margins for improvement, even theoretically, exist.

Exergy analysis results typically suggest improvement efforts should concentrate more on internal rather than external exergy losses, based on thermodynamic considerations, with a higher priority for the processes having larger exergy losses. Note that this statement does not indicate that effort should not be devoted to processes having low exergy losses because simple and cost-effective ways to increase efficiency by reducing small exergy losses should certainly be considered when identified.

In general, application and R&D decisions should not be based exclusively on the results of energy and exergy analyses, even though these results provide useful information to assist in such decision making. Other factors must also be considered, such as economics, environmental impact, safety, and social and political implications.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy, Overview • Entropy • Entropy and the Economic Process • Exergoeconomics • Exergy • Exergy Analysis of Waste Emissions • Exergy: Reference States and Balance Conditions • Thermodynamic Sciences, History of • Thermodynamics, Laws of

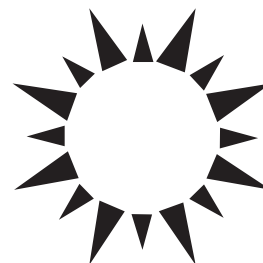
Further Reading

- Brodyanski, V. M., Sorin, M. V., and Le Goff, P. (1994). "The Efficiency of Industrial Processes: Exergy Analysis and Optimization." Elsevier, London.
- Dincer, I., and Rosen, M. A. (2002). Energy and exergy analyses of thermal energy storage systems. In "Thermal Energy Storage Systems and Applications," pp. 411–510. Wiley, London.
- Kotas, T. J. (1995). "The Exergy Method of Thermal Plant Analysis," Reprint. Krieger, Malabar, FL.
- Moran, M. J. (1989). "Availability Analysis: A Guide to Efficient Energy Use." Rev. ed. American Society of Mechanical Engineers, New York.
- Moran, M. J., and Sciubba, E. (1994). Exergy analysis: Principles and practice. *J. Eng. Gas Turbines Power* **116**, 285–290.
- Moran, M. J., and Shapiro, H. N. (2000). "Fundamentals of Engineering Thermodynamics." 4th ed. Wiley, New York.
- Rosen, M. A. (1992). Evaluation of energy utilization efficiency in Canada using energy and exergy analyses. *Energy Int. J.* **17**, 339–350.
- Rosen, M. A. (1999). Second law analysis: Approaches and implications. *Int. J. Energy Res.* **23**, 415–429.
- Rosen, M. A. (2001). Energy- and exergy-based comparison of coal-fired and nuclear steam power plants. *Exergy Int. J.* **1**, 180–192.
- Rosen, M. A., and Dincer, I. (1997). Sectoral energy and exergy modeling of Turkey. *ASME J. Energy Resources Technol.* **119**, 200–204.
- Rosen, M. A., and Horazak, D. A. (1995). Energy and exergy analyses of PFBC power plants. In "Pressurized Fluidized Bed Combustion" (M. Alvarez Cuenca and E. J. Anthony, Eds.), pp. 419–448. Chapman & Hall, London.
- Szargut, J., Morris, D. R., and Steward, F. R. (1988). "Exergy Analysis of Thermal, Chemical and Metallurgical Processes." Hemisphere, New York.
- Yantovskii, E. I. (1994). "Energy and Exergy Currents: An Introduction to Exergonomics." Nova Science, New York.

Exergy Analysis of Waste Emissions

MARC A. ROSEN

University of Ontario Institute of Technology
Oshawa, Ontario, Canada



-
1. Exergy and Exergy Analysis
 2. Energy and the Environment
 3. Exergy and the Environment
 4. Relations between Exergy and Environmental Impact
 5. Environmental Impact of Exergy Losses
 6. Comparison of Exergy and Other Indicators for the Environmental Impact of Waste Emissions
 7. Conclusions

Glossary

chemical exergy The maximum work obtainable from a substance when it is brought from the environmental state to the dead state by means of processes involving interaction only with the environment.

dead state The state of a system when it is in thermal, mechanical, and chemical equilibrium with a conceptual reference environment (having intensive properties pressure P_o , temperature T_o , and chemical potential μ_{ko} for each of the reference substances in their respective dead states).

environmental pollution cost (EPC) The environmental cost of a pollutant, based on such factors as the cost to society for the correction or compensation of environmental damage and the cost incurred to prevent a harmful emission from escaping into the environment.

environomics A methodology for analyzing and improving energy-related systems by taking into account energy, exergy, economic, and environmental factors.

environmental state The state of a system when it is in thermal and mechanical equilibrium with the reference environment (i.e., at pressure P_o and temperature T_o of the reference environment).

exergy (i) A general term for the maximum work potential of a system, stream of matter, or a heat interaction in relation to the reference environment as the datum state. Also known as available energy, availability, essergy, technical work capacity, usable energy, utilizable energy, work capability, work potential, and xergy.

(ii) The unqualified term exergy or exergy flow is the maximum amount of shaft work obtainable when a steady stream of matter is brought from its initial state to the dead state by means of processes involving interactions only with the reference environment.

exergy analysis An analysis technique, based on the second law of thermodynamics, in which process performance is assessed by examining exergy balances, losses, and efficiencies.

exergy consumption The exergy consumed or destroyed during a process due to irreversibilities within the system boundaries. Also known as dissipation, irreversibility, and lost work.

reference environment An idealization of the natural environment that is characterized by a perfect state of equilibrium—that is, the absence of any gradients or differences involving pressure, temperature, chemical potential, kinetic energy, and potential energy. The environment constitutes a natural reference medium with respect to which the exergy of different systems is evaluated.

reference state A state with respect to which values of exergy are evaluated. Several reference states are used, including environmental state, dead state, standard environmental state, and standard dead state. Also known as ground state.

removal pollution cost (RPC) The cost of removing a pollutant from a waste stream prior to discharge into the environment.

waste emission Release to the environment of a material that is not considered usable and is therefore being discarded.

Exergy analysis is a thermodynamic technique for assessing and improving systems and processes, which is similar but advantageous to energy analysis, in large part because it is based on the second law of thermodynamics. The exergy of an energy form or a substance is a measure of its usefulness. Exergy can be applied beyond thermodynamics to understanding waste emissions and reducing their environmental

impact because exergy can also be viewed as a measure of the departure of a substance from equilibrium with a specified reference environment, which is often modeled as the actual environment. Furthermore, exergy is a measure of the potential of a substance to cause change. The exergy of an emission to the environment, therefore, is a type of measure of the potential of the substance to change or impact the environment. The greater the exergy of an emission, the greater its departure from equilibrium with the environment and the greater may be its potential to change or impact the environment. The exergy of an emission is zero only when it is in equilibrium with the environment and thus benign. These points suggest that exergy may be, or provide the basis for, one effective indicator of the potential of an emitted substance to impact the environment.

This article describes both the relation between exergy and the environment and the exergy of waste emissions. Such knowledge can permit the development of a better understanding of, and more rational approaches to mitigating, the environmental impact associated with waste emissions. First, exergy and its use as a tool through exergy analysis are briefly described. Then, energy and the environment and exergy and the environment, and some relations between them, are discussed. Finally, exergy is compared with other indicators for the environmental impact of waste emissions.

1. EXERGY AND EXERGY ANALYSIS

Exergy is defined as the maximum amount of work that can be produced by a flow of matter or energy as it comes to equilibrium with a reference environment. Exergy is a measure of the potential of a flow to cause change, as a consequence of being in disequilibrium with the reference environment. For evaluations of exergy, the state of the reference environment, or the reference state, must be specified completely, normally by specifying its temperature, pressure, and chemical composition. The reference environment used in exergy evaluations is discussed further in Section III.B. Exergy is not subject to a conservation law. Rather, exergy is consumed or destroyed due to the irreversibilities in any process.

Exergy analysis is a method that uses the conservation of mass and conservation of energy principles together with the second law of thermodynamics for the design and analysis of energy systems. The exergy method is useful for attaining

more efficient energy-resource use because it enables the locations, types, and true magnitudes of wastes and losses to be determined. Therefore, exergy analysis can reveal whether, and by how much, it is possible to design more efficient energy systems by reducing the inefficiencies in existing systems. Applications of exergy analysis to many processes and systems are described elsewhere in this encyclopedia.

Exergy concepts can also be applied to disciplines outside of thermodynamics. One area in which applications of exergy are increasing is that of economics. Another area in which applications of exergy are increasing is that of environmental emissions and impact, and this is the topic of the current article.

2. ENERGY AND THE ENVIRONMENT

Energy production, transformation, transport, and end use have significant impacts on the earth's environment, and environmental costs are usually associated with the emissions—whether they are of a thermal, chemical, nuclear, or other nature—from processes and technologies. One means of reducing the environmental impact of emissions is to increase the efficiency of resource utilization. However, increasing efficiency generally entails greater use of materials and energy resources, and the environmental burdens associated with these can offset the environmental gains of improved efficiency.

Data on the current world distribution of per capita energy consumption suggest that a close correlation exists between a country's energy consumption and economic development. However, there does not appear to be a simple correlation between a country's energy consumption and the decay of its natural environment. As environmental concerns became major issues in the 1980s, increasing attention was paid to the relationship between energy use and the environment. However, there are limited works on the relationships between the exergy and environmental concepts even though to appropriately reduce energy consumption and the related environmental impact, engineers, scientists, and policymakers may benefit from such knowledge.

One of the major challenges of designing technologies so as to reduce their environmental impact involves determining an environmentally optimal configuration or selecting the most appropriate one from competing options. This determination or

selection is made difficult by the complex relationship between the technology being considered and the characteristics of the residuals produced. Process or technology changes generally affect the flow rate and composition of all resulting effluent streams. Therefore, evaluating alternatives often requires one to compare the relative environmental merits of distinctly different residual streams. Existing systems for performing such analyses have focused primarily on subjective ranking techniques.

3. EXERGY AND THE ENVIRONMENT

There is agreement among some researchers that a less subjective metric capable of providing greater insight may exist in the form of the thermodynamic quantity exergy. Exergy analysis provides information about energy and material quality that is both descriptive and physically significant. Because of its origins within the thermodynamic community, however, few applications and investigations of the potential of exergy methods for reducing environmental impact have been performed.

Some work is nevertheless being carried out. For instance, the Consortium on Green Design and Manufacturing at the University of California, Berkeley, is carrying out a project titled "Exergy as an Environmental Indicator" in an effort to increase the practical application of exergy analysis to rectifying the problems associated with material and energy flows in industry. This work has focused on developing a generalizable technique to calculate exergy in an industrial setting, exploring the significance of environmental ground states and how they might be utilized to leverage results, and establishing the requisite databases for performing a wide range of highly diverse analyses. In addition, several projects have been carried out at the Delft University of Technology and University of Twente to determine whether exergy analysis can be used as an instrument for environmental policy development, especially for the comparison of alternative production chains. These studies are investigating if the linkages of exergy with phenomena such as pollution and dispersion can be converted into a reliable tool on which policy decisions can be based, and they are exploring how the environmental effects of processes can be linked to or expressed in terms of exergy changes. Also, a methodology known as *environomics* has been developed for analyzing and

improving energy-related systems by simultaneously taking into account energy, exergy, economic, and environmental factors.

The exergy contents of waste emissions are more meaningful than the corresponding energy contents as measures of potential for environmental impact. By definition, material and energy flows only possess exergy when in disequilibrium with a reference environment. The exergy associated with waste emission streams has the potential to cause environmental damage, particularly when it is released in an unconstrained manner into the environment. Some believe that by considering the exergy content of a waste emission, rational and meaningful assessments can be made of the environmental impact potential of the emission.

3.1 Reducing Environmental Impact with Exergy

Efforts to increase energy and exergy efficiency can reduce environmental impact by reducing energy and exergy losses. Increased efficiency frequently contributes to achieving energy security in an environmentally acceptable way, partly by reducing emissions. Increased efficiency also reduces the requirement for new facilities for the production, transportation, transformation, and distribution of the various energy forms; these additional facilities all cause some environmental impacts. To further reduce environmental pollution, efficiency-improvement actions often need to be supported by pollution-control technologies and/or fuel substitution. Regional and national actions to improve exergy efficiency, in particular, can have major impacts on environmental protection.

One study investigated thermodynamics and sustainable development and noted that one tool likely to be important for obtaining sustainable development is exergy analysis. That study stated that whereas energy can never be destroyed, exergy can be destroyed and this exergy destruction (irreversibility) should be appropriately minimized to make development more sustainable. This study also showed that environmental effects associated with emissions and resource depletion can be expressed in terms of one indicator, which is based on physical principles.

Several researchers have suggested that a potentially useful way to link the second law and environmental impact is through exergy because it is a measure of the departure of the state of a system from that of the environment. The magnitude of the

exergy of a system depends on the states of both the system and the environment. This departure is zero only when the system is in equilibrium with its environment. Eminent thermodynamicist Myron Tribus suggested in 1971 that performing exergy analyses of the natural processes occurring on the earth could form a foundation for ecologically sound planning because it would indicate the disturbance caused by large-scale changes.

An understanding of the relations between exergy and the environment may reveal some of the underlying fundamental patterns and forces affecting changes in the environment and help researchers to deal better with environmental damage. Before examining such relations, however, it is useful to discuss the reference environment used in exergy analysis.

3.2 Exergy and the Reference Environment

The exergy Ex contained in a system may be written as

$$Ex = S(T - T_o) - V(P - P_o) + N_k(\mu_k - \mu_{ko}), \quad (1)$$

where the intensive properties are temperature T , pressure P , and chemical potential of substance k , μ_k , and the extensive properties are entropy S , volume V , and number of moles of substance k , N_k . The subscript o denotes conditions of the reference environment. It is evident from this equation that the exergy of a system is zero when it is in equilibrium with the reference environment (i.e., when $T = T_o$, $P = P_o$, and $\mu_k = \mu_{ko}$ for all k).

Exergy analysis requires a reference environment model. Exergy is evaluated with respect to the reference environment model, with the exergy of a flow and system being dependent on the intensive properties of the reference environment. The exergy of the reference environment is zero, and the exergy of a stream or system is zero when it is in equilibrium with the reference environment. There are several theoretical characteristics of the reference environment. The reference environment is in stable equilibrium and has all parts at rest relative to one another. No chemical reactions can occur between the reference environment components. Furthermore, the reference environment acts as an infinite system, is a sink and source for heat and materials, and experiences only internally reversible processes in which its intensive state remains unaltered (i.e., its temperature, pressure, and chemical potentials for each of the components present remain constant).

The natural environment does not possess the theoretical characteristics of a reference environment. The natural environment is not in equilibrium, and its intensive properties vary spatially and temporally. Many chemical reactions in the natural environment are blocked because the transport mechanisms necessary to reach equilibrium are too slow at ambient conditions. Thus, the exergy of the natural environment is not zero because work could be obtained if it were to come to equilibrium. In developing reference environment models for exergy analysis, a compromise is often made between the theoretical requirements of the reference environment and the actual behavior of the natural environment.

Several reference environment models have been proposed. The most significant classes of these are the following:

Natural environment subsystem models: These models attempt to realistically simulate subsystems of the natural environment. A model consisting of saturated moist air and liquid water in phase equilibrium was proposed in 1963. An extension of that model that allows sulfur-containing materials to be analyzed was proposed subsequently. The temperature and pressure of this reference environment (Table I) are normally taken to be 25°C and 1 atm, respectively, and the chemical composition is taken to consist of air saturated with water vapor, with the following condensed phases at 25°C and 1 atm: water (H_2O), gypsum ($CaSO_4 \cdot 2H_2O$), and

TABLE I
Reference Environment Model

Temperature, T_o	25°C	
Pressure, P_o	1 atm	
Composition	(i) Atmospheric air, saturated with H_2O at T_o and P_o , having the following composition:	
	Air constituents	Mole fraction
	N_2	0.7567
	O_2	0.2035
	H_2O	0.0303
	Ar	0.0091
	CO_2	0.0003
	H_2	0.0001
	(ii) The following condensed phases at T_o and P_o :	
	Water (H_2O)	
	Limestone ($CaCO_3$)	
	Gypsum ($CaCO_4 \cdot 2H_2O$)	

limestone (CaCO_3). The stable configurations of C, O, and N, respectively, are taken to be those of CO_2 , O_2 , and N_2 as they exist in air saturated with liquid water at T_o and P_o (the temperature and pressure for the reference environment); that of hydrogen is taken to be in the liquid phase of water saturated with air at T_o and P_o ; and those of S and Ca, respectively, are taken to be those of $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$ and CaCO_3 at T_o and P_o .

Reference substance models: With this model, a reference substance is selected for every chemical element and assigned zero exergy. A model in which the reference substances are selected as the most valueless substances found in abundance in the natural environment was proposed by Jan Szargut. The criterion for selecting such reference substances is consistent with the notion of simulating the natural environment but is primarily economic in nature, and it is vague and arbitrary with respect to the selection of reference substances. Part of this environment is the composition of moist air, including N_2 , O_2 , CO_2 , H_2O , and the noble gases; gypsum (for sulfur); and limestone (for calcium). A related model, in which reference substances are selected arbitrarily, has also been proposed. This model is not similar to the natural environment. Consequently, absolute exergies evaluated with this model do not relate to the natural environment and cannot be used rationally to evaluate efficiencies or environmental impact.

Equilibrium and constrained equilibrium models: A model in which all the materials present in the atmosphere, oceans, and a layer of the crust of the earth are pooled together and an equilibrium composition is calculated for a given temperature was proposed by Joachim Ahrendts in 1980. The selection of the thickness of crust considered is subjective and is intended to include all materials accessible to technical processes. Thicknesses varying from 1 to 1000 m and a temperature of 25°C were considered. For all thicknesses, the model differed significantly from the natural environment. Exergy values obtained using these environments are significantly dependent on the thickness of crust considered and represent the absolute maximum amount of work obtainable from a material. Since there is no technical process that can obtain this work from materials, the equilibrium model does not give meaningful exergy values when applied to the analysis of real processes. Ahrendts also proposed a modified version of his equilibrium environment in which the calculation of an equilibrium composition excludes the possibility of the formation of nitric acid

(HNO_3) and its compounds. That is, all chemical reactions in which these substances are formed are in constrained equilibrium, and all other reactions are in unconstrained equilibrium. When a thickness of crust of 1 m and temperature of 25°C are used, the model is similar to the natural environment.

Process-dependent models: A model that contains only components that participate in the process being examined in a stable equilibrium composition at the temperature and pressure of the natural environment was proposed by Bosnjakovic in 1963. This model is dependent on the process examined and is not general. Exergies evaluated for a specific process-dependent model are relevant only to the process. They cannot rationally be compared with exergies evaluated for other process-dependent models or used in environmental assessments.

4. RELATIONS BETWEEN EXERGY AND ENVIRONMENTAL IMPACT

Three relationships between exergy and environmental impact that may help explain some of the patterns that underlie changes in the environment are discussed in the following sections.

4.1 Order Destruction and Chaos Creation

The destruction of order, or the creation of chaos, is a form of environmental damage. Fundamentally, entropy is a measure of chaos. A system of high entropy is more chaotic or disordered than one of low entropy. For example, a field with papers scattered about has a higher entropy than the field with the papers neatly piled. Conversely, exergy is a measure of order. Relative to the same environment, the exergy of an ordered system is greater than that of a chaotic one.

The difference between the exergy values of the two systems containing papers described previously is a measure of the minimum work required to convert the chaotic system to the ordered one (i.e., in collecting the scattered papers). In reality, more than this minimum work, which only applies if a reversible cleanup process is employed, is required. The exergy destroyed when the wind scatters a stack of papers is a measure of the order destroyed during the process.

4.2 Resource Degradation

The degradation of resources found in nature is a form of environmental damage. A resource is a

material, found in nature or created artificially, that is in a state of disequilibrium with the environment. Resources have exergy as a consequence of this disequilibrium. For some resources (e.g., metal ores), it is their composition that is valued. Many processes exist to increase the value of such resources by purifying them (i.e., by increasing their exergy). This is done at the expense of consuming at least an equivalent amount of exergy elsewhere (e.g., burning coal to produce process heat for metal ore refining). For other resources (e.g., fuels), it is normally their reactivity that is valued (i.e., their potential to cause change or “drive” a task or process).

By preserving exergy through increased efficiency (i.e., degrading as little exergy as necessary for a process), environmental damage is reduced. Increased efficiency also has the effect of reducing exergy emissions, which, as discussed in the next section, also play a role in environmental damage.

The earth is an open system subject to a net influx of exergy from the sun. It is the exergy (or order states) delivered with solar radiation that is valued; all the energy received from the sun is ultimately radiated out to the universe. Environmental damage can be reduced by taking advantage of the openness of the earth and utilizing solar radiation instead of degrading resources found in nature to supply exergy demands. This would not be possible if the earth were a closed system because it would eventually become increasingly more degraded or entropic.

4.3 Waste Exergy Emissions

The exergy associated with process wastes emitted to the environment can be viewed as a potential for environmental damage. Typical process wastes have exergy, a potential to cause change, as a consequence of not being in stable equilibrium with the environment. When emitted to the environment, this exergy represents a potential to change the environment. In some cases, this exergy may cause a change perceived to be beneficial (e.g., the increased rate of growth of fish and plants near the cooling water outlets from thermal power plants). More often, however, emitted exergy causes a change that is damaging to the environment (e.g., the death of fish and plants in some lakes due to the release of specific substances in stack gases as they react and come to equilibrium with the environment).

Emissions of exergy to the environment can also interfere with the net input of exergy via solar radiation to the earth. The carbon dioxide emitted in stack gases from many processes changes the atmo-

spheric CO₂ content, affecting the receiving and reradiating of solar radiation by the earth.

The relation between waste exergy emissions and environmental damage has been recognized by several researchers. By considering the economic value of exergy in fuels, in 1970 Gordon Reistad developed an air pollution rating in which the air pollution cost for a fuel was estimated as either the cost to remove the pollutant or the cost to society of the pollution (i.e., the tax that should be levied if pollutants are not removed from effluent streams). Reistad claimed the rating was preferable to the mainly empirical ratings then in use.

4.4 An Apparent Paradox

The reader may at this point detect a paradox. It was pointed out that exergy in the environment—in the form of resources—is of value, whereas exergy in the environment—in the form of emissions—has a “negative” value because it has a potential to cause environmental damage.

This apparent paradox can be resolved if the word *constrained* is included in the discussions (Fig. 1). Constrained sources of exergy in the environment are of value. Most resources found in nature are constrained. Unconstrained emissions of exergy to the environment have negative value. In fact, if emissions to the environment are constrained (e.g., separating out sulfur from stack gases), then there are two potential benefits: (i) The potential for environmental damage is blocked from entering the environment, and (ii) the now-constrained emission potentially becomes a valued commodity, a source of exergy.

4.5 Qualitative Illustration

The decrease in the environmental impact of a process, in terms of the three measures discussed in this section, as the exergy efficiency increases for the process is

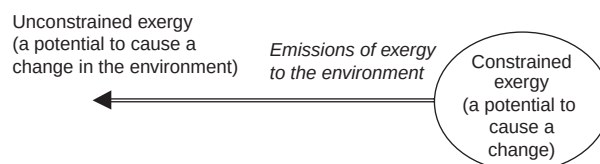


FIGURE 1 Comparison of constrained and unconstrained exergy illustrating that exergy constrained in a system represents a resource, whereas exergy emitted to the environment becomes unconstrained and represents a driving potential for environmental damage.

illustrated qualitatively and approximately in Fig. 2. Exergy methods also correlate with sustainability if, for simplicity, we consider sustainability to indicate the degree to which human activity can continue without facing a lack of resources or negatively changing the environment. To better illustrate the relation between exergy and sustainability and environmental impact, Fig. 2 can be expanded as shown in Fig. 3. In this figure, sustainability is seen to increase and environmental impact to decrease as the exergy efficiency of a process increases. The two limiting efficiency cases in Fig. 3 are significant:

- As exergy efficiency approaches 100%, environmental impact approaches zero since exergy is converted from one form to another without loss (either through internal consumptions or through waste emissions). Also, sustainability becomes increasingly great because the process approaches reversibility (i.e., no losses occur so the process can go forward and backward indefinitely).
- As exergy efficiency approaches 0%, sustainability approaches zero because exergy-containing resources are used but nothing is accomplished. Also, environmental impact becomes increasingly great because to provide a fixed service, an ever-increasing

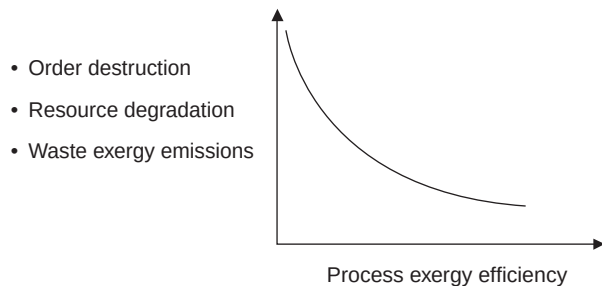


FIGURE 2 Qualitative illustration of the relation between the exergy efficiency of a process and the associated environmental impact in terms of order destruction, resource degradation, or waste exergy emissions.

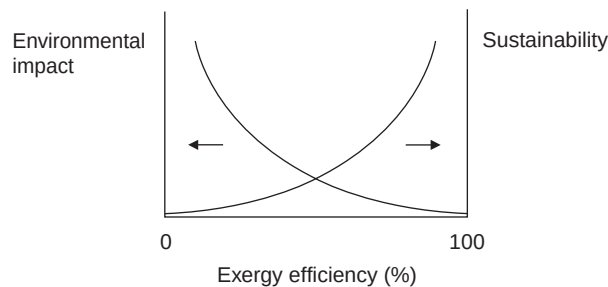


FIGURE 3 Qualitative illustration of the relation between the environmental impact and sustainability of a process and also its exergy efficiency.

quantity of resources must be used and a correspondingly increasing amount of exergy-containing wastes are emitted.

4.6 Illustrative Example

The ideas discussed in this section are demonstrated for a unit of a typical 500-MWe coal-fired electrical generating station. Overall balances of exergy and energy for the station are illustrated in Fig. 4. This station was examined in the article titled “Exergy Analysis of Energy Systems.” For this station, each of the three relationships between exergy and environmental impact described in the previous subsections are demonstrated:

- Waste exergy emissions from the plant occur with (i) the gaseous wastes exiting through the stack, (ii) the solid wastes exiting the combustor, and (iii) waste heat released both to the lake, which supplies condenser cooling water, and to the atmosphere. Contained within each of these emissions is a potential to impact on the environment. The societal

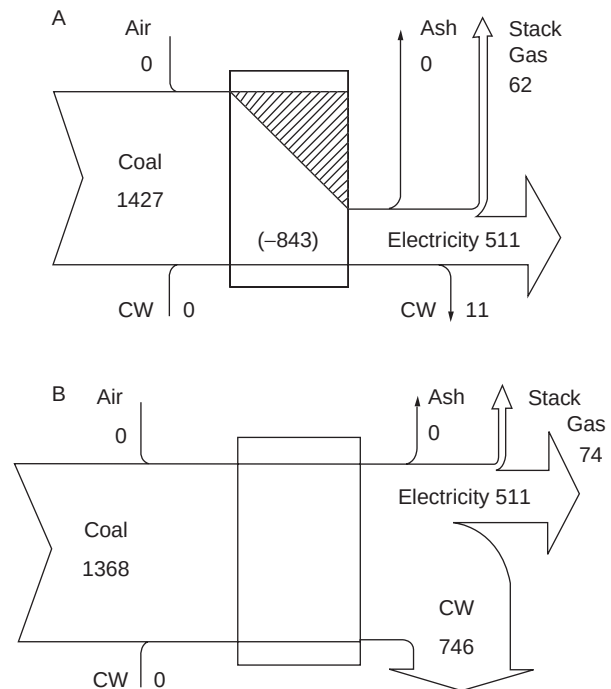


FIGURE 4 Overall energy and exergy balances for a unit of a typical 500-MWe coal-fired electrical generating station. The rectangle in the center of each diagram represents the station. Widths of flow lines are proportional to the relative magnitudes of the represented quantities. CW, cooling water. (A) Exergy balance showing flow rates (positive values) and consumption rate (negative value, denoted by hatched region) of exergy (in MW). (B) Energy balance showing flow rates of energy (in MW).

consensus regarding emissions of harmful chemical constituents in stack gases and “thermal pollution” in local bodies of water indicates that the potential for impact of these emissions is already recognized but not from an exergy perspective.

- During the electricity generation process, a finite resource (i.e., coal) is degraded as it is used to drive the process. Although a degree of degradation is necessary for any real process, increased process efficiency can reduce this degradation for the same services or products. In the extreme, if the process considered in the example were made ideal, the exergy efficiency would increase from 37 to 100%, and coal use as well as related emissions would decrease by more than 60%. These insights are provided by the exergy, but not energy, considerations.

- The more abstract concepts of order destruction and chaos creation are also illustrated. Certainly the degradation of coal to stack gases and solid wastes represents a destruction of order, especially since less exergy is present in the products of the process. The emissions of wastes to the environment increase the chaos of the world by allowing the constrained material and heat products of combustion to be released without constraint into the environment.

5. ENVIRONMENTAL IMPACT OF EXERGY LOSSES

As mentioned earlier, exergy losses, which are made up of exergy destruction and waste exergy emissions, have a significant effect on environmental impact. Exergy destruction, in particular, can be used as one of the most significant criteria for assessing the depletion of natural resources. Exergy analysis can assist efforts to minimize the use of natural resources because it can indicate where the work potential of natural resources in relation to the surrounding environment is lost (i.e., where irreversibility, or exergy destruction, occurs).

The exergy in a flow can only be entirely converted to products in a fully reversible process since there are no exergy destructions within the system and no waste exergy emissions to the environment. In real-world processes, which are irreversible, both destruction of the exergy of resources and waste exergy emissions occur. Much effort is being expended on reducing resource exergy destructions and eliminating waste exergy emissions—often by converting them into useful pro-

ducts. The reversible process noted earlier represents a theoretical ideal that we can strive toward but never actually achieve.

A waste emission possesses exergy as a result of its being in a state of mechanical, thermal, and/or chemical disequilibrium with the reference environment. It is useful here to follow the standard separation of exergy into two components—physical and chemical. In this article of exergy and the environment, the exergy of an emission attributable to mechanical and thermal disequilibrium is not considered significant because the potential impact on the environment of physical exergy is limited. That is, pressure differences between an emission and the environment normally dissipate soon after the emission reaches the environment, and temperature differences are normally localized near the emission source (e.g., thermal pollution in the region of a lake near the cooling water discharge of a thermal power plant) and can be controlled. The exergy of an emission due to chemical disequilibrium (i.e., the chemical exergy), on the other hand, is often significant and not localized. The chemical exergy of emissions is normally of more concern.

6. COMPARISON OF EXERGY AND OTHER INDICATORS FOR THE ENVIRONMENTAL IMPACT OF WASTE EMISSIONS

The exergy of waste emissions has been compared to some other measures to assess or control the potential environmental impact of emissions, including air emission limits established by the government of Ontario, and two proposed quantifications of “environmental costs” for emissions from fossil fuel combustion. These comparisons are carried out to detect trends and patterns that may permit the exergy of a substance to be used as a useful predictor of potential environmental impact and consequently as a tool in establishing emission limits that are rationally based rather than formulated by trial and error.

6.1 Background on Pollution Limits and Costs

Air pollution limits in Ontario, Canada, are covered by that province’s principal legislation governing environmental quality, the Environmental Protection Act. The overall goal of the legislation is to ensure environmental conditions such that human health

and the ecosystem of the earth are not endangered. In Ontario, the Ministry of the Environment is the governing body that oversees the development and implementation of environmental legislation for industry. For the control of air pollution, allowable air emission limits (i.e., allowable mass of pollutant per volume of air averaged over a specified time period) that must be achieved prior to discharge (e.g., prestack) are provided for numerous substances. The potential of a substance to impact on the environment is evaluated by 10 parameters: transport, persistence, bioaccumulation, acute lethality, sublethal effects on mammals, sublethal effects on plants, sublethal effects on nonmammalian animals, teratogenicity, mutagenicity/genotoxicity, and carcinogenicity. With this information, point of impingement (POI) air emission limits are determined through a method known as the best available pollution control technology.

6.2 Environmental Costs of Emissions to Air

Two methods for developing environmental costs for emissions to air are considered here:

- For air emissions resulting from the combustion of fossil fuels, consider the cost of removing the pollutants from the waste stream prior to discharge into the environment. This cost can be related to the exergy of the pollution and is referred to as the removal pollution cost (RPC). The removal costs for waste emissions are evaluated as the total fuel cost per unit fuel exergy multiplied by the chemical exergy per unit fuel exergy and divided by the exergy efficiency of the pollution removal process. The exergy efficiencies for the removal of pollutants from waste streams vary. Some sources indicate that exergy efficiencies are less than 5% when the removal process involves mechanical separation. For simplicity, exergy efficiencies of 1% for all pollutants are used here.
- Estimate environmental costs of pollutant, which are referred to here as environmental pollution costs (EPCs). Such work is most advanced for emissions released into the atmosphere resulting from the combustion of fossil fuels (e.g., carbon dioxide and sulfur dioxide). Environmental costs for some of these emissions have been estimated for Canada (Table II). Values for EPCs are based on quantitative and qualitative evaluations of the cost to society for the correction or compensation of environmental damage and/or the cost incurred to

TABLE II

Environmental Pollution Costs (EPCs) for Selected Pollutants

Pollutant	EPC (1998 Can. \$/kg pollutant) ^a
Particulates ^b	4.11
CO	3.71
NO _x	2.91
SO ₂	2.65
CH ₄	0.97
Volatile organic compounds	0.45
CO ₂	0.030

^aThe EPCs are based on values from 1990, with an adjustment applied to the dollar values to account for inflation in Canada between 1990 and 1998 (as measured by the Consumer Price Index for all products). Statistics Canada reports the adjustment factor as 1.164.

^bIncluding heavy metals (e.g., lead, cadmium, nickel, chromium, copper, manganese, and vanadium).

prevent a harmful emission from escaping into the environment.

6.3 Relations among the Measures

POI air emission limits, standard chemical exergies, RPCs, and EPCs for emissions have been studied and some preliminary relations discerned. For instance, evidence suggests that EPC increases with increasing standard chemical exergy. Also, EPCs generally appear to increase at a decreasing rate with increasing percentages of pollution emission exergy. These general trends are expected and logical.

Overall, the two measures considered here for the environmental cost of pollutants (RPCs, which are based on the cost to remove a pollutant from the environment, and EPCs, which are based on the costs attributable to the negative impact of the pollutant on the environment), although based on different principles and evaluated differently, are of the same order of magnitude for a given pollutant. The RPC methodology is based on a simple, theoretical concept, whereas the EPC methodology relies on subjective interpretations of environmental impact data. Thus, exergy-based measures for environmental impact may provide a foundation for rational environmental indicators and tools.

7. CONCLUSIONS

The relations between environmental impact and exergy, and between the chemical exergy values of

substances found in waste emissions and measures of environmental impact, could contribute to the development of simple, rational, and objective procedures for assessing the harmful effects on the environment and predicting the potential for environmental impact for a substance. This view is primarily based on the general premise that a substance has exergy due to its disequilibrium with respect to a reference environment, and therefore the exergy of unconstrained waste emissions has the potential to impact on the environment. The potential usefulness of exergy analysis in addressing and solving environmental problems is significant, but further work is needed if utilizable, objective exergy-based measures of the potential of a substance to impact on the environment are to be developed.

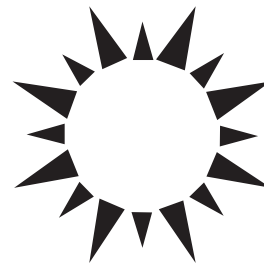
SEE ALSO THE FOLLOWING ARTICLES

Entropy • Exergy • Exergy Analysis of Energy Systems • Exergy: Reference States and Balance Conditions • Thermodynamic Sciences, History of • Thermodynamics, Laws of • Waste-to-Energy Technology

Further Reading

- Ayres, R. U., Ayres, L. W., and Martinas, K. (1998). Exergy, waste, accounting, and life-cycle analysis. *Energy Int. J.* **23**, 355–363.
- Berthiaume, R., Bouchard, C., and Rosen, M. A. (2001). Exergetic evaluation of the renewability of a biofuel. *Exergy Int. J.* **1**, 256–268.
- Connelly, L., and Koshland, C. P. (2001). Exergy and industrial ecology. *Exergy Int. J.* **1**, 146–165, 234–255.
- Creyts, J. C., and Carey, V. P. (1997). Use of extended exergy analysis as a tool for assessment of the environmental impact of industrial processes. In “Proceedings of the ASME Advanced Energy Systems Division” (M. L. Ramalingam, J. L. Lage, V. C. Mei, and J. N. Chapman, Eds.), AES Vol. 37, 129–137. American Society of Mechanical Engineers, New York.
- Edgerton, R. H. (1982). “Available Energy and Environmental Economics.” Heath, Toronto.
- Frangopoulos, C., von Spakovsky, M. R. (1993). A global environomic approach for energy systems analysis and optimization (parts 1 and 2). In “Proceedings of the International Conference on Energy Systems and Ecology, Cracow, Poland,” pp. 123–144.
- Gunnawiek, L. H., and Rosen, M. A. (1998). Relation between the exergy of waste emissions and measures of environmental impact. *Int. J. Environ. Pollut.* **10**(2), 261–272.
- Kotas, T. J. (1995). “The Exergy Method of Thermal Plant Analysis.” Reprint ed. Krieger, Malabar, FL.
- Moran, M. J. (1989). “Availability Analysis: A Guide to Efficient Energy Use.” Rev. ed. American Society of Mechanical Engineers, New York.
- Reistad, G. M. (1970). Availability: Concepts and applications, Ph.D. dissertation, University of Wisconsin, Madison.
- Rosen, M. A. (1999). Second law analysis: Approaches and implications. *Int. J. Energy Res.* **23**, 415–429.
- Rosen, M. A., and Dincer, I. (1997). On exergy and environmental impact. *Int. J. Energy Res.* **21**, 643–654.
- Rosen, M. A., and Dincer, I. (1999). Exergy analysis of waste emissions. *Int. J. Energy Res.* **23**, 1153–1163.
- Rosen, M. A., and Dincer, I. (2001). Exergy as the confluence of energy, environment and sustainable development. *Exergy Int. J.* **1**, 3–13.
- Sciubba, E. (1999). Exergy as a direct measure of environmental impact. In “Proceedings of the ASME Advanced Energy Systems Division” (S. M. Aceves, S. Garimella, and R. Peterson, Eds.), AES Vol. 39, pp. 573–581. American Society of Mechanical Engineers, New York.
- Szargut, J., Morris, D. R., and Steward, F. R. (1988). “Exergy Analysis of Thermal, Chemical and Metallurgical Processes.” Hemisphere, New York.
- Tribus, M., and McIrvine, E. C. (1971). Energy and information. *Sci. Am.* **225**(3), 179–188.
- Wall, G., and Gong, M. (2001). On exergy and sustainable development. *Exergy Int. J.* **1**, 128–145, 217–233.

Exergy: Reference States and Balance Conditions



ROBERT U. AYRES

Center for the Management of Environmental Resources
Fontainebleau, France
Chalmers Institute of Technology
Gothenburg, Sweden

ANDREA MASINI

London Business School
London, United Kingdom

-
1. Standard Formalism for Chemical Exergy Calculations
 2. Sink-Specific Reference Species
 3. Exergy Balance as a Tool for Analysis
 4. Exergy as a Measure of Resource Availability or Losses

Glossary

exergy The maximum amount of useful work that can be extracted from a material as it approaches thermodynamic equilibrium with its surroundings by reversible processes. The second law of thermodynamics implies that it is never possible to utilize all available or potentially useful work for practical purposes. There is some waste or loss of exergy in every process. Lost exergy corresponds to increasing entropy.

Gibbs energy of formation The total amount of energy that is either used up or released during a chemical reaction. Gibbs free energy $(\Delta G) = (\Delta H) - t(\Delta S)$, where (ΔH) is the change in enthalpy; t is the temperature at which the reaction took place; and (ΔS) is the change in entropy, or amount of disorder, that occurs in the molecules involved during the reaction.

reference species One or more of the common chemical forms that characterize the final equilibrium state of a given "mix" of chemical elements, taking into account relative atomic abundance. The equilibrium state is what would finally result from free, spontaneous, and uninhibited interactions with all the other elements in the mix.

The standard chemical exergy, per mole, of any pure compound involving chemical elements can be

computed by means of a simple formula. To calculate the chemical exergies of mixtures and composite materials, such as ores and alloys, it is only necessary to have detailed composition data and do the sums. To compile generic exergy tables, however, it is necessary to adopt a general convention on reference states.

1. STANDARD FORMALISM FOR CHEMICAL EXERGY CALCULATIONS

The standard chemical exergy, per mole, of any pure compound involving these elements can be computed by means of a simple formula, namely

$$B = G + \sum_j n_j b_j, \quad (1)$$

where G is the standard Gibbs free energy of formation of the compound, the n_j are molar fractions of the j th chemical element, and the index j runs over the elements in the compound. The G values are tabulated in standard reference works, such as the *Handbook of Physics and Chemistry*. The b_j values have been calculated for all the important elements. To calculate the chemical exergies of mixtures and composite materials, such as ores and alloys, it is only necessary to have detailed composition data and do the sums.

To compile generic exergy tables, however, it is necessary to adopt a general convention on reference

states. This could be a daunting task if a very local definition were adopted for each case. However, noting that there are three major environmental media—or “sinks”—a reasonably general definition has been proposed. In the case of chemical elements found in the atmosphere as such (oxygen, nitrogen, inert gases) and carbon—because carbon reacts with atmospheric oxygen to yield a gaseous oxide (CO_2)—the composition of the atmosphere is taken to be the reference state. In the case of chemical elements whose oxides or chlorides are soluble in water, the composition of the ocean is taken to be the reference state. For all other elements, mainly metals whose oxides are insoluble solids, the average composition of the earth’s crust down to some specified depth is taken to be the reference state. (The technical problem here is that the earth’s crust is not in chemical equilibrium with either the atmosphere or the oceans. In particular, the earth’s crust contains significant quantities of hydrocarbons and sulfides, of biological origin, that are potentially reactive with oxygen but that are not exposed to the atmosphere. Obviously in the very long run, in the absence of biological activity, these materials would be oxidized and all of the free oxygen in the atmosphere would ultimately be combined with nitrogen, carbon, or sulfur. Thus, the essence of the convention is to assume that the three media [air, oceans, and crust] are in chemical equilibrium in the short run—that is, in terms of the time scales of interest.) The latter definition is somewhat arbitrary, to be sure, but it turns out that calculated values in most cases are not extremely sensitive to the crustal depth assumption.

A procedure has been developed that results in a computed “standard” exergy for each material that is independent of the actual sink. This is done by assuming each element is obtained by successive reactions from a relatively stable chemical form that is chosen once and for all for each element and identified as its reference species. The reference species should be one of the more common chemical forms that would characterize the final equilibrium state of a given “mix” of chemical elements, taking into account relative atomic abundance. (The equilibrium state is what would finally result from free, spontaneous, and uninhibited interactions with all the other elements in the mix. This idealized situation is sometimes called a Gibbs reactor.) The reference state for most elements is taken to be either its most oxidized or chlorinated form, depending on volatility and solubility. To carry out this procedure the final compounds must be identified, along with

their thermodynamic parameters, and specified at standard temperature (298.15°K) and atmospheric pressure (101.325 kPa).

It is normal practice in the thermodynamics of closed systems to assign a zero value for enthalpy of formation of pure elements. However, this convention makes calculation awkward for open systems. Instead, Szargut has proposed that the zero level should be assigned to the (unique) reference species. (Szargut *et al.* have identified 10 gaseous species for 10 chemical elements, 31 ions dissolved in seawater for 31 chemical elements and other reference species for the same number of elements.) Then the standard exergy for any chemical will be defined as the maximum work that can be extracted when that chemical is transformed by successive reactions into the reference species in the standard state of temperature and pressure.

There is also an adjustment for the reference species themselves, since their natural states are not the same as the standard state. For instance, the natural state is often one with a much lower vapor pressure due to low concentration. Thus, the standard exergy of the reference species in the natural state is the work that can be extracted in going from the standard state to the natural state. Thus, the standard chemical exergy of oxygen is bigger than the chemical exergy of atmospheric oxygen, because the latter is at lower (partial) pressure. Calculation of the chemical exergy for reference species depends on whether the species is found in the atmosphere, in seawater or in the earth’s solid crust.

Implementation of the preceding scheme for a compound requires the following steps:

1. Calculation of the standard exergy of each of the elements $b_{ch,el}^0$ from the corresponding equation representing its formation (by a reversible reaction) from the reference species. (These equations are equations of formation of the reference species from elements, expressed in reverse.) Since there is just one reference species per element, the system of equations has a unique set of solutions. As noted already, reference species should be common components of the environment. Szargut has selected 10 reference species found in the atmosphere, 31 in the ocean, and 13 species in the earth’s crust. The equation is as follows:

$$b_{ch}^0 = \Delta G_f^0 + \sum_{el} n_{el} b_{chel}^0, \quad (2)$$

where G_f is the Gibbs energy of formation of the target compound, b_{ch}^0 is the chemical exergy of the target compound per mole, b_{chel}^0 is the chemical exergy

of the component element (previous tabulated), and n_{el} is the number of atoms of the element in the compound.

2. To correct for differences between standard conditions and actual conditions and abundance in the three main media, we use one of the following three equations. For atmospheric species,

$$b_{cb}^0 = -RT_n \ln \frac{p_n}{p_{0n}}, \quad (3)$$

where p_n is the normal pressure (101.325 kPa) and p_m is the partial pressure of the compound in the atmosphere.

For reference species in the earth's crust,

$$b_{cb}^0 = RT_n \ln x_{in}, \quad (4)$$

where x_{in} is the average concentration (mole fraction) of the i th substance in the earth's crust.

In the case of reference species dissolved in seawater, the approach is slightly different. The following formula allows the direct calculation of the standard chemical exergy of the pure element contained in the reference species:

$$\begin{aligned} b_{cb}^0 = & -j\Delta_f G^0 + \frac{1}{2}jz b_{ch, H_2}^0 \\ & - j \sum n_k b_{ch, k}^0 - 2.303RT_n jz(pH) \\ & - jRT_n \ln m_n \gamma, \end{aligned} \quad (5)$$

where $\Delta_f G^0$ is the standard normal free energy of formation of the reference ion, j is the number of reference ions derived from one molecule of the element under consideration, z is the number of elementary positive charges in the reference ion, n_k is the number of moles of additional elements in the reference ion, m_n is the conventional molarity of the reference species in seawater, and γ is the activity coefficient of the reference species in seawater.

2. SINK-SPECIFIC REFERENCE SPECIES

Szargut's simplified approach is convenient for many purposes. However, it may be inadequate for purposes of assessing potential environmental harm from pollutants, where localized sink-specific reactions are typically important. For instance, the standard chemical exergy of sulfur in Szargut's scheme is based on the assumption that the end product is the SO_4^- ion (in water), independent of whether the sulfur was originally emitted into air or water. It is important to recall that the atmosphere is

not in equilibrium with the ocean (nor is the ocean in chemical equilibrium with the earth's crust). The potential harm due to atmospheric emissions must be related to the exergy of the pollutant with respect to atmospheric reference species only. Clearly, the real situation is somewhat more complex than Szargut's simplified scheme, at least for some of the major biologically active elements and compounds found in pollutants.

The starting point for a *gedanken* equilibration process is a homogeneous mixture of pure chemical elements in a reactor. The chemical composition of the final "soup" of compounds depends on the elementary composition of the original mixture. This is the primary difference between the atmosphere, hydrosphere, and lithosphere. Each chemical element in the starting mix must be used up in the final soup of compounds. For instance, hydrogen tends to combine with oxygen forming water (H_2O) whenever possible. Thus, most of the hydrogen in the earth's crust—including the oceans and atmosphere—is in the form of water (including its ionized forms H^+ and OH^-). Similarly, most of the carbon is also combined with oxygen. This is in the form of gaseous CO_2 in the air, carbonic acid H_2CO_3 (in water), or one of its metallic salts such as calcium or magnesium carbonate in the earth's crust. (The only hydrogen and carbon that are normally found on the surface of the earth [i.e. in the atmosphere, oceans, sedimentary, or metamorphic rocks] that are not on one of these forms are in carbohydrates, fats, or hydrocarbons resulting [directly or indirectly] from photosynthesis. Fossil hydrocarbons, including coal, are derived from previously living matter. Graphite is mostly of metamorphic origin, from coal or other sedimentary hydrocarbon deposits. Hydrogen and carbon are very rare in igneous rock, although some hydrogen sulfide and ammonia [presumably dissolved under pressure in magma] are emitted by volcanoes.)

To get a feel for the situation, suppose the elements were mixed together initially in a gaseous soup but at a low enough temperature to permit stable molecules to survive. Assume that various reactions occurred in sequence, rather than simultaneously. The first step in the sequence might be for the two most mutually attractive elements to combine preferentially with each other. This reaction would continue until the supply of one of the reactants is exhausted. Then, the two most reactive of the remaining elements would combine until one of them is used up, and so on. The procedure continues, in principle, until every possible

(exothermic) reaction has occurred. The final result is a chemical equilibrium in which at least one and possibly several of the most reactive elements are found with only one partner, although that partner may combine preferentially with several others. (Thus, both hydrogen and carbon are found almost exclusively with oxygen, except when they combine with each other. Oxygen—being more abundant—is also the preferred partner of many other elements.) In a true equilibrium it makes sense (in most cases) to choose the most abundant compound of each element as the reference species for that element.

However, this simplified mental model of a sequential reactor fails to allow for a number of real-world factors. In the first place, one of the major combinations resulting from the vapor-phase reactions is water, which condenses at a relatively high temperature, whereas quite a few other light molecules, including hydrogen, methane, ammonia, hydrogen sulfide, and carbon dioxide, as well as molecular nitrogen, oxygen, and the rare gases, remain in vapor phase at the same pressures and temperatures. Thus, all known planets have volatile atmospheres as well as solid cores. The presence of liquid water on earth is a result of the specific combination of temperatures and pressures that make it possible. It also depends on the peculiar fact that frozen water is slightly less dense than liquid water and therefore floats rather than sinking. Absent this unique characteristic of water, polar ice would sink and accumulate on the Arctic and Antarctic ocean floors, gradually cutting off ocean currents. This would eliminate the major mechanism for cooling the tropics and warming the polar regions. The result would be hotter tropics, colder polar regions, and much more violent storms resulting from the temperature gradient. The storms would carry water vapor from the tropics to the poles, where it would condense as ice. It would also make the remaining tropical oceans more and more saline (like the Dead Sea). The end result would be a stable world of glaciers and salt.

Water is an extraordinarily powerful ionic solvent for some alkali metals and halogens, as well as sulfates and nitrates. But many other molecular combinations, especially of metals, do not dissolve to a significant degree in water. Nor do some of the major families of hydrocarbons and lipids. In short, the liquid water on the earth's surface constitutes a kind of filter, which efficiently separates soluble from insoluble species. The latter, of course, build up into sedimentary layers.

The composition of earth's atmosphere is unlike that of the dead planets. Both molecular oxygen and molecular nitrogen exist in the earth's atmosphere only because of biological processes. The role of photosynthesis in maintaining the carbon cycle is very well known. The corresponding role of denitrification organisms is less well understood. But without them, the atmospheric nitrogen would gradually combine with oxygen (thanks to electrical storms) and would be dissolved in the ocean as nitric acid.

Evidently the three major sinks (the atmosphere, the oceans, and the sedimentary layers of the earth's crust) are self-organized systems in a stable state that is far from thermodynamic equilibrium. Such systems exist thanks to complex biological and geological processes driven by a continuous flux of solar exergy, via photosynthesis and denitrifying bacteria.

However, there are problems with the simple theory. An example will help to clarify some points. When another element or compound (say sulfur) is injected into the atmosphere, it reacts with oxygen until it is entirely consumed. The usual product is SO_2 . However, the SO_2 either reacts with water vapor to form sulfurous acid H_2SO_3 or it oxidizes to SO_3 and subsequently reacts with water to form sulfuric acid H_2SO_4 . These acids either wash out in rain or react with ammonia to form ammonium sulfite or sulfate, which also washes out in rain. The acid rain causes corrosion on metal surfaces and changes the pH of soil and freshwater lakes. Eventually most of the sulfur does indeed end up as SO_4^- ion in the seawater. (But seawater is not a permanent residence for sulfur: the sulfate ion gradually combines with calcium or magnesium ions and precipitates into sediments.) Yet much of the potential environmental harm occurs in intermediate environments where the sulfur *enters* as sulfate ion and the most likely final (reference) state is probably calcium sulfate.

Curiously enough, the most abundant sulfur compound in the atmosphere is not sulfur dioxide, as one might expect; it is an organic compound carbonyl sulfide (COS), probably of biological origin. Yet there is a possibility that this compound is produced in the atmosphere by a reaction with methane such as $\text{SO}_3 + \text{CH}_4 \rightarrow \text{COS} + 2\text{H}_2\text{O}$. Unfortunately one cannot easily perform the relevant exergy calculations, due to unavailability of data on Gibbs energy of formation of COS .

In the case of sulfur, one might propose not just one but three reference states, depending on the medium of interest. In the atmosphere itself the

reference state would be either SO_2 or possibly COS ; in seawater it would be SO_4^{2-} ion, and in the (sedimentary) lithosphere it would be calcium sulfate, probably in the form of the mineral gypsum $\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$. In the case of sulfur (and other elements) another complication arises, namely that the lithosphere is anything but homogeneous. In particular, there are vast differences between the sedimentary or metamorphic component of the earth's crust (the part that has resulted from weathering and has been exposed to the atmosphere at some stage) and the unweathered igneous rock left over from the earth's original solidification. Most sulfur in nonsedimentary igneous rocks, however, is combined with metals as sulfides (MS , MS_2 , etc.). As these are exposed to water or air (e.g., by mining operations), the sulfides gradually oxidize to sulfuric acid or sulfur dioxide.

2.1 Standard Theory

The standard theory calculates standard chemical exergy of sulfur by adopting SO_4^{2-} dissolved in water as the reference species. Calculations are based on Szargut's analysis and give the following result:

$$b_{cb}^0[\text{S}] = 609.6 \text{ kJ/mole.} \quad (6)$$

2.2 Sink-Specific Approach

Three independent measures may be computed for the atmosphere, the seawater, and the lithosphere as follows:

Atmosphere. Apart from COS (mentioned earlier), the most abundant compound containing sulfur in the atmosphere is sulfur dioxide, SO_2 . Its mean volume concentration in the atmosphere has been estimated as about $2 \cdot 10^{-8}$. From Eq. (2) we obtain for the reference species

$$b_{cb}^0[\text{SO}_2] = 40.25 \text{ kJ/mole.} \quad (7)$$

The formation of sulfur from SO_2 (or vice versa) can be written as follows



Exergy does not have to be recalculated for oxygen since the original reference species selected by Szargut was atmospheric O_2 . Thus, applying an exergy balance one obtains

$$b_{cb}^0[\text{S}] + b_{cb}^0[\text{O}_2] + \Delta G_r^0 = b_{cb}^0[\text{SO}_2], \quad (9)$$

and

$$\begin{aligned} b_{cb}^0[\text{S}] &= b_{cb}^0[\text{SO}_2] - b_{cb}^0[\text{O}_2] - \Delta G_r^0 \\ &= 107.96 \text{ kJ/mole.} \end{aligned} \quad (10)$$

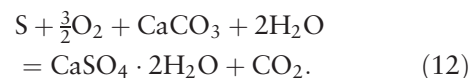
Mean residence time of sulfur dioxide in the atmosphere is fairly low (about 10^{-2} years), thus choosing this substance as a reference species is quite inappropriate. It is reasonable to assume that sulfur-containing compounds easily evolve toward other chemical forms, which produce SO_4^{2-} dissolved in seawater.

Seawater. The chosen reference species is the ion SO_4^{2-} . As noted previously, Szargut calculated $b_{cb}^0[\text{S}] = 609.6 \text{ kJ/mole}$. However, Szargut assumed that the standard chemical exergy for oxygen in seawater was equal to that of atmospheric oxygen. Taking into account the lower concentration of oxygen in seawater, standard chemical exergy of O_2 increases to 21.11 kJ/mole . In principle, the exergy of sulfur should be recalculated. However, for some compounds it seems reasonable to assume (as Szargut does) that reactions actually occur in the vapor phase, even when considering ions dissolved in seawater or solid compounds.

Lithosphere. Szargut's chosen reference species is gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$). Unfortunately, no data on the mass fraction of this mineral are available for calculating its standard chemical exergy as a reference species. However, if an estimated average concentration of 10^{-4} is assumed, we obtain (from Eq. 3)

$$b_{cb}^0[\text{CaSO}_4 \cdot 2\text{H}_2\text{O}] = 22.83 \text{ kJ/mole.} \quad (11)$$

The following reaction can be identified for the formation of sulfur from gypsum:



CaCO_3 is another reference species for the lithosphere and its chemical exergy is 1.0 kJ/mole . Assuming that oxygen and carbon dioxide participate in the reaction in gaseous form and that a liquid reference state may be assumed for water, one obtains

$$\begin{aligned} b_{cb}^0[\text{S}] &= b_{cb}^0[\text{CaSO}_4 \cdot \text{H}_2\text{O}] \\ &+ b_{cb}^0[\text{CO}_2] - \frac{3}{2}b_{cb}^0[\text{O}_2] \\ &- b_{cb}^0[\text{CaCO}_3] - \Delta G_r^0 \\ &= 171.27 \text{ kJ/mole.} \end{aligned} \quad (13)$$

Thus, Szargut calculates the standard exergy for sulfur as 609.6 kJ/mole (assuming sulfate SO_4^{2-} in sea

water as the reference state), but a sink-specific approach yields very different figures, namely 107.96 kJ/mole to atmospheric SO_2 (not a good choice, however, because of its short life) and 171.27 kJ/mole to gypsum in the lithosphere.

The preceding example can be extended to many other elements. For instance, most of the silicon in the sedimentary part of the earth's crust is also combined with oxygen as silica (SiO_2). This is the logical reference species for silicon. In nonsedimentary rock, it is usually in one of the reduced metallic silicate forms MSiO_3 or M_2SiO_3 . These silicates spontaneously react with carbonic acid H_2CO_3 to form silica SiO_2 and corresponding hydroxide MOH or $\text{M}(\text{OH})_2$. The latter reacts further with any acid in the environment to form a metallic carbonate, sulfate, or chloride. Carbonates obviously predominate. Aluminum is always combined with oxygen (as alumina, Al_2O_3), even in igneous rock, so this is the logical reference species. Alumina and silica are sometimes combined in a single mineral (kyanite, Al_2SiO_5).

Sodium, potassium, and calcium in the lithosphere are normally oxidized as Na_2O , K_2O , and CaO . These are often combined in more complex minerals, such as feldspars. These are major rock-forming minerals that constitute up to 60% of many igneous rocks metal. They include potassium feldspars (KAlSi_3O_8), sodium feldspars (albite, $\text{NaAlSi}_3\text{O}_8$), calcium feldspars (anorthite, $\text{CaAl}_2\text{Si}_2\text{O}_8$), and all combinations. The oxides are convenient reference species for all three metals.

Phosphorus in the lithosphere is only found in one mineral, apatite [$\text{Ca}_3(\text{PO}_4)_2$], of biological origin. Apatite is subject to weathering by carbonation, much the same way in which silicates weather. This process releases phosphoric acid H_3PO_4^- or HPO_4^{2-} . These would be natural reference species in water. However, in soil much of this phosphoric acid is bound (occluded) into complexes with iron and aluminum, so the latter would be the logical reference species in the lithosphere. Iron is mostly mined from oxide deposits, also of biological origin, but iron is often partially oxidized (ferrous iron) even in igneous rock. The likely (most abundant) reference species is Fe_2O_3 . Among the other major metals (chromium, copper, lead, manganese, nickel, zinc, etc.), the heavier metals are found mainly as sulfides in igneous rock; however, Szargut chose reference species in the lithosphere that are all oxides.

The halogens (F, Cl, Br, I) combine preferentially with the alkali metals, sodium, potassium, magnesium, or calcium. The halogenated sodium, potas-

sium, and magnesium compounds are quite soluble in water and are mostly accumulated in the oceans, except for evaporite deposits. The logical reference species for the first three are simply the ionic forms; F^- , Cl^- , and Br^- . For iodine it is IO_3^- . In igneous rock, these salts are only found in trace percentages. However, for chlorofluorocarbons (CFCs), which are accumulating in the atmosphere, a specialized sequence of reactions—involving diffusion into the stratosphere and ionization by UV—may suggest a different exergy calculation.

Nitrogen is a special case. Molecular nitrogen in the atmosphere results only from biological processes, namely bacterial denitrification of organic nitrogen compounds. Nitrates are stored in sediments as organic material. They are found in igneous rock only in trace percentages. In principle, molecular nitrogen tends to combine with oxygen as nitrogen oxide (NO), nitrogen dioxide (NO_2), and eventually nitric acid (HNO_3), which reacts with ammonia in the atmosphere or dissolves in water and washes out as acid rain. The acid, in turn, reacts with most metallic hydroxides (in water) to form the corresponding nitrates. However, as mentioned already, the oxidation reactions in air are negligibly slow at normal temperatures and pressures. (They do occur more rapidly at high temperatures and pressures).

In the circumstances, it probably makes sense from an environmental standpoint to assume that one of the alkali metal nitrates [$\text{Ca}(\text{NO}_3)_2$, NaNO_3 or KNO_3] is also the reference species for nitrogen, contrary to Szargut's view, despite the fact that other compounds of these elements are much more abundant in the lithosphere. This may lead to some confusion, but as long as it is understood to apply only to the case of nitrogen compounds (and not to other compounds of the alkali metals, it should be possible to maintain reasonable consistency.

To summarize, it can be argued that a substance with a long—but not infinite—residence time in the atmosphere (or the ocean) might be a candidate for treatment *as if* it were a permanent resident. The argument is that when a compound remains within an environmental medium for a significant time, but subsequently disappears from that medium (whether by transformation, diffusion, precipitation, or wash-out), it may nevertheless be considered as a reference species within that medium. Candidates for atmospheric reference species might include nitrous oxide (150 years), methane (10 years), carbonyl sulfide (5 years), especially in comparison with carbon monoxide (60 days), sulfur dioxide (14 days), or hydrogen sulfide (4 days).

3. EXERGY BALANCE AS A TOOL FOR ANALYSIS

Unlike energy, exergy is not a conserved variable. It can be gained or lost in physical processes. However, exergy can be accumulated or depleted. It can also be stored in mineral ores or fossil fuels, for instance. Finally, as noted previously, exergy inflows and outflows to and from any subsystem are definable and measurable. (Computational aspects are discussed earlier.)

In the case of hydrocarbon fuels and carbohydrates such as cellulose and sugars, the exergy content of a fuel is very closely related to the usual measure (heat-of-combustion or enthalpy). However, as noted already, the term exergy is much more widely applicable. Given the previously stated conventions, it is definable for any material substance whatsoever, whether combustible or not. For instance, the exergy content of a metal ore (i.e., the work required to separate ore of that grade from the earth's crust) reflects the amount of natural "low entropy" that is stored in the ore. This would be lost if the ore were dissipated (mixed) uniformly into the earth's crust or into the ocean. The better the quality of the ore, the greater its exergy content.

Exergy, B as defined previously, is not only a natural measure of the resource inputs to an economic system. It is also a measure of the material outputs. Exergy is lost in all processes, mostly—but not entirely—as low temperature heat. But some exergy is lost in the form of chemically or physically reactive materials. These production or consumption wastes are dissipated into the environment, where the insertion of unfamiliar chemical species (i.e., chemical potentials) in delicately balanced biogeochemical cycles can cause trouble. At the micro-scale, even small trace amounts of some chemicals are enough to disrupt life processes. In fact, there is a general label for such disruptive chemicals: they are toxins.

To calculate the exergy content of waste materials, in practice, there are two approaches. Ideally they should be used together for mutual verification. The first approach requires a detailed knowledge of the chemical composition of the waste stream and the Gibbs free energies of formation of the components. Once these are known, the calculation of exergy proceeds component by component. The difficulty is obvious: for many chemical and metallurgical processes it is difficult to obtain reliable data on the chemical composition of the wastes. Nevertheless, the composition of the waste stream can be estimated

approximately, in many cases, if the basic chemical reactions, temperatures, pressures, and yields are known. Indeed, commercially available software suitable for use in a desktop computer is capable of making such calculations at minimal cost.

The second, alternative approach involves using the exergy "balance" equation:

$$B_{\text{in}} = B_{\text{product}} + \Delta B_{\text{process}} + B_{\text{waste}}, \quad (14)$$

which can be rewritten

$$B_{\text{waste}} = B_{\text{in}} - \Delta B_{\text{process}} - B_{\text{product}}$$

Here B_{process} is essentially a balancing term. It represents internal exergy lost in the process. Evidently, if B_{process} is known, the exergy content of wastes can be determined directly from the composition of the process inputs (including utilities) and that of the main products. The exergies of process inputs and product outputs are computed, as before, using Eq. (13), when Gibbs free energies of formation are known. However, as a practical matter, the exergy loss (or gain) in the process is typically much larger than the exergy content of the waste. Thus, to use Eq. (15) it is necessary to calculate process losses precisely. This is sometimes feasible on a case-by-case basis, but it is rarely easy.

4. EXERGY AS A MEASURE OF RESOURCE AVAILABILITY OR LOSSES

The use of exergy as a measure of resource stocks and flows for purposes of resource accounting was first proposed by Wall. The use of chemical exergy (or its equivalent, potential entropy) as a measure of potential harm from process wastes was a more recent idea. The latter notion also suggests the possibility that chemical exergy may be a useful indicator of sustainability.

However, while the research in this area so far is suggestive, there is more to be done before the application methodology can be regarded as complete and unambiguously successful. The problem is that the computational methodology (developed by Szargut) assumes that the exergy of a given chemical element is independent of the composition of the environmental sink into which it is injected. This implies that the exergy is independent of what specific reactions are induced in the actual sink. This implication seems reasonable enough if, for instance,

exergy is being used for purposes of resource accounting.

However, if exergy is to be used as a measure of potential environmental harm, it is unreasonable to assume that the chemical composition (and other characteristics) of the particular environmental sink are irrelevant. Hence, this simplifying assumption should be replaced by a methodology for calculating sink-specific measures. In particular, critics have emphasized that waste exergy, as calculated in terms of the three gross sinks—atmosphere, oceans, and lithosphere—is not a valid indicator of environmental toxicity. It is possible that a better measure could be developed if the internal chemical environment of living cells were considered as a fourth major sink. However, considerable research would have to be done to verify this conjecture.

However, although waste exergy is not (yet) a reliable measure of environmental toxicity, exergy analysis can be used to evaluate and compare waste flows and aggregate emissions from period to period or country to country. The exergy content of wastes is not necessarily proportional to the potential environmental harm the wastes may cause, but the exergy content of a waste stream is nevertheless a rough measure of its reactivity in air or water (i.e., its tendency to initiate spontaneous uncontrolled chemical reactions in environmental media). In this regard, one can say that although the exergy content of a particular waste stream is not a measure of human or animal toxicity, it is certainly a better

indicator of its potential for causing harm than is its total mass, for instance.

SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Entropy • Exergoeconomics • Exergy • Exergy Analysis of Energy Systems • Exergy Analysis of Waste Emissions • Thermodynamics and Economics, Overview • Thermodynamics, Laws of

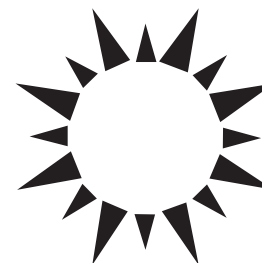
Further Reading

- Ayres, R. U., Martínás, K., and Ayres, L. W. (1998). Exergy, waste accounting and life cycle analysis. *Energy—The Int. J.* 23(5), 355–363.
- Brobst, D. A., and Pratt, W. P. (1973). United States mineral resources. Professional Paper: 820. United States Geological Survey, Washington, DC.
- Glansdorff, P., and Prigogine, I. (1971). “Thermodynamic Theory of Structure, Stability and Fluctuations” (Series). Wiley-Interscience, New York.
- Nicolis, G., and Prigogine, I. (1977). “Self-Organization in Nonequilibrium Systems” (Series). Wiley-Interscience, New York.
- Prigogine, I. (1976). Order through fluctuation: Self-organization and social system. In “Evolution and Consciousness” (E. Jantsch and C. W. Waddington, Eds). Addison-Wesley, New York.
- Szargut, J., Morris, D. R., and Steward, F. R. (1988). “Exergy Analysis of Thermal, Chemical, and Metallurgical Processes” (Series). Hemisphere, New York.

Experience Curves for Energy Technologies

CHRISTINE WOERLEN

Boston University
Boston, Massachusetts, United States



-
1. Definition of Experience Curves for Energy Technologies
 2. Experience Curves
 3. Reasons for Cost Reduction and Consequences for the Formulation of Experience Curves
 4. Problems of Data Identification and Statistical Inference
 5. Some Case Studies
 6. Conceptual Modifications

Glossary

- diffusion** Increase in market share of a new technology after a successful innovation.
- economies of scale** Decrease in unit cost of production through increase in scale of production.
- experience curves** The relationship between unit cost and cumulative deployment as represented by an exponential decay function.
- innovation** First commercial application of a newly developed technology.
- interindustry spillover** Transfer of knowledge, technologies, materials, and the like from one industrial application to another (e.g., the concept of jet engines from aeronautics to gas turbines).
- intraindustry spillover** Transfer of formerly exclusive knowledge and experience from one competitor to another (e.g., by labor mobility).
- learning by doing** Increase in skill level of labor through increasing experience.
- learning curves** The relationship between falling unit cost of production and increasing skill level of labor caused by an increase in experience, represented by an exponential decay function.

Experience curves are used to assess the future cost of a technology as a function of cumulative installations. The underlying rationale is that it is very costly to invent, develop, and prototype a new technology,

process, or product. After an innovation—the first commercial application—a successful technology or product will spread from niche applications to larger markets. In this process of diffusion, unit costs will constantly decrease. This reduction of unit costs over time is described by experience curves.

1. DEFINITION OF EXPERIENCE CURVES FOR ENERGY TECHNOLOGIES

The concept of experience curves was derived from empirically observed learning curves in manufacturing. Learning curves describe the cost reduction effects by learning by doing. Through individual, organizational, and inter- and intraindustry learning, the input of labor necessary to produce a certain product decreases as experience is gained. On the other hand, experience curves describe the cost reduction of a product or technology in general, disregarding the reasons for total cost reduction. Possible reasons for the reduction of unit costs during technology diffusion are economies of scale in all input factors (e.g., intermediates, transaction costs, other overhead), higher rates of production, changes in the production processes, and (particularly in the case of energy technologies) efficiency gains and changes in the product itself.

In the last years, experience curves have drawn considerable attention in the energy field due to a variety of technological challenges. Energy systems in many developing countries are in a phase of expansion and growth. Energy systems in most developed countries go through some restructuring, inducing a change in investor behavior. Both phenomena increase interest in present and future investment costs. Legislators, electricity generators,

and other energy producers are also confronted with environmental demands (e.g., climate change, pollution, nuclear disposal and safety) that require the increased use of new cleanup technologies and new clean energy conversion technologies. Most of these new technologies are in early stages of their diffusion and considerable future cost reduction is expected.

Experience curves also play an important role in energy policy analysis. They are used to assess the “buy-down” cost of a new technology (Fig. 1). To optimize the size of a support program for a specific new technology, experience curves are used to determine how much additional investment is necessary to bring the cost of the technology down to levels at which it is economically competitive with other technologies. To that end, the incremental cost of the new technology and the speed of the cost reduction are determined from empirical data. The integral under the experience curve is equal to the investment required to buy down the cost of the technology between two arbitrary points on the curve.

Experience curves are also widely used in energy scenarios calculations. Investment costs and conversion efficiencies of energy technologies are important driving factors for future energy demand and supply. The most important energy system models, such as those of the U.S. Energy Information Administration and those used in the Intergovernmental Panel on Climate Change’s (IPCC) Special Report on Emission Scenarios, use experience curves for the extrapolation of future energy costs because future investment costs for energy conversion and use are needed as inputs to these models. Because of their low data requirement and relatively robust nature, experience curves are very well suited to generate these inputs.

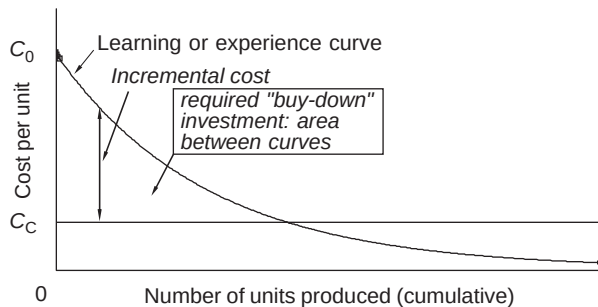


FIGURE 1 Experience curve for an equipment unit of a new energy technology. The unit cost for a piece of equipment falls with increasing cumulative deployment of the technology. C_0 is the cost of a technology at innovation. C_c is the cost of a conventional backstop technology. The integral over the incremental cost up to the point of intersection (“buy-down”) indicates how much additional investment is necessary to make the new technology economically competitive with the established technology.

The more sophisticated energy system models can include dynamic specifications into the model code such that the prices used for optimization of the model are generated during the model runs and depend, in a dynamic way, on the respective deployment figures of the current simulation.

Management consulting firms (e.g., Boston Consulting Group) use experience curves for consulting clients on strategic price-setting behavior. Experience curves describe cost reduction as a “black box.” The actual reasons for the cost reduction are completely ignored in the model. This makes experience curves seemingly simple to use but also imposes some caveats on the use of the concept. The data requirements for an estimation by linear regression are seemingly very low, and extrapolation into the future seems almost trivial. This article first gives more detail on the concept and then discusses the reasons why cost reduction for the whole product or technology might occur. It explains how these reasons should be incorporated into the model and what kind of data problems result. Several examples for cost reduction in various energy technologies and some extensions to the model are discussed.

2. EXPERIENCE CURVES

2.1 The Concept

Experience curves assume a logarithmic relationship between the experience from cumulative installation and cost. Typically, this is represented by the following functional form:

$$C_{\text{cum}} = C_0 * CUM^b, \quad (1)$$

where CUM is the cumulative output of the industry, C_{cum} is the cost of a unit product after output of CUM , b is the “learning index,” and C_0 is the cost at the starting point of the data.

This functional form has some very convenient mathematical properties. For example, the parameter estimates are independent of the starting point of the data used for the estimation. The functional form is completely independent of the stage of development or diffusion of the technology. It implies that the cost reduction per additional unit of output decreases, so decreasing marginal reduction in costs is factored in. However, the functional form also prescribes that there is no limit to cost reduction. Some authors embrace this because they think that human ingenuity and competition will supply the opportunity for continually decreasing cost.

Other authors suggest including a minimum level of cost in the functional form by adding a simple linear term.

The functional form of the experience curve is estimated with linear regression after logarithmic transformation:

$$\log C_{\text{cum}} = \log C_0 + b \log CUM. \quad (2)$$

Empirical studies usually estimate the coefficients of this equation using ordinary least squares and usually report highly significant correlations for the logarithmic form. However, highly significant correlations are to be expected in a logarithmically transformed linear regression of integrated variables and should not lead the investigator to conclusions that are not actually supported by the data. In fact, econometrically tested and statistically significant estimations of the experience curves are hard to find in the literature.

The data requirement for the estimation of Eq. (2) is very small. The data necessary for production cost forecasts (i.e., cumulative installations) are easily available from energy scenario calculations. The high level of aggregation of the data does not require any explicit knowledge about future changes in technological details that are particularly hard to predict. However, some researchers prefer to modify the parameters for time periods in the far future or distant past rather than assuming homogenous cost reduction over several decades.

2.2 The Progress Ratio

The “progress ratio” (*PR*) describes the relative cost after a doubling of cumulative installation. It is a very explicit and obvious benchmark for cost reduction effects. It is calculated from the learning index *b* estimated with the experience curve:

$$PR = 2^b. \quad (3)$$

A progress ratio of 85% means that costs are reduced by 15% with each doubling of installed capacity. This ratio is constant. It is also independent of the rate at which the experience is accumulated, of the developmental stage of the technology, and of possible qualitative changes in the product or technology.

The progress ratio has been calculated for other, non-energy-related industries (e.g., products of the chemical industry or airplane manufacturing). Typically, progress ratios lie between 75 and 85%.

2.3 Scope

Experience curves have been estimated for single firms but are more common for a whole sector or whole industry (e.g., the wind turbine industry, all makers of combined cycle gas turbines) or for cumulative experience in all varieties of one technology (e.g., solar photovoltaics). In a globalized world, and specifically in energy technology markets, it is reasonable to assume that learning processes draw from a worldwide base of experience. Therefore, experience curves for energy technologies should reflect global experience, and global data should be used for their estimation.

3. REASONS FOR COST REDUCTION AND CONSEQUENCES FOR THE FORMULATION OF EXPERIENCE CURVES

Experience curves, as a typical black box model, do not make any assumptions about the reasons for cost reduction. However, for the applications of experience curves, these may not be ignored. This section explains the reasons for cost reduction and implications for the interpretation of progress ratios and cost reduction effects. In energy technologies, there are five reasons for cost reduction and improvements in the competitiveness of a technology: learning by doing, economies of scale in production of the equipment, improvements in the efficiency and quality of the equipment, standardization, and market conditions.

3.1 Learning by Doing

Learning by doing results in cost reductions whenever a firm or any other type of organization improves its productivity of labor through day-to-day operations. Wright studied the development in labor productivity in the U.S. airframe industry and showed that the number of hours required to produce an aircraft declines with the cumulative number of aircrafts produced. Arrow introduced the learning by doing hypothesis into economic mainstream and built a theory of macroeconomic growth on it.

Cost reduction through cumulative experience can be achieved within the firm, but firms can also profit from intra- and interindustry spillover effects. Mobility of “human capital” and copycat behavior are possible causes of intraindustry spillover and can be treated jointly with learning by doing even when looking at a

single firm. Argote and Epple showed that “organizational forgetting,” which would reverse the cost reduction through learning by doing, occurs as well.

Cost reduction through learning by doing has been estimated in the past using the same functional form that later served as the blueprint for the experience curve. Learning by doing reduces the overall cost of a project to the degree that labor is used as a factor of input to production. The share of the total cost for this input factor is not likely to remain constant over time. Thus, if learning by doing is not the only factor contributing to cost reduction, theory does not fully support the use of the same functional form for both curves even if this is common practice.

3.2 Economies of Scale in Production and in Input Factors

The classic reason for cost reduction is economies of scale in the production process and in the procurement of input factors. When the output of an industry grows from a few hand-manufactured units per year to mass production, unit costs can go down by several orders of magnitude. Input materials for large-scale production can be purchased at better prices, and overhead can be reduced. Administrative effort and transaction cost per unit are reduced.

Upscaling a given production facility happens in an incremental way. Thus, even the cost of the aggregate production of a whole industry, particularly a small industry such as any national wind turbine industry, might not follow a smooth decay but rather have kinks. Just like in the case of learning by doing, there are natural limits to the size of factories, after which economies of scale might actually disappear and turn into diseconomies of scale in the form of increases in unit costs. The experience curve is a simplifying and generalizing representation of this growth process.

Economies of scale refer to the scale of current production, that is, to the current output per time. Because the classic experience curve concept does not control for time or the rate of production, estimating the equation can result in the same progress ratios for a given cumulative output at low and high rates of production. The experience curve concept assumes that the experience contribution of the x th unit depends only on how many units of experience have been gained before and not, for example on the cumulative sales in monetary terms. This points to the problem of defining the exact metrics to be used in the curve (e.g., whether the cumulative output

should really be expressed in units produced or in a different metric), discussed in Section 4.

3.3 Improvements in Efficiency and Usefulness of the Product

Energy equipment can also exhibit qualitative changes over time that lead to higher efficiency of conversion without increasing unit cost in proportion. For example, bigger wind energy converters do not usually cost more in proportion as they grow larger, but their specific (marginal) cost per megawatt of capacity decreases. In addition, the efficiency of fuel use typically increases with changes in quality and size of the generator, so that the specific cost of the energy service goes down even more.

Efficiency changes can be taken into account in experience curves, but only if the delivered energy service (e.g., the kilowatt-hours of electricity) are the measure of experience as well as the unit for which costs are considered. Just like in the case of diseconomies of scale, equipment might evolve into an application suited for specific niche markets and, as a result of these qualitative changes, specific costs might be higher in later stages of technology diffusion.

3.4 Diffusion and Standardization

A fourth reason for the reduction of cost with increasing maturity of a technology is the standardization that a successful technology will undergo during its diffusion from niche markets into general markets. Standardization drives down overhead and compatibility costs. A logistic rather than exponential curve seems to be the appropriate representation of this phenomenon.

3.5 Market Conditions

All phenomena discussed so far might not have any effect on unit cost of production or on prices of the technology if the market structure is not sufficiently competitive. Not only might monopolists and oligopolists be able to charge prices that are far higher than the costs of developing, producing, and distributing a product or technology, but a noncompetitive market structure also relieves market participants from the pressure to improve products and production processes. Thus, learning and efficiency gains might be realized more slowly than in competitive situations. Other market conditions, such as the maturity of the market on the demand side and the opportunity of

spillover effects, might affect the speed of learning and improvements of the industry standard in economic or technological terms and, thereby, have indirect effects on the cost reduction over time.

In addition, the market situation in input factor markets (e.g., labor, materials) can play an important role in both improving and reducing cost reduction and experience effects. However, in an abstract and very general model of cost reduction with increased experience, it is very hard to account for these effects.

4. PROBLEMS OF DATA IDENTIFICATION AND STATISTICAL INFERENCE

Considering the reasons for a possible decrease in unit cost of energy technologies, it becomes clear that the variables representing experience as well as cost need to be chosen with care. Some of the changes (e.g., improved efficiency of conversion, improved reliability) only reduce the cost of the delivered output of the technology, whereas others (e.g., economies of scale in production, learning by doing) reduce the cost of equipment manufacturing and will have effects on both possible metrics: the cost of capacity and the cost of the energy service. Other identification problems and problems of statistical inference are discussed in this section.

4.1 The Dependent Variable

4.1.1 Cost per Capacity or per Energy Service

Most of the newer studies of energy technologies, particularly those in the field of renewable and intermittent energy sources, concentrate their analyses on the cost per unit capacity of the technology. These data are readily available for some technologies. Data on the cost of the energy itself (e.g., the power output for some technologies) are harder to obtain for several reasons. The actual energy output depends on local load factors and the supply with fuel inputs. Fluctuations in fuel prices and operation and maintenance (O&M) costs introduce additional uncertainties. Long-term trends in fuel prices or O&M costs might even offset any cost reductions through cumulative experience if measured in prices per energy service. Still, when experience curves are used for a comparison between different energy technologies, the assessment should always refer to the cost per unit of energy (e.g., kilowatt-hour) because this metric accounts for internal changes in

the energy technology as well as in the other conditions that influence the relative costs of different technologies. Only this metric, if any, allows for the assessment of buy-down costs (as defined earlier as the amount of investment needed until the costs reach the level of any backstop technology). To do so, it is recommended to define standard (e.g., average) conditions for load factors and fuel input prices and quantities. To extrapolate cost reduction into the future, the standard situations (e.g., load factors, efficiency, fuel prices) need to be extrapolated into the future as well. This apparent abstraction is a reminder of the character of experience curves as a highly abstract tool for long-term technology assessment rather than as a planning tool for the design of an energy system on short and intermediate time horizons.

Related to this problem is the question of how to include O&M costs. These might be subject to a long-term cost reduction of their own as experiences with use of the new technology are collected and as the design of new equipment is improved for ease of use and reliability.

4.2.2 Costs or Prices

Another problematic aspect of the dependent variable is the question of whether cost or price data are used. Experience curve theory relates to the production cost for equipment or the energy product. Most empirical applications of experience curves to energy technologies use prices because they are easier to obtain than are cost data. When using price data as a proxy for costs, researchers implicitly make very restrictive assumptions on the market structure. Because experience curves include observations from many years, structural changes in the markets for energy technologies should be excluded if price data are used. Another major reason for various trends in prices and costs is changing market transformation policies and subsidy schemes for new energy technologies.

4.2 The Independent Variable: The Measure of Experience

4.2.1 Cumulatively Installed Capacity versus Cumulative Production of Energy Service

The measure of the accumulation of experience should be chosen in close analogy to the measure of cost for which the experience curve is estimated. For the reasons mentioned previously—economies of scale, market fluctuations, and qualitative changes in

the technology—cumulative energy services delivered would be the relevant measure for experience when thinking about the competitiveness of an energy technology, but data access and international aggregation are easier for cumulative capacity in megawatts. Current research investigates the usability of other measures of experience (e.g., cumulative investments, cumulative units of installation).

4.2.2 *The Specification of the Technology*

Whenever spillover effects from other industries are present, the estimation of cost reduction is possibly flawed. This is often problematic when technological concepts originate in other industries (e.g., gas turbine having its roots in jet engines). Also, the design of a technology might change in the course of product diversification and the establishment of a technology in new (e.g., distributed) applications.

The attribution of learning and accumulating experience also depends on which technology one actually examines. For example, the outcome when the experience collected with solar photovoltaic power generation is aggregated over all semiconductor materials that have been used for the purpose is drastically different from that when just the currently prevalent material is included in the experience measure. This phenomenon is highly technology specific, and no general recommendation can be given.

4.2.3 *Global or National Scale*

Typically, in a globalized world, the cumulative experience worldwide should be used as the driver for cost reduction through increased experience. With the exception of nuclear power, the markets for energy technologies are typically global markets. However, if there are specific local dynamics in a technology, the time series could become inconsistent. Furthermore, global data can sometimes be very hard to access, particularly for new technologies that are developed in several countries in parallel.

4.3 The Functional Form: Expression of the Speed of Cost Reduction

The assumption of exponential decay of costs reflects the fact that marginal improvements in the production processes diminish with increased experience. For experience curves, this form has been inferred from the idea of the learning curve. Fast learning with the first batches of production leads to rapid cost reduction initially, which can slow down during later stages of higher industry or business maturity. The

traditional logarithmic shape of assumed cost reduction has very pleasant mathematical properties that technically allow for the estimation by simple linear regression and ordinary least squares methods. However, there is no reason to assume that the reduction in production cost for any piece of energy technology can be indefinitely continued, as implied by this functional form. Common sense suggests that the material cost is a natural lower limit for the cost. This means that the function should at least contain a constant price floor. The smooth logarithmic function also does not account for radical changes, implying that the curve should not be used for monopolies or monopsonies.

Furthermore, as discussed previously, learning and the other cost-reducing phenomena strongly depend on the rates of production. Thus, not all influences are captured by the classic functional form, and a multivariate approach might be much more appropriate even at the expense of the simplicity of the approach. In addition, market conditions and the circumstances of development and production of the equipment are likely to change during the time in which experience is accumulated, so that even at constant rates of deployment or constant rates of growth of production, the velocity of learning and improvements in the technology will not remain constant. Although the simple functional form has its own merits, it might not be able to account for the full story. This once again underlines the black box character of the approach. Experience curves are a tool for generating smooth curves for future projections on a large scale and for the long term, and they are not necessarily suitable for the assessment of historic trends or detailed short-term predictions. Because of this limited ability to reflect the actual processes, experience curves might be suitable only for very coarse assessments.

4.4 Spurious Regression

The experience curve concept implies a statistical relationship between a cumulative variable (e.g., cumulative deployment) and a noncumulative variable (e.g., unit cost). A statistical analysis of such a pair of variables needs to account for the autocorrelation and integration characteristics of the data. Typically, the cost at time t_{x+1} after the deployment of CUM_{x+1} will not be independent of the cost in time t_x after the deployment of CUM_x but can be represented by $C_{x+1} = C_x - \Delta C + \varepsilon$ (i.e., it should resemble a random walk with a downward drift term). Cumulative deployment CUM_{x+1} can be represented by $CUM_{x+1} = CUM_x + \Delta CUM$, where

ΔCUM equals the amount deployed during the period of observation. If the period of observation is a regular interval and the technology exhibits a typical growth pattern of a diffusion phase, the amount deployed during each period ΔCUM itself can be represented by $\Delta CUM_{x+1} = \Delta CUM_x + \Delta(\Delta CUM_x) + \varepsilon$ (i.e., it a random walk process with a upward drift). Thus, CUM_{x+1} does not fulfill the requirements for a stochastic variable. Rather, the classic concept requires an ill-defined regression of a random walk with drift with the integration of a random walk with drift. This vulnerates the standard assumptions of ordinary least squares regression, and in most cases a logarithmic transformation does not change the inherent characteristics of the data, so that the regressions based on the classic concept are most likely to be spurious. Extensive testing of the data and the error term should be conducted to prevent conclusions that are not supported by the data. However, proper statistical treatment and testing is the exception rather than the rule.

5. SOME CASE STUDIES

Recently, the classic experience curve concept has been applied to the area of new energy technologies, mostly renewable energy technologies but also combined cycle gas turbines. In this section, several technology-specific applications are discussed. These case studies demonstrate possible problems with the concept and new energy technologies and offer interesting ideas for improvement. Several large-scale research efforts on experience curves were aggregated in the International Energy Agency (IEA)-led International Collaboration on Experience Curves for Energy Technology Policy (EXCETP). All case studies have in common that where more than one estimate has been made, the estimated progress ratios span wide ranges where, as described previously, the estimate should be robust with regard to the selection of data. This fact is a clear warning to pay close attention to selecting the metrics and the database for the estimation and to testing the estimates for statistical validity, as explained previously. Figure 2 contains a classic display of several examples of estimates of experience curves that have been partly included in the following literature survey.

5.1 Wind Energy Converters

Several studies have applied the concept to various data sets with varying success. Junginger reviewed 16

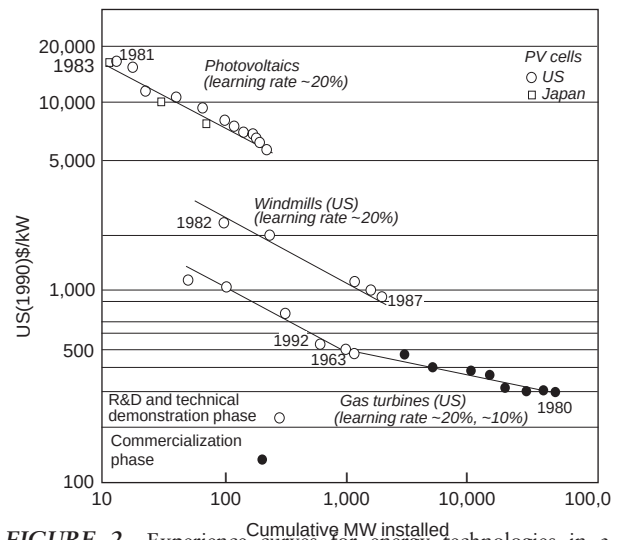


FIGURE 2 Experience curves for energy technologies in a variety of International Institute for Applied Systems Analysis estimates. Figure courtesy of Nakicenovic, N., Gruebler, A., and MacDonald, A. (Eds.) (1998). "Global Energy Perspectives." Cambridge University Press, Cambridge, UK.

studies on cost reduction in wind turbine production and found that they used different cost units and different proxies for the cumulative experience—and got very different results. Most of these studies interpreted the results of the classic function with a very specific set of reasons. These studies hardly ever accounted for the data, identification, or statistical problems mentioned previously. Most studies indicated that progress ratios for wind energy converters have been much higher (the cost reduction has been slower) than the classic rule-of-thumb progress ratios of 75 to 85%. The range of results was 62 to 96%.

5.2 Solar Photovoltaics, Niche Markets, and the Influence of Market Transformation Programs and Policies

Solar photovoltaics has been studied extensively with experience curves. Progress ratios between 68 and 82% have been estimated. The technology was characterized by high initial investment cost and comparatively high maintenance cost due to changing micro-design. The countries with the highest installed capacity of solar panels are Japan and Germany, where the installation was subsidized. Other market transformation initiatives have been launched for developing countries. Using these examples, the effect of market introduction subsidies on the cost of the equipment has been estimated. The induced cost reduction effects have then been

factored into the benefit/cost ratio of the policies to justify the intervention. The experience curves were the basis for the assessment of the necessary amount of subsidy.

Even if the fuel is free and even in good locations, electricity from photovoltaics currently is more expensive than standard grid power. However, in certain niche markets, either sunny rural locations remote from the grid or sunny urban locations where no backup power supply is available, photovoltaics can be a cost-effective solution. From this starting point, the current costs of photovoltaic systems, the market size of the niche markets, and the progress ratio can be used to calculate the point in time at which solar photovoltaics will break even with grid-connected power technologies.

5.3 Gas Turbines and the Influence of the Diffusion Stage of the Technology

Nakicenovic and colleagues estimated a two-stage learning process for gas turbines. After a steep decay in cost with a progress ratio of approximately 80% in the airline industry, a lower rate of cost reduction follows with a progress ratio of approximately 90% (Fig. 2).

According to Colpier and Cornland, who analyzed the economics of the natural gas-fired combined cycle gas turbine (CCGT), the specific investment price for larger CCGTs has decreased by as much as 35% in real terms during the past decade. Compared with experiences in other industries, this is a very high decrease. CCGT is not a new technology in itself but rather a combination of the established technologies of gas and steam turbines that has been used since the 1970s. To find an explanation for the high rates of cost reduction, Colpier and Cornland expanded the classic theory by taking into account strategic pricing behavior. They hypothesized that a strategically pricing firm sells at prices lower than marginal cost during the development phase of a product. The firm develops into a market leader, reduces its own marginal cost with increased experience and can sell under a “price umbrella,” that is, at prices higher than marginal cost and constant until competitors have caught up. In the subsequent “shake-out phase,” prices will drop sharply and approach marginal cost. The authors claimed that this phenomenon has been observed for many different technologies. Their analysis indicated that the prices for CCGT technology have been dropping in a way that looks like a margin shake-out rather than the less drastic decrease of unit

construction cost through experience and economies of scale. This could be supported by the fact that the CCGT market is dominated by only four manufacturers and their licensees. The authors concluded that the decline of the specific investment prices is likely to level off in the future.

More generally, pricing behavior is linked not only to the developmental stages of a technology but also to the rate of deployment and the rate of technological change, both of which affect the rate of learning and of cost reduction in different ways.

5.4 Gas Infrastructure

Zhao investigated historical cost development of gas pipelines. She ignored O&M costs and focused on the construction cost per cubic meter of pipeline capacity. In the United States, these costs increased and decreased twice between 1985 and 1998. Zhao observed no learning effects in onshore pipeline construction and found no significant cost reductions or progress ratios. For offshore pipeline installation, she measured cost in U.S. dollars per cubic meter of pipeline volume and days required to lay 1 km of pipeline. Her data, in fact, indicated that costs go down, although no statistical significance could be achieved. She concluded that because pipeline construction is not a replicative process, learning effects are probably small. Economies of scale can be achieved (e.g., through increased capacities, i.e., pipe diameter), although it is unclear how she derived these results. Experience curves seem to be an inappropriate tool for the analysis of the cost of gas infrastructure.

6. CONCEPTUAL MODIFICATIONS

Quantitative estimates for experience curves should be used with caution because the theoretical foundation, as well as the empirical support, is rather weak. Nevertheless, the occurrence of cost reduction with increasing diffusion and market penetration of a technology is undisputed. To get better estimates for the velocity of cost reduction, more research is under way.

One school of modifications amends the factors that are ignored by the traditional formulations by a simple addition of the logarithm of cumulative expenses for research and development (R&D) in Eq. (2). This formulation implies that cumulative R&D expenses are a perfect substitute for commercial investments in the technology after its innovation

in their effect on cost reduction. Attempts to estimate this specification have not resulted in satisfactory results, due mainly to the statistical and conceptual problems discussed previously.

Other studies have tried to decompose the cost reduction effect into its components, as explained in section II. Here, too, the statistical problems of serially correlated and integrated regression residuals and unit roots have not been solved. However, future research into such multivariate approaches is necessary and promising.

In this respect, it is also useful to work on temporal analyses of economic cost functions. Isoard and Soriat took a first step in this direction when they combined cost functions and learning curves to capture the simultaneous econometric estimation of learning by doing and returns to scale, endogenizing technical progress. A test for Granger causality establishes the causal relationship between deployment and cost reduction.

Other extensions are possible and should be explored. In theory, experience curves can be conceived for aspects other than energy unit costs. Other parameters that improve with cumulative deployment through continuous changes in the average technological design could be the efficiency of energy conversion or the emissions of pollutants from a conversion process. For these, too, the average improvement over time could be modeled with exponential decay functions. Linking these results back to the unit cost can give additional insights into the nature of cost reduction processes.

For future research, it is important to strengthen the foundation of the concept. One such avenue is linking the concept to the life cycle analysis of technologies. The life cycle of a technology starts with the invention of a new technological concept. The concept is simulated and prototyped before the actual innovation. The innovation itself consists of the first commercial application of the technology. For energy technologies, this often happens in niche markets (e.g., aeronautics). The innovation is followed by a diffusion phase, when the market penetration of the technology increases, often along

a similar (logistic) path as followed by the sinking cost of the technology. The forces effecting such a diffusion pattern can theoretically be subdivided into supply-push and demand-pull forces. Therefore, market forces and the economic environment should be included in deriving a better founded concept for the cost reduction of new energy technologies.

SEE ALSO THE FOLLOWING ARTICLES

Cost-Benefit Analysis Applied to Energy • Environmental Kuznets Curve • Resource Curse and Investment in Energy Industries • Technology Innovation and Energy

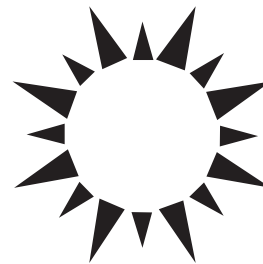
Further Reading

- Argote, L., and Epple, D. (1990). Learning curves in manufacturing. *Science* 247, 920–924.
- Arrow, K. J. (1962). The economic implication of learning by doing. *Rev. Econ. Studies* 29, 155–173.
- Colpier, U. C., and Cornland, D. (2002). The economics of the combined cycle gas turbine: An experience curve analysis. *Energy Policy* 30, 309–316.
- Duke, R., and Kammen, D. M. (1999). The economics of energy market transformation programs. *Energy J.* 20(4), 15–64.
- International Energy Agency. (2000). “Experience Curves for Energy Technology.” Organization for Economic Cooperation and Development/IEA, Paris.
- Isoard, S., and Soriat, A. (2001). Technical change dynamics: Evidence from the emerging renewable energy technologies. *Energy Econ.* 23, 619–636.
- Junginger, M. (2000). “Experience Curves in the Wind Energy Sector: Use, Analysis, and Recommendations, report NWS-I-2000-29.” Department of Science, Technology, and Society, Utrecht University, Netherlands.
- Woerlen, C. (2003). Economic and technological aspects of the market introduction of renewable power technologies. Ph.D. dissertation, Boston University.
- Wright, T. P. (1936). Factors affecting the cost of airplanes. *J. Aeronautical Sci.* 3, 122.
- Zhao, J. (2000). “Diffusion, Costs, and Learning in the Development of International Gas Transmission Lines, IR-00-054.” International Institute for Applied Systems Analysis, Laxenburg, Austria.

External Costs of Energy

DARWIN C. HALL

California State University
Long Beach, California, United States



-
1. Economic Theory of Externalities
 2. Categories of Externalities: National Security and Pollution
 3. Estimates of External Energy Costs
 4. Policy Choices
 5. Conclusions

Glossary

British thermal unit (btu) The average amount of energy per degree Fahrenheit needed to raise the temperature of 1 pint of water from freezing to boiling at sea level, which is equal to the energy needed to raise the temperature by 1°F, from 60° to 61°.

common property It has exclusions on use and allocation of a benefit stream determined collectively by community agreement or through government.

e-folding time This is for exponential decay what doubling time is for exponential growth. Suppose that a stock can be described over time by exponential decay to zero. At initial time, the amount is X_0 and over time the amount is given by $X_t = X_0 \exp(-rt)$. The time derivative is given by $dX/dt = -rX_t$. At $t = 1/r$, $X_t = X_0/e$, where e is the exponential number, approximately 2.71828. The ratio $1/e$ equals approximately 37%, so the e-folding time is the time at which the stock declines by 63% of the original value X_0 . For an annual decay rate, for example, if $r = 2\%$, the e-folding time is 50 years.

NO_x Oxides of nitrogen, such as nitrogen dioxide (NO₂) or nitrogen monoxide (NO).

private property This has exclusions on use and enjoyment of a stream of benefits that are determined by a private owner or owner's agent.

public goods or bads These are not depletable by use and not excludable.

Vickery auction named after Noble Laureate William Vickery, this refers to an auction designed to avoid strategic bidding and to encourage small incremental bids that reveal the marginal cost curves of the bidders. The total amount is specified in advance. Each bidder may submit multiple bids for incremental amounts, and the lowest bids are accepted up to the point at which the

total amount is reached. All winning bidders are paid the same price, not the amount they bid, equal to the lowest of all the losing bids.

willingness to accept (WTA) A schedule of prices and quantities that gives the minimum amount a consumer would accept in exchange for giving up an additional unit, as a function of the number of units the consumer holds.

willingness to pay (WTP) A schedule of prices and quantities that gives the maximum amount a consumer would pay in exchange for obtaining an additional unit, as a function of the number of units the consumer holds.

The external costs of energy are the costs not reflected in the market price, well over 100% for some energy sources. This large divergence between the social cost of energy and the price occurs because of environmental externalities associated with conventional energy sources and national security subsidies for imported oil and for nuclear power. As we shift to imported energy supplies, the remaining investment in domestic energy supply—extraction, transportation, and transformation—is 8% of all investment in the U.S. economy (measured by fixed assets in 2001). Whether domestic or international, many of these energy investments (power plants, pipelines, refineries, and oil and coal extraction) last for 50 years or more. Market prices that omit externalities are encouraging consumption of fossil fuels, investments in military preparedness, international investments in fossil fuel supply, and domestic investments in low-fuel-economy vehicles. At the same time, the same market prices discourage investment in alternative energy technologies, discourage investment in high-fuel-economy vehicles and vehicles that use alternative fuels, and discourage investment in energy efficiency. External costs of energy provide incentives for long-term investments resulting in gross and lasting economic inefficiency.

External costs occur at each stage, from exploration to production, transportation, conversion, consumption, and waste disposal. National security

external costs accrue to oil because of national defense expenditures to secure oil that are paid by general taxes rather than taxes on oil imports. National security risks from the use of nuclear power range from the risk of attacks on power plants to attacks on waste transport and storage and the spread of nuclear technology and material. External costs of fossil fuels result from pollution through environmental media—soil, oceans, freshwater, the troposphere, and the stratosphere. Damage occurs to human health, there is a reduction of ecosystem services to the economy, and there is loss of the intrinsic value of nature. Economic theory helps to focus on marginal (rather than total) external costs and to separate external costs from costs that have been internalized through the market or regulation. Economic theory faces formidable challenges due to joint pollutants, nondepletable catalytic pollutants, and nonconvex, subtle, cascading effects among physical, biological, and social systems. Overly narrow applications of economic theory, particularly marginal benefits and marginal costs of externality abatement, are harmful to the formulation of well-designed policy options. The limits of the economic theory of externalities justify policies, such as the Clean Air Act, that do not require a balancing of marginal benefits and cost to set ambient air quality standards. The challenge is to design, implement, and improve policies that address unintended, harmful effects of energy development, production, and consumption.

1. ECONOMIC THEORY OF EXTERNALITIES

Economic efficiency is defined as occurring when it is not possible to make someone better off without making others worse off, which occurs when the marginal social benefit of consumption equals the marginal social cost of production. In theory, with a specific definition of “perfect competition,” markets achieve economic efficiency. Some of the assumptions are the absence of market power, markets that provide costless and accurate information, rational behavior narrowly defined, the absence of public goods (or public bads), and the absence of externalities. Externalities occur whenever decisions of economic agents cause incidental costs or benefits borne by others that are not reflected in market prices. In particular, external costs drive a wedge between social and private marginal benefits and costs (Fig. 1).

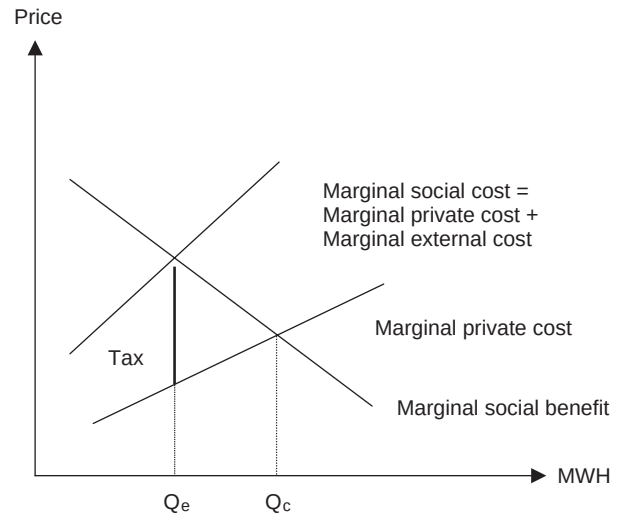


FIGURE 1 Marginal social benefit and cost.

If accurate information were costless, markets would be unnecessary. Markets provide imperfect information at a cost. All markets, whether for pollution credits, future prices of wholesale natural gas or electricity, or commodities sold to consumers, require government regulation to establish and enforce property rights. Property rights specify the right to pollute or, conversely, the right to a clean environment. Both markets and government enforcement of property rights that regulate markets are costly. Broadly, the idea of economic efficiency can inclusively consider the costs of markets and their regulation, including the costs of providing public goods and regulating externalities. In this sense, economic efficiency is the criterion most economists favor to evaluate policy changes. Although there is reason to be sympathetic toward the criterion of economic efficiency, the fundamental assumptions of economic efficiency are inconsistent with most applications to energy externalities. Finally, the criterion of economic efficiency is not adequate to assess strategies to reduce macroeconomic risks or technology forcing policies designed to alter the course of technological change.

1.1 Pigouvian Tax on Output versus Pollution Tax

The earliest analysis of externalities focused on external costs of production, for example, megawatt-hours (MWH) of electricity, where the marginal external costs are added to the marginal private costs to obtain the marginal social costs (Fig. 1). The amount consumed and produced by a competitive

market equals Q_c , and the economically efficient amount is Q_c . In order to achieve economic efficiency, government imposes a tax on output equal to the external cost at Q_c . This approach erroneously assumes that the externality is produced in fixed proportions so that, for example, a tax to account for NO_x emissions from a natural gas-fired power plant would also be applied—in error—to a solar thermal power plant.

Modern environmental economics poses the externality problem in terms of the marginal benefit and cost of externality reduction (Fig. 2). The marginal cost of reducing NO_x is given by listing all methods of pollution abatement in ascending order of incremental cost. The marginal cost is negative until the curve crosses the abscissa at N_f because investments in electricity conservation per megawatt-hour are cheaper than the incremental cost of building and operating new power plants. N_f is the level of air pollution abatement that is actually a free lunch. The marginal benefit of externality abatement depends on the property right—whether the polluter has the right or the public trust establishes a right to a level of a clean environment. The marginal benefit is given either by the damaged party's willingness to pay (WTP) for abatement or the damaged party's willingness to accept (WTA) payment in exchange for enduring the damage (Fig. 2).

The WTA is greater than the WTP due to an income effect and a substitution effect. If the polluter holds the right, the damaged party's budget constraint limits the WTP, but if the right is held by the public there is no budget constraint on the WTA; this is the income effect. The substitution effect can be understood using the example of air pollution; there

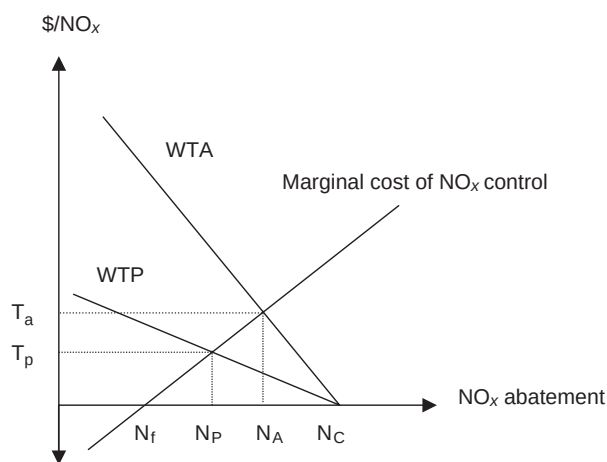


FIGURE 2 Marginal benefit and cost of externality reduction.

are no close substitutes for cleaner air, so the WTA is much larger than the WTP. If the correct marginal benefit curve is WTP, then the optimal amount of abatement equals N_p and the optimal pollution tax equals T_p . If the correct marginal benefit curve is WTA, then the optimal amount is larger, equal to N_A , and the tax equals T_A . There is no unique solution.

Both the WTA and the WTP curves emanate from the point N_c in Fig. 2. The maximum possible reduction in NO_x emissions is equal to the total emitted so that N_c corresponds to zero emissions to the environment. Consequently, whatever the property right, the social demand for externality abatement reaches the abscissa at the point of abatement where the negative externality is zero. In order to simplify the figures and discussion that follow, a single marginal benefit curve (MB) replaces the WTA and WTP curves.

1.2 Demand for Externality Abatement over Time

Demand for externality abatement shifts from MB_1 to MB_2 to MB_3 in Fig. 3 for two reasons—public goods and concentration of the externality. Public goods, such as increased national security or air pollution abatement, have one key property: They are nondepletable in consumption. The social demand for nondepletable goods is found by vertically adding individual demand, in contrast to

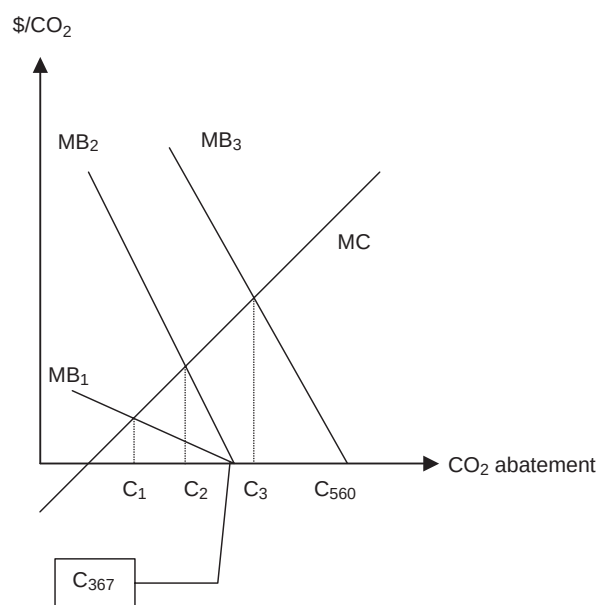


FIGURE 3 Optimal abatement over time.

the horizontal addition for private goods. Consequently, as population grows over time, the demand for externality abatement shifts from MB_1 to MB_2 , holding pollution concentration constant.

An important air pollutant from fossil fuels is CO_2 . CO_2 emissions have increased the ambient concentration of CO_2 since the industrial revolution. The ambient concentration was approximately 280 parts per million (ppm) in the late 1800s, and increased to 367 ppm by 1999. In the next 50 to 60 years, CO_2 concentration is forecast to reach 560 ppm, doubling the concentration since the pre-industrial revolution. Within 200 to 400 years the concentration will increase 4 to 10 times unless we significantly abate emissions and sequester carbon from the atmosphere. In Fig. 3, let C_{367} represent the abatement necessary to return to 280 ppm in year 1999. Over time, the intersection of the demand for abatement and the abscissa will shift to the right, from C_{367} to C_{560} by the year 2055, resulting in the marginal benefit curve MB_3 .

Over time, due to the public good nature of CO_2 abatement and population growth, and the growing concentration of the externality, the optimal externality abatement will increase from C_1 to C_2 to C_3 .

1.3 Regulation, Internalization, and Incremental Change

Regulation internalizes the negative externality from pollution. For example, the local air pollution regulatory authority in the Los Angeles area established a set of stationary-source air pollution regulations requiring controls for power plants and oil refineries, among other pollution sources, to reduce particulates and precursors to ozone. The pollution control cost of electricity and refined petroleum products in southern California is reflected in the prices of these products. In 1990, the air basin was significantly out of compliance with the federal ambient air pollution standards—a point such as O_{1990} in Fig. 4. The federal standards for ozone, 0.12 ppm, exceed the state standards, 0.08 ppm, shown as $O_{0.12}$ and $O_{0.08}$, respectively. If the situation is correctly depicted in Fig. 4, the marginal benefit of abatement, B_{1990} , is greater than the marginal cost of abatement, C_{1990} , and additional abatement is efficient.

It is not possible, however, to estimate the marginal benefit curve for reasons discussed later. In the peer-reviewed literature, a team of economists, chemists, airshed modelers, and health experts estimated the incremental benefit to the Los Angeles region of attaining the federal standards for both ozone and

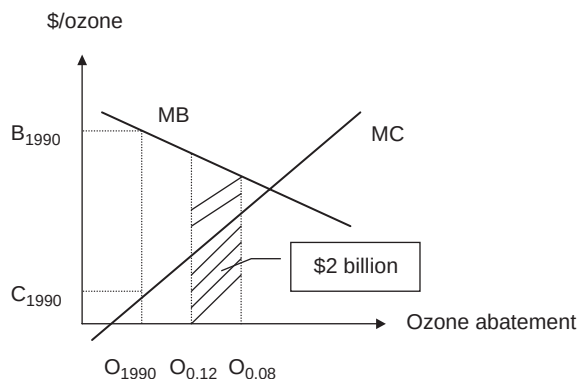


FIGURE 4 Partially internalized by regulation.

particulate simultaneously. The estimate of incremental benefit equals approximately \$10 billion for federal attainment and \$12 billion (1990 \$) for state attainment, so the incremental benefit of moving from the federal to state standards was \$2 billion. If ozone and particulate were a single, homogeneous pollutant, and pollution abatement benefits were convex and met other technical assumptions in economic analysis, the incremental benefit of \$2 billion could be shown as the area under the marginal benefit curve in Fig. 4. Unfortunately, the standard economic theory of externalities shown in Fig. 4 is incorrect.

1.4 Joint Pollutants, Nonconvexities, and Nondepletable Catalytic Pollutants

Chemical reactions do not follow the assumptions of economic models. Ozone (smog) is created by the reaction of hydrocarbons and NO_x . Unburned hydrocarbons are released when we produce, process, and consume petroleum to get energy. Nitrogen, approximately 78% of the atmosphere, is stable at low temperatures, but fossil fuel consumption releases heat that creates NO_x . The ozone isopleths in Fig. 5A, similar to isoquants, give the combinations of reactive organic gases (ROGs) and NO_x that create a given amount of ozone. Because the isopleths include regions with positive slopes, in these regions decreases in NO_x actually increase the amount of ozone, holding ROGs constant, as shown by the arrow in Fig. 5A. The marginal benefit of NO_x abatement first rises and then falls, as shown in Fig. 5B, which is an example of nonconvexity.

Optimal abatement does not occur where the curves cross in Fig. 5B. First, if area A is less than B, it is better to do nothing (*ceteris paribus*). Second, emission control options typically reduce both NO_x and ROGs (*mutatis mutandis*), shifting the marginal

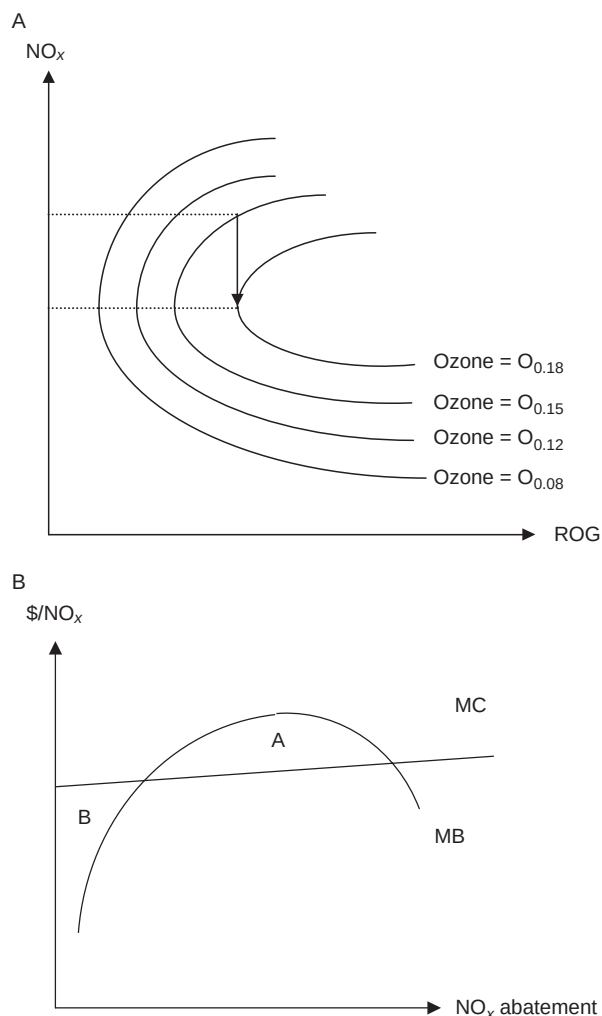


FIGURE 5 (A) Ozone isopleths. (B) Marginal benefit curve (MB) derived from isopleths.

benefit curve. Third, nitrate compounds from the nitrogen oxides are particulates with an aerodynamic diameter $<10\mu\text{m}$, small enough to slip through the nasal barriers and lodge into the lungs. The human body's immune system reacts in ways that cause irritation and exacerbate respiratory disease. Consequently, NO_x causes damage directly as well as indirectly through ozone. Since NO_x acts as a catalyst in the formation of ozone, the same molecule can later contribute to particulate. The analogy in economic theory is a public "bad", which is nondepletable in use. Hence, the marginal benefit of abatement is found by adding the demand for particulate abatement vertically to the demand for NO_x abatement derived from the demand for ozone abatement. These complexities render obsolete the economic theory of externalities for application to energy from fossil fuel.

1.5 Nonconvex, Subtle, Cascading Effects among Physical, Biological, and Social Systems

Consider the policy of planting shade trees and painting surfaces white to lower the urban heat island effect, reducing the demand for air-conditioning to conserve electricity. Electric utilities change the dispatch order, altering the amount of electricity generated from power plants within and outside the urban airshed. The location of power plants and the emission control equipment determine the change in emissions of ROGs and nitrogen oxides, each of which separately causes damage to human health and the environment, that together form ozone in the presence of sunlight and an inversion layer. Surface roughness, affected by planting trees, alters mixing and transport within the airshed. Different types of trees have different emissions of ROGs, further adding to the complexity of impacts. Financial impacts include the dollar values of fuel, pollution credits, electric system reliability and associated timing of power plant investments, tree maintenance, and painting, in addition to the external costs of changes in emissions of greenhouse gases and carbon sequestration. This level of complexity further renders useless the modern economic analysis of externalities depicted in marginal benefit and marginal cost curves for abatement.

1.6 Measuring Policy Options: Partial vs General Equilibrium

The portrayal of economic efficiency in Figs. 2–4 is based on partial equilibrium analysis. General equilibrium analysis models the interrelationships among markets, for example, accounting for impacts of changes in the price of oil that shift the demand for natural gas, shift the demand for vehicles, and shift the supply of transportation services, all of which in turn have effects that ripple throughout the economy.

Petroleum is a minor energy source in the residential and commercial sectors in the United States. Petroleum supplies 24% of the industrial sector, with ample room for substitutes, but petroleum supplies 97% of the transportation sector. Fuel cell technology is close to commercial viability and could be developed as a substitute for petroleum for transportation. Fuel cells would also substantially reduce urban air pollution. A policy to require that a fraction of sales be ultra-low-emission vehicles would speed the development of the infrastructure needed to commercialize fuel cells. A partial equilibrium analysis would deem a successful policy as a failure.

Whether based on hydrogen or natural gas-derived methanol, suppose a successful fuel cell program in 2002 results in transportation costs comparable to the price at that time—\$30 per barrel of oil. Falling oil consumption would likely result in the collapse, or at least the diminishment, of the OPEC cartel market power. The resulting decline in oil prices could easily reach less than \$10 per barrel, placing the fuel cells program politically at risk. Oil interests could commission partial equilibrium studies that show the fuel cells program to be an economic disaster at \$30 per barrel, when in fact oil is at \$10. Of course, the correct counterfactual for evaluating the fuel cells program would hypothesize oil prices in the context of a transportation sector without a substitute, requiring a computable general equilibrium (CGE) model with a cartel sector.

CGE models are routinely applied to environmental regulations to gauge the economic cost of policies to abate pollution. However, CGE models have not been modified to measure lost productivity from mortality, and ill-health effects of pollution causing worker absences or impairing human development, nor have they been modified so that increased medical expenses are measured as counterproductive rather than as adding to economic activity. Of course, the correct counterfactual for evaluating pollution abatement policy would correct CGE models to account for harm to human health, damage to agriculture and fisheries and forests, damage to vehicles and structures such as buildings and bridges, recreational losses, public good losses measured by contingent valuation studies, and so on.

2. CATEGORIES OF EXTERNALITIES: NATIONAL SECURITY AND POLLUTION

Although national security and pollution are the two most important negative externalities from energy production and consumption, neither should be estimated in isolation from the other. Most energy policies affect both national security and pollution.

2.1 Oil and National Security

Of the 68 million barrels of oil produced per day (mbd) in 2000, 30% were from the three top producers: Saudi Arabia (8.4 mbd), Russia (6.5 mbd), and the United States (5.8 mbd). This is changing because Iran, Iraq, Kuwait, Qatar, Saudi

Arabia, and the United Arab Emirates hold 65% of the world's proven reserves, whereas the United States holds 2%. The United States consumes 19.7 mbd while producing 5.8 mbd, on net importing 70%. Of the total petroleum used in the United States, transportation use equals 68%. The transportation sector has virtually no substitutes for petroleum, which supplies 97% of all energy in that sector.

If externality abatement is subsidized through general taxes, the marginal cost of abatement is not necessarily equal to the optimal tax on the externality. As shown in Fig. 4, optimal policy design requires knowledge of both the demand for externality abatement and the cost of abatement. At the federal standard for ozone of 0.12 ppm, Fig. 4 depicts a marginal cost of abatement less than the marginal benefit. Another example is the national security costs of importing oil.

Broadman and Hogan identify three parts of an optimal oil import tariff: high oil prices that reflect monopsony power, adverse macroeconomic effects of higher oil prices on terms of trade, and the adverse impact of oil price shocks. Because the United States consumes 25% of world oil, the price depends on the amount consumed (monopsony power). The U.S. macroeconomy does not perform as well when oil prices are high, changing the terms of trade and requiring greater exports for the same amount of imports. The adverse impacts of oil price shocks are the basis of a 'security tariff' that can be calculated as the expected loss, equal to the product of the probability of an oil price shock from disruption of supplies multiplied by the negative impact of an oil price shock. The negative impact of an oil price shock can be estimated from the magnitudes experienced since 1970. The security tariff is interpreted as the willingness to pay for military protection of oil supplies, estimated in 1985 dollars to be equal to \$7.00 per barrel (conditioned on the price of oil at \$15/bbl). The axes in Fig. 2 could be relabeled so that the ordinate is in dollars per barrel and the abscissa is reduction in national security; thus, point N_C is complete security. Then the optimal tax, T_B is \$7.00 per barrel in 1985 dollars. The marginal cost curve in Fig. 2 would be the cost of additional security, including the cost of additional military preparedness and the cost of the strategic petroleum reserve.

Prior to the invasion of Kuwait by Iraq, the defense expenditures per barrel of imported oil plus the cost of the strategic petroleum reserve (SPR) were approximately \$8.50 to \$9.00 (in 1985 dollars) per barrel. I calculated the lost interest from storing oil, equal to the interest that would have accrued by investing the

amount of money paid for the oil that was stored in the SPR; in 1985 \$ the lost interest equals \$1.15 per barrel of imported oil, and there is an additional \$0.60 per barrel for the increase in the size of the optimal reserve from its inception to the mid-1980s. I also performed a time series analysis of defense spending and oil imports during the period from 1968 to 1989. After removing time trends and random walks, the data reveal a correlation between defense spending in each year to oil imports lagged two years, where one year of the lag accounts for the time delay between publication of oil import data and Congressional appropriations, and the second year of the lag accounts for the delay between appropriations and actual budget years. The result is an econometric estimate equal to \$7.30 of defense spending per barrel of imported oil in 1985 dollars. Consequently, the marginal benefit calculated by Broadman and Hogan is close to the marginal cost I calculated, and one may conclude that in the 1980s the United States provisioned close to the optimal amount of national security related to importing oil. The price of oil did and does not include the national security cost, as it was and still is paid through general taxes rather than a tax on oil imports, and this national security subsidy for oil has the unfortunate consequence of economic inefficiency.

Certainly, the national security cost of oil imports is higher today. Macroeconomic effects of oil price shocks include less productivity, greater unemployment, and an increase in general inflation. Some oil revenue flows to terrorists and states acquiring weapons of mass destruction. Recessions have reverberating impacts on presidential elections. Presidents Carter and George H. W. Bush and Vice President Gore lost reelection following oil price spikes. In all three cases, the Federal Reserve Board responded to general inflation caused by oil price spikes by increasing the discount rate. Whether the recessions were caused directly by the oil price spikes or indirectly by the misdirected response of the Federal Reserve Board, the effect is the same: a recession during a presidential election.

2.2 Nuclear Power and National Security

Three types of terrorism are plausible. First, nuclear power plants and research reactors in the United States are potential targets for terrorists determined to release radioactivity. Second, increased trade in nuclear fuel and waste products provides more opportunities for sophisticated or government-sponsored terrorists to obtain the materials for a nuclear device. Third, dispersion of radioactive material

throughout a city requires little sophistication and may not cause a large amount of much human and physical damage but would have a substantial psychological impact.

2.2.1 Attacks on Nuclear Power Plants

Attacks from the inside and outside are technically feasible, and the risk is growing. For example, Iran threatened on Tehran radio, "U.S. centers and nuclear reactors can be more vulnerable than the missile bases of the Islamic Republic of Iran." Approximately one-half of terrorism incidents occur in the Middle East, followed by Europe and Latin America, particularly Colombia (drug related) and Peru (the Marxist Sendero Luminoso). A substantial portion of international terrorist attacks are on U.S. interests. More than 300 metric tons of plutonium is generated annually, 45 of which is separated and suitable for weapons. In comparison, 6 kg of plutonium was used in the bomb dropped on Nagasaki. Plutonium separated from spent fuel is shipped to waste disposal sites. Some of this transport occurs on commercial airlines and by truck, train, and ship.

2.2.2 Diverting Nuclear Material: Nuclear and Dirty Bombs

Nuclear material is a threat if used to make a nuclear bomb or if spread by a conventional bomb. Nuclear bombs can be made from plutonium or highly enriched uranium. Plutonium is produced in plants that reprocess waste from conventional reactors. Natural uranium contains approximately 0.7% of the fissile isotope U-235. Conventional reactors use uranium slightly enriched to approximately 5% U-235. Low-enriched (to 20%) U-235 is used in research reactors. Highly enriched U-235 is a category higher than 20%, and weapons-grade U-235 is enriched to 90%.

India and Pakistan both produce enriched uranium; India also produces plutonium. These countries and others publicly embarked on programs to produce nuclear power, transferred the technology, and developed parallel programs to develop nuclear weapons. North Korea and Iran are openly pursuing this path.

2.2.3 Diverting Nuclear Waste: Poisoning Cities

Nuclear waste from breeder reactors includes plutonium-239, which is extremely toxic in minute quantities and has a half-life of 24,400 years. Low-level nuclear waste from conventional nuclear power plants includes cesium-137 and strontium-190, which

are less toxic and have half-lives of approximately 30 years. Twenty half-lives is equivalent to 600 years, the time during which the waste poses a risk.

Unsophisticated terrorists, willing to accept exposure and death, could use nuclear waste to poison a city. The level of health risk from low-level waste may not be as great as that from an attack on a nuclear plant or from a bomb. However, the terror of exposure to nuclear radiation, no matter how low the dose, would still be considerable.

2.3 Pollution

Pollution occurs at every link in the chain from production to consumption: exploration, extraction, transportation, conversion, processing, consumption, and disposal. Pollution travels through environmental media, including tropospheric and stratospheric air pollution, surface and ground freshwater pollution, ocean pollution, and soil pollution. Damages vary across energy sources and uses, including human mortality and morbidity, reduction of ecosystem services to the economy, immiserization of the human spirit, and diminution of the intrinsic value of nature.

The objective of estimating the benefits of lower emissions is to determine, in economic terms, what damage society avoids by selecting one set of policy options over others. To determine this, the potential damages of emissions must be calculated. This involves the following:

1. Estimate the changes in emissions for all pollutants affected by policy changes.
2. Estimate the changes in ambient pollutant concentrations in all relevant environmental media predicted to result from changes in emissions.
3. Estimate the pollution transformation and transport to target.
4. Establish the exposure–dose/response relationships between ambient changes and health effects and impacts on horticulture, visibility, buildings/structures, etc. For human mortality and morbidity, this requires the additional intervening step of estimating changes in exposure and dose in the damaged population that would result from changes in ambient concentrations, where dose depends on body size, age, activity, and location affecting intake and internal concentration.
5. Calculate the level of effects projected.
6. Attach dollar values to the estimated effects.

7. Attribute dollar values to the pollutants that caused the damages or to the policy changes that reduced the damages.

There are various methods for accomplishing each of these steps, but each must be completed to calculate the economic value of a policy change. Each step represents a complex calculation that often violates economic assumptions, such as differentiability and convexity. All of the links in this chain present challenges and pose obstacles to consensus, as the following examples of urban air pollution and global warming illustrate.

2.3.1 Estimating Reductions in Emissions from Policies

Step 1 requires an inventory of all the types of pollutants affected by the policy options and the appropriate counterfactual with which to compare the policy. An example is regional air pollution caused by electricity generation. A seemingly straightforward policy is to hold a Vickery auction for renewable electricity generation of a certain number of megawatts, such as the auction held by the California Public Utilities Commission in 1994 (and rescinded the following year on the grounds that deregulation was impending and markets would best decide how much and what type of additional capacity to build). An electricity dispatch model is needed to calculate the change in emissions.

Power plants are typically dispatched in increasing order of short-run marginal cost until the system demand is met. As demand changes from off-peak to peak periods, the least expensive of the remaining plants is dispatched to meet load, subject to technical constraints, spinning reserve requirements to meet fluctuations in load, and a reserve margin to cover maintenance outage and unscheduled outages. Some plants cannot be shut down during the weekend and some have a minimum operating level. Some are slow-start plants and others are quick-start plants. Technical constraints determine the type of load each can serve.

The addition of power plants to a system alters all of the previous calculations. By accounting for the location of the units, one can calculate emissions for each airshed. The calculation of emissions is a direct function of the calculated dispatch order. Consequently, the emission estimates depend on the reordering of dispatch due to a change in the combination of power plants.

2.3.2 Estimating Changes in Pollution Concentration

Step 2 requires the application of a model to project how emission changes will alter ambient pollution concentration. This really has two parts; the baseline pollution concentration is forecast in the absence of the policy, and the altered pollution concentration is forecast conditional on the change in emissions. Fig. 6 provides an example of a baseline forecast of ambient CO_2 , showing historical levels during the past 600 million years and projecting what could happen during the next 400 years based on consumption of economically available fossil fuels.

The Intergovernmental Panel on Climate Change uses 120 years as the e-folding time atmospheric residence for a pulse increase of CO_2 sufficient for $\text{RCO}_2 = 2$. RCO_2 is the ratio of CO_2 to the ambient concentration prior to the industrial revolution in the late 1800s. The e-folding time is the number of years at which the biosphere and ocean removes 63% of the increase from pre-industrial levels; so that if CO_2 were to instantaneously double and no additional emissions occurred, after 120 years ambient CO_2 concentrations would still be 37% higher. For increases greater than $\text{RCO}_2 = 2$, the e-folding time increases dramatically. Carbon uptake by plants is limited because biological material decays, releasing carbon. For example, at the Triassic–Jurassic boundary approximately 200 million years ago, CO_2 was so high that climate effects led to the disappearance of most species. Almost all models used by economists make the assumption of linearity, seriously underestimating ambient concentrations for a tripling ($\text{RCO}_2 = 3$) or more. At three times preindustrial concentrations, the e-folding time would be between 380 and 700 years. To model the atmosphere beyond a doubling of CO_2 , the models used by economists are misspecified. For a tripling ($\text{RCO}_2 = 3$) or larger, the natural science literature assumes that 50% of the emissions remain in the atmosphere, as presented in three alternative baseline forecasts in Fig. 6B.

Three economic models, combined with the previously discussed assumption, predict baseline RCO_2 increasing to 9, 11, and 16. To get a sense of the scale involved, Fig. 6A shows atmospheric concentrations measured in RCO_2 during the past 600 million years. The macro models forecast that economic forces will cause, within the next 325–350 years, the atmosphere to revert to an era never experienced by most plants and animals living today—levels not experienced by Earth at any time during the past 375 million years.

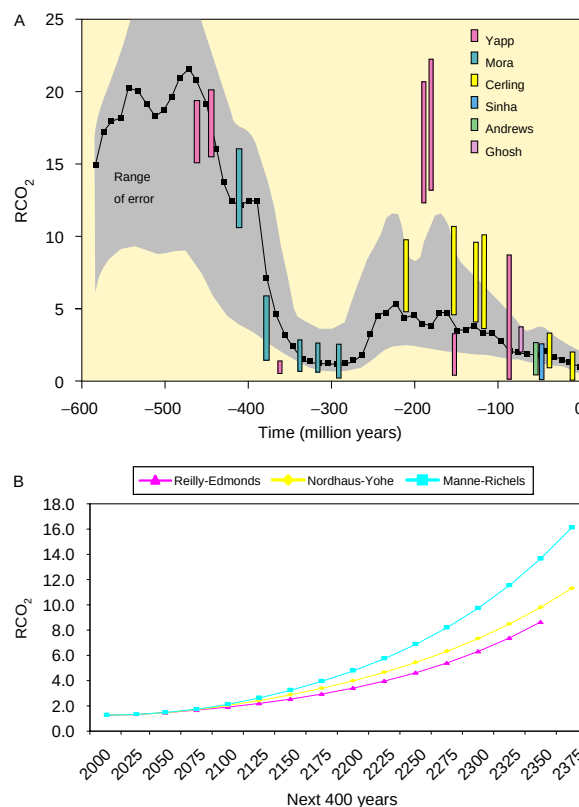


FIGURE 6 (A) Past atmospheric CO_2 . Atmospheric CO_2 versus time for the Phanerozoic (past 550 million years). The parameter RCO_2 is defined as the ratio of the mass of CO_2 in the atmosphere at some time in the past to that at present (with a preindustrial value of 300 parts per million). The heavier line joining small squares represents the best estimate from GEOCARB II modeling, updated to have the effect of land plants on weathering introduced 380–350 million years ago. The shaded area represents the approximate range of error of the modeling based on sensitivity analysis. Vertical bars represent independent estimates of CO_2 levels based on the study of paleosols. Reprinted with permission from Berner, R. A. (1997). The rise of plants and their effect on weathering and atmospheric CO_2 . *Science* 276(25), 544–546. Copyright American Association for the Advancement of Science. (B) Baseline forecasts of RCO_2 . The rate of emissions over time is based on three macroeconomic models that link forecasts of future economic activity with the potential use of fossil fuels. In descending order of the growth rate for CO_2 emissions, the macro models are labeled MR, NY, and RE. MR refers to Manne, A., and Richels, R. (1990). *Energy J.* 11(2), 51–74. NY refers to Nordhaus, W., and Yohe, G. (1983). In “Changing Climate,” pp. 87–153. National Research Council, National Academy Press, Washington, DC. RE refers to Reilly, J., Edmonds, J., Gardner, R., and Brenkert, A. (1987). *Energy J.* 8(3), 1–29. Accounting for all fossil fuels, the economically available fossil fuels vary between 8 and 17 metric teratons. The MR forecast assumes 17 metric teratons, NY assumes 12, and RE assumes 8. Reproduced with permission from Hall, D. C. (2001). In “The Long-Term Economics of Climate Change: Beyond a Doubling of Greenhouse Gas Concentrations” (D. C. Hall and R. B. Howarth, Eds.). Elsevier, Amsterdam.

2.3.3 Estimating Pollution Transformation and Transport

To continue the example of global warming, ultra-violet sunlight passes through CO₂, warming the earth's surface. Infrared heat radiates back into the atmosphere, warming greenhouse gases (GHGs) such as CO₂ and radiating back to the earth (the greenhouse effect). As the concentration of GHGs rises, so does the mean global temperature. Coupled general circulation models of the atmosphere and oceans project increases in temperature and changes in rainfall frequency and intensity as GHGs rise.

Natural scientists expect the ambient concentration of GHGs to double within the next 100 years (most likely by 2060). Natural scientists use a doubling as a convenience in the thought experiment: If concentrations double, what will be the impact on global average temperature? The answer, from coupled general circulation models (GCMs), is 1.5–4.5°C. For easier cross-discipline comparison, economists use the same benchmark, $RCO_2 = 2$.

If we confine the baseline forecast to a doubling of concentration, it may be reasonable to assume that Earth's systems return to equilibrium, as is done in one famous model (DICE, RICE, and derivatives). Most economists assume that the ocean captures a significant amount of heat, with an e-folding time of 500 years, so increases in deep ocean temperature decay back toward the preindustrial revolution level outside the time frame of any model. This specification is inconsistent with findings of ocean surface water temperatures. More problematic for analysis of $RCO_2 > 2$, the Atlantic Ocean thermohaline circulation may end, resulting in a new stable equilibrium for ocean and atmospheric circulation.

Figure 7 presents baseline forecasts of mean global temperature from a model that specifies an ocean thermal lag, where one-half of the radiative forcing is stored in the ocean for 50 years and then released. The three baseline forecasts show how global temperature forecasts depend on key assumptions regarding the sensitivity of the climate to increases in CO₂ and the amount of economically available fossil fuels. The ocean thermal lag is consistent with findings that from 1948 to 1998, the heat content of the top 3000 m of the oceans increased by 2×10^{23} J, corresponding to a warming rate of 0.3 w/m² or approximately one-half of the warming predicted by GCMs for the past century.

Based on a comparison among three GCMs, for a doubling of CO₂ equivalent gases, the models predict 12% greater warming for the United States than the global mean because the United States is at a high

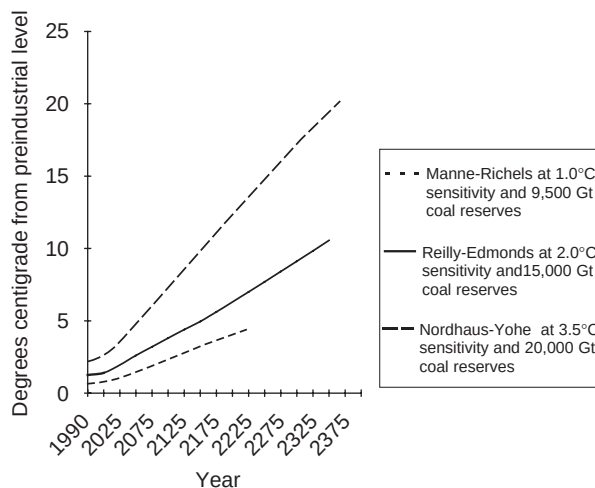


FIGURE 7 Mean global temperature increase.

latitude. As the earth warms, the largest temperature changes will occur near the poles.

2.3.4 Establishing Exposure–Dose/Response Relationships

Models of exposure–dose/effect relationships for air pollution are increasingly complex. Much of the past economics literature simply correlates air pollution concentrations with morbidity and mortality. One recent model calculates dose taking into account indoor, outdoor, and in-transit exposure; activity level that increases respiration; and body mass that changes with age. Exposure is measured in the REHEX model by time-weighted and location-specific ambient concentrations relative to indoor, in-transit, and outdoor activity for each age group and work status. Doses are integrated measures of pollutant intake. REHEX translates exposure into dose by accounting for activity and age. Dose is dependent on activity (e.g., work type) and age (children have higher metabolism and elderly people have lower metabolism).

Individual ill effects are identified in the health effects literature, with dose–response estimates that vary by age and health status. Effects range from eye irritation to hospitalization and death.

2.3.5 Calculating the Level of Effects

For some examples, such as climate change, the level of effects depends on the time frame for the analysis. Moderate climate change, based on a doubling of GHG concentrations, benefits midlatitude and upper latitude countries with milder weather; milder weather benefits agriculture, forestry, shipping, and transport; milder weather lowers heating bills and makes surroundings more conducive to human activity.

Damage to midlatitude countries includes more frequent and severe droughts and storms, hurricanes, cyclones, rapid snowmelt, and flooding; increased costs for water transport, storage, and flood control; migration of pests and increases in agricultural pest damage and control costs; migration, spread, and ill effects of disease and increases in health care costs; higher cooling bills; damage to coastal cities; reduced aquaculture yields; damage to coastal ecosystems and wetlands; and beach erosion. Damages to equatorial countries will be severe, including immersion of island nations and coastal properties, river delta flooding and damage to wetlands, damage to agriculture, loss of life, the destruction of coral reefs worldwide and the fisheries and tourism dependent on them, and severe droughts and floods.

Tripling or quadrupling of GHG concentrations would have severe consequences for the global economy, human health, and just about every aspect of human endeavor. Frozen deposits of methane in ocean sediments could thaw, instantaneously releasing massive amounts of GHGs into the atmosphere. There may be several global equilibria of GHG concentrations in the atmosphere and oceans. Tripling or quadrupling GHG concentrations could cause Earth's climate to shift to a new equilibrium and to never return to the climate of today. A tripling of GHG concentrations within 150 years is consistent with several models of economic activity and concomitant fossil fuel consumption. The range of uncertainty for a doubling of GHG concentration is between 1.5 and 4.5 degrees warming of mean global temperature, but we are ignorant of the range of temperature increase if GHG concentrations were to triple, quadruple or quintuple. We have modeled the range of temperature increase that would result in a shut-down of the Atlantic Ocean currents that keep the climate mild in northern Europe, but we are ignorant of the GHG concentration necessary to warm intermediate ocean water and release catastrophic quantities of methane to the atmosphere.

2.3.6 Attaching Dollar Values

Valuing the adverse effects associated with energy involves estimation of the value people place on avoiding the identified effects. Many of the effects are nonmarket goods, such as damage to human health and loss of life. Dollar values for nonmarket goods can be derived indirectly from hedonic studies or directly through surveys using contingent valuation methods or their derivatives. Hedonic studies use market data to obtain values such as differences in wages to measure wage premiums paid in exchange

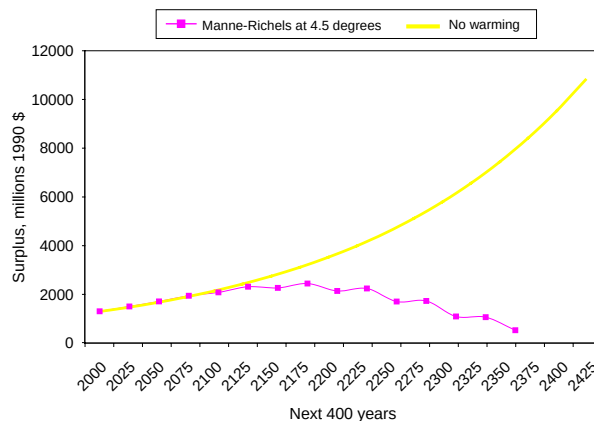


FIGURE 8 Agricultural surplus for wet climate.

for higher risk jobs. From these studies, economists calculate the value of statistical life (VSL). Contingent valuation surveys ask respondents how much they are willing to pay to reduce risks or avoid ill effects.

VSL studies have recently been used to estimate the value of one life year. Assuming that the elderly are primarily at risk means that premature death from particulate matter (PM) is heavily discounted. When these biased estimates are corrected by including the risk of death to children, the estimated value of reducing particulates will increase, unless of course we use income stream over a lifetime and discount it. Then, the value of a child's life is much smaller than that of an adult at a prime earning age.

Valuation also requires the specification of a discount rate. With a rate greater than 3% (which is the minimum that many economists use), costs and benefits 100 years or more into the future are discounted to virtually zero in today's dollars and thus have virtually no impact on a benefit-cost calculation. For example, Fig. 8 shows an estimate of the time path of U.S. agricultural producer and consumer surplus in a scenario of climate change in which initially there are minor benefits from warming but ultimately agriculture in the United States collapses. With a discount rate of 5%, the present value of warming is greater than that of avoiding warming because the early minor benefits count more than the eventual collapse. If the discount rate is set to 1%, however, then economic valuation of the future years has significant weight to conclude in favor of avoiding collapse.

2.3.7 External Costs of Pollutants and Benefits of Policies

Step 6 requires knowledge of the precise relationships between ambient levels of all pollutants that

derive from the initial emissions and the ability to separate the contribution of each pollutant to ambient changes.

A study estimated the value of meeting federal ozone and PM₁₀ standards in the South Coast Air Basin of California. For this purpose, steps 2–6 were completed. Step 1 was omitted because the question was what the benefits would be of meeting the standards, so the assumption was that emissions would be reduced enough to meet that objective. Step 7 was also omitted because there is no scientific or economic foundation for allocating inputs among joint products or for attributing products to non-depletable inputs. Another study based on the same methodology followed steps 2–5 for the purpose of projecting the pattern of exposure as emissions are reduced. These studies recognized the fundamental difficulties of evaluating a change in emissions relative to eventual damages, including the difficulties presented by the need to estimate air quality changes from emissions changes and to allocate damages back to emissions, and by the vast gap between what is known about the health effects of ozone and other pollutants and what can be quantified and valued in economic terms.

3. ESTIMATES OF EXTERNAL ENERGY COSTS

To understand the magnitude of negative external costs of energy, it is useful to aggregate and average over energy sources, comparing fossil fuels (coal, crude oil, and natural gas) with nuclear power, with solar power separately, and with solar power combined with natural gas to be operated as a

baseload power plant. Aggregation in this sense can be easily misinterpreted as a set of estimated pollution taxes that would result in economic efficiency. This is not the case, as explained previously.

In Table I, the estimated externality values are in Year 2000 dollars, updated from previous studies using the implicit price deflator. The national security costs and carbon emissions can be directly attributable to specific energy sources. The national security costs of oil are the addition of national defense expenditures attributable to importing oil and the cost of the strategic petroleum reserve. Note that the national security cost of oil has not been updated to reflect events since 1990, so this estimate is low. The estimates for coal and nuclear power are specific to those energy sources. Numerous externalities have been omitted, particularly for coal, including water pollution, waste disposal, and wholesale destruction of ecosystems at mining sites. The air pollution damages from natural gas and oil, however, represent a sum total that has been attributed to the energy sources linearly, based on the proportions of specific air pollutants from each source.

Table I does not provide optimal taxes that will internalize the externalities. The economic theory of externalities, depicted in Fig. 1, shows a tax on output that coincides with the units of measurement in Table I, whereas Fig. 2 shows a tax on externalities that would have units of measurement in dollars per pollutant. Even if the values were reported in dollars per pollutant, such calculations are illustrative only because of nonconvexities (e.g., NO_x is a nondepletable catalytic pollutant that converts into both ozone and particulate) and the inability of models to capture subtle effects that cascade between natural and social systems.

TABLE I

External Costs from Pollution and National Security^a

Energy source	Private cost market price 2000 \$	External marginal cost					Social marginal cost
		Subsidies	Carbon emissions	Air pollution	Other externalities	National security	
Natural gas (\$/mmbtu)	5.80	—	0.24	0.39	—	0	6.46
Crude oil (\$/bbl)	29.50	—	1.92	14.88	—	12.70	59.20
Coal (\$/ton)	31.70	—	9.05	17.98	—	0	58.73
Nuclear (\$/MWh)	6.75	1.80	0	0	1.80	0.12	8.68
Solar thermal (\$/MWh)	9.51	0	0	0	—	0	9.51
Baseload solar and gas (\$/MWh)	5.80	0	0.14	0.22	—	0	6.16

^aSource: Calculated from Hall (1990, 1992a,b) using the implicit price deflator for the Year 2000 relative to previous estimates.

Table I demonstrates that external costs of coal and oil are at least double the price; natural gas and nuclear power have much lower external costs than coal and oil; and the major problem with nuclear power is the private cost of building, maintaining, upgrading, and operating power plants and waste disposal. It is also clear from Table I that in the southwest United States thermal solar power with natural gas backup as a baseload unit has a lower social cost than nuclear power.

4. POLICY CHOICES

There are five aspects to good economic policy analysis. One is to consider all aspects of economic efficiency that policies affect. The second is to recognize when static economic efficiency arguments are not applicable due to misspecification of the alternatives, nonconvexities, nondepletable pollutants, uncertainty, timescale, and ignorance. The third is to incorporate the dynamic efficiency criterion in the context of state dependency of technology. The fourth is to consider policy strategies that increase flexibility and diversity to reduce macroeconomic risk. Finally, the fifth is to analyze wealth distribution effects of policies that determine political viability.

4.1 Evaluating All Market Failures That Policy Choices Affect

Externalities include benefits and damage from climate change. Many policies that would reduce emissions of GHGs would also reduce other forms of pollution caused by energy production and consumption, and some policies would also alter national security risks. Similarly, many policies that reduce regional air pollution also reduce water pollution, national security risks, and externalities from global warming. Policies to encourage investment in energy efficiency also address other forms of market failure, such as asymmetric information. Policies with joint benefits require benefit estimation that is inclusive and not exclusive; this is obvious, but it is rarely done in the economics literature.

Standard incentive-based policies include pollution taxes; tradable permits; hybrid combinations of long-term tradable permits with repeated, short-term permit sales; and permits tied to direct regulation of emission limits. These policies are evaluated with the static economic efficiency criterion but do not

address some of the market failures associated with underinvestment in energy-efficiency technologies.

Studies show internal rates of return $>20\%$ on commercial applications of energy-efficient lighting ballasts, whereas interest rates are much lower. These market imperfections—adverse selection, high transaction costs, principle agent problems, and intrafirm information flows—result in economic inefficiency. Specifically, asymmetric information exists between ballast sellers and potential buyers for improvement in quality, resulting in a lemons market and adversely selecting the lower quality (at the lower price). Because separate meters do not exist for each use of electricity, and a single meter is used to calculate electricity bills, businesses have a high cost of evaluating the efficiency of appliances. One principle agent problem occurs when commercial buildings are leased, the owner is the agent for the lessee (the principle) in the purchase of fluorescent lighting, and asymmetric information exists between the principle and the agent regarding the replacement costs. Another principle agent problem occurs when the renter is not separately metered, but rather the owner pays the electric bill and the renter and purchases the lower priced and inefficient appliance.

There are three additional sources of market failure for energy efficiency investments. First, electricity customers face higher borrowing rates for energy efficiency investments than the interest rates faced by electricity producers to invest in additional supply. Second, retail rate design for electricity and natural gas is based upon short-run marginal costs rather than long-run marginal cost rate, so that rates encourage investment in energy efficiency only up to the point that the capital and operating cost of the energy efficiency investment is cheaper than the operating cost of energy production. Third, there are negative externalities in energy production that are avoided by energy conservation. If the borrowing costs are lower for electricity producers than electricity efficiency investments by electricity consumers, or if the electricity is priced equal to short-run marginal cost of production, then at the margin electricity consumers will not invest in efficiency even when it is less expensive than the marginal cost of building and operating additional power plants.

4.2 Policy Options and the Static Efficiency Criterion

Policy options include direct regulation (or pejoratively, command and control regulation) and incen-

tive-based approaches, including pollution taxes and tradable permits allocated to polluters. Direct regulation includes requirements for pollution permits that polluters install the best available control technology when rebuilding power plants. Pollution taxes have fallen out of political favor, and may have constitutional challenges because the optimal pollution tax may vary over time and each tax adjustment could legally require congressional approval. Tradable permits give the holder the right to pollute a certain amount, to buy permits if emissions increase, and to sell permits if emissions decrease. One version of tradable permits is to lease the permit at a market auction. The so-called property rights solution is to allocate tradable permits in accordance with past levels of pollution, in essence appropriating common property and giving the property right to the polluter, without a corresponding property tax or any recognition of the economic rent.

If the property right to cause the externality resides with the party causing the damage, then the WTP is the correct marginal benefit curve. If the property right resides with the damaged party, the WTA is the marginal benefit curve (see Fig. 2). Solutions proposed by economists imply different allocations of property rights. Pollution taxes retain the property right for the public. Tradable permits, whose allocation is based on past levels of pollution, allocate the property right to the polluter. If tradable permits are leased at auctioned prices, the public retains the property right. When the U.S. Environmental Protection Agency changed direct regulations so that dirty power plants no longer had to be retrofitted to reduce pollution, the property right shifted from the pollutee to the polluter, and the marginal benefit curve was reduced by fiat, from the WTA to the WTP in Fig. 2. There is no unique optimal level of externality abatement without assuming who should own the property right, a normative issue.

Modern economic analysis favors the property rights solution to negative externalities. This selection is based on the criterion of static economic efficiency applied to Fig. 2. The solution requires setting the permitted amount of externalities equal to $N_C - N_P$, and polluters will trade permits if they can reduce pollution at a marginal cost less than the price of the permit. The result will be N_P abatement and the equilibrium price of the permit equal to the optimal pollution tax, T_P . This solution avoids the cost of direct regulation inherent in having regulators determine the cost of technology—a job best left to the market.

The property rights solution is an obvious misnomer since the other solutions—taxes, direct regulation, and

auctioned leases of tradable permits—all have property rights retained by the public. Because the tradable permits are not subject to property taxes, it is also a misnomer to refer to the right to pollute as private property. An alternative is to lease tradable permits, retaining the property right for the public; this is referred to as the common property solution, favored by a small but growing group of influential economists. The common property solution—leasing tradable permits to pollute $N_C - N_A$ —results in N_A abatement and an equilibrium price equal to T_A in Fig. 2.

Both the property rights tradable permits and the common property tradable permits options do not result in efficiency unless the amount of permitted pollution depreciates over time so that abatement increases in accordance with the analysis in Fig. 3, with the marginal benefit of abatement increasing over time due to the public good nature of abatement and population growth, and the increase in the pollution concentration shifting the intercept of the marginal benefit.

More problematic for the tradable permit approaches is that they do not acknowledge the joint pollutants, nonconvexities, and subtle cascading effects among natural and economic systems that render marginal benefit and marginal cost curves irrelevant.

4.3 Dynamic Efficiency and State Dependence of Technology

Some new technologies show promise for reducing externalities from energy. If CO_2 can be economically captured, it could be stored in several ways. One idea is to release CO_2 at the bottom of the oceans, but ocean currents could recirculate it. A second approach is to inject CO_2 into depleted oil and gas wells, raising pressure to capture remaining oil and natural gas and leaving the CO_2 in an impermeable reservoir; Ontario Power Generation has purchased the rights to GHG emission credits of 9 million tons of carbon and plans to inject the gas into old oil fields in the United States. Another idea is to inject CO_2 into coal mines, displacing and capturing methane gas. A Norwegian oil firm is injecting CO_2 into a saline aquifer under the North Sea, where the carbon is sequestered by reacting with the salt.

The integrated gasifier combined cycle (IGCC) is a technology that might lead to economically capturing CO_2 by burning pulverized coal in the presence of steam, generating CO and hydrogen gas. The hydrogen can be used to operate fuel cells or burned

to generate steam for electricity, and the CO can be catalytically converted to CO₂ and captured.

Fuel cells can chemically generate electricity from either hydrogen gas or methane gas. Some buses, trucks, and smaller vehicles are powered by fuel cells. Fuel cells can also be strategically placed to stabilize a transmission and distribution grid, avoiding urban air pollution emissions.

Infrastructure investments for energy supply and transportation—dams, power plants, pipelines, roads, and transportation networks—all have lifetimes in decades, even centuries. We can identify irreversible environmental damage of staggering magnitude lasting centuries, but natural science is unable to pinpoint how much pollution the earth's systems can sustain prior to triggering cataclysmic events. What is certain is that eventually we need to redirect the path of technological change.

In the 1960s, economists argued correctly that as rising market prices reflected scarcity of fossil fuels, the result would be substitution, new discoveries of fossil fuels, and technological change that reduced the cost of discovering, extracting, transporting, and converting fossil fuels to useful purposes. It is the success of these arguments that makes it clear that policies that encourage technologies such as the fuel cell, IGCC, solar thermal power, solar electric power, biomass, geothermal energy, ocean thermal energy conversion, wind power, and other substitutes for polluting fossil fuels will also lead to further improvements in fossil fuel technology because owners of fossil fuels have extraordinary wealth that can only be protected by research and development that furthers fossil fuel supply and use. Economic forces and institutional and political infrastructure lead to technological path dependency.

Direct regulation has an advantage over tradable pollution permits when evaluated by the dynamic efficiency criterion for selecting among policy options. Direct regulation can force technological change and can do so quite flexibly. An example is to require a fraction of sales to be zero- or ultra-low-emission vehicles, allowing vehicle manufacturers to trade the requirement among each other. Another example is to require a fraction of the power purchased by a utility to be generated by zero or lower polluting technologies, called a "set-aside" for renewable energy; this can be efficiently accomplished with a Vickery auction—such as the 1994 auction held by the California Public Utilities Commission and discussed above. Energy efficiency standards for building design, electric motors, appliances, heating, ventilation, refrigeration, and air-conditioning are also examples.

Technology-forcing policies include pricing and tax subsidies of renewable electricity technologies, and renewable energy tax credits; rebates, low-interest loans, and subsidized electricity rates for efficient appliance purchases; and mileage standards for vehicles (CAFE standards).

4.4 Flexibility and Diversity to Reduce Risk

Policies that reduce macroeconomic risk can be designed by applying the concepts of flexibility and diversity to hedge against adverse outcomes. Consider that an electric utility sets spinning reserve requirements at least as large as the amount supplied by the largest power plant in operation. Grid operation derates transmission capacity to account for loop-flow feedbacks that would cause the entire grid to fail if the largest transmission line were to short. Backup capacity is always on call in the amount equal to the largest single link on which the system relies. Our macroeconomic performance has hinged on a single supplier of oil, Saudi Arabia, with no backup capacity except the strategic petroleum reserve.

Solar thermal with natural gas backup, wind power, and other renewable technologies provide both flexibility and diversity. The small size of renewable electricity generators, coupled with the speed with which they can be built, permits a close matching of supply and demand, reducing the need for substantial reserve margins for electricity supply. Technology-forcing policies for energy efficiency and alternative energy sources result in more available substitutes in the event that fossil fuel prices increase.

4.5 Political Viability and Wealth Redistribution

The argument against pollution taxes is that the size of wealth redistribution makes taxes politically unfeasible. This argument is simply wrong and has been proven so. The argument against marginal cost water rates was that it is politically unfeasible—witness the Tucson, Arizona, city council voted out en masse 1 year after implementing such rates in the 1970s. However, a two-part increasing block tariff for water rates, based on long-term marginal cost, is politically feasible for the city of Los Angeles. Similarly, a two-part pollution tax could be levied with a low or zero tax up to some amount, and a higher tax thereafter, providing three policy tools: the initial tax, the breakpoint, and the higher tax. Furthermore, pollu-

tion taxes could be designed to be revenue neutral, potentially leading to greater efficiency in energy use and in other sectors of the economy. It is easy to imagine pollution taxes offsetting tax credits for technology development, for example.

Other pollution-reducing policies meet similar resistance. Auctioning pollution permits would be more efficient than allocating them to existing polluters, the latter model being the most often followed in the United States. Any effort to implement such an auction scheme is met with effective political resistance. Ironically, allocation with no auction is a large-scale wealth-distribution scheme that favors polluters.

Wealth redistribution clearly affects the political feasibility of policy options. Dating back at least to "The Logic of Collective Action," economists have recognized that the interests of small, well-organized groups regularly outweigh the larger public interest. The argument against pollution taxes is often couched in terms of the regressive effect on the average family. This is a disingenuous means of diverting attention from the very substantial economic rents that fossil fuel producers and marketers currently collect while placing health and other environmental risks on the families.

5. CONCLUSIONS

Economic analysis is useful for informing policy choices but not capable of proving that any single policy choice is more efficient than another. To do so would require the analyst to have all the information about costs and benefits that precludes central planners from replacing markets in the allocation of resources. The laws of chemistry and physics result in violations of assumptions in the theory of externalities. Furthermore, economic efficiency as measured in benefit-cost analysis is a narrow measure of net human welfare that indicates nothing about the equity of one option compared to another. It is the most inclusive method economists have developed to date, but it still is not appropriate as the primary means of choosing among policies that have far-reaching implications for this and future generations.

This should not be taken as an argument to abandon economic analysis but rather to work closely with other social and natural scientists to improve our ability to more accurately, transparently, and fully represent and compare alternative energy futures.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Energy, Overview • Cost-Benefit Analysis Applied to Energy • Electric Power Generation: Valuation of Environmental Costs • Geopolitics of Energy • Goods and Services: Energy Costs • Greenhouse Gas Abatement: Controversies in Cost Assessment • Motor Vehicle Use, Social Costs of

Further Reading

- Broadman, H. G., and Hogan, W. W. (1988). Is an oil tariff justified? An American Debate: The numbers say yes. *Energy J.* 9(3), 7–29.
- Cifuentes, L., Borja-Aburto, V. H., Gouveia, N., Thurston, G., and Davis, D. L. (2001). Hidden health benefits of greenhouse gas mitigation. *Science* 293, 1257–1259.
- Daily, G. (ed.). (1997). "Natures Services: Societal dependence on natural ecosystems." Island Press, Washington, DC.
- Hall, D. C. (ed.). (1990). Contemporary policy issues [Special edition]. *Social Private Costs Alternative Energy Technol.* 8(3). [Reviewed in *Rev. Gen. Thermique* 350, 1991]
- Hall, D. C. (1992a). Social cost of CO₂ abatement from energy efficiency and solar power in the United States. *Environ. Resour. Econ.* 2, 21–42.
- Hall, D. C. (1992b). Oil and national security. *Energy Policy* 20(11), 1089–1096.
- Hall, D. C. (ed.). (1996). Advances in the Economics of Environmental Resources: Marginal Cost Rate Design and Wholesale Water Markets. JAI, Greenwich, CT. [Reviewed by G. L. Carriker (1998). *Am. J. Agric. Econ.* 80(4).]
- Hall, D. C. (1998). Albedo and vegetation demand-side management options for warm climates. *Ecol. Econ.* 24, 31–45.
- Hall, D. C., and Hall, J. V. (1997). Estimating the benefits of emissions reductions in complex atmospheres. *Int. J. Global Energy* 9(4–6), 286–298.
- Hall, D. C., and Howarth, R. B. (eds.). (2001). The long-term economics of climate change: Beyond a doubling of greenhouse gas concentrations. *Adv. Econ. Environ. Resour.* 3, JAI
- Hall, D. C., Win, M. S., and Hall, J. V. (1995). Air pollution impacts from demand side management. *Energy Int. J.* 20, 27–35.
- Hall, J., Winer, A., Kleinman, M., Lurmann, F., Brajer, V., and Colome, S. (1991). Valuing the health benefits of clean air. *Science* 255(5046), 812–817.
- Hamilton, J. D. (1996). This is what happened to the oil price-macroeconomy relationship. *J. Monetary Econ.* 38, 215–220.
- Hanemann, W. M. (1991). Willingness to pay and willingness to accept: How much can they differ? *Am. Econ. Rev.* 81(3), 635–647.
- Koomey, J. G., Sanstad, A. H., and Shown, L. J. (1996). Energy-efficient lighting: Market data, market imperfections, and policy success. *Contemp. Econ. Policy* 14(3), 98–111.
- Kothavala, Z., Oglesby, R. J., and Saltzman, B. (1999). Sensitivity to equilibrium surface temperature of CCM3 to systematic changes in atmospheric CO₂. *Geophys. Res. Lett.* 26, 209.
- Levitus, S., Antonov, J. I., Boyer, T. P., and Stephens, C. (2000). Warming of the world ocean. *Science* 287(5461), 2225–2229.

- McElwain, J. C., Beerling, D. J., and Woodward, F. I. (1999). Fossil plants and global warming at the Triassic–Jurassic boundary. *Science* 285, 1386–1390.
- Rosenfeld, A., Atkinson, C., Koomey, J., Meier, A., Mowris, R., and Price, L. (1993). A compilation of supply curves of conserved energy for U.S. buildings. *Contemp. Policy Issues* 11, 45–61.
- U.S. Environmental Protection Agency (1999). “Benefits and Costs of the Clean Air Act 1990–2020.” U.S. Environmental Protection Agency, Washington, DC.
- U.S. Environmental Protection Agency (2001). “Benefits and Costs of the Clean Air Act 1990–2020: Draft Analytical Plan for EPA’s Second Prospective Analysis.” Office of Policy Analysis and Review, Washington, DC.
- Viscusi, W., Magat, W., Carlin, A., and Dryfus, M. (1994). Socially responsible energy pricing. *Energy J.* 15(2), 23–42.
- Walton, A., and Hall, D. (1990). Solar power. *Contemp. Policy Issues* 8(3), 240–254.

Fire: A Socioecological and Historical Survey

JOHAN GOUDSBLOM

University of Amsterdam
Amsterdam, The Netherlands



-
1. The Original Domestication of Fire
 2. Agrarianization
 3. Industrialization
 4. Symbolic and Religious Aspects
 5. Conclusion: The Eclipse of Fire?

Glossary

agrarianization The formation and spread of a socio-ecological regime structured around the domestication of plants and animals.

anthroposphere That part of the biosphere that has been affected by humans.

domestication The “taming” and incorporation into human society of originally “wild” nonhuman sources of energy.

industrialization The formation and spread of a socio-ecological regime structured around the exploitation of fossil fuels.

The ability to handle fire is, along with language and the use of tools, a universal human attainment: no human society in historical times is known to have lived without it. The ability to handle fire is also exclusively human: whereas other animals use rudimentary signals and tools only humans have learned to handle fire. From the outset, mastery of fire has strengthened the position of humans in the biosphere. At the same time, it has deeply affected the relations among humans themselves. Appreciation of the control over fire is therefore essential for an understanding of both the external expansion and the concomitant long-term internal transformation of the anthroposphere. The control over fire has been essential in the emergence of agriculture and stock raising, as well as the rise of modern industry. Its

implications, including both benefits and costs, are so far-reaching as to often remain unnoticed and taken for granted. This entry intends to convey the enormous importance of the fire regime for human history and its profound impact on the biosphere.

1. THE ORIGINAL DOMESTICATION OF FIRE

1.1 Origins

Human life, like all life, consists of matter and energy structured and directed by information. All life is part of an ecosystem, and all ecosystems together constitute the biosphere—the total configuration of living things interacting with each other and with nonliving things. Every form of life continuously affects, and is affected by, its ecosystem.

As participants in the biosphere, the early hominids and their only surviving species, *Homo sapiens*, have gradually strengthened their position—at first slowly and almost imperceptibly, later at an increasingly more rapid pace with ever more striking consequences. In the process, they engrossed increasingly more terrain and incorporated increasingly more non-human resources into their groups: first fire, then, much later, certain selected plants and animals, and, later again, fossil fuels. As they incorporated more energy and matter into their societies, those societies grew in size, strength, and productivity, while at the same time also becoming more complex, more vulnerable, and more destructive.

During the first stage in human history and “prehistory,” known in archaeology as the Paleolithic or Old Stone Age, which lasted for thousands of millennia, the overall pace of social and cultural development was slow in comparison with later

stages. Yet some momentous changes took place, with great consequences for the relationships between humans and the natural environment. Humans began not only to use but also to make tools of their own, and they learned to control fire. The combination of tools and fire enabled groups of humans to leave their original habitat, the savannas of East Africa, and to migrate into other parts of the world, penetrating first into remote corners of Eurasia and then also into Australia and the Americas. The Paleolithic can thus be seen as a long run up, which was later followed by an enormously accelerating sprint of which our present age is the latest episode. It was the scene of the incipient expansion of the human realm, the anthroposphere.

A basic trend in all human history, and certainly during its earliest phases (often designated as prehistory), has been the increasing differentiation between humans and all closely related animals in terms of their behavior, power, and general orientation or attitude—their *habitus*. Thanks to the flexibility acquired in the course of evolution, humans were able to learn a large repertory of new forms of behavior. Especially successful were those innovations in behavior that added to human power vis-à-vis other large animals, both predators and competitors. Transmitted by learning from generation to generation, those innovations entered into the human *habitus* and became second nature.

The primary condition for the process of differentiation in behavior, power, and *habitus* has been, and continues to be, the inborn human capacity for culture as manifested in technology, organization, and civilization, each of which represents the results of social learning. Indeed social learning is the crux of culture: gathering information and passing it on to others—at first in direct interaction, at a later stage in history through written texts and today by other audio-visual means as well. The stores of cumulated information (or “cultural capital”) have enabled people to tap increasingly larger and more varied flows of matter and energy and to integrate those flows into their societies.

The domestication of fire was the first great human interference with natural processes. It had numerous far-reaching implications, stretching from the first hesitant beginnings to our contemporary fuel-driven economy. It demands our attention because it reveals something of the socioecological infrastructure of the contemporary world.

In several ways, the original fire regime may be seen as a paradigm for the socioecological regimes that were developed later. It presents a paradigm in a

double sense. First, in actual practice the regime by which humans learned to extend their care for and control over fire could serve as a model for subsequent forms of care for and control over other forces in nonhuman nature such as plants and animals. Second, we may regard the domestication of fire as a model case in a more theoretical fashion, since it brings out the strong link between such apparently contradictory tendencies as increases in control and dependency, in robustness and vulnerability, and in the potential for production and for destruction.

1.2 First Impact

Fire, like all natural forces, has a history. Chemically fire is a process of highly accelerated oxidation of matter (fuel) induced by heat (ignition). Three conditions are therefore necessary for it to occur: oxygen, fuel, and heat. During the first eons in the history of the earth, at least two of these, oxygen and fuel, were absent. Oxygen did not become available until, after at least a billion years, life emerged. And it was only less than half a billion years ago, during the Devonian geological age, that life assumed the form of plants, providing matter suitable for burning. From then on, most places on earth with seasonally dry vegetation were regularly visited by fire, ignited on rare occasions by falling rocks, volcanic discharges or extraterrestrial impacts, but mostly by lightning.

Its domestication by humans opened an entirely new episode in the history of fire. Humans thoroughly altered the frequency and intensity of fires. They brought fire to regions of the planet where it seldom or never burned spontaneously, and they tried to banish it from places where without human interference it would have burned repeatedly. Thus, increasingly, “natural” fire receded and made way to human or, more precisely, anthropogenic fire.

Wherever humans migrated, they took their fire along. The presence of humans with fire deeply altered the landscape, including flora and fauna. The human impact is amply documented (though still controversial) for Australia—a continent that was colonized by humans rather late. Everywhere on the planet, areas such as rain forests, deserts, and the polar regions, which were not receptive to fire proved to be hard to penetrate for humans too.

Humans are the only species that has learned to manipulate fire. Control over fire has become a species monopoly, with an enormous impact on other species, both animals and plants. It provides us with

an excellent example of how new forms of behavior could change power balances—in this case, between humans and all other animals, ranging from primates to insects—and how shifts in power balances could engender changes in habitus, both among the humans who gained greater self-confidence from the presence of fire in their groups and among animals that might be bigger and stronger than humans but learned to respect and fear their agility with fire.

Control over fire, in addition to having become exclusively human, has also become universally human. We know of no human society of the past 100,000 years that lacked the skills needed to control fire.

The original domestication of fire was a dramatic transition. Palaeoanthropologists are still debating when exactly it took place. The estimates range from as many as 1.5 million to a mere 150,000 years ago. It is still an open and fascinating question whether the first steps to control over fire coincided with other changes in early human development.

In retrospect, the initial domestication of fire was an event of momentous import. A wild force of nature—blind, capricious, and hazardous—was now tended, cared for, and supplied with fuel. Our early ancestors went to all this trouble, not altruistically, but because it served them well. They put the potentially destructive and essentially purposeless force of fire to work for their own productive purposes. They managed to make fire regularly available. They no longer had to hunt for it, hoping to find somewhere the smoldering remains of a natural blaze; they made it a part (and even the center) of their own group, and they revered it as a symbol of eternal life.

The domestication of fire made humans less directly dependent on natural forces that continued to be beyond their control such as the alternation of day and night or the cycle of the seasons. It made the contrast between dark and light, between warm and cold, or between wet and dry more amenable to manipulation, and thus gave humans a greater margin of freedom from the grip of nature. It increased their power, defined as the capacity to influence the outcome of an interaction. Armed with fire, humans were able to open up impenetrable tracts of bush and to drive away animals much fiercer and stronger than they were. The gain in power made their lives more comfortable and secure. The possibilities of heating, lighting, and cooking all contributed to what we would now call a higher standard of living.

1.3 Long-Term Consequences

“Wherever primitive man had the opportunity to turn fire loose on a land, he seems to have done so from time immemorial.” This statement by the American geographer Carl Sauer may sound exaggerated; but it still fails to convey the full impact that the domestication of fire has had, both on the larger biosphere and, within the biosphere, on human society itself.

The most immediate effect of the domestication of fire on the biosphere in general was an increase in the frequency with which fires occurred. Prior to its human mastery, fire was ignited mostly by lightning. From now on another source was added: even before they learned to make fire themselves, humans were able to preserve it in their hearths and to apply it wherever they saw fit. Consequently, as the number of anthropogenic fires increased, the proportion of natural fires diminished. The geologist Peter Westbrook has suggested, in an admittedly speculative manner, that the earliest human fire use may have affected the planetary atmosphere and caused some of the major climate changes in the Pleistocene. Substantive evidence indicating modification of the landscape by human foragers equipped with fire has been brought forward for Australia, where most of the indigenous forests were burned down in the millennia following the arrival of the first Aborigines.

From the very beginning, humans used fire in two basic forms: the hearth and the torch. The hearth was the original site at which a fire was kept, usually at a cave entrance where it could be protected against the rain and still receive some air circulation. Since it had to be tended and since fuel had to be brought to it, it served almost naturally as a center for group life, providing heat, light, and a common focus. From the hearth developed in the course of time a variety of fire containers such as crucibles, stoves, kilns and furnaces, and, in our day, the mobile engines of cars and airplanes.

The hearthlike uses of fire have always had two kinds of environmental side effects. First of all, fuel had to be supplied. As long as human communities were small and living in areas with abundant wood, this did not cause much of a problem. When greater numbers of people started living in large urban concentrations, however, the need for fuel became a strong contributing factor to deforestation over big areas and, in our own age, to depletion of fossil resources. The second side effect of hearthlike fire consists of its waste products: ashes and smoke. Although smoke might be useful in driving away

insects and other insidious creatures, it has always been considered mostly as a nuisance to be got rid of. As long as people lived in isolated caves or huts, this was relatively easy. Urbanization and industrialization have seriously aggravated the problem.

While the functions of the hearth were originally primarily turned inward, the torch was a more outward directed implement. It was used not only to provide light at night, but also by day to set fire to shrubs and grasses—an effective way to destroy obstacles for foraging and to rout animals, both predators and prey, and thus to extend the human domain. Torches undoubtedly contributed to deforestation: wood was burned wholesale, regardless of its possible value as timber or fuel. In the age of agriculture, the torch was used for slash and burn and other techniques for clearing land, and it served as the model for a whole array of fire weapons culminating in our own time in rocket-propelled missiles.

Surveying the entire trajectory of the human use of fire from its earliest beginnings, we can distinguish three stages. During the first stage, there were no groups possessing fire; there were only groups without fire. Then, there must have been a second stage, when there were both groups with fire and groups without fire. We do not know how long that stage lasted, nor do we know how often it may have recurred. All we know is that it has come to an end. It was a transitional stage, leading up to the stage in which humankind has now been living for thousands of generations: the stage when there are no longer any groups without fire. All human groups are groups with fire.

Although we lack empirical evidence for the first two stages, that very lack leaves us no choice but to accept the conclusion that societies with fire were in the long run obviously more fit to survive than societies without fire. If we then ask why it was that societies without fire disappeared, there seems to be only one plausible answer: because they had to coexist with societies with fire, and apparently in the long run such coexistence proved impossible.

This may sound like a dismal conclusion suggesting fierce contests ending in the elimination of the losers. If such contests did indeed take place, they have left no trace of empirical evidence; we only have the highly imaginative evocations of what might have happened in books and films like *The Quest for Fire*, directed by Jean-Jacques Annaud. However, we can also view the fact that possession of fire has become a universal attribute of all human societies as an important example of the general rule that changes

in one human group lead to changes in related other groups. If group A had fire, and the neighboring group B did not, group B had a problem. It could either try to minimize contact with group A and perhaps move away, or do as group A had done and adopt a fire regime, which should not pose insurmountable difficulties as long as the capacity to learn from the others was sufficient. In the latter case, instead of a zero-sum elimination struggle there would have been what American freelance author and scientist Robert Wright calls a “nonzero” situation, with an outcome from which both parties benefited.

The rule that changes in one human group lead to changes in other related groups may sound like a rather tautological explanation for social change, but it is not. It is a highly generalized empirical observation, similar to an observation we can make about fire: fire generates fire and, in a similar fashion and more generally, change generates change, and social change generates social change.

Such is the nature of the dynamics of human society and culture. After the original domestication of fire, it was never humans alone who interacted with other human groups and with nonhuman nature. It was always humans with fire, equipped with fire and with the products of pyrotechniques: cooked food, pointed spears and arrows, earthenware, metal tools and weapons. Their presence put an end to humans without fire.

Another general conclusion to be drawn from these observations is the following: changes in climate and precipitation have never ceased to be important causes for humans to change their way of life. Humans are no different from other species in that they will always have to accommodate the basic conditions of earthly nature, such as the alteration of day and night or of monsoons and seasons. In the course of human history, however, in addition to these overridingly powerful extra-human conditions, conditions brought about by humans themselves have become increasingly more important to the extent that, in our contemporary world, humanity has become a major agent of ecological change.

1.4 Regimes

The domestication of fire meant that people tamed a strong and potentially destructive natural force and made it into a regularly available source of energy. In so doing they initiated changes in the natural environment, in their social arrangements, and in their personal lives. These three aspects (ecological,

sociological, and psychological) are all part of the changing human relationship with fire.

In its ecological aspects, the domestication of fire affected the relations between humans and the nonhuman world so deeply that we may call it the first great ecological transformation brought about by humans, which was followed much later by the second and third of such transformations, generally known as the agricultural and industrial revolutions and better characterized by the long-term processes of agrarianization and industrialization.

Each of the three transformations spelled the formation of a new socioecological regime: the fire regime, the agrarian regime, and the industrial regime, marked by, respectively, the utilization of fire and elementary tools, the rise and spread of agriculture and animal husbandry, and the rise and spread of large-scale modern industry. The later regimes have not made the earlier regimes obsolete; rather, they have absorbed them and, in so doing, transformed them. Each new regime brought an expansion of the anthroposphere within the biosphere.

Defining the three regimes jointly in similar terms is helpful in order to better understand each of them separately as well as in their interrelations. A common conceptual model invites and facilitates comparison. The comparison allows us to explain the sequence in the emergence of the regimes and to perceive not only their similarities and differences but also their interlocking.

2. AGRARIANIZATION

2.1 Continuities

The history of the past ten thousands years may be read as a series of events accompanying the process of agrarianization of humankind, a process in the course of which humanity extended the domain of agriculture and animal husbandry all over the world and in so doing made itself increasingly more dependent on this very mode of production.

The initial transition from gathering and hunting to agriculture and animal husbandry was not necessarily abrupt. A group that started to cultivate a few crops would not have to give up its older ways altogether. There would have been only few, if any, agrarian societies from which gathering and hunting disappeared completely at once. However, the proportion of products acquired in the older way inevitably diminished as agriculture and animal husbandry advanced.

From the very beginning, the process of agrarianization was linked closely to the domestication of fire. It is hard to imagine how people could have begun to cultivate plants and to domesticate animals had the art of handling fire not already been thoroughly familiar to them. For one thing, they needed a hearth fire to cook on. The first crops cultivated on any large scale were cereal grains such as wheat, rice, and maize, which, owing to their high nutritional value and their capacity to withstand storage for long periods, formed very appropriate staple foods for a human community; to serve this purpose, however, they had to be made more easily digestible with the help of fire.

A second and very different reason why the control of fire formed a precondition for agrarianization was the human predominance over all other mammals, which was grounded partly in the use of fire. The human species' monopoly over fire was so solidly established by the time agriculture began, and is today so easily taken for granted, that it is seldom given separate attention in this context. Yet it deserves mention. Their hegemony in the animal kingdom enabled people not only to bring certain species, such as goats and sheep, under direct control, but also—at least as important—to keep most of the remaining “wild” animals at a distance from their crops and herds.

Third, experience in controlling fire may have furthered plant and animal domestication in another, even more intangible way, which our distance in time makes it difficult to assess precisely but which we are also, for that very reason, likely to underestimate. The time-honored practice of handling fire could not have failed to prepare humans for the many chores involved in agriculture and animal husbandry. It taught them that extending care upon something nonhuman could be well worth the trouble and thus made it more acceptable to them to accommodate the strains of an agrarian life, full of self-imposed renunciation for the sake of a possible future yield.

The most immediately visible link between early agriculture and ancient use of fire lay in the custom of burning off land with an eye to food production. Of old, foraging peoples were wont to apply their torches in order to keep the land open for gathering and hunting. Even in recent times, those firing practices were continued in some parts of the world, as in Australia where the Aborigines' judicious use of fire in keeping their land open for kangaroos and humans has become known as “firestick farming,” a term suggesting a form of proto-agrarianization.

2.2 Sequences

The rise of agriculture was in many respects remarkably similar to the domestication of fire. Again, humans added new sources of energy to their own, this time by adopting certain plants and animals and integrating them into human societies. Plants that were formerly wild now began to be cultivated, wild animals were tamed and used for food or other purposes such as traction, and all these species were made part of the human domain, of the anthroposphere, which thus increased in size and complexity.

The transition from foraging to agriculture did not automatically make people happier and healthier. Agrarian life brought new hardships. Diets became more monotonous. Sedentary life in villages increased the susceptibility to disease. It allowed for a rise in fertility, but it also caused higher mortality. The result was large numbers of children, many of whom never reached adulthood. Not surprisingly, research on skeletons reveals that after agrarianization, human lives tended to become briefer, and bodies shorter. It is equally understandable that the texts of the great religions of Eurasia all exhibit a nostalgia for the lost paradise of a pre-agrarian era.

The very nature of agrarian life foreclosed a return to foraging, however. There are a few known cases of such a return, but those are exceptions. Almost everywhere, with so many people, and so little land, the only option for the survival of an agrarian population was cultivation. "Work" was writ large over the agrarian world.

If, at a very early stage, the domestication of fire had made human groups more productive but also more vulnerable (because from now on they had to rely on their fire), the rise of agriculture had the same twofold effect. Being able to grow more food, human groups grew more populous and became more dependent on their crops, and thus more vulnerable to failure or loss of their harvests. As the expansion of agriculture and pastoralism left increasingly less land available for foraging, the opportunities to escape from this vicious circle dwindled.

The first stage in the process of agrarianization necessarily involved clearing the land, removing any existing vegetation that would compete with the planted crops. In many cases, the most efficient procedure to accomplish this was by means of fire. As long as land remained plentiful, continued recourse was often taken to fire in a system practiced throughout the world and known under various regional names that are usually subsumed in the standard literature under the label "shifting cultivation."

Shifting cultivation implies that a piece of primary forest is first cleared by the method of slash and burn and is then used for one or more harvests of crops. When, after a while, crop nutrients in the soil become exhausted and undesired plants (weeds) begin to dominate, the farmers temporarily abandon the land and turn to an adjacent lot, burn the vegetation down, and bring it under cultivation, again until they find harvesting unrewarding. Eventually they return to their first plot, which by then has become secondary forest or secondary bush, and resume their activities of burning, planting, and harvesting there. The duration of the entire cycle may vary as to time and place, but the principle remains the same.

In many parts of the world, in the course of time, burning the land and letting it lie fallow for a number of years or seasons ceased to be a regular practice and was replaced by more intensive methods of working the soil, requiring greater investments of labor but yielding a larger output per acre and thus making it possible to feed more mouths. The most common means of accomplishing higher yields by harder work were irrigation and ploughing.

2.3 Fire in Settled Agrarian Societies

In all agrarian societies with permanent settlements, new uses of fire and new attitudes toward fire developed. During the long first stage of human fire use, the main concern always was to keep the communal fire burning. From now on, however, the main concern became almost the opposite: to keep the fires that were lit for numerous purposes from spreading and from running out of control.

The uses of fire became increasingly more varied. Specialized pyrotechnic crafts emerged, such as blacksmiths and potters. Among the growing urban populations, fire was regarded with greater anxiety, for several reasons: with the proliferation of fires, the risks of conflagrations increased, and with the accumulation of property, people had more to lose.

Of course, it was fire as a natural force with its self-generating destructiveness that was feared. But more and more, this natural force manifested itself almost exclusively in the guise of anthropogenic fire. With all the fires burning in a city, one moment of carelessness might cause a conflagration. People had to rely on other people's caution. They had to oppress attempts at deliberate fire setting. And they had to reckon with the very worst danger: the organized form of murder and arson known as war.

A common problem in all advanced agrarian societies was the prevention of uncontrolled fire in

cities. This problem may have been less urgent where the major construction material was stone or brick; but throughout the agrarian era we find major cities in which most houses were built from timber. Rome, Constantinople, Moscow, London, and Delhi as well as the capitals of China, Korea and Japan all suffered conflagrations; all of them faced problems of fire prevention.

Today we may tend to conceive of problems of fire prevention primarily in terms of technology: of building materials and technical equipment. However, the problems were (and still are) at least as much civilizational, or social. The crucial issue was human behavior, care, consideration. According to a well-known story, a temple in ancient Greece burned down because the servant was careless; similar events must have occurred all over the world. Everywhere it was the human factor that counted most of all.

This was also reflected in the presence or absence of fire brigades. Today, it may seem self-evident that a city should have a fire brigade. However, the power relations in earlier societies could be such that the authorities would forbid the organization of fire brigades by citizens. A dramatic example of a ban put on urban fire brigades is provided by two letters that were exchanged in the second century CE between the governor of a province in the eastern part of the Roman empire, Pliny the Younger, and his emperor, Trajan. In the first letter, Pliny informed the emperor that the capital city Nicomedia was completely ruined by a fire; people had watched the blaze passively, unable to do anything. He therefore asked permission to establish a fire brigade of citizens in order to prevent such disasters in the future. The emperor denied the request bluntly, for—as Pliny should have known—such an organization of citizens could easily turn into a club with subversive political activities.

It is still unknown to what degree Trajan's reply was typical of rulers in agrarian empires. The scattered evidence provided by local histories suggests that the situation was very different in the towns of medieval Europe, where the members of various guilds were required to take a share in fighting fires. This issue is a promising subject for comparative historical research.

3. INDUSTRIALIZATION

3.1 The Industrial Regime

While the basic theme of human history over the past 10,000 years has been the agrarianization of the

world, the history of the past 250 years has unfolded under the aegis of industrialization. In the process, the anthroposphere has become one global constellation, extending all over the planet, and its impact on the biosphere has become more and more intense.

In line with the approach to the domestication of fire and agrarianization, industrialization can be defined as the formation of a socioecological regime, structured around a new source of energy: fossil fuel—first in the form of coal, later also of oil and gas. The nature of this new energy source has made the industrial regime different in a basic way from the earlier socioecological regimes. Unlike plants and animals, and even wood, fossil fuel is not directly connected to the continuous flow of solar energy. It is a residue of solar energy from a remote past contained in geological formations. The energy stocks are not diffuse like sunlight but concentrated in particular locations from where they can be extracted through concerted human effort. They have two seemingly contradictory properties: they are abundant and finite.

The abundance is indeed great. Coal, oil, and gas represent the remains of the nonoxidized biomass—in other words, unburned fuel—of over 3 billion years. When people began exploiting those enormous reserves they entered as it were, in the words of the environmental historian Rolf Peter Sieferle, a “subterranean forest” of inconceivably large dimensions, which, moreover, in the course of 250 years of exploitation and exploration proved to contain far more riches than was originally expected.

Yet no matter how large, the hidden stocks are also finite. In contrast to plants, which are the direct autotrophic products of the photosynthetic conversion of solar energy, and to animals, which are the heterotrophic consumers of plant and animal food, geological stocks of fossil fuel do not partake in any living metabolism. They are incapable of growth or reproduction, irreplaceable, unrenovable and, as is becoming more and more evident, their use generates combustion products that enter the biosphere.

3.2 Origins and Preconditions

The genesis of industrialization raises a whole array of intriguing questions concerning its preconditions and its immediate causes as well as its functions and side effects. These are in principle the same questions that can be raised about the domestication of fire and the origins of agriculture. Because industrialization began only recently, we have the benefit of far more and far more precise empirical evidence; never-

theless, the problems remain puzzling. Any answer that comes to mind can only be tentative and subject to caveats.

One thing is certain: industrialization was not triggered by a major change in climate. At most, its beginnings more or less coincided with the end of the most recent secular dip in temperature in Europe, from 1550 to after 1700, known as the little Ice Age; but no relationship with this climatological event was evident. Industrialization was a completely anthropogenic transformation, brought about by humans in societies that were fully equipped with fire and agriculture.

The beginning of industrialization heralded a new era in human history. Yet it was clearly a continuation of the fire regime, for every industrial process rested on controlled use of fire. Industrialization also presupposed agrarianization: it could never have started without a strong agrarian basis that was able to feed numerous people who were not engaged in agriculture themselves and who could therefore be recruited for work in industry and trade. From the outset, the industrial world coexisted and coevolved with the agrarian world, in close interdependence and symbiosis, culminating in our time in forms of industrial farming or agroindustry that lean as heavily on fossil fuel as any other form of industrial production.

3.3 Further Parallels

Just as agrarianization must have begun in small farming enclaves, carved out in an environment that continued to be the domain of foragers, industrialization started with single steam-powered factories—often called mills as if they still were driven by wind or water—standing apart in the agrarian landscape. To sketch their initially semi-isolated position, Sieferle's image of industrial archipelagos is appropriate.

From the very start, however, even as islands in an agrarian landscape, the factories had an ecological impact stretching beyond the land on which they stood. To begin with, considerable amounts of matter and energy went into building them. The usual construction material was partly timber, but to a larger extent brick, the manufacturing of which involved huge quantities of fuel. The engines were made largely of metal, also in a highly fuel-intensive fashion. Then, brick and timber were needed for the houses to accommodate the workers. So even before the factories began to operate, their tentacles were already reaching into the environment.

Once in operation, the factories had to be supplied with a continuous input of material, such as iron and cotton to be processed and of fuel to keep the engines going. The need for fuel explains the location of the early industrial plants: close to coal mines and, in the case of heavy industry, sites of iron ore deposits. Of course, a nearby sea or river port facilitated transportation; failing that, canals were built, and later railways, connecting the islands of industrial production with the areas where the raw materials were found and with the markets where the products were sold. In 19th-century Britain, an ever more extensive and intricate network of canals and railways was formed through which the industrial regime spread its reach over the entire country.

3.4 Industrialization as Intensified Land Use

Our earliest human ancestors treated the land they lived in as essentially all of the same nature—a territory for both collecting and hunting, for foraging food as well as fuel. In agrarian societies, most of the land was parceled up into three distinct parts with distinct functions: fields or arable planted with crops; pastures or meadows where cattle and horses grazed; and heath and wood land providing fodder for pigs and sheep and, even more important, fuel and timber. All three areas were controlled by humans, who were able to make the land thus divided up more productive for their own specific purposes, but who also found themselves faced with a continuously decreasing wild acreage.

Industrialization offered an escape from this predicament. The exploitation of coal was a highly intensified form of land use, which relieved the pressure to use large areas extensively for growing wood for fuel. The virtual expansion of land attained through the use of coal helps to explain the unstoppable advance of industrialization, once it was underway. Like the agrarian regime, the industrial regime made offers that were disagreeable in many ways to a great many people, but in the long run for all of them impossible to ignore or to refuse. The huge gains in energy could be transformed into economic, political, and military advantages that proved to be irresistible. All over the world, societies without fossil energy industry made way for societies with fossil energy industry, a process which could not fail to have profound consequences for the biosphere.

In our own age, each year more than 100,000 square kilometers of remaining wild land in Africa, Asia, and South America are subjected to human

exploitation. One of the causes is the need for fuel in the Third World; in many densely populated areas, fuel for the fire to cook on has become as proverbially scarce as the food to be cooked. In addition, there is a rising demand for timber and pulp from the most highly industrialized countries. The greatest threat to the continued existence of natural forests does not lie in felling the trees, however, but in the indiscriminate burning down of entire tracts in order to clear the ground for raising crops and cattle for commerce. The world economy generates a rising demand for products of tropical agriculture. To meet this demand, garden plots are converted into plantations, and the small farmers have to leave. At the same time, population continues to grow, so that pressure on land becomes even greater. As a result, the contradictory situation arises that in a world suffering from severe shortage of wood, each year many tens of thousands of hectares of forest are set on fire.

In Australia, exceptionally, anthropogenic fire has diminished since the establishment of a European colonial regime. From the time of their first arrival, the British had a sensation of the Aborigines going about burning the land incessantly. The British took every measure they could to repress these burning practices, with such success that rainforest has been regaining ground again in Australia over the past 200 years. An unintended side effect has been the increasing incidence of large wildfires during dry seasons, for the many fires of limited scope lit by the Aborigines prevented the vast fires that now, when they are raging, constitute a serious threat to the metropolitan suburbs.

3.5 Current Uses and Manifestations of Fire

The 20th and 21st centuries have seen a common shift in preference for energy sources all over the world. Wherever electricity is now available, it is the favorite means of generating light, and combustion engines are, along with electricity, the favorite means of generating motion.

The shift from steam engines to electricity and internal combustion is reflected in the changed industrial landscape, which in the Western world is no longer dominated by endless rows of factory chimney stacks. Yet the methods of production in modern industry, and in agriculture as well, continue to be highly fuel-intensive. Most of the energy consumed, including most electricity, is derived from the fossil fuels coal, oil, and gas. Combustion

processes thus still play a central role, but they are relegated to special containers so that most people are not directly confronted with them. Soot, smoke, and fire risks are reduced to a minimum. The furnaces and combustion chambers in which enormous heat is concentrated remain cool on the outside.

Typical products of modern fuel-intensive industry are cars, with engines designed to be propelled by finely tuned and minutely controllable combustion processes. Indeed, the car may almost serve as a symbol of the highly complex and differentiated ways in which, in our day, thermal energy is being used. Cars are set in motion by burning fossil fuel. They are made of steel, plastic, and glass—materials that are produced and processed at high temperatures. Yet no one who gets into his or her vehicle and turns on the electrical ignition to start the engine needs be consciously aware of using fire and products of fire. When driving, people do not perceive the processes of combustion that keep their car going: they do not see the petrol gas burning under the bonnet, nor have most of them even remotely sensed the fire in the factories and power plants without which their cars would never have been produced at all.

A very different example to the same effect is farming. In the early 19th century, when Britain was already beginning to industrialize, practically all the energy consumed on the farm was still produced within the confines of the farm and its fields in the form of human and animal labor; the open fire that was burning in the hearth was fuelled with wood from the immediate surroundings. By the turn of the 21st century, the situation has become very different, with practically all the energy used now brought in from outside the farm, in the form of fertilizer, oil and petrol, and electricity.

A major advantage of the new sources of energy is their flexibility. The fuels are easier to transport and to distribute than wood or coal, and the combustion can be regulated more precisely. Given the technical facilities, gas, oil, and electricity provide for very even and accurately controllable flows of energy. Electricity has the additional advantage of being totally “clean” at the place of destination. Domestically, a few simple actions and a negligible risk suffice to provide people with an immense array of services: some substituting for old chores such as cleaning the floor and washing dishes, others based on entirely new appliances such as television sets and computers. Industrially, the same advantages apply at a much larger scale, permitting a far more diversified use of power than was possible with steam power.

The impact of electricity and internal combustion engines makes itself felt in every sector of social life: in agriculture, industry, traffic and transportation, domestic work and leisure. Everywhere it is possible to mobilize large quantities of energy with very little physical effort. The result is to make life more comfortable in many respects, enhancing the sense that physical processes can be mastered and also, at times, fostering the illusion of independence.

An illusion it clearly is. Regardless of whether people can avail themselves of energy in the form of a petrol engine, a battery, or a connection to an electric circuit or a gas main, in each case they are able to do so only because they are part of a complex and far reaching network of social interdependencies, connecting them eventually to the energy stored in fossil fuels. As long as the supply lines are functioning, and as long as people are able to meet their financial obligations, they do not need to bother much about the entire constellation. They are immediately confronted with it, however, the moment something goes wrong with any of the conditions.

In this way the exploitation of the new sources of energy clearly continues a trend that has begun with the first domestication of fire. Dependence on the forces of nature has become more indirect (which is not to say less important!), and by the same token dependence on cultural and social resources has increased. A complicated technical and organizational apparatus is needed to continuously maintain the supply of energy. Most of this apparatus is located behind the scenes of industrial society, invisible to the ordinary consumer. Energy is made available in such a convenient fashion that it is easy to forget the social effort required to produce it.

That social effort is spent, first of all, at the drilling wells and in the mines where the energy is won and, next, during the operations of processing it into consumable gas, oil, or electricity, and of transporting and distributing it. While the many provisions needed for the undisturbed flow of energy are often taken for granted, they cannot fail to exert permanent pressures on those who benefit from it as customers. The bills have to be paid, financially and otherwise.

The permanent availability of electricity, at every hour, in many parts of the world, has led to a diminution of the contrast between day and night. By the middle of the 19th century, the large investments made in their factories impelled many owners, under the pressures of competition, to let the engines run day and night. Gaslight illuminated the workplace. In the 20th century, night life has steadily extended,

especially in the cities. Water mains, sewage, gas, electricity, telephone, fax, internet, radio, police, fire brigade, hospitals are all generally expected to operate day and night. International interdependencies never come to a halt. This is one of the reasons why many people turn on the news as soon as they wake up in the morning; before resuming their daily activities they wish to learn what has happened while they were asleep—in their own country, where it was night, and elsewhere, where it was day time.

Once in a while there is a hitch. Sometimes the local supply of electricity breaks down, as happened for a number of hours in the blackout in New York on July 13, 1977, and for longer periods at a regional scale in California in 2001 and in the northeastern United States and southern Canada in 2003. In New York the failure was due to technical causes; in California its primary causes lay in disturbances in the financial sector. A combination of economic and political complications brought about the international oil crisis of 1973, when the majority of oil-producing countries jointly managed to enforce a drastic increase in the world price of crude oil.

Disturbances are remarkably rare, in view of the prodigious growth of energy consumption since 1950. The industrial economy is a fuel economy, revolving around the regular supply of fuel that can be easily converted into energy. Light, warmth, motion, and even coolness are produced with fuel. The rising supply of all these fuel-intensive amenities in turn constantly stimulates demand from customers, who are eager to enhance both their physical comfort and their social status.

In a large part of the world, access to the benefits of modern technology is still restricted to a small upper crust of society. Moreover, while in the western world and parts of Asia effective measures have been taken to reduce the polluting effects of industrial and domestic fuel use in densely populated areas, the quality of the air has only deteriorated in the rapidly growing megacities of Africa, Asia, and Latin America. As a consequence, according to an estimate of the World Health Organization in 1997, air pollution killed about 400,000 people worldwide annually.

Still, economic growth has continued in the poor parts of the world, and this is bound to affect the rich countries as well. Wealth attracts poverty; history abounds with examples of this general rule. When the opportunity offers itself, many people from poorer regions will try to migrate to regions with a higher standard of material comfort.

Meanwhile, global fuel consumption continues to rise. In rich countries, advanced technologies permit

customers to enjoy the use of their appliances and vehicles without any physical inconvenience caused by the combustion processes. Poorer people generally are prepared to put up with noise and smell to profit from fossil fuel energy.

3.6 Fire and War

With growing interdependence among people, dependence on natural forces has become more indirect: longer and more ramified social chains are formed between the production of things and their use. Even the threat of violent destruction of lives and property comes far less often from natural forces than from forces unleashed by one human group against another. The most powerful groups are those that command the organizational and technical means to mobilize huge amounts of energy and matter against their enemies.

In our day, the search for the most effective means of violence has led to the exploitation of a new source of energy: nuclear fission. For the first time in human history, a new source of energy was first applied at a large scale in war, with the sole intent of massive destruction. It was the result of an enormously concentrated accumulation of technical, scientific, and economic capital, invested in a single one-purpose enterprise: the production of an atomic bomb.

But no matter how single-focused the effort, the invention of the atomic bomb followed a general rule in the history of technology and produced unintended side effects. One of these was that the development of the bomb by the United States became the first step in a process of escalation and proliferation, which has become the greatly enlarged version of the paradigmatic case *par excellence* of an arms race. Another unintended consequence has been the rising level of risk of fatal consequences in case of a breakdown in any of the nuclear plants that were built in the second half of the 20th century.

War has always been an important factor in the relations between humans and their physical environment. In most cases, it was directed at destruction of a part of the anthroposphere: especially that part in which the organizational basis of the enemy was supposed to be most vulnerable. The more investments a group made in controlling its environment, the more susceptible it became to losses through violence. Advanced agrarian communities could suffer a severe setback in case their rice fields, vineyards, or terraces were destroyed. Many cities in agrarian societies underwent drastic reduction or were even totally annihilated after military surren-

der: the greater the concentration of physical wealth, the more irreparable the destruction.

In industrial society, enormous means of destruction were developed even before the invention of the atomic bomb. During World War II, air raids brought devastation to a great many cities. Global industrial society proved to be sufficiently resilient and affluent, however, so that after the war every bombed city was rebuilt at its original site. Such were the powers of recuperation of the industrialized parts of the anthroposphere.

4. SYMBOLIC AND RELIGIOUS ASPECTS

4.1 The Blessings of Fire

We may safely assume that fire, having played such a crucial role in the development of human group life, has always kindled the imagination and given rise to rites and beliefs. Unfortunately, for the earliest stages we have only material remains to go by, and it is very difficult to reconstruct the symbolic and religious meanings attached to fire solely on that basis. Since the invention of writing, however, many facts relating to these intangible aspects have been recorded. More recently, anthropological fieldwork has yielded a profusion of data and insights.

Thus, from all over the world myths have been collected relating how man first came to master fire. Typically these myths represent the initial conquest of fire not as a gradual long-term process but as a single adventurous event, often marked by cunning and deceit. According to most myths, fire was at first possessed by the gods until, at a given moment, they gave it to humans or, alternatively, it was stolen from them, either by some animal who then passed it on to humans or by a semidivine Promethean culture hero.

All these myths convey a sense of the preciousness of fire. Fire is regarded as a unique possession, which separates humans from the animal world. As a corollary of this conception, many peoples used to tell that other peoples did not have fire and, consequently, could not cook their food so that they had to live like animals. Needless to say, fire has been a universal element of culture ever since the emergence of *Homo sapiens*, and all these stories are spurious and slanderous.

Directly related to such beliefs is the idea that fire steered by human hands has purifying effects. The domain of human culture is seen as coinciding with the range of domesticated fire. Thus, cooked meat is

regarded as edible for humans, raw meat as only fit for beasts. Similarly, land cleared by burning is trusted as belonging to the human domain, whereas uncleared bush and forest are feared as wild and dangerous, a hideout for predators, serpents, insects, and evil spirits.

The sense of preciousness of fire also found expression in rites. The communal fire would be the center of group life, and people would treat it as sacred, surrounding it with ceremony and taking great care that it would never cease burning. The Roman cults of the hearth of Vesta, the fire goddess, show how such ancient rites could persist in an urban setting. In many societies the perpetual burning of the communal fire would be interrupted once a year; all fires were then ritually extinguished and replaced by new, pure fire that had to be solemnly ignited by the chief priest.

4.2 The Curses of Fire

When, after the rise of agriculture, people increasingly came to live in villages and cities, the number and variety of fires also increased, and so did the risk that any one of these fires would cause a conflagration. Fire continued to be a cherished necessity, but since it was so abundantly available, the sense of its preciousness was often superseded by anxiety and fear of the dangers it posed.

The transition was reflected in some of the great religions that emerged in Eurasia during the past 3000 years. The veneration of fire was most strongly preserved in Hinduism, with its cult of the fire god Agni, and in Zoroasterism, with its cult of Atar. Both religions kept alive the age-old worship of fire as a primeval force of vital importance to humans; but this was combined with awe and anxiety for its destructive powers.

Christianity and Islam stand much further away from the ancient worship of fire. Some old traditions persist, as in the burning of incense and candles, including the renewal at Easter in Roman Catholic churches. Fire no longer plays any part in beliefs about the origins of human society, however. Instead, it is strongly associated with evil. Sinners and unbelievers are threatened with a life after death in which they will be tormented by eternal burning. It seems as if the agonies of city conflagrations as described by several pagan authors from ancient Rome have found a religious symbolization in the vision of hell.

Fire festivals, such as those held over Europe at Saint John's day when things considered impure

would be committed to the flames, were occasions of combined joy and horror. Modern literature and art abound with evocations of the terrors of fire. When novelists describe the nightmare of a conflagration, they attribute the disaster, quite realistically, not to natural forces such as lightning but to human action—either to individual negligence, to madness or malice, or to collective hostilities in war.

4.3 Fire as an Element

A more neutral way of regarding fire, not in terms of its good or bad effects on human beings but as a natural phenomenon in its own right, was first proposed by philosophers of ancient Greece. They developed, out of various antecedents, a generally accepted doctrine according to which the universe consisted of four elements: earth, water, air, and fire. In China and India, similar systems of thought were worked out. With only minor modifications the worldview in which fire constituted one of the major elements remained the dominant cosmology for natural scientists and physicians well into the modern era.

Alchemists and chemists, as the direct successors of the metallurgists and fire masters of a previous age, made the study of fire into their central concern. Just as modern mechanical industry in its initial stages owed a great deal to the steam engine, the rise of modern physical science would have been inconceivable without the numerous efforts of investigators experimenting with and theorizing about fire.

5. CONCLUSION: THE ECLIPSE OF FIRE?

The controlled use of fire has been a human species monopoly since time immemorial. It constitutes an integral part of the apparatus with which humans have established and continue to maintain their dominance over other species.

Ever since the earliest stages of its domestication, the same phenomenon has recurred again and again in the human relations with fire: deliberately sought advances in control lead to unintended increases in dependency, and as the dependency sinks deeper into the social and economic infrastructure, it tends to become increasingly less clearly perceptible. Like agrarianization, industrialization began with conspicuously huge applications of fire. As industrial production became more specialized and highly organized, so did the use of fire. In everyday life in highly industrialized contemporary societies, flames

are visibly present only in such highly domesticated guises as cigarette lighters, candles for ceremonial use, or wood fires intended to create an atmosphere of comfort and relaxation. When, on the other hand, a fire is shown on television or in the press, it almost always spells war and disaster. The regularly controlled combustion processes on which industrial production largely rests are now mostly relegated away from public view.

In accordance with this apparent eclipse, natural scientists and philosophers no longer regard fire as one of the four elements out of which the world is composed. The very concept of fire has disappeared from scientific literature, to be replaced by the more abstract concept of energy—something that cannot be directly seen, heard, smelled, or felt. Similarly, many a day may pass in the life of the average citizen of a modern industrial state during which he or she does not perceive any physically observable fire. Yet it continues to be an integral part of the anthroposphere.

SEE ALSO THE FOLLOWING ARTICLES

Cultural Evolution and Energy • Early Industrial World, Energy Flow in • Forest Products and Energy • Hunting and Gathering Societies, Energy Flows in • Women and Energy: Issues in Developing Nations • Wood Energy, History of • Wood in Household Energy Use • World History and Energy

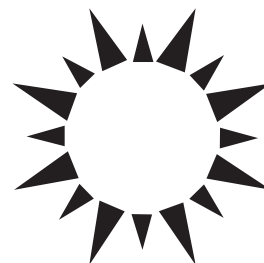
Further Reading

- De Vries, B., and Goudsblom, J. (eds.). (2002). "Mappae Mundi. Humans and Their Habitats in a Long-Term Socio-Ecological Perspective." Amsterdam University Press, Amsterdam.
- Elias, N. (2000). "The Civilizing Process. Sociogenetic and Psychogenetic Investigations." Rev ed. Blackwell, Oxford.
- Goudsblom, J. (1992). "Fire and Civilization." Allen Lane, London.
- Goudsblom, J., Jones, E., and Mennell, S. (1996). "The Course of Human History. Economic Growth, Social Process, and Civilization." M. E. Sharpe, Armonk, NY.
- Grübler, A. (1998). "Technology and Global Change." Cambridge University Press, Cambridge, UK.
- McClellan, J. E. III, and Dorn, H. (1999). "Science and Technology in World History." Johns Hopkins University Press, Baltimore.
- Pyne, S. J. (1991). "Burning Bush. A Fire History of Australia." Henry Holt and Co., New York.
- Pyne, S. J. (2001). "Fire. A Brief History." University of Washington Press, Seattle, WA.
- Rossotti, H. (1993). "Fire." Oxford University Press, Oxford, UK.
- Sieferle, R. P. (2001). "The Subterranean Forest. Energy Systems and the Industrial Revolution." The White Horse Press, Cambridge, UK.
- Simmons, I. G. (1996). "Changing the Face of the Earth. Culture, Environment, History," 2nd ed. Blackwell, Oxford.
- Turner, B. L., II, Clark, W. C., Kates, R. W., Richards, J. F., Marhews, J. T., and Meyer, W. B. (eds.). (1990). "The Earth as Transformed by Human Action Global and Regional Changes in the Biosphere over the Past 300 Years." Cambridge University Press, Cambridge, UK.
- Wright, R. (2000). "Nonzero. The Logic of Human Destiny." Pantheon, New York.
- Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money and Power." Simon and Schuster, New York.

Fisheries and Energy Use

PETER TYEDMERS

Dalhousie University
Halifax, Nova Scotia, Canada



1. Introduction
2. Overview of Global Fisheries
3. Energy Inputs to Fisheries
4. Energy Performance of Fisheries for Human Consumption
5. Energy Performance of Fisheries for Reduction
6. Conclusions

Glossary

active fishing gear Equipment used for fishing in which the aquatic organisms being targeted encounter the gear primarily through movement of the gear itself. Examples include cast nets, dredges, harpoons, jigs, and all forms of seine and trawl fishing.

animate energy Energy dissipated through the application of human or animal muscles to do work.

artisanal fisheries Small-scale fisheries that are typically executed by members of fishing households (as opposed to commercial companies) in which relatively small amounts of capital and energy and relatively small, if any, fishing vessels are employed.

culturally mediated energy All forms of energy dissipated through the application of technologies in the process of human activities.

energy intensity The amount of culturally mediated energy required to provide a given quantity of a product or service of interest. In the current context, energy intensity is expressed either as the total joules of energy required to land a live weight or “round” tonne of fish or shellfish harvested, or in terms of fuel use intensity, the total liters of fuel burned directly on fishing vessels per tonne of fish or shellfish landed.

energy return on investment (EROI) ratio A dimensionless ratio calculated by dividing the amount of useful energy provided by a given activity by the culturally mediated energy dissipated in providing it. In the case of food production systems, a common energy output used to calculate the EROI is the edible protein energy yield from the system being evaluated.

fishing Any activity that results in the catching, collecting, or harvesting of fish, and or aquatic invertebrates for

any purpose other than scientific research. In the current context, the harvesting of aquatic plants and mammals is not considered as fishing, nor is aquaculture.

industrial fisheries Those fisheries typically undertaken by commercial companies in which relatively large quantities of capital and energy are deployed.

net primary productivity (NPP) The difference between the total amount of carbon taken up by plants via photosynthesis and the amount of carbon lost by living plants via respiration.

passive fishing gear Equipment used for fishing in which the aquatic organisms being targeted encounter the gear primarily through their own movement or as a result of the movement of the surrounding waters. Examples include all forms of hook and line gears, drift nets, gillnets, traps, and weirs.

Fishing is one of the most ancient and widespread of human endeavors. Contemporary fisheries harvest an enormous variety of aquatic organisms from virtually every aquatic environment on the planet using a diverse array of technologies. Reflecting the diversity of global fisheries, culturally mediated energy inputs also vary widely in both form and magnitude. At one extreme, traditional artisanal fisheries typically rely on relatively small inputs of exogenous energy, mostly in the form of wind, and animate energy to propel vessels and haul nets. Indeed, among subsistence fisheries in which animate energy inputs predominate, the nutritional value of the catch must routinely exceed the energy expended by the human muscles engaged in the fishery if it is to remain viable. Although such traditional, low-input fisheries persist in many parts of the world, high-input, industrialized fisheries now account for the majority of global landings. Among these fisheries, particularly those targeting high-value species, it is now common for direct fossil fuel energy inputs alone to exceed the nutritional energy embodied in the catch by at least an order of magnitude.

1. INTRODUCTION

As with all human activities, fishing entails the dissipation of energy in support of its primary activity, the harvesting of aquatic organisms. While the energetic cost of fishing is less obvious, and consequently receives much less attention than the direct impact that fishing has on targeted stocks and associated marine ecosystems, it is precisely the availability of abundant energy, in particular fossil energy, that enables many contemporary fisheries to continue even when stocks are in decline. Consequently, analyses of the forms and quantities of energy dissipated in fisheries, and in particular changes in energy use over time, can provide a powerful measure of the biophysical scarcity of exploited populations.

2. OVERVIEW OF GLOBAL FISHERIES

2.1 Scope and Extent

Data compiled by the Food and Agriculture Organization (FAO) of the United Nations indicate that, globally, approximately 90 million tonnes of fish and shellfish, representing more than 130 genera and 750 species, are harvested from the wild each year (Fig. 1). Although in excess of 200 countries report landings of fish and shellfish, the top 10 fishing nations account for over half of the total (Fig. 2). Similarly, while fishing is conducted in virtually every aquatic environment on the planet, over 90% of global landings are derived from marine waters, with most of this coming from highly productive coastal shelf and upwelling ecosystems.

2.2 Importance of Fisheries

One of the oldest forms of food production, fishing currently provides two-thirds of the fish and shellfish consumed by people worldwide, with aquaculture accounting for the balance. As a result, fishing directly contributes approximately 10% of the total animal protein intake by humans. In addition, although wild-caught food fish tend to contain less fat than terrestrial animals, the chemical composition of marine-sourced lipids is of significant nutritional importance. In particular, many species of wild-caught fish are relatively rich sources of Omega-3, or n-3, fatty acids, compounds that are believed to

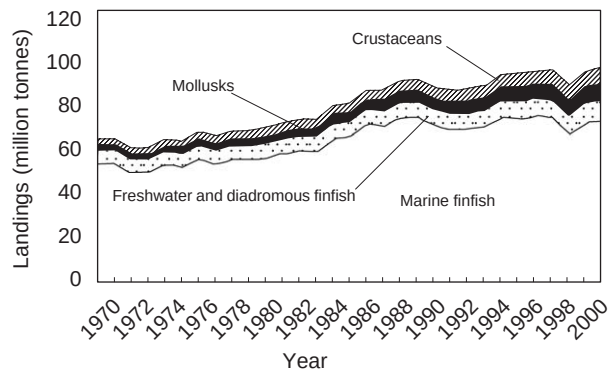


FIGURE 1 Global capture fisheries landings: 1970–2000. Data from FAO statistical database Fishstat+.

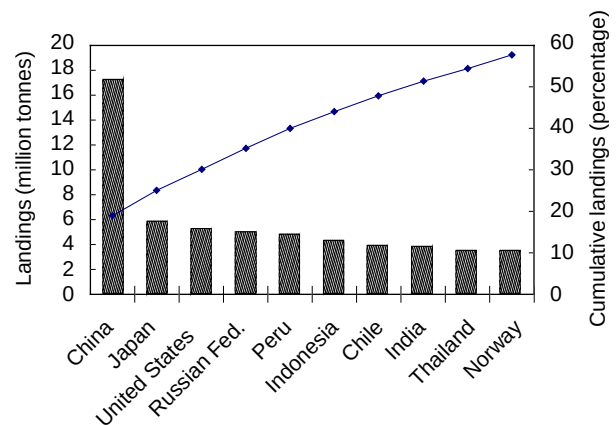


FIGURE 2 Absolute and cumulative fisheries landings by the top 10 fishing nations in 1998. Data from FAO (2000).

substantially decrease the risk of heart disease, inflammatory processes, and certain cancers.

It is important to note, however, the role of fish in human diets varies widely among countries and regions (Fig. 3). This variation not only reflects differences in resource availability and cultural preferences with respect to the seafood consumption but also, to a certain extent, development status. Of the 30 countries whose citizens are most dependent on fish as a source of protein, over 80% are less developed countries.

Of the approximately one-third of global landings not consumed directly as human food, the vast majority is destined for reduction to fish meal and oil. Currently, both of these reduction coproducts are directed overwhelmingly to livestock and aquaculture feeds, with smaller volumes of fish oil being consumed directly by people, either as edible oil products or nutritional supplements. Traditionally, however, fish oils have also been used in a wide

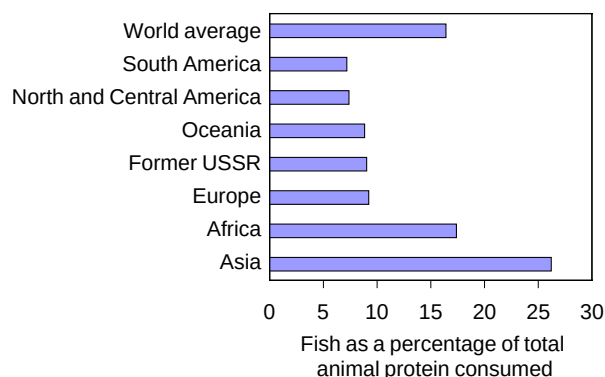


FIGURE 3 Importance of fish and shellfish from all sources to human diets.

range of industrial applications or burned as a source of fuel and illumination. While representing relatively small total landings, other nonfood fisheries are conducted to provide an array of high-value medicinal, decorative, and luxury products including shark cartilage, pearls, and live fish for the aquarium trade.

Reflecting their scale and diversity, it is not surprising that fisheries can play an important role in national and regional economies. Globally, the FAO estimates that as of the late 1990s, fisheries directly employed approximately 25 million people and yielded products whose first sale value exceeded \$80US billion annually. The socioeconomic impact of fisheries is most pronounced, however, in many less industrialized countries and rural coastal communities throughout the world, where alternative employment opportunities are often limited.

2.3 Fishing Technologies

Based on archaeological evidence from central Africa, humans have been using dedicated fishing technologies, in this early instance a fishing spear, for at least 90,000 years. As the scope and range of fisheries has expanded over the intervening millennia, so too has the diversity of methods used to harvest fish and shellfish. While it is impractical here to explore the immense range of fishing techniques used around the world, a few broad generalizations are useful: fisheries in which some form of fishing vessel is used account for the vast majority of global landings, and most of the world's catch is taken using active fishing gears, particularly trawling and purse seining techniques in which large nets are respectively either towed behind fishing vessels or used to encircle schools of fish.

3. ENERGY INPUTS TO FISHERIES

Energy flows have been used to evaluate the performance of food production systems for over a hundred years. However, it was not until the first oil price shock of 1973 and the resulting concern regarding the dependency of industrialized food production systems on fossil fuels that systematic analyses were undertaken. One of the first major products of this period of research was Gerald Leach's handbook "Energy and Food Production" published in 1976. While it dealt primarily with agricultural systems, it included, for the first time, data on the major culturally mediated energy inputs to six fisheries from four continents. Since then, a small number of researchers in various parts of the world have continued to evaluate the energetics of fisheries from a variety of perspectives.

From an energetic perspective, fishing is a process in which a variety of culturally mediated forms of energy are dissipated in order to capture edible chemical energy embodied in the carcasses of fish and shellfish. The ultimate source of both the chemical energy embodied in the harvested organisms and most forms of culturally mediated energy is the sun. Consequently, the amount of solar energy required to sustain global landings and fishing activities is enormous. However, to date very few fishery-specific energy analyses have systematically attempted to account for solar energy inputs. As a result, other than a brief discussion of the ecosystem support associated with growing the biomass of animals harvested by global fisheries, solar energy inputs are not a focus of this review. Instead, this work deals primarily with the major forms and quantities of culturally mediated energy inputs to fishing.

3.1 Ecosystem Support of Fisheries

While not a direct measure of the solar energy flux underpinning fish and shellfish harvests, estimates have been made of the scale of ecosystem support appropriated by global fisheries. Typically, this is done by estimating the fraction of global aquatic net primary productivity (NPP)—or the amount of carbon fixed by plants via photosynthesis minus that lost through respiration—required to sustain the mass of fish and shellfish harvested. In 1995, Daniel Pauly and Villy Christensen, fisheries scientists currently based at the University of British Columbia, provided the most robust analysis of this measure to date. Their work indicates that as of the

mid-1990s, global fisheries landings and discards appropriated approximately 8% of the total NPP available in the world's aquatic ecosystems. More striking, however, were their findings regarding those highly productive freshwater and near-shore marine ecosystems that sustain nearly 85% of total fisheries landings yet spatially represent less than 10% of the planet's total aquatic environment. Within the context of these most productive aquatic ecosystems, fisheries alone appropriate between 24 and 35% of the NPP available.

3.2 Forms of Direct Energy Inputs

Culturally mediated energy inputs to fishing activities can be categorized into direct and indirect types. Common indirect inputs, often referred to as embodied energy inputs, are those associated with building and maintaining fishing vessels and providing fishing gear, bait, and ice. In contrast, in most fisheries, direct energy inputs are typically those required to propel fishing vessels and deploy fishing gears. The three dominant forms of energy dissipated to these ends include animate, wind, and fossil fuel energy, each of which are briefly described next.

3.2.1 *Animate Energy*

Animate energy is the one form of culturally mediated energy common to all of the world's fisheries regardless of their technological sophistication. In many traditional artisanal fisheries, human muscles still provide the bulk of the energy used to deploy gears, handle the catch, and, where they are employed, propel vessels. However, while animate energy inputs may represent the dominant form of energy applied in a fishery, in absolute terms, the quantities involved are typically small. For example, among subsistence fisheries in which animate energy inputs predominate, the nutritional value of the catch must routinely exceed the energy expended by the human muscles engaged in the fishery if it is to remain viable. In the case of most contemporary fisheries, however, while human animate energy inputs remain part of the production equation, they are generally dwarfed by the inputs of wind or fossil fuel energy.

Unlike preindustrial agriculture wherein a wide variety of animals were domesticated to provide important secondary sources of animate energy, relatively few fisheries have systematically employed animals. Examples, however, include the traditional use of trained fishing cormorants in parts of China

and Japan and the now largely outmoded European practice of using otters and diving ducks to drive fish into traps.

3.2.2 *Wind*

For as long as people have sailed, it is likely that wind energy has been used to support fishing activities. The integration of wind energy into fisheries not only allowed fishing vessels to be propelled farther and faster than would otherwise be possible, it facilitated the development of fishing techniques that would otherwise have been impractical, if not impossible, to conduct from a rowed boat. Specifically, various trawl or dragger fisheries in which nets or dredges are towed either through the water or along the seafloor were almost all first developed within the context of sail fisheries.

Although little if any research has been conducted to quantify the wind energy inputs to sail-assisted fisheries, prior to the introduction of fossil fuels, it likely accounted for a large proportion of the culturally mediated energy inputs to global fisheries.

3.2.3 *Fossil Fuels*

As with most other food production sectors, fossil fuels have become the dominant form of energy used in fishing in a relatively short period of time. The process began in England in the late 1800s, when coal-fired steam engines were first installed on trawl fishing vessels to provide power for both propulsion and net hauling. With the advantages of increased speed and power, together with the ability to operate regardless of the wind, steam trawling expanded rapidly. Although coal- and oil-fired steam engines remained in use on fishing vessels throughout much of the 20th century, their significance was ultimately eclipsed by internal combustion engines.

Gasoline- and diesel-fueled internal combustion engines were first adapted for use on fishing boats in the early 1900s. Technological advances made during and after the Second World War, however, greatly accelerated their integration into fisheries. Consequently, over the past 50 years not only has the size of the global fishing fleet increased but so has its power. The trend to larger, more powerful fishing vessels is exemplified by the emergence of so-called supertrawlers, vessels that can exceed 100 m in length with propulsive engines well in excess of 10,000 horsepower. At the same time, more and more relatively small engines are introduced annually into small-scale fisheries around the world. Given these twin trends of increasing prevalence and size of engines, it is not surprising that the bulk of the

world's catch is now harvested using vessels propelled by fossil fuels.

3.3 The Relative Importance of Direct and Indirect Energy Inputs

Among modern industrial fisheries in which the major direct and indirect culturally mediated energy inputs have been systematically analyzed using either process analysis or input-output techniques, a consistent pattern has emerged. Direct fuel energy inputs typically account for between 75 and 90% of the total culturally mediated energy inputs, regardless of the fishing gear used or the species targeted (Fig. 4). Depending on the character of the fishery and the scope of the analysis conducted, the remaining 10 to 25% is generally composed of energy inputs associated with vessel construction and maintenance, and the provision of labor, fishing gear, bait, and ice if used.

On most fishing vessels, direct fuel inputs are used primarily for vessel propulsion. In some fisheries, however, secondary energy-consuming activities, including onboard processing, refrigeration, and freezing, can account for a nontrivial portion of the fuel burned. Squid jig fisheries are interesting and fairly extreme examples of fisheries in which a relatively large proportion of fuel inputs are used for activities other than vessel propulsion. In these fisheries, vessels typically employ batteries of high-

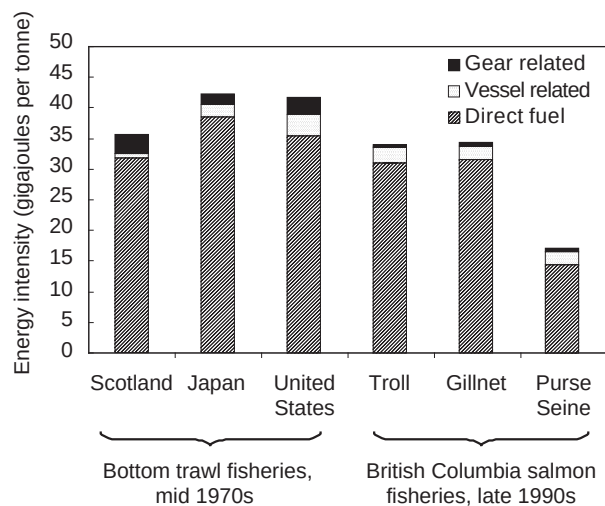


FIGURE 4 Relative contribution of direct and indirect energy to various fisheries. Data sources: Scotland: Edwardson, W. (1976). The energy cost of fishing. *Fishing News Intl.* 15(2), p. 36–39. Japan: Watanabe and Uchida (1984); United States: Rawitscher (1978); British Columbia salmon fisheries: Tyedmers (2000).

intensity lamps, automated jigging machines, and freezers, all powered by diesel-fuelled generators to attract, hook, and preserve the catch while at sea. As a result, these nonpropulsion energy demands can, in combination, account for as much as 40% of the total fuel burned (Fig. 5).

Among indirect energy inputs, those associated with building and maintaining the fishing vessels themselves regularly account for the largest fraction (Fig. 4). This is particularly the case when a vessel's major components (i.e., its hull, superstructure, decks, and holds) are fabricated primarily from relatively energy-intensive materials such as aluminum and steel as opposed to wood or fiberglass. For example, an analysis of the direct fabrication and major embodied energy inputs to build and equip two equivalent 10 m long salmon gillnet fishing vessels, one constructed largely of fiberglass and wood and the other of aluminum, found that the embodied energy inputs were over 50% higher in the case of the aluminum-hulled vessel (Table I). Moreover, when the electricity required to fabricate the aluminum-hulled vessel is included in the analysis, the total embodied energy costs of this vessel rise to 2.5 times that of its fiberglass counterpart.

3.4 Comparing the Energy Performance of Fisheries

Most energy analyses of fisheries have focused largely, if not exclusively, on evaluating the direct fuel inputs to fishing. This relatively narrow focus

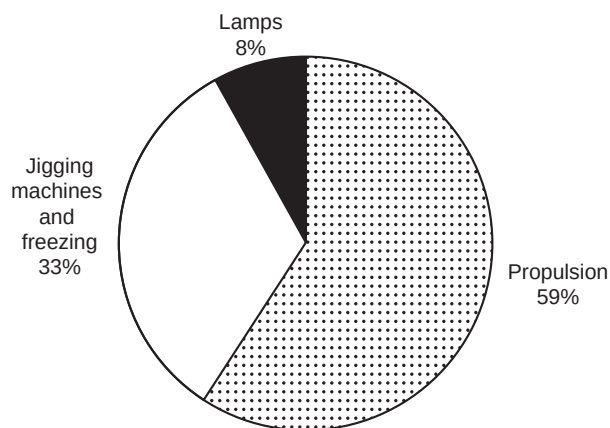


FIGURE 5 Functional fuel utilization on a Japanese long-distance squid fishing vessel. Data from Ishikawa, *et al.* (1987). A case study of energy analysis of long distance squid angling. *Bull. Jap. Soc. Sci. Fish.* 53(9), 1525–1531.

TABLE 1

Direct and Indirect Energy Inputs to Build a 10-Meter-Long Salmon Gillnet Fishing Vessel Fabricated from Either Fiberglass or Aluminum

Inputs	Material and embodied energy inputs			
	Fiberglass hull		Aluminum hull	
	kg	GJ	kg	GJ
Glass				
Matting	1500	15		
Sheet	100	1	100	1
Wood (m ³)	5	21		
Fiberglass resin	1200	90		
Steel				
Engine and transmission	1110	28	1110	28
Fishing equipment	400	10	400	10
Anchor, winch and chain	200	5	200	5
Hardware and other	1050	26	550	14
Lead (batteries)	100	2	100	2
Mixed metals (controls, etc.)	390	10	390	10
Aluminum				
Fuel and water tanks	540	75	540	75
Fishing equipment	330	46	330	46
Hull, decks, holds, and cabin			2250	315
Total embodied energy		329		506
Electricity to weld hull, etc.				324
Total		329		830

not only reflects the fact that these inputs dominate the energy profile of fishing, but detailed analyses of indirect inputs tend to be labor intensive and time consuming. Therefore, in order to facilitate the comparison of as wide a range of fisheries as possible, throughout the balance of this review only the direct fuel energy inputs to various fisheries is considered. Consequently, the energy intensities of various fisheries is expressed in terms of the liters of fuel burned per "round" or live weight tonne of fish or shellfish landed.

However, as food fisheries can also be thought of as food energy "producing" systems, it is also useful

to calculate their energy efficiency. Traditionally within analyses of agriculture, aquaculture, and fisheries systems, this has been done by calculating either their industrial energy input to edible food energy output ratio or its inverse, the edible energy return on investment (EROI) ratio. As the nutritional importance of fish and shellfish is primarily a function of their protein content, edible protein energy output is the most meaningful basis of comparison. Therefore, edible protein energy return on investment (protein EROI) ratios for various fisheries are also presented.

4. ENERGY PERFORMANCE OF FISHERIES FOR HUMAN CONSUMPTION

To date, most energy analyses of fisheries for human consumption have concentrated on relatively large-scale industrial fisheries that are conducted for commercial gain. Moreover, few analyses have been undertaken on fisheries based in the Southern Hemisphere. As a result, much of the discussion that follows reflects these biases.

Given the diversity of fisheries for human consumption, both in terms of species harvested and fishing gears used, it is helpful to focus on broad subsets when considering their energy performance. Consequently, the following discussion is organized into four major divisions: fisheries for demersal species, or those fish that spend much of their life on or near the ocean floor; fisheries for pelagic species, or those fish that spend most of their life swimming in the water column; fisheries for invertebrates; and subsistence fisheries.

4.1 Fisheries for Demersal Species

Fisheries for demersal finfish species are some of the largest, most economically important fisheries in the world. Each year, upwards of a hundred species of demersal finfish may be harvested in fisheries globally. Of these, the most important include the various gadoid, or codlike species, and flatfish species. Given that, by definition, demersal species live on or near the ocean floor, fisheries for these species are undertaken in waters ranging from under 10 m to many hundreds of meters deep. As a result, many different types of fishing gears are used in their capture. By far the largest tonnages, however, are

harvested each year using trawl and to a lesser extent, long-line fishing gear.

As a result of their importance, fisheries for demersal species have been the subject of numerous energy analyses over the last three decades (Table II). Some of the most recent work to systematically evaluate the fuel consumption of a wide variety of demersal finfish fisheries examined 29 fisheries involving more than 3000 vessels based in Atlantic

Canada, Iceland, Norway, and Germany. As of the late 1990s, these fisheries together harvested approximately 1.2 million tonnes of fish annually from ecosystems around the North Atlantic. While fuel use intensities varied by an order of magnitude among these 29 fisheries, from lows of just under 300 l/t to more than 2000 l/t, on average they consumed just over 500 liters of diesel per tonne of fish landed.

TABLE II
Energy Performance of Industrial Fisheries for Direct Human Consumption

Main fishery targets	Gear	Time frame	Location of fishery	Fuel use intensity (liters/tonne)	Edible protein EROI
Demersal fisheries					
Redfish spp.	Trawl	Late 1990s	North Atlantic	420 ^a	0.11
Cod/flatfish spp.	Danish seine	Late 1990s	North Atlantic	440 ^a	0.10
Cod/haddock	Longline	Late 1990s	North Atlantic	490 ^a	0.091
Cod/saithe	Trawl	Late 1990s	North Atlantic	530 ^a	0.084
Alaskan pollock	Trawl	Early 1980s	North Pacific	600 ^b	0.052
Flatfish spp.	Trawl	Early 1980s	NW Pacific	750 ^b	0.066
Croakers	Trawl	Early 1980s	NW Pacific	1500 ^b	0.029
Flatfish spp.	Trawl	Late 1990s	NE Atlantic	2300 ^a	0.019
Pelagic fisheries					
Herring/mackerel	Purse seine	Late 1990s	NE Atlantic	100 ^a	0.56
Herring	Purse seine	Early 1990s	NE Pacific	140 ^c	0.36
Herring/saithe	Danish Seine	Late 1990s	NE Atlantic	140 ^a	0.35
Salmon spp.	Purse seine	1990s	NE Pacific	360 ^c	0.15
Salmon spp.	Trap	Early 1980s	NW Pacific	780 ^b	0.072
Salmon spp.	Gillnet	1990s	NE Pacific	810 ^c	0.068
Salmon spp.	Troll	1990s	NE Pacific	830 ^c	0.067
Herring	Purse seine	Early 1980s	NW Pacific	1000 ^b	0.051
Skipjack/tuna	Pole and line	Early 1980s	Pacific	1400 ^b	0.053
Skipjack/tuna	Purse seine	Early 1980s	Pacific	1500 ^b	0.049
Swordfish/tuna	Longline	Late 1990s	NW Atlantic	1740 ^a	0.042
Salmon spp.	Gillnet	Early 1980s	NW Pacific	1800 ^b	0.031
Swordfish/tuna	Longline	Early 1990s	Central Pacific	2200 ^d	0.027
Tuna/billfish	Longline	Early 1980s	Pacific	3400 ^b	0.022
Shellfish fisheries					
Abalone/clams	Hand gathering	Early 1980s	NW Pacific	300 ^b	0.11
Crab	Trap	Late 1990s	NW Atlantic	330 ^a	0.057
Scallop	Dredge	Late 1990s	North Atlantic	350 ^a	0.027
Shrimp	Trawl	Late 1990s	North Atlantic	920 ^a	0.058
Shrimp	Trawl	Early 1980s	North Pacific	960 ^b	0.056
Norway lobster	Trawl	Late 1990s	NE Atlantic	1030 ^a	0.026
Crab	Trap	Early 1980s	NW Pacific	1300 ^b	0.014
Spiny lobster	Trawl	Early 1980s	NW Pacific	1600 ^b	0.017
Squid	Jig	Early 1980s	NW Pacific	1700 ^b	0.033
Shrimp	Trawl	Late 1990s	SW Pacific	3000 ^d	0.019

Notes to sources: ^aTyedmers (2001); ^bWatanabe and Okubo (1989); ^cTyedmers (2000); ^dunpublished data.

Much of the variability in fuel use intensity among the 29 North Atlantic fisheries appears to be a function of two factors. Not surprisingly, the relative abundance and catchability of various targeted species has a major influence on resulting energy performance. Evidence of this is can be seen by comparing average fuel use intensities of various trawl fisheries targeting very different species assemblages in different parts of the North Atlantic (Table II). Specifically, Canadian and Icelandic trawl fisheries targeting redfish species experienced average fuel use intensities of just over 400 l/t while trawl fisheries for cod and related species from across the North Atlantic averaged 530 l/t. At the other extreme, German trawl fisheries targeting high value flatfish species in the North Sea consumed approximately 2300 liters of diesel per tonne of fish landed.

A second factor that affects the energy performance of fisheries is the type of fishing gear employed. This can be illustrated by focussing on those North Atlantic fisheries that targeted Atlantic cod and related species (Table II). On average, cod fisheries in which Danish seine nets were used burned approximately 440 liters per tonne of fish landed. In contrast, long-line fisheries averaged 490 l/t, while trawl fisheries for cod averaged 530 l/t.

Looking beyond the North Atlantic, a comprehensive analysis of the fuel use intensity of all Japanese fisheries as of the early 1980s suggests that North Pacific fisheries for demersal species may be more energy intensive than many of their North Atlantic counterparts (Table II). Specifically, Japanese trawl fisheries targeting either Alaskan pollock, various flatfish species or yellow croaker burned, respectively 600, 750, and 1500 liters of diesel per tonne of fish landed. It is important to keep in mind, however, when comparing the performance of these Japanese fisheries with those described previously from the North Atlantic, that almost 20 years has elapsed between when the two data sets were acquired. Over that period of time much can have changed that might influence the energy performance of any specific fishery. As we shall see, however, evidence from a variety of fisheries suggests that there has been a general tendency for the energy performance of industrial fisheries to decline over time.

From an energy efficiency perspective, the performance of fisheries targeting demersal species also ranges widely. As measured in terms of their edible protein EROI, the 29 North Atlantic fisheries discussed previously together averaged 0.095. This means that the edible protein energy content of all the fish that they landed amounted to just under 10% of

the fuel energy that they burned. More generally, edible protein EROI values for individual fisheries targeting demersal species can range from approximately 0.02 to over 0.1 (Table II). While efficiencies in this range may seem low, they are broadly in keeping with the energy performance of other industrial fisheries and, in fact, compare favorably with other animal protein production systems (Table III).

4.2 Fisheries for Pelagic Species

Pelagic fish species harvested for human consumption encompass a diverse variety of animals. They range from the small, densely schooling, largely planktivorous herrings, to large, top-level carnivores including various tuna and billfish species. What they generally have in common, however, is that all are typically harvested using fishing gears deployed at or near the ocean's surface.

TABLE III

Edible Protein EROI values from Livestock and Aquaculture Production Systems

Production system (locale)	Edible protein EROI
Carp—unspecified culture system (Indonesia)	0.70
Tilapia—unspecified culture system (Africa)	0.11
Mussel—longline culture (Scandinavia)	0.10
Carp—unspecified culture system (Israel)	0.084
Turkey (United States)	0.077
Milk (United States)	0.071
Tilapia—unspecific culture (Israel)	0.066
Tilapia—pond culture (Zimbabwe)	0.060
Swine (United States)	0.056
Eggs (United States)	0.038
Catfish—intensive pond culture (United States)	0.030
Chicken (United States)	0.029
Tilapia—intensive cage culture (Zimbabwe)	0.025
Atlantic salmon—intensive cage culture (Canada)	0.025
Shrimp—semi-intensive culture (Colombia)	0.020
Chinook salmon—intensive cage culture (Canada)	0.020
Lamb (United States)	0.020
Atlantic salmon—intensive cage culture (Sweden)	0.020
Beef—feedlot (United States)	0.019
Seabass—intensive culture (Thailand)	0.015
Shrimp—intensive culture (Thailand)	0.014

For sources, see Tyedmers (2001).

Reflecting the diversity of their quarry, fisheries for pelagic species display the widest range of energy intensities of fisheries for human consumption (Table II). At one extreme are those that target herring and associated species using purse seine or Danish seine nets. For example, recent analyses of herring fisheries from both the Atlantic and Pacific oceans indicate that they typically consume between 100 and 150 liters of diesel per tonne of fish landed. It is interesting to note, however, that an analysis of a herring fishery based in Japan, conducted in the early 1980s, reported a fuel use intensity of 1000 l/t. It is unclear from the data available why such a large difference should exist between this fishery and comparable fisheries in other parts of the world.

In sharp contrast, analyses of fisheries for relatively high value, large pelagic species like tuna and billfish report fuel use intensities ranging from 1400 to 3400 l/t (Table II). Among these fisheries, those that use long-line fishing gear are often the most energy intensive.

Between these two energy performance extremes, lie fisheries for Pacific salmon. Reported fuel use intensities of salmon fisheries range from just under 400 l/t to 1800 l/t (Table II). In a comparative analysis of the three fishing gears used to harvest salmon in British Columbia, Canada, it was found that purse seiners burned, on average, just under half the amount of fuel, per tonne of salmon landed, than did either gillnet or troll fishing boats.

Largely as a result of the major differences in the amount of fuel that they burn, the energy efficiency of fisheries for pelagic species, measured in terms of their edible protein EROI ratios, range from as low as 0.02 (or 2%) to over 0.5 (or 50%).

4.3 Invertebrate Fisheries

While fisheries for invertebrates account for a relatively small proportion of global landings (Fig. 1), they are some of the most lucrative fisheries in the world. It is not surprising therefore that in many instances they can be highly energy intensive. For example, fisheries targeting various species of shrimp and lobster often burn upwards of 1000 liters of fuel per tonne of landings (Table II). Interestingly, however, not all relatively high value invertebrate fisheries are this energy intensive. Some, including the valuable North Atlantic scallop and crab fisheries, consume between 300 and 350 liters of fuel per tonne of shellfish landed.

As measured in terms of their edible protein EROI ratios, many invertebrate fisheries are relatively

energy inefficient. This not only reflects rather high fuel use intensities, but the fact that many species of invertebrates yield relatively small quantities of edible meat. For example, the adductor muscle in scallops generally only represents 10 to 12% of the live weight of the animal. Similarly, the edible portion of many crab and lobster species normally does not exceed 30% of their live weight. In contrast, edible muscle usually constitutes between 50 and 60% of most finfish species.

4.4 Subsistence Fisheries

To date, little attention has been paid to the energy performance of small-scale fisheries. However, an analysis of a northern Canadian aboriginal subsistence fishery targeting primarily pike and pickerel in rivers and lakes provides some intriguing and potentially worrying insights into their performance. From data collected through the late-1970s, human ecologist Fikret Berkes estimated that Cree fishermen, based in the community of Fort George, Quebec, burned, on average, approximately 1400 liters of gasoline per tonne of fish landed in the course of tending their fishing nets. Remarkably, this level of fuel use intensity is on the same scale as some of the most energy intensive industrial fisheries for human consumption in the world. It also highlights the extraordinary vulnerability of fossil fuel-dependent artisanal fisheries, and ultimately the communities that depend on them, to disruptions in the supply or increases in the price of fuel.

4.5 Temporal Trends in Energy Performance

Numerous analyses have been undertaken to evaluate how the energy performance of fisheries for human consumption has changed through time. Although the time periods and types of fisheries analyzed differ, when their results are expressed in a consistent format, say in terms of their edible protein EROI ratios, a regular and troubling pattern emerges (Fig. 6). Over time, the energy performance of virtually every fishery analyzed has declined, and in some cases quite dramatically. For example, the edible protein EROI ratio of the entire New Bedford, Massachusetts-based fishing fleet experienced a five-fold decline from just over 0.15 (or 15%) to under 0.03 (or 3%) over the period from 1968 to 1988 (Fig. 6A). Similarly, over the course of a decade, beginning in the late 1960s, the edible protein EROI

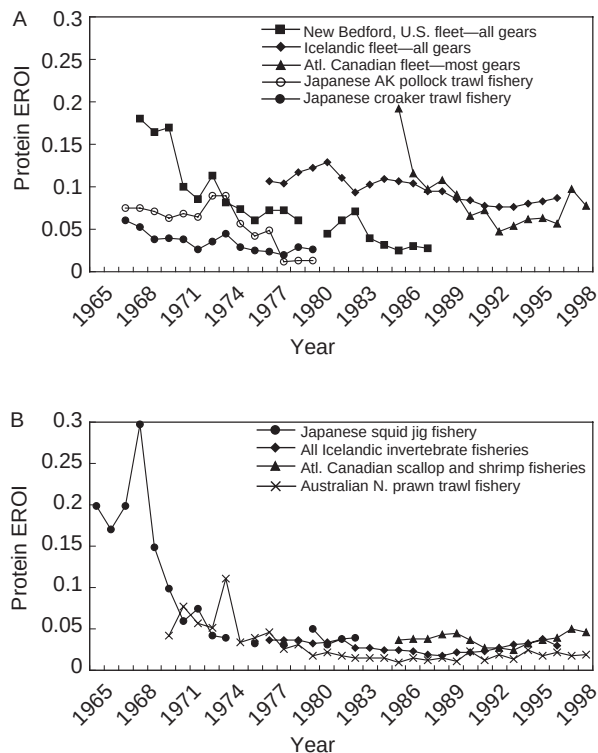


FIGURE 6 Changes in edible protein EROI ratios of fisheries for human consumption: 1965–1999. (A) Demersal finfish fisheries. (B) Shellfish fisheries. Data compiled from Watanabe and Uchida. (1984); Sato, K. *et al.* (1989). Chronological change of energy input for squid angling in Japan 1956–1983. *Bull. Jap. Soc. Sci. Fish* 55(11), 1941–1945; Mitchell and Cleveland (1993). Resource scarcity, energy use and environmental impact: A case study of the New Bedford, Massachusetts, USA Fisheries. *Environ. Management* 17(3), 305–317; Tyedmers (2001); and unpublished data.

ratio of the entire Japanese-based squid fishing fleet declined from over 0.20 (or 20%) to consistently below 0.04 (or 4%) (Fig. 6B).

While their relative importance may vary between fisheries, the two factors that likely contribute most to the general trend to poorer energy performance over time include decreases in relative abundance or proximity of targeted fisheries resources, and increases in the size, power, and technical sophistication of both fishing vessels and fleets.

5. ENERGY PERFORMANCE OF FISHERIES FOR REDUCTION

As many as 50 countries routinely report some production of fish meal and oil. However, less than a dozen account for the lion's share of total annual

TABLE IV

Fuel Energy Intensities of Reduction Fisheries as of the Late 1990s

Main fishery targets	Gear	Location of fishery	Fuel use intensity (liters/tonne)
Capelin/herring	Purse seine	NE Atlantic	20
Menhaden	Purse seine	NW Atlantic	32
Capelin/herring	Trawl	NE Atlantic	80
Blue Whiting	Purse seine	NE Atlantic	85
Sand Eels/herring	Trawl	NE Atlantic	95
Herring/sand Eels	Purse seine	NE Atlantic	100
Herring/mackerel	Trawl	NE Atlantic	110

Data source: Tyedmers (2001).

global output. The most significant of these include Peru, Chile, Iceland, Norway, Denmark, Japan, the United Kingdom, the United States, and the countries of the former Soviet Union.

Although they account for 30% of global landings, most of the world's fish meal and oil production is derived from large tonnage landings of a relatively small number of species. In the waters off Peru and Chile the major species harvested for reduction include the Peruvian anchoveta, Inca scad, and the South American pilchard. In the northeast Atlantic, the main species destined for reduction include capelin, blue whiting, Norway pout, European sprat, various species of sand eel together with smaller quantities of Atlantic herring and Atlantic mackerel. In U.S. waters, Atlantic and Gulf menhaden account for virtually 100% of landings destined entirely to reduction.

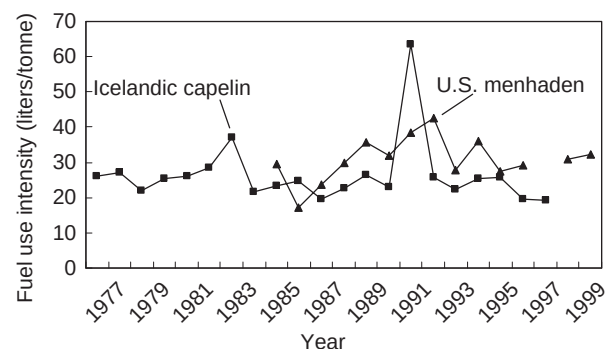


FIGURE 7 Changes in fuel use intensity of fisheries for reduction, 1977 to 1999. Data from Tyedmers, 2001, and unpublished data.

Recent analyses of various North Atlantic fisheries for reduction indicate that in most instances, they consume less than 100 liters of diesel per tonne of fish landed (Table IV). Remarkably, it is not unusual for some of the largest of these fisheries to routinely burn well below this mark. For example, time series data from both the Icelandic and American purse seine fisheries, targeting capelin and menhaden respectively, indicate that in recent decades, they have seldom consumed more than 40 liters of fuel per tonne of fish landed (Fig. 7). Such low and seemingly stable fuel use intensities would seem to suggest that in both instances, fish stocks and fleet effort are being carefully managed.

While the data available do not support direct head-to-head comparisons, from the spectrum of fuel use intensities associated with various North Atlantic fisheries for reduction, it would appear as if those employing purse seine nets are routinely less energy intense than those using trawl fishing gear (Table IV).

6. CONCLUSIONS

Global fisheries are highly varied. They harvest an enormous diversity of aquatic animals using a variety of mostly vessel-based fishing gears. Not surprisingly, therefore, the major forms and absolute quantities of culturally mediated energy inputs dissipated in their execution also vary widely. However, in the case of most fisheries, particularly large-scale industrial fisheries over the past half-century, fossil fuel inputs have come to dominate their energy profile.

Of concern, though, is evidence from a number of fisheries from around the world of a general tendency to poorer energetic performance over time. This pattern of ever-increasing apparent biophysical scarcity reinforces the oft expressed concerns regarding both the declining status of many of the world's fish stocks and the growing size and power of the global fishing fleet. In spite of the ongoing declines in their energy performance, however, many industrial fisheries for human consumption remain energetically competitive with other animal protein-producing systems.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Diet, Energy, and Greenhouse Gas Emissions • Ecosystem Health: Energy Indicators • Entrainment and Impingement of Organisms in Power Plant Cooling • Food System, Energy Use in • Industrial Agriculture, Energy Flows in • Livestock Production and Energy Use • Marine Transportation and Energy Use • Ocean, Energy Flows in • Wetlands: Impacts of Energy Development in the Mississippi Delta

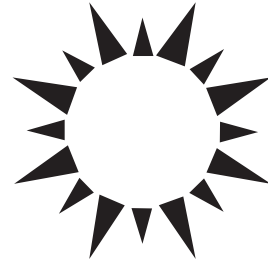
Further Reading

- Food and Agriculture Organization of the United Nations. (2000). The state of the world fisheries and aquaculture 2000. FAO, Rome, Italy.
- Leach, G. (1976). "Energy and Food Production." IPC Science and Technology Press, Surrey.
- Mitchell, C., and Cleveland, C. J. (1993). Resource scarcity, energy use and environmental impact: A case study of the New Bedford, Massachusetts, USA Fisheries. *Environ. Management* 17(3), 305–317.
- Pauly, D., and Christensen, V. (1995). Primary production required to sustain global fisheries. *Nature* 374, 255–257.
- Pimentel, D., Shanks, R. E., and Rylander, J. C. (1996). Bioethics of fish production: Energy and the Environment. *J. Agricult. Environ. Ethics* 9(2), 144–164.
- Rawitscher, M. A. (1978). "Energy Cost of Nutrients in the American Diet." Unpublished doctor of philosophy thesis. University of Connecticut, Storrs, United States.
- Tyedmers, P. (2000). "Salmon and Sustainability: The Biophysical Cost of Producing Salmon through the Commercial Salmon Fishery and The Intensive Salmon Culture Industry." Unpublished doctor of philosophy thesis. University of British Columbia, Vancouver, Canada.
- Tyedmers, P. 2001. Energy consumed by North Atlantic Fisheries. In "Fisheries Impacts on North Atlantic Ecosystems: Catch, Effort and National/Regional Datasets" (D. Zeller, R. Watson, and D. Pauly, Eds.), Fisheries Centre Research Reports 9(3), 12–34.
- von Brandt, A. (1984). "Fish Catching Methods of the World." Avon Litho Ltd., Stratford-upon-Avon, Warwickshire.
- Watanabe, H., and Okubo, M. (1989). Energy input in marine fisheries of Japan. *Bull. Jap. Soc. Sci. Fish.* 53(9), 1525–1531.
- Watanabe, H., and Uchida, J. (1984). An estimation of direct and indirect energy input in catching fish for fish paste products. *Bull. Jap. Soc. Sci. Fish.* 50(3), 417–423.
- Wilson, J. D. K. (1999). "Fuel and Financial Savings for Operators of Small Fishing Vessels." FAO Fisheries Technical Paper 383. FAO. Rome, Italy. 46.

Flywheels

JOHN R. HULL

Argonne National Laboratory
Argonne, Illinois, United States



-
1. Introduction
 2. Rotor
 3. Bearings
 4. Power Input and Output
 5. Containment
 6. Applications

Glossary

flywheel Rotating object, such as a wheel, often used for opposing and moderating by its inertia any fluctuation in speed in the machinery with which it revolves; also capable of storing mechanical energy.

high-temperature superconductor (HTS) A superconductor with a critical temperature above the boiling point of nitrogen.

motor/generator (M/G) Combined motor and generator used to convert between the mechanical energy of the flywheel and electrical energy.

rotational kinetic energy Energy in the form of rotational motion given by the quantity $KE = 1/2 I\omega^2$, where KE is the kinetic energy (units of joules), I is the moment of inertia (units of kgm^2), and ω is the rotational speed (units of rad/s).

specific energy Energy stored divided by mass of object (units of m^2/s^2).

The flywheel, often in the form of a spinning disk or cylinder, stores rotational kinetic energy. As an integral part of the spindle whorl and the potter's wheel, the flywheel is among the most ancient mechanical devices known to humans. The use of the wheel for storing energy seems to be older than its use as a means of supporting vehicles. The flywheel is a fundamental element of many machines, opposing and moderating by its inertia any fluctuation of speed in the machinery with which it revolves. The flywheel principle is also used in several toys such as the top and the yo-yo. More recently,

flywheels have been used as mechanical batteries to store energy for electrical applications. If a motor/generator (M/G) is attached to a flywheel, rotational kinetic energy may be transformed easily and efficiently into electrical energy and vice versa.

1. INTRODUCTION

The essential components of the flywheel are the rotor, bearings, power conversion, and containment. The rotor is the rotating mass in which the energy is stored. The bearings connect the rotating part of the flywheel to a nonrotating platform. Conversion of the stored energy to mechanical power may be accomplished with a mechanical linkage such as a transmission gear. When the flywheel is used as a mechanical battery, a motor/generator (M/G) converts energy between rotating kinetic and electrical. Flywheels are capable of attaining high specific energy, and tip speeds of high-performance flywheels are several times the speed of sound in air. Such flywheels are contained in a vacuum vessel, and an additional safety-related structure often surrounds the flywheel to contain possible debris in case of failure.

The kinetic energy E of a rotating rigid body is

$$E = 1/2 I\omega^2,$$

where I is the moment of inertia of the body about the axis of rotation and ω is the angular frequency of rotation. The flywheel increases its rotational speed as energy is accumulated. The flywheel slows when it discharges energy for use.

A distinctive feature of all flywheel energy storage (FES) systems is their very high power density. They can be charged at very high rates and deliver very high powers, with the only limit being the rating of the transmission system and the torque the flywheel itself can withstand. Power densities greater than 500 W/kg are commonplace. The specific energy of a

flywheel is comparable to electrochemical batteries in the range of 10 to 50 Wh/kg.

The discussion in this article is limited to the flywheel used for energy storage. Flywheels have several advantages over chemical batteries. First, flywheel life is nearly independent of the depth of discharge. The power is limited by the power electronics and the size of the M/G rather than by electrochemistry. Second, the state of the charge of a flywheel battery is readily determined from its rotational velocity, whereas determination of the state of charge for an electrochemical battery is more difficult. The flywheel generates no emissions, makes little noise, and is made of materials with negligible disposal problems.

2. ROTOR

The rotor is the rotating part of the flywheel. It consists of components of the bearing, the M/G, and material whose only purpose is to provide the moment of inertia.

The maximum amount of kinetic energy that can be stored in a rotating solid is determined by its geometry and the strength of the materials that compose the solid. For a rotor composed of a single isotropic material, which is the case for most metals, the maximum specific energy e is given by

$$e = K\sigma_u/\rho,$$

where σ_u is the ultimate burst strength, ρ is the density, and K is a geometric factor of unity or less. The geometric factor K depends only on the shape of the rotor, so the specific energy is independent of size. Because specific energy is proportional to the square of the velocity, a consequence of this equation is that, for a fixed geometric proportion, the rim speed at which a rotor will burst is independent of size. Thus, rim speed is a better figure of merit for comparing rotor materials than is rotational speed because the rotational speed at which the rotor bursts is inversely proportional to size; that is, smaller rotors can attain higher rotational speeds.

In general, when calculating the burst speed, one must take into account tangential hoop stress and radial stress independently. For most geometries with isotropic materials, hoop stress dominates. For operating flywheels, a factor of safety is introduced so that none of the components ever approaches burst speed. Moderate-strength steel, such as 4130, is often used for metallic rotors because of low cost and manufacturing experience. Maraging steel and tita-

nium alloys are used for metallic rotors with higher specific energy. Circumferential stress must also be taken into account when the flywheel is accelerating or decelerating. However, in nearly all cases, the stress induced by acceleration is negligible when compared with the stress induced by centrifugal force.

The specific energy of flywheels can increase dramatically if anisotropic materials such as fiber composites are used to create the rotor. Although many combinations and designs of composites are possible, the most common designs contain long strands of high-strength filaments embedded in an epoxy matrix. The strands are wound predominantly in the circumferential direction. An example of such a rotor is shown in Fig. 1. E-glass, S-glass, graphite, and Kevlar are typically used fibers. Graphite fibers are the strongest and may be 3 to 10 times stronger than, and approximately one-fifth as dense as, steel. Graphite is a crystalline form of carbon. The diameter of an individual graphite fiber ranges from approximately 25 to 800 μm , and the strength of a fiber can range from approximately 2 to 8 Gpa. Usually, the smaller the diameter, the stronger the fiber.

In addition to providing higher specific energy, fiber composites provide an advantage over metals in the area of confinement. When a metallic rotor breaks, its debris usually includes several large pieces moving at high speed. A fiber composite rotor tends to fail with small shard-like pieces or even small grains.

Fiber-based composites exhibit low shear stress, which is related to delamination or separation of the epoxy from the fiber. Thus, hoop-wound designs that contain fiber composites are very strong in the hoop

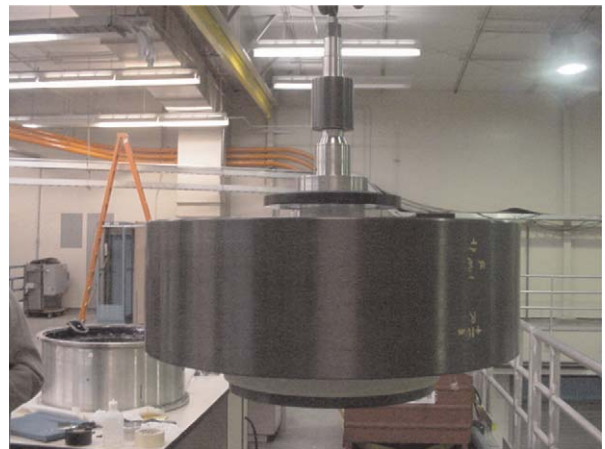


FIGURE 1 Photo of a 1-kWh flywheel with graphite fiber rotor. (Courtesy of Boeing.)

direction but relatively weak in the radial direction. Radial stress is the result of the tendency of the faster moving outer layers to expand more under centrifugal force than the slower moving inner layers. Radial strength can be increased by introducing fibers oriented in the radial direction in a small fraction of the rotor volume.

The radial growth problem also manifests itself in designs that position bearings and power conversion equipment on shafts or hubs at relatively small radii. Connection to a rim with a much larger radius must be made in a manner that takes into account the difference in radial expansion. This is often accomplished by using spokes, flexibly shaped hubs, or connectors that accommodate the radial expansion with a contraction along the axial direction of rotation.

With composite rotors, the kinetic energy in a (large) frisbee-sized flywheel would be approximately 1 kWh, and a person-sized flywheel could store 20 to 40 kWh. Metallic rotors are typically limited to rim velocities of 300 to 500 m/s, whereas composite rotors can easily achieve speeds above 1000 m/s. The highest reported rim velocity at the time of this writing was 2087 m/s. A commonly used rotor design calls for lower strength, lower cost glass fibers for the inner section of a wheel and higher strength, higher cost graphite fibers for the outer section.

It is desirable that the rotor remain rigid over all of its operating speed. Bending modes of the rotor may become resonant vibrations if the rotational speed is at the same frequency as the bending mode. Fibers may be strategically placed along the axial direction of the rotor or in some other direction to change the frequency of a bending mode.

Flywheel design must also take into account the influence of rotor dynamics. A prominent example of this influence is the use of flywheels as mechanical batteries on vehicles. A spinning flywheel has a gyroscopic moment that is proportional to the rotational velocity and acts to keep the axis of rotation always in the same direction. If the vehicle changes direction in a way that changes the axis of rotation, a large torque can be exerted on the vehicle. To overcome this tendency, the flywheel can be suspended in a special gimbaled mounting, or two counter-rotating flywheels can be paired so that the net gyroscopic moment is zero.

3. BEARINGS

The bearings connect the rotating part of the flywheel with a stationary housing and must support

the weight of the rotor, keep it in position, and allow it to rotate freely. These functions must accommodate any motion associated with both the normal rotation of the flywheel and any change in balance of the rotating components. The bearings must resist any external influence that exerts a torque on the wheel that is resisted by the wheel's gyroscopic moment. Bearings may also be divided into thrust bearings and journal bearings. A thrust bearing provides most of its force along the axis of rotation, whereas a journal bearing provides most of its force perpendicular to the axis of rotation.

The bearings of a flywheel must provide stability against motion in any of the 5 degrees of freedom associated with a rotating rigid body. From a mechanical viewpoint, the bearings must provide levitation force, stiffness, and damping. The bearings provide stiffness, which may be thought of in terms of a set of springs that connect the rotor to a massive foundation. A suspended system of this sort will exhibit resonant frequencies at which oscillation amplitudes can be large if an exciting force exists at this frequency. If the rotor is rotating at a speed close to one of these resonant frequencies, it can excite this oscillation. Such rotational speeds are called critical speeds or criticals. The three fundamental modes of oscillation for a rotor with an axis of rotation in the vertical direction are shown in Fig. 2. The oscillation along the axis of rotation (vertical mode in Fig. 2) is generally not harmful. However, a mode transverse to this axis (radial mode) can lead to large imbalance forces as the center of mass rotates at some radius about the system's geometrical center, defined by the bearings. This imbalance force increases with oscillation amplitude and with the square of the rotational frequency. If the criticals are insufficiently damped, the imbalance forces may be large enough

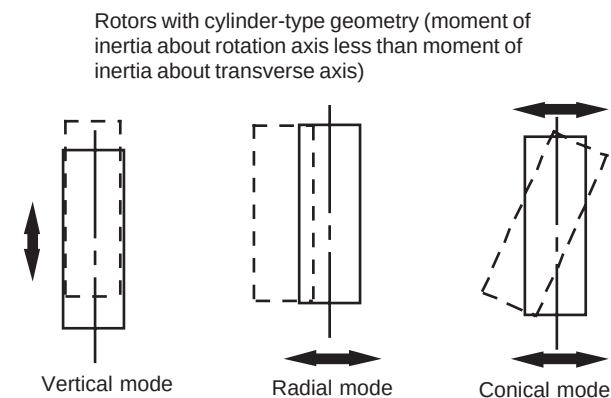


FIGURE 2 Schematic of rigid body critical modes for rotor with vertical rotational axis.

TABLE I

Influence of Bearing Type on the Design and Performance of Flywheels with Stored Energy in the 1- to 10-kWh Range

Bearing type	Levitation pressure (MPa)	Intrinsic loss fraction of $KE_{\max}/h$	Parasitic loss fraction of $KE_{\max}/h$	Total loss fraction of $KE_{\max}/h$	Criticals	Sensitive to rotor imbalance
Mechanical	> 10	0.1	0.02	0.12	Design out of operating range	Most
Magnetic	0.6	<0.001	0.01	0.01	Tunable	Some
Superconducting	0.3	<0.0001	0.001	0.001	Far below operating range	Least

to overcome the restoring force of the bearings and destroy the flywheel.

If the rotor runs in vacuum, some of the bearing components must run in vacuum. Typically, an auxiliary water cooling system is used for mechanical bearings and active magnetic bearings (AMBs). Superconducting bearings require a cryogen. The energy that is needed to operate the refrigerator of the pumps for the coolant is a parasitic loss.

Three general types of bearings are used in flywheels: mechanical, magnetic, and superconducting. Table I provides information for an approximate comparison of the influence of bearing type on flywheel design and performance of flywheels with stored energy in the 1- to 10-kWh range.

Table I shows that the losses of mechanical bearings render these bearings suitable for FES for storage times on the order of minutes to perhaps an hour. Losses of magnetic bearings render these bearings suitable for storage times up to perhaps a day. Superconducting bearings are suitable for FES with storage times up to many days.

FES systems with mechanical bearings have been sold commercially. Prototypes of FES devices that contain magnetic bearings and superconducting bearings are being field-tested. Development continues in all of these bearing technologies.

3.1 Mechanical Bearings

Mechanical bearings can be divided into several subcategories: rolling-element bearings, hydrodynamic journal bearings, gas bearings, squeeze-film bearings, and hydrostatic bearings.

Rolling-element bearings are probably the most common type of bearing used in machinery. They are relatively compact, can transmit heavy loads of various forms, and can be installed and serviced easily. Several specific design types with a range of stiffness are available. Damping in rolling-element

bearings is very small but may be enhanced by squeeze-films.

Journal bearings consist of a circular section of shaft (the journal) rotating inside a bearing “bush,” which is nominally circular. The clearance space between the two is partially filled by the lubricating fluid, which is pressurized by the motion. In addition to reducing the friction in rotational motion, this hydrodynamic film behaves like a complicated arrangement of springs and dampers and so influences the critical speeds and imbalance response. If designed improperly, the film can contribute to rotor instability, which is the result of self-excited vibration. Hydrostatic bearings differ from hydrodynamic bearings in that the lubricant pressure required to separate the bearing surfaces is supplied from some external pressure source and is not a result of journal rotation.

Gas bearings are similar to oil-lubricated bearings, but because the gas is compressible, the behavior differs. Compared with hydrodynamic bearings of similar surface area, gas bearings carry much smaller loads and generally operate with smaller film clearances. The advantage of gas bearings is that they tend to have much smaller rotational losses because the viscosity of a gas is much smaller than that of a liquid. Several types of gas bearings are available: self-acting, externally pressurized, porous wall, and flexible vane.

Squeeze-film bearings are a special case of hydrodynamic bearings. This type of bearing consists of a conventional rolling-element bearing and an external structure for damping radial motion. The outer race of the rolling-element bearing is surrounded by a damper casing. A small radial clearance between the outer race and the inner bore of the casing accommodates a thin lubricating film. The fit is so close that the outer race of the conventional bearing does not rotate. However, the squeezed film is pushed from one region to another by the radial force of the

outer race. The lubrication pressure is built up purely by the action of squeezing the film in this clearance. Friction between the lubricant and the races, and inertia in the lubricant itself, produces a radial stiffness.

If the rotor of the flywheel is rotating in air, any of the preceding bearings can be used. If the rotor operates in vacuum, the number of bearing choices is limited and the vapor pressure of the lubricant must be low. Solid-lubricated ceramic bearings (e.g., those made of silicon nitride [SiN] and lubricated with molybdenum disulfide or other proprietary solid films) are often used in this situation.

3.2 Active Magnetic Bearings

In AMBs, levitation is achieved by the attractive force between an energized magnet and a ferromagnetic body. Typically, the magnet is a stationary electromagnet and the rotor is a ferromagnet. Such a system is inherently unstable, so active feedback is used to modulate the field of the electromagnet depending on the position of the rotor. Such bearings have been used for decades in industrial applications and have been used recently in flywheel prototypes.

In the basic setup of an AMB, position signals from gap sensors are used by a controller power amplifier to set the appropriate currents and voltages of the electromagnets in such a way that stable levitation takes place. In the example of an iron rotor below a ferromagnet, if the rotor starts to fall so that the gap becomes too large, the current in the magnet is increased to increase the attractive force. Similarly, if the gap becomes too small, the current is decreased. Control is effective on bearing activation and continues for all rotational speeds.

A distinctive advantage of the AMB is that it can adapt itself to operating conditions and can communicate with its environment. Within certain limits, the AMB can be set to operate with arbitrary stiffness and damping, independent of the rotor position. The air gap (i.e., the position of the rotor) can be controlled in real time, independently of the stiffness. As a consequence, it is possible to maintain a fixed rotor equilibrium position for various loads. A rigid rotor can be made to rotate about its principal axis without transmitting vibrations to the foundation. By changing the stiffness and damping, the rotor is capable of easily crossing the critical speeds of the bearings. In addition, bending vibrations of an elastic rotor can be significantly reduced through an appropriately designed active control loop. Damping and stiffness are usually mutually dependent.

Early AMBs were powered by analogue controllers; however, the tendency now is toward digital controllers. The design of early AMB systems was based on proportional–integral–differential control algorithms. In general, the feedback law is chosen to be linear and is developed on the basis of the system linearized about the static equilibrium position. Recent investigations have shown that there is performance merit in the use of certain nonlinear feedback laws.

Typically, switching amplifiers are used to power the coils. These amplifiers use transistor switches that connect the coils to a DC voltage. The transistors are either on or off and produce on the coil a square wave voltage, which yields a sawtooth current. By controlling the on and off times, most current waveforms can be produced in the coils. The disadvantage of the switching amplifiers is that the oscillations in the current cause remagnetization loss in the magnetic bearing. However, the shorter the switching period, the weaker are these oscillations.

One of the fundamental limitations of an AMB is that the bearing force slew rate is limited because the magnet coils are inductive and the power supply to the driving amplifier is subject to a maximum voltage. The slew rate limitation causes the bearing force to change more slowly than the control signal demands and, thus, introduces a phase lag. Another way in which to state this effect is that the inductance of the coils and the maximum voltage provide an upper limit to the stiffness of the system.

The position sensor typically uses a high-frequency carrier and provides a linear signal versus the rotor location. Optical sensors, inductive displacement sensors, capacitance sensors, Hall effect sensors, and eddy current sensors are commonly used. The sensors must be contact free and able to measure a rotating surface. The surface quality of the rotor and the homogeneity of the material at the sensor will influence the measurements. A bad surface will produce noise disturbances, and geometry errors may cause disturbances with the rotational frequency or with multiples thereof. “Training” of the sensors in place may alleviate some of these noise problems. In this technique, the rotor is rotated at some fixed speed, the surface is measured over many revolutions, and the average of the measurements for a given set of angles is stored in a computer memory so that the sensor “remembers” the surface. The data set serves as a background that is subtracted from the measurement during regular operation.

A new type of AMB, which is under development, does not use sensors. In place of the usual current

amplifier, a voltage amplifier is used. The current is measured in the amplifier, and a control algorithm can be developed so that the rotor can be stabilized from this measurement.

In principle, the current requirements in the coils of an AMB can be made arbitrarily small if resolution of the sensors is high enough and the sensor and power electronics are fast enough to respond. That is, it takes less force (and therefore less current) to return a rotor from a 1-micron deviation from equilibrium than it does to return it from a 1-mm excursion. As sensors become faster and their resolution becomes higher, and as power electronics becomes faster, one may expect that the losses in AMBs will decrease. At some point, the homogeneity of the rotor and the magnet system will provide the determining limit of these losses.

3.3 Superconducting Bearings

In its simplest form, a superconducting levitation system includes a permanent magnet (PM) levitated in a stable position over a superconductor. Because the superconductor is not a magnet and exhibits special properties, its presence enables the system to bypass Earnshaw's theorem, a general theorem in magnetism that states that any magnetic system will be statically unstable. One of the fundamental properties of superconductors is their tendency to exclude magnetic flux from their interiors. This exclusion of magnetic flux, called the Meissner effect, makes them behave like a strong diamagnet. Accordingly, a superconductor with a PM positioned close above it develops a shielding current, which excludes flux in such a way that the actual magnet "sees" its mirror image.

More specifically, if the magnetization of the PM is in the vertical direction with its north pole down, the image will also be vertically magnetized but with its north pole up, exerting a repulsive force on the real magnet. The closer the magnet gets to the superconductor, the stronger the repulsive force; the farther away the magnet, the weaker the force. This setup can yield a levitation that is stable in the vertical direction.

The levitation phenomenon that is the result of the diamagnetic response is greatly enhanced by additional phenomena that are a result of flux pinning. A flux pinning center is a nonsuperconducting region such as an inclusion, a crack, or some other crystalline defect. Because the superconducting region that surrounds the nonsuperconducting center is strongly inclined to exclude magnetic flux, a mag-

netic flux line through the center tends to become trapped (pinned) there. With a sufficient number of flux lines trapped in the superconductor, the PM will remain stably levitated in position, even over a flat surface. The flux lines between the PM and the superconductor act in an imperfect analogy to mechanical springs with attachments on the PM and in the superconductor. If the magnet is moved vertically or horizontally, the "springs" will tend to pull the magnet back to its equilibrium position. If the flux pinning is sufficiently strong, the magnet may be stably suspended below the superconductor or even along its side.

The potential for superconductor levitation in various applications received greatly increased attention with the discovery of high-temperature superconductors (HTSs) that exhibited critical temperatures exceeding the boiling point of nitrogen. The current material of choice for superconducting bearings is Y-Ba-Cu-O (YBCO). This choice is due to its large superconducting current density at liquid nitrogen temperatures and the ability to grow large grains. In bulk YBCO, the grains can grow to diameters of several centimeters when made by a process called melt texturing. In the current state-of-the-art, the upper limit to this process appears to be a diameter of approximately 10 cm.

If the suspended PM is pulled hard enough to the side or vertically, it is possible to move one or more of the trapped flux lines into new pinning centers and so change the equilibrium position. Such a change leads to hysteresis in the levitation force and an associated energy loss if it occurs in a cyclic pattern.

So long as the magnetic flux distribution in the superconductor does not change, rotation of the PM encounters no resistance. In reality, hysteresis loss in the superconductor will act to decrease the rotation rate. If one picks a radius on the surface of the PM and travels in a circle (azimuthally) on the surface at that radius, one sees a large DC magnetic field with a small AC component. The AC component produces a changing magnetic field at the surface of the superconductor and produces a hysteresis loss. The ideal PM would exhibit a magnetization that is perfectly symmetric about its axis of rotation; that is, it would exhibit a high degree of azimuthal homogeneity. Real PMs contain defects that are a result of the manufacturing process. These defects are mainly a misorientation in individual grains of the easy magnetization from the intended direction. Typically, in PMs with the best homogeneities, at a fixed radius above the rotating surface, the amplitude of the AC component of the magnetic field is of the order of 1%

of the average field at that radius. This inhomogeneity, although small, is sufficient to cause a detectable decay in rotation rate when the magnet spins in vacuum.

Even if the PM were perfectly homogenous, hysteresis loss would still occur due to horizontal and vertical vibrations that are transmitted to the system from the outside world. However, in practical superconducting bearings, the loss due to PM inhomogeneity is much larger than that due to vibration transmission.

As discussed in section II, the bearing gives rise to oscillation modes with corresponding critical frequencies associated with these resonances, with the most important being radial resonance. (If the rotor is a long cylinder, conical resonance is also important, but the discussion is essentially identical to that for radial resonance.) Accordingly, one may divide the loss behavior into three regions: below the resonance, at the radial resonance, and above the resonance. Below the resonance, the losses are due to the inhomogeneity of the magnetic field of the PM. The resonance region typically occurs at low frequency, that is, much lower than the operating range of FES systems. Above the resonance, the losses are influenced by an additional factor, caused by the rotation of the magnet about its center of mass rather than about its center of geometry or center of magnetism. Above and below the resonance, the losses are velocity independent.

Because of the size limitation imposed during production of high-performance bulk superconductors, a large bearing system will need an array of superconductors. In the case of a single magnet levitated over an array of superconductors, an additional loss term arises due to the magnetization of the individual superconductors on levitation. The magnetization of the array leads to an AC magnetic field that is seen by the rotating PM and results in eddy current losses that occur in the PM and depend on the electrical conductivity of the PM.

The low stiffness of the HTS bearing has profound consequences in rotor dynamic behavior of flywheels that is not exhibited with other types of bearings. Rigid rotor bearing resonances, due to the magneto-mechanical interactions in the PM/HTS system, occur at very low frequencies that are on the order of several hertz. Thus, the operating speed regime of disk or ring-type rotors is in the supercritical range, that is, well above the bearing resonant frequencies. Therefore, the operating speed range between the bearing resonances and the frequencies at which bending modes start to appear in the rotor structure is usually large. Furthermore, the low-stiffness

bearing allows the rotor to self-center at low speeds, with relatively low out-of-balance forces at these resonances, so that rotor balancing is less critical than with other bearing designs. The HTS bearings often contain large gaps between rotating and stationary surfaces and can tolerate radial oscillations of several millimeters in amplitude.

3.4 Magnetic Bias

In any of the bearing systems discussed in the preceding sections, considerable advantage in terms of reduced losses can be achieved by using a magnetic bias to take up most of the bearing load. For example, augmentation may take the form of an Evershed-type design, in which a PM pair, one on the stator and one on the rotor, are in attractive levitation (Fig. 3). The spacing of this magnet pair is adjusted so that, at the equilibrium position, the attractive force is between

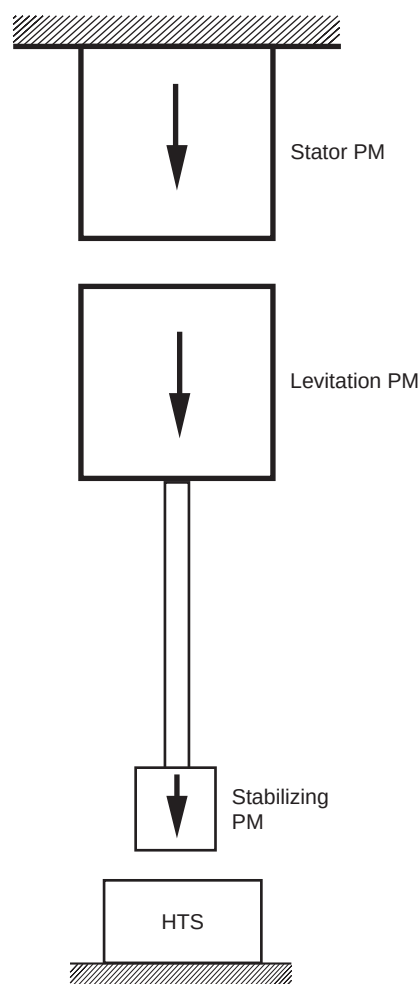


FIGURE 3 Evershed-type magnetic bearing with HTS bearing providing stability.

99 and 100% of the rotor weight. With this magnet pair, the system is stable radially but unstable vertically. An auxiliary bearing system (mechanical, active magnetic, or superconducting) is used to provide the remaining less than 1% of the rotor weight plus the required stiffness. The rotational losses of a PM pair generally consist of eddy currents induced in one PM by the azimuthal field inhomogeneity of the other PM. These losses are typically smaller than those associated with the auxiliary bearing that makes up the stabilizing part of the system. Low-loss magnetic biasing can also be accomplished with a PM and ferromagnet combination.

Because superconducting bearings provide a stabilizing stiffness in all directions, the magnetic bias can be designed with low stiffness. Because an AMB typically provides stable stiffness in only one direction, it may be appropriate to design the magnetic bias with a rather large stiffness that provides stability in all directions but one. Then, only one AMB controller is required to stabilize against movement in the unstable direction.

4. POWER INPUT AND OUTPUT

When the flywheel is used as a mechanical battery, energy must be converted from rotational kinetic energy to usable electrical power and vice versa. This conversion usually involves two major components: an M/G and power electronics. The M/G converts energy between mechanical and variable-frequency electrical. Power electronics then converts between variable-frequency AC and DC electricity. In some cases, the DC is the output of the flywheel. More commonly, additional power electronics converts between the DC and either the external AC electrical grid or a critical load.

Many types of M/Gs are possible. In one commercial flywheel design, the flywheel is on a coaxial shaft that is separate from, and spins faster than, the main generator. When power is required, a clutch is engaged to transfer torque to the constant-speed generator. The motor and generator could be separate components, but typically the M/G is one unit, with the function of operating as motor or generator determined by the power electronics.

A common design for the motor is the brushless synchronous type. In this design, one or more PMs, attached to the rotating part of the flywheel, move in close proximity to stator coils. The power electronics equipment supplies a voltage to the stator coils that is timed synchronously with the rotational angle of the

rotor. When operated as a generator, the rotating PMs induce a voltage on the stator coils, and the power electronics regulate how much current flows. Both the output frequency and the output voltage of the generator depend on the angular velocity of the flywheel; both increase as velocity increases.

The primary function of the M/G is to provide torque to the flywheel. An undesirable side effect in most M/G designs is the production of radial and axial forces that tend to destabilize the flywheel. These forces must be accommodated by the bearings, and for magnetic or superconducting bearings, which are of inherently low electromechanical stiffness, the destabilizing tendency can be significant. Special M/G designs, with slightly lower specific power, have been used in flywheels with these bearings. In one method, a toothless back iron is used in the stator coils. A second method uses a special PM geometry called a Halbach array, which produces an internal dipole magnetic field and very little radial or axial force, even when the rotor part of the M/G is not perfectly centered relative to the ironless stator coils.

Progress in power electronics, particularly in high-power insulated gate bipolar transistors and field effect transistors, underlies higher power flywheel operation. Variable-speed motor drives can be used for the power input section. Several motor operating modes are possible, with a common one providing fixed-voltage pulses to the stator coils at various pulse widths. As the flywheel speed increases and the back electromotive force (EMF) on the stator coils builds, the pulse lengths become longer. Because the generator output voltage decreases with decreasing speed, there is a lower limit on the speed at which power can be withdrawn at the rated value. As speed and voltage decrease, if power stays the same, the current must increase. This means the power electronics must be designed for the lowest operating speed of the flywheel, leading to larger equipment than would otherwise be required for the higher speeds. As an example, a flywheel might be designed to provide full power anywhere from maximum to half-maximum speed. Because stored energy is proportional to the square of the velocity, the available stored energy would be only 75% of the maximum rotational kinetic energy.

5. CONTAINMENT

Flywheels are often contained in some type of housing to maintain the flywheel in a vacuum and to provide personal safety from contact or, in the case

of failure, from debris. Early flywheels were not enclosed, and this is true today of some modern low-performance flywheels such as those used in presses.

Housing flywheels in evacuated containers reduce aerodynamic drag, with the level of vacuum being determined by whatever value of windage loss is tolerable in a particular application. Today's systems typically operate somewhere between atmospheric pressure (100 kPa) and 1 mPa, with the highest speed devices operating at the lowest pressure. Lower speed flywheels are larger but can be operated in air with acceptable losses. Reduction of aerodynamic drag is important for high-performance flywheels not only to reduce the loss rate of stored energy but also to prevent the flywheels from overheating.

A potential parasitic loss is the power required to run the vacuum pumps. It is believed that the rotor containment chambers of commercial versions of high-performance flywheel systems will be exhausted and then hermetically sealed. Vacuum will be maintained over time by a getter system such as barium or activated charcoal. Considerable experience in this technology has been gained from the fabrication of vacuum-insulated containers such as thermos bottles and superconducting magnet dewars. If the rotor is metal, few problems should be encountered in accomplishing this. A properly cured composite should exhibit little outgassing at ambient temperature, but this remains a technical issue for high-performance rotors.

Good design makes the probability of a flywheel failure extremely small. However, for high-performance flywheels, where operating experience is still limited, a safety containment structure is often part of the vacuum housing. For systems that use magnetic bearings, the containment often includes a set of mechanical touchdown bearings. The safety containment of a flywheel is otherwise similar to that of aircraft gas turbines. It must avoid any fragment penetrations and avoid excess deformation; limit the forces, moments, and vibrations transferred to the outside; and avoid the emission of dangerous or pollutant substances.

When a flywheel breaks up, the total angular momentum of the debris cloud remains the same as before the breakup, and this moment must be accommodated by the containment structure. The motion of individual debris fragments is translational and rotational. For very small particles, the energy is nearly entirely translational. Composite material rims usually produce many tiny fragments. Depending on the type of failure, they may take the form of long needle-like shards or small grains. Any PM or

metallic component usually forms larger fragments in the debris cloud. These fragments may provide strong local loading when they hit the containment shell.

Containment structures typically employ a suite of strategies. The energy of the particle pieces is often absorbed in plastic deformation of a collapsible structure inside the containment wall. Turnaround rings and deflection shields are often used to redirect particle momentum. Rotatable liners are sometimes employed to accommodate the angular momentum of the debris cloud.

6. APPLICATIONS

Flywheels have been used in numerous applications. They have provided the prime energy for turning the propellers of torpedoes and for numerous transportation applications in electric buses and railways, both onboard and as wayside power providers. They have been used in mine locomotives where explosion risk is present, and they have been used to power city buses. Flywheels are under consideration to replace chemical batteries in the International Space Station and for use as combined energy storage and momentum wheels in earth-orbiting satellites. Tiny flywheels are used as the universal prime movers in some toys. Flywheels with very high discharge rates, called compulsators, have been used to power electromagnetic rail guns.

Flywheels continue to be used in all reciprocating engines and in all machines that require very high power for a small part of their working cycles. All mechanical presses use flywheels, usually of the disk or rim type; these flywheels are quite massive for the quantity of energy they store. The ability of flywheels to deliver very high power for a short time has been used in inertial friction welding and in some machines such as inertial starters. Other applications of flywheels can be found in devices that must maintain their angular velocity in case of a power failure or during the shutoff of a plant; typical examples include lubrication or cooling pumps and emergency systems.

Flywheels used as mechanical batteries are often integral parts of hybrid electric vehicles. Here, the flywheel is used for times when high power is required, thereby allowing the internal combustion engine to run at a fixed rate where efficiency can be optimized and exhaust pollutants can be minimized. The engine runs an alternator that continuously charges the flywheel. The vehicle uses both an

internal combustion engine and an electric motor for propulsion. The basic idea is that the average power needed to propel the vehicle should be supplied by the engine, which then can operate at a nearly constant optimal speed, thereby reducing fuel consumption, air and noise pollution, and engine maintenance requirements as well as extending engine life. Short bursts of power, for climbing hills and acceleration, are taken from an energy store (a battery of one sort or another) that is replenished directly by the engine or by regenerative braking when the vehicle is slowed down.

Electric utilities have a great need for inexpensive energy storage systems, such as flywheels, because their inexpensive baseload capacity is typically underused at night and they must use expensive generating sources to meet their peak loads during the day. The availability of a distributed network of diurnal storage could also make use of underused capacity in transmission lines at night as well as add robustness to the electric grid. Efficient energy storage would also be beneficial to renewable energy technologies such as photovoltaics and wind turbines.

The most practical FES option assumes arrays of many individual flywheel units. Although FES units with stored energy as low as 2 to 10 kWh may have some applicability, electric utilities are most likely to desire units with approximately 5 MWh of storage

capacity. Units with such capacity, composed of modules of individual 10-kWh flywheels, are approximately the largest that can be factory assembled and truck mounted for delivery to substations or energy storage facilities. With a fleet of FES units that handle real-time automatic grid control, a utility could manage generation variability as well as load variability. Such units could also provide uninterruptible power for factories.

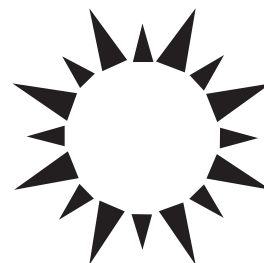
SEE ALSO THE FOLLOWING ARTICLES

Batteries, Overview • Electrical Energy and Power • Electric Motors • Hybrid Energy Systems • Mechanical Energy • Storage of Energy, Overview

Further Reading

- Ehrich, F. F. (1992). "Handbook of Rotordynamics." McGraw-Hill, New York.
- Genta, G. (1985). "Flywheel Energy Storage." Butterworths, London.
- Goodwin, M. J. (1989). "Dynamics of Rotor-Bearing Systems." Unwyn Hyman, London.
- Hull, J. (2000). Superconducting Bearings. *Supercond. Sci. Technol.* 13, R1–R14.
- Schweitzer, G., Bleuler, H., and Traxler, A. (1994). "Active Magnetic Bearings." Hochschulverlag, Zurich, Switzerland.

Food Capture, Energy Costs of



JERRY V. MEAD

State University of New York, College of Environmental Science and Forestry
Syracuse, New York, United States

-
1. Introduction
 2. Bioenergetics
 3. Food Capture, Maximum Power, and Evolution
 4. Environmental Impact on Energy Cost of Food Capture
 5. Feeding Strategies and the Costs of Food Capture
 6. Summary

Glossary

- abiotic** Pertaining to nonliving environmental factors, including light, temperature, water, nutrients, soil, and atmospheric gases.
- assimilation efficiency** The fraction of ingested food that an organism is able to absorb.
- bioenergetics** The study of the flow and transformation of energy within and between living organisms and their environment.
- biotic** Of or having to do with life or living organisms.
- ectotherm** An organism that regulates its body temperature largely by exchanging heat with its surroundings.
- endotherm** An organism that generates heat from food energy to maintain its body temperature, typically above the temperature of its surroundings.
- homeotherm** An organism, such as a mammal or bird, that maintains a constant body temperature, largely independent of the temperature of its surroundings.
- maximum power hypothesis** A concept stating that there will be selection for rate of energy use that yields the most useful work per unit time.
- poikilotherm** An organism, such as a fish or reptile, having a body temperature that varies with the temperature of its surroundings.

The amount of energy expended by an organism to collect a unit of energy is referred to as the energy costs of food capture. The ability of a species to capture and utilize energy is key to its survival. Therefore, all organisms expend energy to maintain basic body functions and require additional energy to feed, grow, and reproduce. The energy costs for

capturing food include energy expenditure for finding, handling, digesting, and assimilating the food. Both biotic and abiotic factors influence the profitability of a given method of feeding, and the energy gained from feeding minus the energy costs of capturing that energy is an organism's net gain of energy that can be used to increase the fitness of a population.

1. INTRODUCTION

The influence of the environment on the success of food capture can be mediated by behavioral and physiological adaptations. However, such adaptations are not made by the organism, but are products of natural selection. The theory of evolution through the process of natural selection states that traits in organisms that produce the most descendants have been selected for over time. For an organism to create descendants, an organism must have a surplus of energy to invest in reproduction and/or the rearing of offspring. In a sense, organisms are players in a game in which there are fees, in terms of energy and nutrients, to remain a player and additional fees that must be invested into activities that may win a player surplus energy. Surplus energy can then be used to perpetuate their genes. In this game, the rules are determined by physical, chemical, and biological processes. Successful organisms are those that, by making profitable investments, acquire surplus energy per unit time of playing the game and then use this energy to perpetuate their genes.

Why are there so many species on Earth? This is one of the most fascinating questions asked by ecologists. The main reason for the richness of species may be that a diverse set of organisms is more effective than a limited set is at capturing available food energies. This idea is illustrated in the following discussions by describing the current understanding, for both plants and animals, of the

energy costs of maintaining life and capturing food. The focus is on net energy gain, although it is recognized that resources other than energy, such as nutrient availability, can influence the profitability of a foraging strategy. How the biota minimize the energy costs of food capture using physiological and behavioral adaptations is also described.

2. BIOENERGETICS

Understanding the costs related to food capture requires a basic knowledge of an organism's demand for energy and the profitability of captured energy in terms of growth and reproduction. Bioenergetics is the study of the flow and transformation of energies used within and by living systems. A thorough review of bioenergetics is not the purpose in this section; instead, the focus is on an organism's basic needs for energy and how these basic needs relate to the environment, followed by a discussion of how an organism can invest its energy into capturing food to yield a net profit.

2.1 Energy Flow in Organisms

The pathways of energy flow in biota are similar among species. Species vary primarily in terms of rate or efficiency of processes, method of food capture, and types of foods consumed. Once food has been collected, it is defined as consumed. A common misconception is that everything consumed enters an organism. Sometimes an organism loses pieces or fluids of the food being consumed before it is utilized. The portion of food consumed that passes into an organism is defined as ingested material. Ingested food then passes through the internal energy pathways of the organism.

Ingested food must be absorbed or assimilated into the organism for food to be useful to do work. For plants, light energy that has excited a photosynthetic pigment is converted into internal available energy. In animals, food in the digestive system is assimilated through absorption across the digestive tract. Saprophytic organisms such as fungi assimilate foods by absorbing it directly into cells through diffusion. The efficiency of assimilation varies among biota and within species (Table I). These differences in efficiencies are due to many factors, such as rate of food ingestion (related to temperature, incidentally), food quality, and organism physiology. Energy that is not assimilated is lost through egestion in animals, through light reflectance and heat loss in plants, and through leaching in saprophytes. In animals, there is

another loss associated with moving ingested materials through the body, i.e., specific dynamic action. Other organisms have smaller costs for internal transport of foods and rely mainly on diffusion and chemical processes.

Assimilated food energy is the actual quantity of energy available to fuel various life processes. In most cases, assimilated energy is used first to fuel metabolism. Total metabolism refers to the sum of energy used to rebuild and fuel the basic functions of life (basal metabolism) and the energy used to perform activities (active metabolism) such as swimming, flying, or walking.

2.2 Energy Costs and Demands for Food Capture

2.2.1 Energy Cost of Metabolism

The energy and time costs of capturing and finding food include both basal and active metabolism. Basal metabolism increases exponentially as a function of environmental temperature until the temperature reaches the optimal temperature for the organism and then declines with increasing temperature. Optimal temperatures for basal metabolic rate can differ among species and among individuals of a given type of species; this is primarily due to enzyme reactivity. In addition, basal metabolism decreases per unit body weight as body weight increases, and varies depending on whether an organism regulates its internal body temperature (Fig. 1). Time spent searching or waiting for food requires energy for basal metabolism. The increased physical energy used in active metabolism can be one of the largest costs in food capture and is difficult to measure or predict because it depends largely on behavior and, like basal metabolic rate, can be controlled by environmental conditions. For example, a tuna fish can increase its metabolic rate when foraging and when the tuna seeks prey in warmer surface waters.

2.2.2 Energy Costs for Locomotion

Time spent searching for food can be costly in energy for an individual. These costs can be large due to energy-intensive activities such as walking or the time required locating prey. The energy costs of various activities are often difficult to quantify. Scientists have developed elaborate techniques to measure active metabolism, including assessing metabolic tracers (e.g., oxygen), using sensors that measure muscle contractions, and analyzing movement using video equipment or mechanical devices

TABLE I
Assimilation Efficiencies for Poikilothermic and Homeothermic Species^a

Species	Diet	Assimilation efficiency (%)	Production efficiency (%)
Aquatic poikilotherms			
<i>Megalops cyprinoidea</i> (fish)	Fish	92	35
<i>Nanavax</i> sp. (predatory gastropod)	Invertebrates	62	45
<i>Hyallolela azteca</i> (amphipod)	Algae	15	15
<i>Ferrissia rivularis</i> (stream limpet)	Algae	8	19
<i>Tegula funebris</i> (herbivorous gastropod)	Algae	70	24
<i>Hydropsyche occidentalis</i> (filter-feeding caddis fly)	Algae	55	10
<i>Littorina</i> sp. (periwinkle)	Detritus	45	14
Terrestrial poikilotherms			
Saltmarsh grasshopper	<i>Spartina</i>	27	37
<i>Melanoplus</i> species (grasshopper)	Lespedeza	37	37
Mites (Oribatei)	Detritus	25	22
Sugarbeet leaf (plant)	Light	16	—
Terrestrial homeotherms			
Least weasel	Mice	96	2.3
Little brown bat (<i>Myotis lucifugus</i>)	Insects	88	—
Marsh wren	Insects	70	0.5
Grasshopper mouse	Insects	78	5.7
Kangaroo rat	Seeds	87	6
Vole	Vegetation	70	3
Jackrabbit sp. (<i>Lepus</i>)	Vegetation	52	5.5
Jackass penguin (<i>Spheniscus demersus</i>)	Anchovies	78	—
Cottontail rabbit (<i>Sylvilagus</i> sp.)	Vegetation	52	6
Cattle	Vegetation	38	11
African elephant	Vegetation	32	1.5

^aCompiled by J. Mead using data from Colinvaux (1986) and other sources.

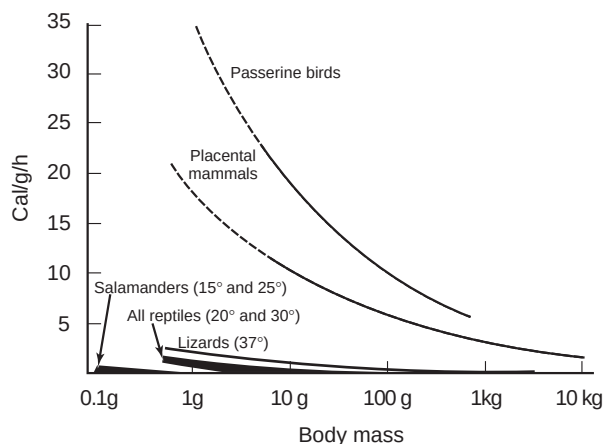


FIGURE 1 Resting metabolism as a function of body size and temperature (in °C). Ectothermic animals can be small because their energy consumption for maintenance is low. A bird weighing 1g would use about 30 times the energy used by a reptile of the same weight. Not surprisingly, 1-g reptiles exist but 1-g birds do not. Reproduced with permission from Pough (1980), with permission from the University of Chicago Press.

(e.g., treadmills). For example, the field metabolic rate (FMR), the total metabolism for an organism in an environment, can be estimated for both terrestrial and amphibious animals by injecting “heavy” water into an animal’s body. Heavy water has more neutrons than protons; the deuterium isotope, which replaces many of the hydrogen atoms in the water molecules, is detectable using laboratory techniques. Because the metabolic processes of most plants and animals depend on water, the rate of metabolic processes such as basal and active metabolism can be quantified by tracing the rate of uptake and release of the heavy deuterium isotope in the injected water. The field metabolic rate for many organisms has been estimated (Table II).

Movement is often the largest energy cost of food capture. Swimming, flying, and running are the basic types of movement. The energy efficiency of movement (distance traveled/unit weight/unit energy expended) differs with the medium that the

TABLE II

Measured Field Metabolic Rate and Energy Cost of Food Capture for Selected Ectotherms and Endotherms^a

Species	Locomotion used	Feeding strategy	FMR ^b (kJ/g/day)	Cost of food capture (kJ captured:kJ cost)
Ectotherms				
Alder tree (<i>Alnus viridis</i>)	Nonmotile	Capture light energy	0.0005	10:1
Norway Spruce tree (<i>Picea abies</i>)	Nonmotile	Capture light energy	0.00275	9.1:1
Mussels	Nonmotile	Filter feeder		5:1
Komodo dragon (<i>Varanus komodoensis</i>)	Walking	Sit-and-wait predator	0.054	
Sand lizard (<i>Eremia lineocellata</i>)	Walking	Sit-and-wait predator	0.166	1.6:1
Black and yellow sand lizard (<i>Eremia lugubris</i>)	Walking	Stalking predator	0.210	2:1
Endotherms				
Anna's hummingbird (<i>Calypte anna</i>)	Flying	Collector	7.1	1.7:1
Little brown bat (<i>Myotis lucifugus</i>)	Flying	Stalking predator	2.2	1.9:1
Jackass penguin (<i>Spheniscus demersus</i>), breeding	Swimming	Stalking predator	0.64	2.1:1
Jackass penguin (<i>Spheniscus demersus</i>), nonbreeding	Swimming	Stalking predator	0.64	6.8:1
Commercial fishing boat (shrimp, scallops, herring, cod)	Propeller	Filter feeding		0.03:1–0.14:1
Modern agriculture	Tractor	Omnivore	5.0–60	1.4:1–60:1
Humans, foraging	Walking	Omnivore	0.001	3:1–6:1
Humans, pastoralism	Walking	Omnivore	0.01	3:1–5:1
Shifting agriculture	Walking	Omnivore	0.4–1.5	17:1–25:1
Traditional farming	Walking	Omnivore	0.5–2.0	17:1–20:1

^aData compiled by J. Mead.^bFMR, field metabolic rate.

organism is traveling through (such as water or air) and the morphology, body weight, and speed of movement of the organism. The rate of movement that is most energy efficient is defined as the minimum cost of transport. Most organisms have a rate of movement that has the highest energy efficiency, but there are some exceptions (such as starlings), when energy efficiency increases with speed.

For each mode of locomotion, the energy cost of transport decreases per unit weight of organism with increasing body weight (Fig. 2). Large animals such as blue whales may expend much energy per unit time for movement, but the energy cost to move a gram of whale is relatively low compared to the energy cost of moving a gram of mouse the same distance. The relation between body weight and the cost of transport helps explain the profitability of

movement in foraging energetics. The high cost of locomotion for small organisms is the reason why most migratory animals have body weights greater than 1 kg. It is simply too expensive for small animals to migrate long distances without additional forces such as wind and ocean currents. The present explanation for the increase in energy efficiency in locomotion with size is that, compared to smaller organisms, larger organisms have greater inertial forces that aid in locomotion, and frictional forces (such as gravity) are much less due to lower ratios of surface to body weight.

Swimming by fish is the most energy-efficient method of locomotion on Earth, followed by powered flight and then crawling and running (Fig. 2). Each species must have structures and physiologies suited to the type of environment in which they live. Swimming and flying are more

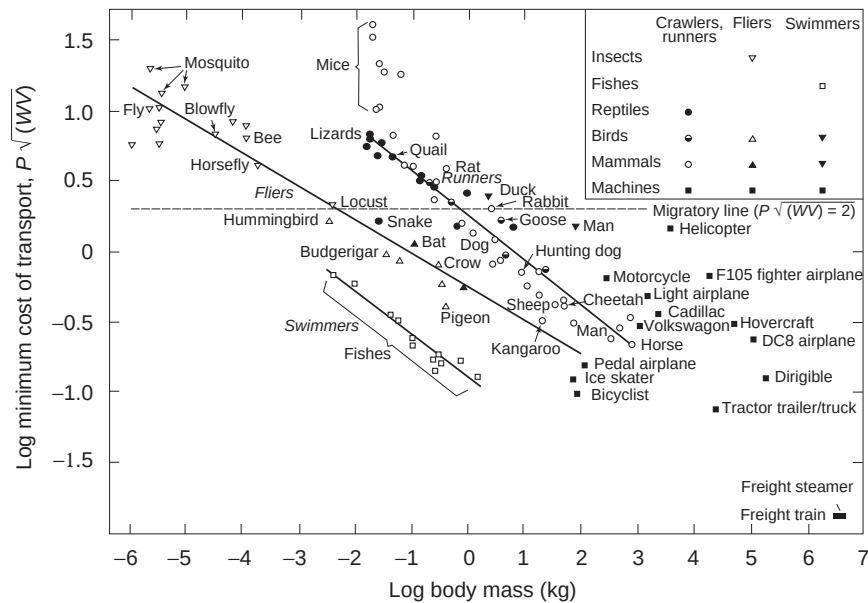


FIGURE 2 Energy costs of swimming, flying, and running, illustrating the energetic costs for searching for food. The data suggest that animals suited to swimming are more energy efficient at locomotion compared to flying and running species. Reproduced with permission from Tucker (1975).

efficient methods of movement than is running or walking because muscle efficiency is less in running or walking, compared other methods of transport. In running or walking, a muscle must contract and stretch to complete a cycle of movement such as swinging a leg. Contracting a muscle pulls on tendons that move parts of the body, and, in walking, muscles contract in the leg to push the body forward. Once the leg has contracted and performed work by pushing the body forward, the muscle must be stretched to maintain the body's center of gravity and to move the leg back into position for contraction. The energy expended to stretch a muscle is lost as heat in walking or running and performs negative work in that the stretch of the muscle must displace the appendage used in locomotion in a direction that is not the direction of travel. The main reason swimming and flying are a more energy-efficient means for locomotion compared to walking or running is that muscles are stretched less often. As a means for movement, however, swimming, not flying, is more energy efficient; until recently, scientists could explain only one-seventh of the difference in efficiency between these two methods of locomotion. Several aspects of swimming explain its efficiency: (1) muscles are typically stretched little in swimming; (2) most fish can regulate buoyancy using a swim bladder and/or their body density is similar to that of the surrounding medium, decreasing the energy used in maintaining a vertical position

(flyers and runners must invest energy in maintaining their vertical position in the air); (3) vortices generated from the thrust of a foil, such as a tail, allow a swimmer to recover some of the energy lost in the movement of the tail; and (4) swimmers can utilize vortices produced by the environment, giving them the ability to achieve energy efficiencies greater than 100%. Machines such as submarines or boats have low energy efficiencies in movement compared to fish because their rigid hulls restrict the vessels' ability to utilize energy found in vortices in the environment. In addition, human mechanisms for converting mechanical energy into a jet or propulsive force do not generate vortices that are useful for recovering energy lost during thrust. Propeller-produced propulsion generates vortices that are perpendicular to the direction of movement and offer little potential for recovering energy. In this case, natural selection still seems to be a more efficient engineer than humans are.

Birds, bats, and insects are the only animals that have the ability for self-powered flight. Flying and swimming are similar in that lift and thrust are the two principal forces important to locomotion. Thrust is defined as the force applied by the organism to generate forward movement, whereas lift is the force exerted by an organism to change or maintain vertical position in a fluid. At a given body weight, flying is a less energy-efficient means of locomotion compared to swimming because more energy is

needed by flyers to generate lift. Both swimmers and flyers that use foils, such as wings and tails, generate vortices that can produce efficient jets of fluids. The jet produced by movement of a foil produces motion from thrust. Without a medium such as water or air, the movement of a tail or wing would not create a propulsive jet.

Walking and running organisms can recover some of the energy lost in stretching muscles by having elastic structures that store the energy used in stretching or by having many legs (e.g., centipedes), which reduces the energy expended in maintaining the center of gravity. The low energy efficiency of locomotion for small-bodied species has yet to be explained, but larger organisms are believed to have elastic tendons that can recover some of the energy lost in stretching muscles. As body size increases, animals have more tendon biomass to absorb energy used in stretching muscle. The recovery of energy used in stretching muscle is similar to the deformation of a bouncing ball. A ball bounces because the energy in the force of the ball hitting the floor is stored briefly in the deformed ball and then is released when the ball structure reforms, which pushes the ball away from the floor. Like a bouncing ball, biota can store in the animal's body some of the energy in the force of the animal's appendage hitting the ground, and then use this stored energy to maintain the center of gravity or for movement, by redirecting the energy into a force that can perform work.

The wheel has increased energy efficiency in locomotion from that of walking and running by reducing the number of muscles that are stretched. Imagine a person walking on a level surface and a person riding a bicycle next them; the walker's body dips during stride and then is pushed upward, but the bicyclist's body does not dip or bob during motion. A wheel maintains a person's center of gravity during stride and eliminates the bobbing action of a walker and the associated stretching of muscles. Each force applied to a pedal by bicyclist results in a perpendicular forward force using a gear. The ability of biota to perform certain types of locomotion has a large influence on the energy costs of finding and capturing food, but the environment sets the stage for the need for movement.

Once the energy cost of total metabolism has been paid, surplus energy and materials are used in foraging, growth, or reproduction. If surplus energy does not exist, life functions can be performed only by using the finite energy supply stored by the organism.

3. FOOD CAPTURE, MAXIMUM POWER, AND EVOLUTION

3.1 Optimal Foraging Theory and Maximum Power

Optimal foraging theory states that traits of an individual will be selected for that maximize the surplus energy gained per unit time from foraging. Optimal foraging theory is used to predict the most profitable feeding strategy for an organism given the energy costs for capturing food versus the energy gained from food capture. For example, as shown in Eq. (1), Holling's disk equation (so named because C. S. Holling used paper disks in his model of functional response) can be used to predict an optimal feeding strategy:

$$\frac{\text{Surplus energy}}{\Delta \text{Time}} = \frac{(\text{Energy gain} - \text{Cost of capture})}{(\text{Search time} + \text{Capture time})} \quad (1)$$

where surplus energy is food energy captured that is available for maintenance, activities, growth, or reproduction; energy gain is energy assimilated by capturing food; cost of capture is energy expended while searching for and capturing food; search time is time spent searching for food; and capture time is time spent capturing food. Equation (1) illustrates several components of a feeding strategy that can be altered to maximize the amount of surplus energy gained from feeding. An organism can reduce the time hence the energy costs to search for or capture food. In addition, the energy gained from consuming food can be maximized by selecting for the highest quality food items.

Optimal foraging theory fits neatly within a more universally applicable concept, Howard Odum's maximum power concept. The maximum power concept predicts that the optimal strategy would be to invest the amount of potential energy that yielded the largest quantity of useful work per unit time, which for organisms is surplus energy per unit time. An organism must invest an amount of potential energy into capturing food that yields the most work in terms of growth or reproduction per unit time, which may not be the most energy-efficient food capture strategy available to them.

3.2 Influence of Time and Space on Foraging Energetics

Factors that influence the cost of food capture can be spatially and temporally complex. Food supply may be patchy in distribution and these patches of food may move, shrink, or grow. In addition, physical conditions

may vary over space and time, influencing the cost of food capture. An organism can increase the profitability of a foraging strategy by investing energy and time into migration. For example, zooplankton production in the North Pacific Ocean changes over a season because the spatial and temporal variability of the environment affects the profitability of foraging strategy. Fish that remain in one location and consume the zooplankton may not maximize net energy gain because the food abundance shifts with the season. Even if animals find a good patch of food to feed on, it may be more profitable to move to a new location after a period of feeding than to remain in one place. Movement to new areas for foraging can be risky, in that it is uncertain whether an animal will find better foraging conditions, the animal may expose itself to predation risks, and there is an energy cost for locomotion.

Average-rate-maximizing foraging models use patch models to predict optimal feeding strategies by determining the most profitable investment of energy and time spent searching for food in a patch versus moving between patches. If the entire region available for foraging is thought of as a mosaic of patches of feeding conditions that vary in profitability, the mean profitability of patches can be used as the benchmark to determine the relative profitability of feeding in any given patch. Patch models predict the amount of time an organism should feed in a given patch by determining the profitability of foraging in a patch relative to the mean profitability of foraging in all patches and the amount of time required to move to a new patch. In conditions in which the relative profitability of a patch and the time required to move between patches are high, the model predicts that the forager will remain in the patch for a long period of time. Foragers are thought to move between patches often when foraging conditions are relatively poor when compared to other patches, and energy and time costs for travel are small. Foraging models aid in understanding the role of migration in foraging energetics, but these models are only formalizations of knowledge about these systems and hence are incomplete.

4. ENVIRONMENTAL IMPACT ON ENERGY COST OF FOOD CAPTURE

4.1 Environmental Factors Influencing the Energy Cost of Food Capture

Many factors in the environment can influence the energy cost of food capture, and understanding their

individual effects is difficult because many factors interact. Abiotic and biotic conditions that influence foraging energetics include the quantity and quality of food, environmental temperature, light, fluid dynamics, nutrient availability, and risk of injury or mortality.

4.1.1 Quantity and Quality of Food

Quantity of food influences the amount of time and energy spent searching for food; food quality relates to the profitability of a food item for growth and reproduction. In many foraging systems, the amount of food available for foraging is not directly proportional to the energy cost of food capture. As food abundance increases, the amount of food captured per unit time reaches an asymptote (Fig. 3) and the energy costs of food capture decrease.

4.1.2 Environmental Temperature

Temperature is considered as the master regulator of metabolic rate. Knowing how a species responds to temperature is useful in understanding the ability of an organism to perform feeding activities and the associated energy costs. The response of an organism to environmental temperature is described as homeothermic, poikilothermic, ectothermic, or endothermic, based on the physiology of the organism. Homeotherms are species with internal body temperatures that vary little over time; poikilotherms have varying internal body temperatures. Humans are homeotherms that maintain internal body temperature at 98.6°F. Insects are examples of poikilotherms because their body temperatures may vary

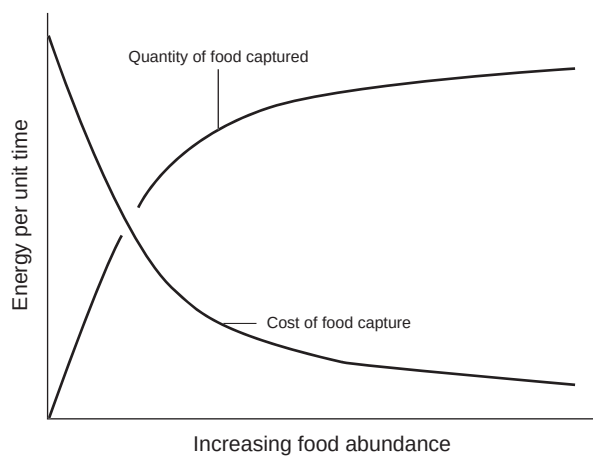


FIGURE 3 Energy cost of food capture as a function of food abundance, illustrating that in many feeding systems, the ability to capture food reaches an asymptote with increasing food supply.

greatly over time. Ectotherm body temperature is regulated by the external environment, whereas endotherms regulate their internal body temperature using internal energy. A homeotherm is a specialized endotherm that has a relatively constant internal body temperature, whereas endotherms do not necessarily need to have constant body temperatures. Each of these factors influences the efficacy and efficiency of energy capture.

Endotherms have a feeding advantage over ectotherms in cold climates, because the metabolism of ectotherms, and hence feeding, is restricted in a cold environment. Despite this restriction on feeding ability, ectotherms may not starve because the basal metabolic demand for energy is also less in cold weather. The endotherms have an advantage over ectotherms in these cold climates because they can generate internal body heat, increase their metabolism, and increase their foraging rate at an energy cost. Large size decreases heat loss through decreasing surface-to-volume ratios, thus polar bears in the Arctic reduce energy costs of maintaining elevated body temperature by being large and have higher metabolisms for foraging. A potential disadvantage to the homeothermic design is that, when compared to poikilotherms, it is a high-energy demand, low-energy efficiency system that requires more food for survival compared to a poikilotherm of similar body weight (Fig. 1; Tables I and II). If food is in low supply, the homeotherm may starve quickly because it must use its energy reserves to maintain an internal body temperature.

Robins (*Erithacus rubecula*) are a good example of the utility of increasing metabolism in a cold environment. The robin is homeothermic, but its prey (earthworms and insects) is both poikilothermic and ectothermic. The prey of the robin has a much lower total metabolism compared to a robin in a cold environment; the higher active metabolism of the robin enables it to catch prey easily because it can move much faster than the prey. The difference in active metabolism between prey and predator can decrease the predator's cost of food capture, as in the robin's case, or it can increase the cost of prey capture when the prey, compared to the predator, has much higher active metabolism.

The Komodo dragon (*Varanus komodoensis*) found in Indonesia is an example of a predator having difficulty running down and capturing its prey because it has a much lower active metabolism. The Komodo dragon preys on birds, rodents, and large mammals (such as deer). Deer, the primary prey item of dragons, are homeotherms, whereas the Komodo

is a poikilotherm. It is difficult for the Komodo to run down a deer and physically overpower it because of the Komodo body structure and relatively low active metabolism compared to a deer. However, the dragon does have a good chance at injuring or subduing a deer through surprise attacks. Even if the prey escapes in a surprise attack, if the dragon has bitten the prey, the dragon often captures it eventually because the prey has been poisoned with a deadly bacteria (*Staphylococcus*) found in the Komodo's saliva.

4.1.3 Light

The quality and quantity of light affects the growth rates of plants and also the success of many foragers at capturing food. Light intensity influences prey detection, rate of photosynthesis, and indirectly several other factors that impact foraging systems. The amount of energy received from light per unit time is related to the frequency of photons that reach a surface. Frequency of light is inversely proportional to the wavelength traveled by photons. Wavelength also defines the type of light energy present in an environment. Wavelengths of visible light have certain colors. Light energy used for plant growth—photosynthetic active radiation (PAR)—ranges in wavelength from 400 to 700 nm. Red light, at a wavelength around 700 nm, has the lowest energy for photosynthesis. Blue light, at a wavelength of around 400 nm, is the highest energy PAR.

In aquatic communities, the amount of energy in various wavelengths determines how far light is transmitted through water and hence the availability of light to plants. If a photon of light is conceptualized as an arrow shot from a bow, the arrow that has the most initial energy, in momentum, will travel deepest into a body of water before frictional forces deplete the arrow's momentum. Thus green light is absorbed or reflected near the surface, whereas blue light, which has the most energy, will travel further than green light. The clarity of a gas volume or a body of water, which relates to the amount of materials in suspension, will influence how deeply light penetrates. As result of differences in the photons' ability to be transmitted through water, relative abundances of photons traveling across different wavelengths change with distance from the surface of a body of water.

Plants and fungi must invest energy to capture food energy by building structures and chemicals (e.g., chlorophyll). Adaptations in the structure of a forest canopy and in the composition of an ocean phytoplankton community appear to maximize surplus

energy captured from the sun. The structure of a forest canopy represents layer upon layer of vegetation in space whereby upper layers shade lower layers from sunlight. A fraction of the light intercepted by the overhanging canopy is reflected. Sunlight that penetrates into the forest canopy has a larger relative abundance of high-energy frequencies of light, but is lower in total energy compared to light at the top of the forest canopy, but still contains energy for photosynthesis. In a dense forest, the amount of leaves will increase per unit area the closer the canopy layer is to the forest floor, until the addition of leaf biomass does not yield a net gain from energy capture.

Natural selection has even used ultraviolet and infrared light for food capture. Rattlesnakes and blowflies are just two examples of the ways species sensing infrared or ultraviolet light can reduce the energy costs of finding food. Rattlesnakes are poikilotherms that have low active metabolisms in cold conditions (Table II). Thus, like the Komodo dragon, rattlesnakes have difficulty chasing and then subduing their prey and are more effective at capturing food as sit-and-wait predators that use poisons. The pit organs on the rattlesnake's head are receptors of infrared radiation that assist in locating food. Blowflies, in contrast, forage on carrion. Carrion is an ephemeral source of food and is patchy in distribution. Carrion is often a good source of energy and nutrients, creating a strong selection pressure for organisms that can find and utilize this resource quickly. Ultraviolet radiation is emitted at certain stages in carrion decay. If a light frequency matches that emitted from carrion decay, the fly investigates; otherwise, the fly conserves its energy for more likely sources of food. The blowfly's energy cost of food capture would be much larger without the ability to sense the presence of such an ephemeral food source.

The direct relation between light and the success of predator prey systems is apparent, but other indirect effects of light on these systems must not be overlooked. Most of life on Earth is fueled by solar energy. Temperature of the surrounding environment depends on energy in light, which in turn not only regulates the metabolism of organisms but determines other physical properties of the environment that affect the energy cost of food capture, such as fluid dynamics.

4.1.4 Fluid Dynamics

Fluid dynamics influence the cost of food capture and movement by controlling food availability. For

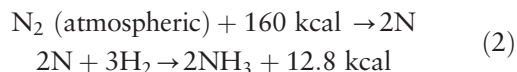
example, caddis fly larvae and black flies feed on food transported in fluids. Species of caddis fly larvae live on stream bottoms and depend on food that has deposited in the stream bottom. Water flow may move food away.

On a larger scale, fluid dynamics generates hurricanes, tornadoes, and flooding that can greatly alter environments. Changes in foraging energetics for a species associated with a disturbance can be beneficial or nonbeneficial for growth and reproduction and the impacts on foraging may be long term or short term. Many foraging organisms have a high "energy returned on investment" (EROI) because they are exploiting the large energy flow of their environment.

Fluid dynamics also impacts the thermodynamics of biota. The movement of fluids increases the rate of thermal conductivity, causing phenomenon such as the wind-chill effect. The profitability to a homeotherm of foraging can be altered by changes in thermal conductivity because of changes in energy costs of maintaining body temperature. Similarly, by influencing thermal conductivity, fluid dynamics alters the ability of poikilotherms to perform foraging activities.

4.1.5 Nutrient Limitation

In plants, nutrient limitations increase the cost of capturing light energy by limiting the plant's ability to build the necessary structures and chemicals for photosynthesis. Nitrogen fixation is an energy cost for food capture that can compensate for nutrient limitation. Bacteria and blue-green algae are the only organisms capable of nitrogen fixation and are either free living or form symbiotic associations with plants or other organisms (e.g., termites, protozoa, and legumes). Nitrogen fixation is an energy-intensive method for fixing atmospheric nitrogen to compensate for nutrient limitation [Eq. (2)]. Nitrogen fixation is only profitable in nitrogen-depleted environments.



An interesting example of the energetic constraint that relative nutrient availability can have on energy cost of food capture is found in a trout pond ecosystem located below a pasture in western Manitoba, British Columbia. Several of the eutrophic Manitoba ponds studied by David Schindler (of the University of Alberta) had phytoplankton populations composed mainly of nitrogen-fixing, blue-green algae that were encouraged by the high concentration of phosphorus

in water from cow manure. Blue-green algae are a very poor source of food for upper trophic levels and in addition are poisonous, so trout died. Nitrogen was added to the trout ponds, which changed the algal assemblage to predominately green algae, because nitrogen fixing was no longer advantageous. Green algae are a high-quality food source for zooplankton and resulted in an increase in trout production.

4.1.6 Foraging Energetics in Relation to Risk of Mortality

The energy cost of food capture can be influenced by the risk of mortality by restricting the time and/or location that an organism spends foraging. For example, Barbara Peckarsky (of Cornell University) found that the presence of insectivorous fish (trout) gave a chemical signal for mayflies to forage mainly at night, rather than when trout could see them. The hiding of mayflies under stones during the day in response to the presence of trout increased the energy cost of food capture for both predator and prey.

Disease or injury can also increase the energy cost of food capture for plants. The tar spot fungi (*Rhytisma* sp.) infect the leaves of trees maple trees, creating large, black, tarlike spots on leaves. Sometimes the tar-spot infection is so profuse that the tree will drop all of its leaves and then grow new ones. Growing a new set of leaves costs the tree much energy and nutrients and the energy contained in leaf biomass is large enough that it is energetically profitable for a tree to reabsorb leaf energy and nutrients before shedding them. If the tar-spot disease affects the tree late enough in the growing season, the energy cost of developing new leaves is larger than the energy returned from photosynthesis in the remaining part of the growing season, and thus a tree will often not foliate until next spring.

5. FEEDING STRATEGIES AND THE COSTS OF FOOD CAPTURE

5.1 Individual Strategies

5.1.1 Net-Building Filter-Feeding Caddis Flies

In ecosystems that are made up of flowing water, organisms can minimize energy investment in capturing food by constructing nets. Filter feeding can yield low energy costs for food capture, but there are several physical and biological constraints. These constraints include hydraulic forces, variability in the physical conditions, abundance of drifting food, quality of drifting food, filter net design, general

lack of net portability, and the ability and energy costs for an organism to maintain its position in a flow regime. Within these constraints are trade-offs when caddis flies exploit different habitats and types of food.

Often the most abundant and consistently available source of drifting food is in high flow-rate environments. The process of natural selection has produced several species of caddis flies that can utilize these environments. Adaptations include terminal hooks for securing the invertebrate to the substrate, silken safety cords to prevent drift when dislodged, and portable cases that act as ballasts in high water flow. Additional adaptations are found in net design because the fly must build a net that has a size, shape, and mesh size that functions most effectively at capturing the particle sizes available. As a result, filter feeding by caddis flies does not always yield high-energy returns. Net design is important in that hydraulics regulates the size of food that is transported in water column. As flow rate increases, the size of particles transported by water increases through increased kinetic energy. Larger food items are typically of higher quality (animal in origin) and are usually preferred by filterers, but the combination of large particles and high flow can destroy the net. Once energy has been invested in a net, the environmental conditions may leave a fly with a net that does not capture food effectively. This fly would then have to take an energetic loss and either abandon the habitat for better conditions, create a new net, or continue filtering with a net of marginal quality. Filter feeders have a large potential to minimize the cost of food capture, but the fly may not reach this potential in an environment that is temporally and spatially dynamic.

5.1.2 Filter-Feeding Mussels

Filter feeding using energy to pump water into and out of clams and mussels can also be a very profitable feeding strategy. To examine this, the basal metabolic rate of mussels can be estimated in the laboratory during a time when they are not feeding. Small plastic beads are then added to an aquarium container containing mussels; the mussels start to filter out the particles, increasing their oxygen consumption. Real food particles added to the chamber in place of plastic beads allow measurement of growth rate and food consumption. The quantity of food energy consumed per unit energy invested in feeding is then calculated by difference. The highest energy return in investment into feeding is estimated at about 500% (Table II).

5.1.3 Drift Feeding and Fishes

Some fish feed on suspended food in the water column. Alewives and other herrings have netlike or filter systems whereas other fish have an evolutionary modification in the buccal cavity that reduces energy cost in food capture. Energy costs for foraging by fish and by mussels and caddis flies are different in that fish have a much lower cost for powered locomotion compared to a mussel or caddis fly; this enables fish to adjust to environments with varying conditions.

Many fish have specialized buccal cavities that can literally suck up their food. This modification in the cavity increases capture success and reduces the amount of energy invested in movement of the fish to swallow prey. Sliding cartilaginous mouth structures underlie the mechanism that allows these fish to suck up their prey; the volume of the buccal cavity can be expanded when the mouth structures slide into position. When the cavity is expanded, the negative pressure created inside the mouth causes external water and food to flood inside the cavity.

A classic study by Kurt Fausch (of Colorado State University) on territoriality and foraging in salmonines shows how stream fishes use adaptations in the buccal cavity to minimize the cost of food capture. Fausch observed salmonines in a stream consume drifting food items. Stream hydraulics over space can be very complex. Rocks on the bottom generate shielding to water flow, creating pockets of slow-moving water that are in close proximity to water moving at a high rate of speed. A fish downstream of such a structure invests little energy to remain in position while having access to a nearby environment that has a good supply of drifting food. The mosaic of food availability and locations of such slow-water feeding habitats creates feeding locations of differing quality. Fausch found that fish defend these feeding zones, resulting in a hierarchy of fish occupying habitats of differing quality.

5.1.4 Territoriality Reduces Energy Cost of Food Capture

Territoriality within a species' population can maximize surplus energy yield from feeding according to what is called Allee's principal. At first, this principal appears to conflict with optimal foraging theory in that it would be seen that the increased cost for active metabolism in defending a territory would decrease surplus energy yielded from feeding. However, studies have shown that the energy cost in defending a territory is small in relation to the surplus energy gained by protecting the availability of resources. Additionally, Allee's principal says that

the population would gain from an established "pecking" order, so that energy is not wasted in conflict. If populations grew beyond available space and resources, the mean individual fitness would decrease. A large population of trout may have such a high demand for energy that the available resources are used mainly for metabolism, leaving little or no surplus energy for growth and reproduction. An optimal number of fish relative to their food supply is one that creates a maximum amount of surplus energy.

5.1.5 Phytoplankton Migration and Nutrient Limitation

Some types of phytoplankton, such as dinoflagellates, decrease the energy cost of capturing light energy by investing energy into movement. A stationary phytoplankton can experience a decrease availability of local concentrations of nutrients needed for photosynthesis and growth. Phytoplankton that invest energy into movement are able to exploit water with higher nutrient concentrations.

5.1.6 Olfaction and Predation

Having the ability to detect prey through chemoreception or olfaction increases the probability of food capture by giving the predator the element of surprise over its prey or by preparing the predator for an encounter. Thus, this strategy reduces the energy costs for finding prey.

5.1.7 Cottrell's Analogy: Humans, Machines, Fossil Fuels, and Agriculture

Machines using fossil fuels have enabled humans to pay the energy costs of food capture. The use of machines to exploit resources is not exclusive to humans. Cottrell gives an analogy between the human invention of the steam engine and the evolution of the water pump in filter feeders. Both the steam engine and the filter-feeder's water pump have increased the power to perform food-collecting activities, the result being an increase in populations of humans and filter feeders. The fundamental difference between the humans and filter feeders is that humans gained the ability to subsidize the costs of food capture using an external source of energy, and the filter feeders costs are subsidized by individual active metabolism.

Subsidizing the costs of food capture using fuels can lead foragers to foraging strategies that deviate from what optimal foraging theory would predict, because the energy cost for food capture is no longer important. Modern ocean fishing techniques

illustrates how humans have increased their power for capturing food using fossil fuels. The quantity of ocean fish humans harvest in a year is increasing while the energy cost for capturing fish increases even as fish abundances decline. Humans continue to invest more energy into fishing, even though the energy profitability of catching fish is declining.

Cuban offshore fishing industry and agriculture are clear examples of the level of human dependence on paying the energy costs of food capture with fossil fuels. Cuban offshore fishing collapsed when cheap fossil fuel supplied by the former Soviet Union was no longer available (Fig. 4) to run large commercial fishing boats. In addition, the production of important crops such as sugar cane and tobacco decreased because the country could not support the energy cost of running farming equipment.

As human population has increased, so has food production. Some attribute this increase in food production to large technical leaps in food harvesting efficiency (Table II), but most of the increase is due to elevated rates of production made possible by fossil fuel subsidies. In fact, data on the energy cost of food capture for various methods used in agriculture indicate that as farming practices evolved from shifting agriculture to traditional farming to modern agriculture, the rate of production increased but energy efficiency decreased (Table II). Earlier methods used for food capture, such as foraging and pastoralism, have much lower energy returned on investment compared to other methods and cannot be performed at high rates. Although shifting agriculture is one of the most profitable investments

of energy into food capture for humans, it is insufficient to feed all of the human population worldwide because yield per hectare, when averaged over the entire area, is very low. Modern humans have subsidized the energy costs of food capture to the point where the EROI is far less important than yield per unit time or area.

The foraging strategies described here are clearly a subset of the many others found in the world. For example, species have developed the use of sonar to find food (e.g., bats) to exploit food resources, and other species produce light to attract their prey (e.g., angler fish). The utility of studying energy costs in understanding feeding strategies is quite apparent.

5.2 Group Behaviors That Increase EROI in Food Capture

Foraging as a group of individuals can decrease the energy cost of food capture. Food may be located more effectively by groups than by individuals (e.g., dolphins, wolves), and food may be herded into locations or conditions that make capture much easier than otherwise, or the group can recover energies that are normally lost by an individual.

5.2.1 Dolphins and Whales: Herding Preys Using Bubbles

Herding of prey by a group of predators is not exclusive to humans. Whales and dolphins decrease the energy cost of food capture by herding their prey close to the water surface, where capture is easier. Small to medium-sized fish such as sardines school into tight groups of moving fish called “bait balls” when they detect a predator. Bait balls act as defense mechanism for many single-predator encounters because predators are typically much more effective at capturing individual fish. However, groups of cetaceans (whales, dolphins, etc.) take advantage of this defense mechanism by swimming under the bait ball, separating out individual fish from the outer perimeter of the ball and herding them up, toward the surface, where they are easy to capture.

5.2.2 Honeybees and the Waggle Dance

The time and energy needed to find food can be reduced if a population can focus its feeding on regions where food is abundant. Communication among members of a population allows an individual to instruct or lead other members of its population to a food source it has located. Honeybees have evolved communication skills to do just that. When a

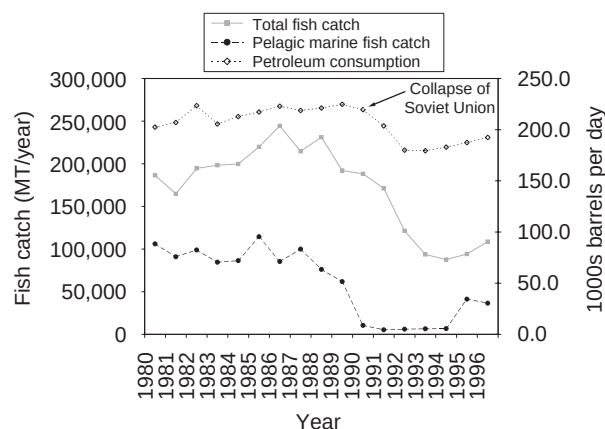


FIGURE 4 Fish catch in relation to quantity of oil consumption in Cuba. Cuban fisheries dependent on oil subsidies collapsed when the low-cost Soviet oil supply was no longer available. Compiled by J. Mead and P. Thompson (SUNY-ESF) using data from the Food and Agricultural Organization.

foraging honeybee finds a good source of nectar, it collects a full load and then returns to the hive. On arrival at the hive, the bee performs an elaborate “dance” and emits a series of sounds using its wings, informing the other bees in the hive where the food is located. The movements and sounds the bee communicates with its wings are known as the waggle dance. Researchers have found that the communication of the foraging bee is very detailed, and includes direction, distance, and elevation of the food source.

6. SUMMARY

Thinking about foraging species in terms of energetics, environmental gradients, and foraging theory helps to comprehend and synthesize the performance of species in the world around us. Many factors influence the energy cost of food capture, but to understand the foraging strategy of an organism requires integration of these factors. The world has a fascinating array of different living systems, all of which depend on energy. There are low-energy, high-energy efficiency ectothermic organisms and high-energy, low-energy efficiency endothermic organisms. Organisms vary greatly in body size and physiological and behavioral adaptations; when considered collectively, these adaptations enable a strategy for capturing food that ultimately attempts to further the persistence of a species’ traits. Based on the data in Table II, plants seem to be well adapted to capture a little more energy per unit energy invested in food capture compared to other nonhuman species. Also, humans have the ability to gain more energy from their energy investment in food capture compared to other species, but often sacrifice energy

efficiency for rate of production, following Odum’s maximum power concept.

SEE ALSO THE FOLLOWING ARTICLES

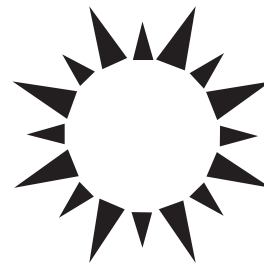
Conversion of Energy: People and Animals • Food System, Energy Use in • Heterotrophic Energy Flows • Human Energetics • Migration, Energy Costs of • Photosynthesis and Autotrophic Energy Flows • Reproduction, Energy Costs of • Thermoregulation

Further Reading

- Alexander, R. M. (1999). “Energy for Animal Life.” Oxford Univ. Press, Oxford.
- Colinvaux, P. (1986). “Ecology.” John Wiley & Sons, New York.
- Hall, C. A. S. (1995). “Maximum Power: The Ideas and Applications of H. T. Odum.” University Press of Colorado, Boulder.
- Keragossian, M., Ringler, N. H. (2004). Foraging ecology of fishes: In “Fishes and Forestry—Worldwide Watershed Interactions and Management.” Blackwell Publ., Oxford, in press.
- Nagy, K. A. (1987). Field metabolic rate and food requirement scaling in mammals and birds. *Ecol. Monogr.* 57(2), 11–128.
- Peters, R. (1983). “The Ecological Implications of Body Size.” Cambridge Univ. Press, Cambridge.
- Pough, F. H. (1980). The advantages of ectothermy for tetrapods. *Am. Natural.* 115, 92–112.
- Schmidt-Nielsen, K. (1997). “Animal Physiology: Adaptation and Environment.” Cambridge Univ. Press, Cambridge.
- Smil, V. (1991). “General Energetics: Energy in the Biosphere and Civilization.” John Wiley & Sons, New York.
- Stephens, D. W., and Krebs, J. R. (1986). “Foraging Theory.” Princeton Univ. Press, New Jersey.
- Tucker, V. A. (1975). The energetic cost of moving about. *Am. Sci.* 63(4), 413–419.

Cost–Benefit Analysis Applied to Energy

J. PETER CLINCH
University College Dublin
Dublin, Ireland



-
1. Definition of Cost–Benefit Analysis Applied to Energy
 2. History of Cost–Benefit Analysis
 3. Rationale for Cost–Benefit Analysis Applied to Energy
 4. Cost–Benefit Methodology
 5. Valuation of Energy Externalities
 6. Further Issues in the Application of Cost–Benefit Analysis to Energy

Glossary

- benefits transfer** A process of taking environmental valuation estimates from existing studies and applying them outside of the site context in which they were originally undertaken.
- discounting** Adjusting a cost or benefit in the future so that it may be compared to costs and benefits in the present.
- energy externality** A cost or benefit that arises from an energy project but is not borne or appropriated by those involved in the energy project (e.g., pollution from energy generation).
- environmental valuation** The process of placing monetary values on environmental effects.
- revealed preference methods** Environmental valuation methods that use data from actual behavior to derive values for environmental assets.
- sensitivity analysis** A presentation of various possible outcomes of a future project or policy under conditions of uncertainty.
- shadow pricing** The process of adjusting a distorted market price to better reflect the true opportunity cost to society of using that resource.
- stated preference methods** Direct environmental valuation methods that ask people in a survey to place a value on an environmental asset.
- total economic value** A categorization of the various elements of value of a resource, project, or policy.
- value of statistical life** The translation of the value of a small reduction in the risk of death into a value for a

project that would involve the equivalent of one life saved or lost.

welfare criterion The criterion in a cost–benefit analysis to assess whether a project or policy benefits society.

welfare measure A measurement tool to assess the strength of consumers' preferences for or against a project or policy in order to assess its costs and benefits.

Cost–benefit analysis applied to energy is the appraisal of all the costs and all the benefits of an energy project, policy, or activity, whether marketed or not, to whomsoever accruing, both present and future, insofar as possible in a common unit of account.

1. DEFINITION OF COST–BENEFIT ANALYSIS APPLIED TO ENERGY

Cost–benefit analysis (CBA), also known as benefit–cost analysis, is rooted in applied welfare economics. It is a way of organizing and analyzing data as an aid to thinking. It provides a set of procedures for comparing benefits and costs and is traditionally associated with government intervention and with the evaluation of government action and government projects. The underlying rationale for CBA is rational choice; that is, a rational agent will weigh the costs and benefits of any proposed activity and will only undertake the activity if the benefits exceed the costs.

A confusion that sometimes exists relates to the difference between social cost–benefit analysis and cost–benefit analysis. However, to an economist, social impacts and environmental impacts are economic impacts, so any analysis purporting to be a CBA must address social and environmental impacts.

In this way, CBA differs from financial analysis in that it assesses the costs and benefits to the whole of society rather than to a portion of society. For example, a financial analysis of a proposed new power plant by a private energy supplier will not take into account the costs imposed on wider society (such as environmental impacts), whereas a CBA, by definition, will do so. In addition, there are various other subtle but important differences between financial analysis and CBA related to the fact that market prices do not always reflect the true value of goods and services to society. This provides the basis for a technique known as shadow pricing, which adjusts distorted market prices such that they may be used to value costs and benefits appropriately.

As the previous definition implies, a CBA must include all the costs and all the benefits to whomsoever accruing. Thus, the definition instructs that whether the costs or benefits are marketed or not makes no difference. In other words, even if some of the costs or benefits are not exchanged in markets at all and so do not have a price, they may still be of value. This provides the basis for the inclusion of environmental and wider social impacts in CBAs of energy projects and policies. The common unit of account in CBA is money. Therefore, insofar as is possible and reasonable, all impacts, including environmental and social, should be evaluated in monetary terms so that they can be compared with one another. This provides the basis for a technique known as nonmarket or environmental valuation—that is, placing monetary values on goods and services not exchanged in markets, such as environmental improvements and damages.

In many cases, the costs and benefits of a project or policy will arise at different times. This is the case in relation to most energy activities, where the consequences of actions may not be known until far into the future, particularly in the case of environmental impacts such as global warming. This provides the basis for a technique known as discounting, which allows future costs and benefits to be compared with those in the present.

CBA has been used *inter alia* to estimate the environmental costs of power generation, the net social benefit of different power supply options, the net benefit of the use of renewable energy sources (e.g., solar, wind, and hydroelectric power), the net benefits of alternative modes of transport, the benefits of energy efficiency programs in the commercial and domestic sectors, as well as for policy analysis of energy efficiency regulations such as energy rating and labeling.

2. HISTORY OF COST-BENEFIT ANALYSIS

Detailed histories of CBA can be found elsewhere, but the essence is as follows. In 1808, U.S. Secretary of the Treasury Albert Gallatin gave what is believed to be the first recommendation for the employment of CBA in public decision making when he suggested that the costs and benefits of water-related projects should be compared. In 1936, the U.S. Flood Control Act deemed flood control projects desirable if “the benefits to whomsoever they may accrue are in excess of the costs.” In 1950, after a debate on the foundations of CBA in welfare economics, the U.S. Federal Inter-Agency River Basin Committee published a guide to CBA known as the “Green Book.” From then on, there was considerable interest in CBA and the development of cost-benefit rules, with seminal contributions by a number of authors, particularly during the 1970s. Although the distortions created by government intervention in the 1970s provided much room for the application of CBA, the technique went through a lull in the early 1980s when structural reform replaced piecemeal intervention. However, CBA has made a comeback since then, mainly as a result of increased public awareness of environmental degradation and the associated demand that this cost be accounted for in analyses of projects and policies. However, for a number of years, the economist’s view that environmental impacts could be assessed in monetary terms had little impact on policy, with many environmentalists having a particular problem with this concept. For example, in the U.S. Clean Air Act (1970), it was forbidden to use CBA, and the U.S. Clean Water Act (1972) set the objective of the elimination of all discharges into navigable waters. However, in the 1980s, it became a requirement that the U.S. Environmental Protection Agency perform CBAs on all environmental directives to assess their economic efficiency and significant resources were devoted to researching cost-benefit techniques and methods of placing monetary values on the environment. Although it was not until the Single European Act was introduced in 1987 that a legal basis for European Union (EU) environmental policy was provided, the nature of environmental policy reflected the origins of the EU being an economic one. EU policy that “the potential benefits and costs of action or lack of action must be taken into account in the preparation of policy on the environment” provides an obvious platform for the use of CBA.

3. RATIONALE FOR COST-BENEFIT ANALYSIS APPLIED TO ENERGY

The need for CBA applied to energy results from the existence of market failure. If there were no market failure and all energy suppliers were privately owned, there would be no need for CBAs of energy projects and policies. If market signals were appropriate, those activities and projects that are in the best interests of society would be implemented automatically and financial analyses by firms would simply replace any need for CBAs. However, the market does not work perfectly. Market failure necessitates CBA of energy projects, policies, and activities as a result of the existence of externalities, the associated need for environmental policy instruments, state ownership of energy generation and supply, assessment of sustainability, and “green” national accounting.

3.1 Energy Externalities

One of the most significant areas of environmental concern today relates to the environmental impacts of energy generation from fossil fuels, particularly acidification precursors, greenhouse gases, and localized air quality impacts. Much of these emissions result from power generation and the transport sector. These environmental impacts occur at local, national, regional, and global levels.

The environment can be defined as those parts of our physical and psychological endowment that we somehow share, which are “open access.” An economic analysis shows that it is precisely this shared nature of many environmental endowments that threatens their quality and character. In a market system, goods are allocated by the price mechanism. In a free market, the price of a good is determined by the demand for, and supply of, the good. Price is therefore a reflection of society’s willingness to pay for, or the valuation it places on, the good in question. However, the shared nature of many environmental assets, such as air, the atmosphere, and water resources, means that they are not owned and so do not have a price. When goods are seen as free they are overused. Thus, the market fails to protect environmental assets adequately.

This failure of the market can be best demonstrated with a simple example: If a power plant produces SO₂ emissions resulting in acid rain, which causes forest damage and damage to buildings, in a free market these costs will not be reflected in the cost of power generation. Thus, the price of the energy

will not reflect the true cost to society of generating it. In this case, more energy will be generated and more of the environmental asset depleted than is in the best interest of society. When a cost is imposed on those other than the person or business that produces the cost, this is known as an external cost. Of course, many energy projects will reduce such externalities, such as through renewable energy sources and improved heating and insulation systems.

When assessing the social efficiency of energy activities using CBA, it is necessary to include these externalities and other wider societal impacts. However, because by definition such costs and benefits are not traded in markets, there is no price to reflect their value. This provides the basis for the tools of environmental valuation.

3.2 Calibration of Environmental Policy Instruments

Unless some mechanism is found to internalize these externalities, they will cause a misuse of society’s scarce resources. This provides the basis for government intervention in the market. The notion of externalities was addressed by economists as far back as the 1930s. Pollution was seen as the consequence of an absence of prices for scarce environmental resources and so the prescription for solving the problem was to introduce a set of surrogate prices using environmental (Pigouvian) taxes and charges. A large literature has developed on these so-called market-based instruments. One of the more controversial concepts in environmental economics is that there is an optimal level of pollution. Pollution is usually the by-product of a useful activity, such as the production of goods and services; in other words, it is an input into the production process. Reducing pollution usually involves a cost in terms of output forgone and/or the cost of the technical process of abatement. Therefore, there is a trade-off between the environmental improvement and the costs of pollution reduction. The optimal level of pollution occurs where the benefit of another unit of abatement (the marginal benefit) equals the (marginal) cost, with the (marginal) cost being higher after that point. In the case of an extremely hazardous pollutant, the optimal level of pollution will be zero or close to it, but in most cases the optimal level of pollution will be positive. CBA provides the tools to assess these trade-offs. In reality, few economic instruments are actually based on such valuations. Research on so-called target setting using

environmental valuation is still in a preliminary stage. However, it could be used to assess the optimal level of an energy tax.

3.3 State Ownership of Energy Generation and Supply/State Support of Energy Efficiency Initiatives

In some countries, energy supply companies are owned by the state. In some cases, the original rationale for such ownership was that no private company would bear the large fixed costs of developing an energy supply system. In addition, it is sometimes suggested that state companies can perform as natural monopolies and thereby exploit economies of scale while producing at minimum average cost and subsidizing energy for certain sections of society. CBA has a role to play in assessing the merits of such state initiatives and is particularly relevant when assessing the merits of government intervention to improve energy efficiency, for example, by subsidizing wind power or a domestic energy efficiency program.

3.4 Sustainability Assessment and Green National Accounting

CBA is useful in assessing sustainability; that is, what is the value of our resource stock, how is it being depleted, and is man-made capital that is replacing it of equal value? Green accounting adjusts traditional measures of economic performance such as gross national product to reflect environmental damage and the depletion of resources. In relation to energy, CBA may be used to assess the social costs and benefits of depleting energy resources such as fossil fuels.

4. COST-BENEFIT METHODOLOGY

4.1 Welfare Criterion

In using CBA to assess the costs and benefits of a project or policy to society, it is first necessary to define the social efficiency criterion. The most frequently used criterion in microeconomic theory is Pareto optimality. Using this criterion, a project would be deemed to pass a CBA if it made at least one person better off without making anybody else worse off. For an investment to pass the Pareto optimality test, compensation must be paid to those bearing the costs so as to leave them indifferent

between their welfare *ex post* and *ex ante*. Most policies and investments leave somebody worse off and thus most would fail to meet the Pareto criterion. Frequently in CBA, the Kaldor-Hicks test is used. This test embodies the potential compensation principle. Kaldor categorized an action as efficient if those who gain from the action could compensate the losers and still be better off. Hicks framed the question slightly differently and labeled the action efficient if the potential losers from the action could compensate the potential beneficiaries of the action for not initiating the action. In both cases, the action is an improvement regardless of whether the compensation is actually paid. Thus, if the CBA shows the project to result in a welfare improvement under the Kaldor-Hicks test, this is essentially a potential Pareto improvement since there is the potential, in theory, to pay compensation to the losers such that nobody would be worse off. This will be the case as long as the benefits of the project outweigh the costs (i.e., the benefit:cost ratio is greater than 1).

Whether or not this potential payment is enough to show that a project is good for society has been the matter of much debate. Some authors have suggested that despite the shortcomings of the Kaldor-Hicks criterion, it is often useful to measure the effects of a change in the total value of output independently of the distribution of output; however, in judging the social efficiency of a project, equity aspects should also be considered. A two-part test may be used in which a project or policy is put to the Kaldor-Hicks test and then tested to determine if it improves (or at least does not disimprove) the distribution of income. Approaches that consider questions of equity therefore allow the possibility of acceptance of a project or policy that returns a negative figure for the sum of individual welfare changes if the gains from income distribution outweigh these losses. However, when considering equity the difficulty is in the assignment of weights to individual welfare changes. Consideration of equity issues and distributional consequences are particularly relevant in the case of projects that affect the price of energy and/or health. Energy price increases may be regressive unless poverty-proofing measures are put in place because energy is a necessity and therefore requires a larger proportion of the budget of less well-off people. On the other hand, improvements in domestic energy efficiency may bring significant health and comfort benefits to the elderly, who in some countries have been shown to be more likely to live in energy-inefficient homes and therefore more likely to be subject to winter morbidity and mortality.

4.2 Welfare Measures

In order to assess the costs and benefits of a project or policy, it is necessary to examine the strength of consumers' preferences for or against the project. The standard approach is to measure the benefit an individual obtains from a project or policy by the maximum willingness to pay on his or her part for such a benefit or the maximum willingness to accept compensation for not receiving such a benefit. Similarly, a cost can be measured by the maximum willingness to pay on the part of an individual to avoid such a cost or the maximum willingness to accept compensation for such a cost being imposed on the individual. Such welfare measures may be approximated using shadow pricing and/or direct valuation methods from surveys.

4.3 Temporal Considerations

Account must be taken of the fact that costs and benefits arise at different times. Thus, future costs and benefits must be discounted so that they can be compared with present costs and benefits. The rate of return for an investment can be strongly affected by the discount rate if there is a long time span involved. Therefore, the choice of discount rate is very important. In some estimates of damages from global warming resulting from the burning of fossil fuels, a small change in the assumed discount rate will have significant implications for climate change policy. However, despite the enormous literature on discounting, it is a difficult task to calculate the social rate of discounting. The discount rate is measured by some combination of a social rate of time preference and a measure of the opportunity cost of capital. Economists have been unable to provide a definitive figure for the social rate of discount, and so a rate tends to be chosen rather than measured.

In an application of CBA to any project or policy, a range of discount rates should be used. Often, for a government project, a published government test discount rate is used for making policy recommendations. These vary by jurisdiction. For example, UK treasury guidelines for the opportunity cost of public investment take the rate of return to marginal private sector investment as 6%. For pricing output sold commercially, public sector agencies are expected to achieve a required average rate of return of 8% to be consistent with the private sector. In the United States, recommended rates for investment projects vary. The Office of Management and Budget recommends a rate of 10%, whereas the General

Accounting Office prefers a rate of 2.5% based on government bond yields. The World Bank has traditionally used an 8% rate in project analysis.

4.4 Uncertainty and Risk

In *ex ante* CBAs, investments with long time horizons are subject to uncertainty, particularly in relation to future prices. In addition, when evaluating costs and benefits *ex ante*, there may be considerable uncertainty regarding the magnitude of various parameters. This is particularly the case in relation to potential environmental impacts and their valuation. In cases in which there is a high degree of uncertainty in relation to the coefficients being used to give the monetary value of outputs of a project, lower bound estimates are sometimes used because, if the results are to be the basis for the recommendation of a change in policy, it is thought by some authors that it is better for the benefits to be underestimated rather than overestimated. It is important to note that lower bound results do not necessarily imply the lowest possible results. Rather, they denote the lower limits of a range of reasonable estimates.

Sensitivity analysis, whereby results are calculated for various scenarios, should be used in all CBAs. In addition, confidence intervals should be presented where possible. A general rule is that the more results the better while giving guidance and reasons for the particular estimates chosen by the author(s) for the purposes of recommendations based on the CBA results. Risk assessment may be incorporated by including probability estimates for various scenarios.

4.5 Shadow Prices and Wages and Community Stability

Shadow prices must be used when there are distortions within an economy that render certain prices invalid as measures of the social opportunity costs and benefits of using resources in the project in question. In the real world, market prices may not reflect the social value of goods or services because of distortions in the market, such as monopoly power (which inflates price), indirect taxes and subsidies (unless they are correcting some other distortion), and unemployment. Thus, when calculating each value of a project, it is necessary to assess the appropriateness of using the market value and, where necessary, these values should be adjusted for market distortions. This can be a problem in energy projects for which a particular piece of technology may be supplied by one

company and/or estimates must be made regarding how price would decline should such a piece of technology be widely adopted, for example, in a national domestic energy efficiency retrofit.

The approach to valuing labor is worth detailing. In reports purporting to be CBAs, labor or “employment” is sometimes treated as a benefit rather than a cost. Labor is an input into a project and, therefore, a cost even though it may have some additional benefits such as community stability, as discussed later. However, that cost may be less than the market wage would suggest. If labor markets clear, the shadow price of labor will equal the market wage. Any increase in employment in one sector of the economy will merely reduce the availability of labor to another sector. This is likely to be the case when the jobs are highly skilled and/or there is a shortage of such skills. However, for a country with a high unemployment rate, it could be argued that an increase in the demand for labor in one sector of the economy will not necessarily displace a job in another sector; that is, there may be employment additionality whereby if the new job is filled by a person who was previously unemployed, a very low cost (not zero, unless it is considered that the unemployed’s spare time is totally unproductive) in terms of output forgone is imposed on society. In this case, the shadow price of labor would be close to zero. It has been the practice of some government CBAs to assume a shadow price of labor of zero. However, this is against the international practice of setting the shadow price of labor at most a fraction below the market wage. For example, in Canada, the shadow wage is usually 95% of the market wage, and in the United Kingdom they are usually considered equal. In many energy projects, specialized skills are required and so a shadow wage close to the market wage will usually be appropriate. However, in some household energy efficiency improvement schemes, long-term unemployed people have been trained how to fit various sorts of insulation, for example. In this example, a case could be made for using a lower shadow wage rate.

A related value that should be considered is whether an energy project may contribute to community stability. For example, would the provision of a new power plant in a rural area provide jobs such that depopulation of that area would halt? This value is not the value of the job per se. Rather, it is the added value of a job being in a particular place. These values are difficult to assess but are real. For example, it is often observed that employees will be willing to accept lower salaries to live in their hometown.

4.6 Valuing Changes in Energy Use

When calculating the benefits of reductions in energy use, the price of energy may be used to value such savings. However, it may be necessary to shadow price such savings. If the state supplies energy, an assessment must be made of whether there is any form of subsidy involved; if so, this should be removed. Where there is monopoly power in the energy market, either locally or, for example, through the monopolistic behavior of the Organization of Petroleum Exporting Countries, the price of energy is unlikely to reflect the true cost to society of its generation.

Energy supply is considered so essential that a premium may be placed on its supply from domestic sources if this reduces the risk of supply interruption. Most countries have assessed the importance of such supply, and this information may be used to assess any premium that should be applied. If the increase in supply that would be generated by the energy project under consideration is marginal to the total supply in a country, it is unlikely that such a premium would be significant.

In an *ex ante* analysis, a sensitivity analysis should be undertaken for alternative scenarios regarding the possible future changes in energy prices. These are difficult to predict because there are a number of offsetting effects. Possible deflationary pressures include the development of more energy-efficient machinery and technology, improved availability (and therefore reduced prices) of non-fossil fuel energy sources and systems, and any change in the regulation of the energy sector. Possible inflationary pressures include rapid economic growth and a consequent increased demand for energy, the exercising of market power in oil production, future scarcity of fossil fuels, and the introduction of carbon taxes and/or tradable permits in light of the Kyoto Protocol. However, if the price of energy is inflated by a carbon tax/emissions-trading system set to reflect the external costs of energy use, it is important to avoid summing external costs estimated by the techniques described later with the energy savings using this price because if the price already reflects the external effects, it would be inappropriate to include them a second time.

4.7 Externalities

Shadow prices adjust for one form of market failure. However, a CBA must also take into account the potential existence of externalities and public goods, as explained previously. The consideration of such

costs and benefits is extremely important in CBAs because all of the costs and all of the benefits to the whole of society are being considered. The difficulty is that these external costs and benefits are not exchanged within markets, so there is no price to reflect their value. For this reason, it is necessary to use nonmarket or environmental valuation approaches to place monetary values on these externalities.

5. VALUATION OF ENERGY EXTERNALITIES

Most energy projects have some impact on the environment that must be factored into a CBA. For example, a new coal-fired power plant will have implications for climate change, localized air pollution, and the production of acidification precursors. A cost-benefit analyst assessing the merits of a renewable energy project or an energy efficiency program will wish to factor in the value of reductions in such environmental effects.

5.1 Economic Value Concepts

A useful framework for assessing the effects of environmental impacts on humans is known as the total economic value approach. This expands the traditional concepts of economic value to include option value and existence value. In essence, an environmental asset is considered to have both use and nonuse value.

Use value can be broken down into actual use value and option value. Suppose we are estimating the value of damage caused to woodland from acid rain resulting from power generation. Actual use value is derived from the actual use of the asset and is composed of direct use value, such as timber from the woodland, and indirect use value, such as the value that would be yielded to inhabitants downstream should a forest reduce flooding.

Option value can be broken down into a number of components: the value individuals place on the preservation of an asset so that they may have the option of using it in the future (value of future options for the individual) for recreational purposes, for example; the value they attach to preserving the asset because it can be used by others (vicarious value); and the value they attach to preserving the asset so that future generations have the option of using it (bequest value). Thus, total use value is

defined as

$$\begin{aligned}\text{Total use value} = & \text{direct use value} \\ & + \text{indirect use value} \\ & + \text{option value,}\end{aligned}$$

where

$$\begin{aligned}\text{Option value} = & \text{value of future options for the} \\ & \text{individual} \\ & + \text{vicarious value} \\ & + \text{bequest value.}\end{aligned}$$

Nonuse values are existence or passive-use values. Existence values reflect the benefit to individuals of the existence of an environmental asset, although the individuals do not actually use the asset. These values may exist due to sympathy for animals and plants and/or some human approximation of the intrinsic value of nature (i.e., this value reflects a sense of stewardship on the part of humans toward the environment). Existence value is classed as a nonuse value such that total economic value (TEV) is defined as

$$\begin{aligned}\text{TEV} = & \text{actual use value} + \text{option value} \\ & + \text{existence value.}\end{aligned}$$

It is important to note that the TEV approach attempts to encapsulate all aspects of a resource that enter into human's utility functions (i.e., it is anthropocentric). Thus, it does not attempt to find an intrinsic value for the environment, where intrinsic means the value that resides in something that is unrelated to humans. However, it does attempt to include the intrinsic value that humans bestow on nature. There is no agreement on the exact breakdown of the values in the TEV framework; however, the important point in CBA is to account for all these values as best as possible.

5.2 Environmental Valuation Methods

There are two broad categories of nonmarket valuation methods: revealed preference methods and stated preference methods.

5.2.1 Revealed Preference Methods

Revealed preference studies use data from actual behavior to derive values for environmental assets. Suppose the value of a marketed good depends in part on an environmental good. These methods try to isolate the value of the environmental good from the

overall value of the marketed good. Such techniques include the following:

Hedonic pricing method: The value of a house depends, in part, on its surrounding environment. This method examines the determinants of the house's value and then isolates the effect that the environmental feature (e.g., a nearby park) has on the house. This is used as an estimate of the value of the environmental asset in question.

Travel cost method: The value of the park can be estimated by the willingness to pay to attend the park in terms of costs of travel and entrance to the park.

Production function approaches: These value a decrease (increase) in environmental quality as a result of some activity by the loss of (addition to) the value of production due to that activity. For example, the cost of acid rain resulting from pollution from fossil fuel-based power plants might be measured by the lost output of timber from forests (dose-response functions), the cost of restoring the damage to buildings (replacement cost), or the cost of defensive expenditures.

The production function approaches may seem more applicable to estimating externalities of energy generation, but the hedonic pricing method and travel cost method may be used to estimate the value of an amenity such as a forest park. This value could then be used to calculate the cost of damage to the park from air pollution.

5.2.2 Stated Preference Methods

Direct valuation methods ask people in a survey to place a value on the environmental asset in question by rating (contingent valuation), ranking (contingent ranking), or choosing trade-offs between various policy alternatives (choice procedures). These direct methods via questionnaires are the only approaches that can estimate existence value. The best known and most controversial of these methods is contingent valuation.

The contingent valuation method (CVM) collects information on preferences by asking households how much they are willing to pay for some change in the provision of an environmental good or the minimum compensation they would require if the change were not carried out. Since the mid- to late 1980s, this method has become the most widely used approach for valuing public goods and there has been an explosion in the number of CV studies.

The CVM approach is to choose a representative sample, choose an interview technique (face-to-face,

telephone, or mail), and then construct a questionnaire containing:

- A detailed description of the good being valued and the hypothetical circumstance in which it is made available to the respondent.
- Questions that elicit the respondents' willingness to pay for the good being valued: open ended (How much would you pay?), bidding (Will you pay x , if not, will you pay $x-y$), payment card (Which of these amounts would you choose?), and binary choice (If it costs x , would you pay it? Yes/no).
- Questions about the respondents' characteristics (e.g., age, education, and income), their preferences relevant to the good being valued (e.g., how concerned or interested are they in the environment), and their use of the good.

The mean or median willingness to pay from the sample of households surveyed can be multiplied up to the population level to give a total valuation.

5.3 Valuing Environmental Damage and Improvements from Energy Projects

The most significant environmental effects of energy projects that are included in CBAs tend to be the contribution of such projects to global warming from changes in CO₂ emissions and acid rain resulting from the production of NO_x and SO₂. In addition, many energy projects have implications for the production of particulate matter that may have consequences for human health and visibility.

The impacts of climate change are as follows: storm and flood damage (increased in some regions, such as the Caribbean and Pacific, but not others), increased mortality and morbidity, effects on small islands (sea levels rise due to heating water and melting ice caps), increasing salinity of estuaries, damage to freshwater aquifers, and effects on agriculture and forestry. It is estimated that a substantial portion of the existing forested area of the world will undergo major changes.

There are three approaches to placing a monetary value on the benefits of greenhouse gas reductions. The damage-avoided approach values 1 ton of carbon not emitted by the cost of the damage that would have been done by global warming in the event that it had been emitted. The offset approach measures the value of not emitting 1 ton of carbon using one method compared to the next cheapest alternative method. The avoided cost of compliance

approach measures the 1 ton of saved carbon by the avoided cost of compliance with a global/regional CO₂ emissions reduction agreement. A variety of estimates have been produced regarding the benefits of reducing CO₂, including studies by the United Nations Intergovernmental Panel for Climate Change.

Several major research projects have endeavored to place monetary values on all the environmental effects related to energy. Two EU-funded projects, one on the external costs of power generation (ExternE) and the other on European environmental priorities, are very useful sources of information on such values.

The ExternE project used an “impact pathway” methodology. Under this methodology, the fuel cycle, or life cycle of the fossil fuel used for generation, is first divided into stages, such as extraction, transport, combustion, and waste disposal. The environmental burdens imposed at each stage and the likely impact of these burdens on the environment and the public are isolated. The impacts are then valued using a variety of economic valuation methods, such as contingent valuation, direct cost of damage, or cost of mitigation. These costs are then normalized to give a cost per unit energy output and a value per ton of a pollutant reduced.

The “priorities” project endeavors to derive values for various environmental externalities suitable for use in a European context. The impacts of SO₂ on acute mortality, morbidity, materials, and the fertilization effect are all included; however, some impacts are omitted, such as forest damage, adverse effects on ecosystems and cultural heritage, and the dampening effect that SO₂ has on global warming. The main areas of uncertainty relate to the value of mortality. NO_x emissions are dealt with under two headings, acidification and tropospheric ozone. The fertilization effect of NO_x is examined under acidification. For damage to health and crops, the effects of NO_x on tropospheric ozone are used. Again, the bias is toward underestimation because damage caused to forests, biodiversity, non-crop vegetation, and materials is excluded.

Care should be taken when using unit damage values such as those in the studies mentioned previously. First, the confidence intervals are large (and should always be quoted in a CBA). In addition, there is confusion about which emissions cause which adverse effects. For this reason, it is very easy to double count when using estimates derived from different studies. For example, in some studies negative health effects are included in SO₂ damage

estimates, whereas in others the most significant health effects result from particulate matter emissions. Finally, the damage caused by 1 ton of most air pollutants (excluding CO₂) will vary depending on the geographical origin of the emission. However, most estimates are derived from an exercise known as benefits transfer.

5.4 Benefits Transfer

To reduce costs and where there is no alternative procedure, it is standard practice in CBAs to apply existing monetary valuation studies outside of the site context in which they were originally undertaken. This is known as benefits transfer. There are a number of rules when using this method, the most obvious being that the contexts of the studies should be very similar. In addition, appropriate statistical techniques should be used, including meta-analyses of best practice studies; the studies should have similar population characteristics; and property rights should be similarly distributed. The method is most accurate when the impacts are not affected by local characteristics (e.g., income and population size). The most comprehensive studies of energy externalities referred to previously involve significant benefits transfer.

5.5 Valuing Mortality and Morbidity

Placing monetary values on life and death is perhaps the most controversial area of economics. There are two major approaches for carrying out this task. The first values life directly through livelihood. This is known as the human capital method and is the oldest and most easily applied. It equates the value of life of an individual as the present value of future lost output of the individual, which is usually estimated using income. However, it is important to note the substantial drawbacks of this approach. Most obvious, it values livelihood rather than life per se. It is quite improbable that there would be a 1:1 relationship between individuals’ earnings and their demand for risk reduction. In other words, the human capital approach makes no reference to how humans value risks to their own lives. One of the most glaring flaws of this approach, however, is the fact that it places a zero value on the retired population and those who do not “produce” marketed output (e.g., those who work in the home).

The appropriate way of valuing reductions in mortality is not to value a life per se but rather to value people’s willingness to pay for a reduction in

the risk of mortality. These data derive primarily from one of three study formats: the wage risk model (also known as the hedonic wage model), the averted behavior method, and the contingent valuation method. Although space does not permit a detailed explanation of these methods, the general approach is to value a small reduction in risk and then translate this into the equivalent of a life saved. This is the value of a statistical life (VSL).

There has been controversy regarding this approach in relation to the estimation of the costs of power generation. The VSL estimates would be expected to be considerably lower for a developing country because the willingness of people to pay to reduce risk of death is partially dependent on income. Thus, the VSL used for a developing country (e.g., for deaths from coal mining) would be expected to be lower than the VSL used for a developed area (e.g., for deaths from air emissions in Europe by the burning of such coal). It is notable that the ExternE project steered clear of controversy by using the same VSL for each area.

The value of morbidity may be assessed using a cost of illness approach. In the case of air pollution from energy generation, this would be the sum of the cost, both public and private, of hospital stays and drugs for the treatment of associated illnesses. More positively, the health benefits of improving domestic energy efficiency would be the associated avoided costs of illnesses. Other costs avoided by wider society include the loss of productivity from those who are ill and unable to work. The avoided cost to the individual beyond drugs and hospitalization direct costs (if any) is more difficult to calculate. One approach is to estimate individuals' willingness to pay to avoid "restricted-activity days." The priorities study referred to previously contains such estimates.

6. FURTHER ISSUES IN THE APPLICATION OF COST-BENEFIT ANALYSIS TO ENERGY

When undertaking or evaluating a CBA of an energy project, there are a number of additional issues that must be considered. The tool is not without its controversies. These arise because of its use of money as the common unit of measurement. First, assessing welfare changes using willingness to pay and/or willingness to accept is perfectly valid in theory if one believes that income is distributed optimally. How-

ever, care must be taken when the population(s) whose welfare is being assessed has a high income variance. Poor people will be willing to pay less, all else being equal, simply because of their budget constraint. In addition, because the values and market prices used in a CBA of an energy project or policy are in part a product of the existing income distribution, if that distribution changes, prices will change, and so will market and nonmarket values. Second, distributional effects tend to be ignored in CBAs and yet energy policies often have very different consequences for different socioeconomic groups. Because energy is a necessity, energy efficiency schemes may benefit those on lower incomes more than average. However, without poverty-proofing measures, energy taxes will comprise a much higher proportion of their income than average. Third, not all support attempts to place monetary values on environmental assets, and the value of the statistical life concept is particularly controversial. Some find it unpalatable to consider that the environment might be traded with other goods and services let alone "lives." In addition, as in all public policy or human actions, the environmental values used are by definition anthropocentric.

Finally, the environmental valuation methods are limited in their ability to assess values. Many environmental outcomes assessed by natural scientists are subject to large confidence intervals even before economists endeavor to value them. Also, the valuation methods vary in their ability to assess environmental values. Economists have identified a number of criteria by which environmental valuation methods should be evaluated: validity (the degree to which it measures the total economic value of a change in the environment), reliability (the extent to which the variability in nonmarket values is due to random sources), applicability (assesses where the techniques are best applied), and cost (the practicality and viability of carrying out the valuation study). In general, revealed preference methods (e.g., productivity loss and replacement cost) are widely accepted valuation methods that provide valid and often reliable estimates of use values. However, they are limited in that they only provide estimates of use values, and it is often difficult to find suitable and reliable links between market goods and environmental amenities. The estimated values are often sensitive with respect to modeling assumptions, reducing their reliability in such cases. In general, however, revealed preference methods are applicable for valuation of policies that primarily affect use values and can provide very reliable results. Stated

preference methods such as contingent valuation are capable of measuring both use and nonuse (existence/passive-use) values. They are applicable to a larger array of environmental issues than are revealed preference methods, and they are often the only applicable methods for benefit estimation. However, they do have a number of limitations and may be subject to a number of biases.

There are some simple rules for using cost-benefit techniques. Simpler environmental valuation methods should be used where possible (e.g., production function approaches) while recognizing these are primarily of interest when the policy changes will mainly influence use values. A high degree of reliability in stated preference methods may be achieved by adhering to sound principles for survey and questionnaire design and if the method is applied to suitable environmental goods. The methods work best for familiar environmental goods and for salient environmental issues. The methods are likely to be least reliable where the environmental good being valued is complex (e.g., in the case of biodiversity).

Finally, in CBAs of energy initiatives, it is important to note that the estimates of values are made at the margin. The plausibility of the techniques described depends in part on the fact that most things remain the same, with the implicit assumption that there is no paradigm shift in values, power,

income, and the like during the time period of assessment.

SEE ALSO THE FOLLOWING ARTICLES

Decomposition Analysis Applied to Energy • Ecological Risk Assessment Applied to Energy Development • Energy Development on Public Land in the United States • Environmental Change and Energy • Equity and Distribution in Energy Policy • External Costs of Energy • Green Accounting and Energy • Value Theory and Energy

Further Reading

- Clinch, J. P., and Healy, J. D. (2001). Cost-benefit analysis of domestic energy efficiency. *Energy Policy* 29(2), 113–124.
- Commission of the European Communities—DG Environment (2003). European environmental priorities Web site: An integrated economic and environmental assessment. Available at europa.eu.int/comm/environment. ExternE Web site: <http://externe.jrc.es>.
- Hanley, N., and Spash, C. (1993). “Cost-Benefit Analysis and the Environment.” Elgar, Aldershot, UK.
- Layard, R., and Glaister, S. (eds.). “Cost-Benefit Analysis,” 2nd ed. Cambridge Univ. Press, Cambridge, UK.
- Zerbe, R., and Dively, D. (1994). “Benefit Cost Analysis in Theory and Practice.” HarperCollins, New York.

Food System, Energy Use in

CHRIS E. DUTILH

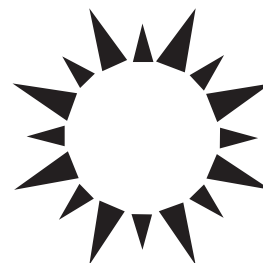
Unilever

Rotterdam, The Netherlands

ANITA R. LINNEMANN

Wageningen University

Wageningen, The Netherlands



-
1. Definition and Boundary of the Food System
 2. On-Farm Production
 3. Transport
 4. Processing, Packaging, and Storage
 5. Consumption
 6. Conclusions

Glossary

conversion efficiency The efficiency with which energy from one component (feed) is converted into another component (e.g., meat, milk) expressed as a kg kg^{-1} .

direct energy The energy requirement for a specific process, expressed as megajoules per kilogram (MJ kg^{-1} or 10^6J kg^{-1}).

embodied energy Sum of the energy requirements in all preceeding production steps required to generate a food product or intermediate.

energy requirement The energy input required in a certain step in the production of food.

indirect energy All energy requirements not considered as “direct”; for example, auto fuel is usually considered direct energy, whereas energy to build roads is considered indirect.

nutritional value A measure of the amount of energy that becomes available to the body after ingestion of a food product, expressed as kilojoules per kilogram (kJ kg^{-1} or 10^3J kg^{-1}); the nutritional value of a product is not related to the energy use in food production.

Energy use of food production systems starts with photosynthesis. Incoming sunlight is absorbed by chlorophyll in plant leaves. Water and carbon dioxide, which are required for the reaction, are taken up by the cell wall, and oxygen is emitted to the environment. The $(\text{CH}_2\text{O})_n$ (“sugar”) produced is

the building block that plants use to make carbohydrates, proteins, and fats. This conversion process requires chemical energy.

1. DEFINITION AND BOUNDARY OF THE FOOD SYSTEM

Energy use of food production systems starts with photosynthesis, that is, at the interception of sunlight by plants and the subsequent transformation of $\text{CO}_2 + \text{H}_2\text{O}$ into $(\text{CH}_2\text{O})_n$ and O_2 . The incoming sunlight is absorbed by chlorophyll in plant leaves. Water and carbon dioxide, which are required for the reaction, are taken up by the cell wall, and oxygen is emitted to the environment. The $(\text{CH}_2\text{O})_n$ (“sugar”) produced is the building block that plants use to make carbohydrates, proteins, and fats. This conversion process requires chemical energy, which is obtained by the respiration of some of the products formed during photosynthesis: $\text{CH}_2\text{O} + \text{O}_2 \rightarrow \text{CO}_2 + \text{H}_2\text{O} + \text{energy}$. For example, from 1 g glucose ($\text{C}_6\text{H}_{12}\text{O}_6$), a plant can produce 0.9 g other carbohydrates.

The transformation of sugars into other plant components is accompanied by more weight loss. For the production of amino acids, this loss depends on the source that supplies the nitrogen. For example, 1 g of glucose can be transformed into 0.7 g of amino acids when nitrogen is available as ammonia (NH_3). However, the usual source of nitrogen is NO_3^- , and in that case relatively much energy is needed for the reduction of nitrate into ammonia, according to the following pathway: $\text{NO}_3^- + 8 \text{H} \rightarrow \text{NH}_3 + 2\text{H}_2\text{O} + \text{OH}^-$. This pathway results in the production of 0.5 g of amino acids from 1 g of glucose. The production of fat by plants is most energy demanding. For example, 1 g of glucose is required to

generate 0.4 g of fat. The low conversion rate from sugars into fats on a weight basis is one of the reasons why the dry matter yield of oil-producing crops such as soybeans are significantly lower than the yields of starch-producing crops such as potatoes, rice, and grain, even though the lengths of their growing periods are not much different (Table I).

On average, plants use 1 g of glucose to produce 0.7 g of plant material that is made up of 25% protein, 55% cellulose, 6% fat, 6% sugars, and 8% minerals. Carbohydrates, proteins, and fats from plants can be used for human nutrition, but they may also be used as a feed for animals to produce eggs, dairy products, and meat.

Besides sunlight, various other sources and forms of energy are required during the subsequent stages of the food production system. That energy is directly needed to convert raw materials into new products or ingredients; therefore, it is called a direct energy requirement. Apart from energy input, there also is a loss of raw materials during each of the stages. Because raw material represents an amount of embedded energy, the material loss can also be described as an indirect energy requirement for the production stage. The total energy requirement for each stage is the sum of direct and indirect energy. For example, during the consumption stage at home, the direct energy is used to buy, store, and prepare the food. The indirect energy requirement at home is represented by the substantial amount of food (up to 15% of all food purchased) that is not consumed and ends up in the waste bin.

For a complete assessment of the embodied energy of food products, the entire life cycles of those products must be considered. These life cycles start at the farm, where the on-farm production takes place. Crops are sown, grown and harvested, whereas animals are bred, fed, and slaughtered. Food

products end their life cycles at the moment of consumption. Some food products are consumed after a short and simple life cycle, for example, an apple that comes from a consumer's own backyard garden. On the other hand, shops offer a large variety of food products that are the result of sometimes long and complex life cycles. Such food products may contain exotic ingredients or ingredients that have undergone various processing steps, for example, a pizza with a topping of tropical fruits and processed meat.

Irrespective of the complexity of the life cycles of food products, several general handling steps can be discerned. The first step, already mentioned, is on-farm production. Following the harvest, agricultural products are transported either directly to a shop or to a factory for further processing (e.g., washing, cutting, cooking, mixing, packaging). Processed products are stored (either chilled or at ambient temperatures) and subsequently are transported to a retailer, where they are offered for sale. After purchasing the food, a consumer will carry the products home from the shop. At home, the products may be stored for a while, after which time they will be prepared for consumption. Alternatively the consumer can go to a restaurant to eat the food there. In that case, the restaurant owner/employees will have bought, stored, and prepared the food. Each of these steps, for which various alternatives exist, requires energy. In the end, the sum of all energy requirements in each step determines the total embodied energy of a certain food product. In what follows, this article looks in some detail at each step in the production chain and indicates potential differences in energy requirements between eating out and eating at home.

2. ON-FARM PRODUCTION

The embodied energy of food products (as distinct from nutritional or caloric value) differs among various food categories, and substantial differences also exist within a specific food category. Table II presents average embodied energy for some fresh ingredients from plant and animal origin under Dutch farm production circumstances. Crops grown in heated glasshouses, as is common in several northern countries, require substantial higher amounts of energy. That increase may be up to 40 MJ kg⁻¹. In some regions, glasshouses (or plastic equivalents) are used without heating just to make sure that heat from the sun is being collected

TABLE I
Yield per Hectare for Various Industrial Food Crops

	Yield (tonnes/ha)	Dry matter content (percentage)
Tomatoes	40–75 ^a	5
Potatoes	30–70 ^a	25
Vegetables	1–90 ^b	5–20 ^b
Grain	3–10 ^b	95
Oil-producing crops(e.g., soybeans)	1–4	98

^a Depending on agricultural conditions.

^b Depending on the type of crop.

during the day and not lost too easily during the night. That practice saves energy and enhances the growth rate.

To raise the yield per hectare, fertilizers can be used. The production of fertilizers requires approximately 10 to 25 MJ kg⁻¹. The application of fertilizers may add up to 5 MJ kg⁻¹ to the total energy requirement of the on-farm stage. However, in most cases, it is substantially lower than that. For crop protection, various chemical compounds are used (e.g., herbicides, pesticides, fungicides). The production of these chemicals requires between 100 and 1000 MJ kg⁻¹, and the applications vary substantially. The energy requirement for plant protection may rise up to the same level as for fertilizers. However, that figure can vary from year to year and from crop to crop because the application of crop

protection aids depends very much on weather and local circumstances.

The embodied energy for products from animal origin is approximately a factor 10 higher than that for products from plant origin (Table II). Within the category of products from animal origin, fish-meat has a relatively low embodied content. The low energy requirement can be explained by the fact that fish are cold-blooded and, therefore, require less energy to produce body warmth (thermoregulation). There are two main sources of fish: sea fish and fish from aquaculture. The overall energy requirements for the production of these two do not differ much, but the buildups are rather different. For sea fish, a lot of energy is required for the ships at sea—between 2 and 50 MJ kg⁻¹ depending on the type of fish. There are many variables that make an accurate calculation of the energy requirement for fishing very complex (e.g., the amount of by-catch, the efficiency of fish filleting, the use of by-products). In aquaculture, the fish are fed with fish meal, which originates from dry sea fish residues.

Feed-to-meat conversion ratios partly explain the differences in energy requirements between products from plant and animal origin. Table III shows the conversion efficiencies for chicken, pork, and beef based on long-term records of feeding efficiency ratios by the U.S. Department of Agriculture (USDA). Cumulative feed/gain ratios for chicken are between 1.8 and 2.0 for chickens in the most common 2.0- to 2.5-kg range. That ratio is raised by at least 10% to account for the feed requirements of breeder hens and cockerels that do not directly contribute to the meat production, as well as for the fact that some animals die prematurely. Ratios between 2.0 and 2.2 represent the standard of recent good performance. Typical rates from long-term statistics are somewhat lower (Table III).

TABLE II

Embodied Energy for Some Fresh Foods from Plant and Animal Origin Produced in The Netherlands

Product	Embodied energy (MJ kg ⁻¹)
Plant origin	
Potatoes	1
Vegetables	1–4
Fruit	2–5
Pulses	6–10
Animal origin	
Milk	10
Eggs	20
Fish (caught at sea)	20–40
Butter and cheese	50–60
Meat (chicken, pork, beef)	30–70

TABLE III

Typical Efficiencies for Meat Production

	Chicken	Pork	Beef
Feed (kg kg ⁻¹ live weight)	2.5	3.5	3–5 ^a
Edible weight (percentage of live weight)	55	55	40
Feed (kg kg ⁻¹ edible weight)	4.5	6.4	7.5–12.5
Nutritional value (kcal kg ⁻¹)	1800	1700–3300	1200–3300 ^b
Average energy conversion efficiency ^c	11.5	11.2	6.0–9.5

^a Depending on the inclusion of milk production.

^b Depending on fat content of the meat (average nutritional value 2500 kcal kg⁻¹).

^c Percentage of the embodied energy in feed that is retained in the embodied energy of meat, assuming the average embodied energy for cattle feed of 3500 kcal kg⁻¹.

Pork is the least feed-intensive red meat due to pigs' inherently low basal metabolism. The overall feed/gain ratio for North American pigs from weaning to slaughter ranges between 2.5 and 3.5 with *ad libitum* intake of feed. In The Netherlands, on average 320 kg of feed is required for 115 kg of live weight pork. The addition of feeding costs of the breeding stock and adjustments (e.g., for premature mortality) can raise the overall feed/gain ratio significantly. From the early 1900s onward, the nationwide feed/gain ratio for pigs has declined only slightly, in spite of the substantially improved regimes, due to the selection of less lardy pigs. The efficiency of metabolizable energy conversion in pigs peaks at approximately 45%, whereas conversion to fat can be as efficient as 75%.

Cattle are the least efficient converters of feed to meat. Their basal metabolism is higher than that in pigs, and their large body mass and long gestation and lactation periods mean that feed requirements of breeding cattle claim at least 50% more energy than those of pigs and nearly three times as much as those of chickens. The USDA's historic feed/meat data for all of the nation's cattle and calves imply that as little as 3% of gross energy in feed is converted into energy in food.

From the energy makeup point of view, there is a rather poor energy conversion efficiency for meat products, although the situation is not as bad as the conversion efficiency factors in Table III may suggest. A large part of the cattle feed consists of materials that are left over from the human food chain, and cows can convert plant material that humans cannot digest (e.g., grass) into highly valuable food components such as milk and meat. However, we can certainly say that eating animal protein requires significantly more energy and requires more land than does eating vegetable food.

3. TRANSPORT

Crops that are not cultivated under controlled environmental conditions, such as those in heated glasshouses, can be grown only in a certain region during a limited period of the year. However, consumer demand for food is not restricted to these periods. This discrepancy can be overcome in two ways: conservation of the harvested produce and import of products from other regions. The latter option requires energy for transportation, which varies largely with the means and distances of transport (Table IV). As a consequence, major

TABLE IV
Energy Requirement for Transportation

Means of transport	Energy requirement ^a (MJ kg ⁻¹ 1000 km ⁻¹)
Seagoing vessel	0.1–0.2
River vessel	0.4–0.7
Train	0.8–1.0
Truck	2–5
Airplane	10–25
Car	50–250 ^b

^aThe energy requirement depends on size and load factor.

^bDepending on petroleum use as well as the amount of products carried (20–40 kg).

TABLE V
Energy Requirements for Some Industrial Preservation Techniques

Preservation technique	Energy requirement (MJ kg ⁻¹ end product)
Drying ^a	0–15
Freezing	5
Heat treatment ^b	5–10

Source. 1996 data from Unilever operations.

^aThe energy requirement ranges from nearly zero for sun drying to approximately 15 MJ kg⁻¹ for hotair drying.

^bBlanching, pasteurization, sterilization.

differences in the energy requirements will exist for products that are transported over a long distance: In general, the energy requirement for transport by airplanes is very high, whereas the energy requirement for transport by vessels is relatively low.

4. PROCESSING, PACKAGING, AND STORAGE

Many agricultural products undergo some industrial processing before they reach the consumer. Advantages of processed foods over fresh agricultural products include convenience and prolongation of shelf life. In general, industrial food production includes cleaning, mixing, preservation, and packaging operations. Preservation can be achieved by drying, freezing, or heat treatment (Table V). All processing steps require energy, but mechanical handling (interfactory transportation, mixing, and shaping) generally constitutes only 10 to 20% of the total energy requirement. The bulk of the energy

requirement in food processing is caused by heating and cooling required for food preservation.

Several options exist for the packaging of food products. Some food products may require specific packaging materials not only to safeguard food integrity and food safety but also for marketing purposes. Table VI presents data on the energy requirement for the production of the most commonly used packaging materials. When interpreting these data, one should take into account that the packaging for a particular product will require different amounts (weights) of the various packaging materials, thereby complicating a direct comparison of the data in the table. Moreover, the overall energy requirement related to the use of packaging depends on the waste treatment in a particular country due to the energy recovery potential during waste incineration. In case all waste is incinerated with energy recovery, the total energy requirement is reduced as shown in Table VI. Ranges for the overall embodied energy for the ready-to-eat meals, preserved and packed in different ways, are presented in Table VII.

Ready-to-eat meals are prepared in the factory. This preparation requires relatively little energy per portion because the production is large scale. The final product needs to be stored and transported at low temperatures; hence, it requires more energy. Preparation of food at home usually starts with fresh ingredients. In the chain from farm to home, there usually is a significant amount of product loss that

TABLE VI

Energy Requirements for the Production of Some Packaging Materials

Material	Energy requirement (MJ kg ⁻¹)	Recovery potential (MJ kg ⁻¹)
Glass		—
Virgin	13	—
Recycled (100%)	10	—
Steel		—
Virgin	35	—
Recycled (100%)	20	—
Paper/Cardboard		—
Virgin	45	18
Recycled (75%)	35	18
Plastic	85	45
Aluminum		—
Virgin	195	—
Recycled (50%)	100	—

TABLE VII

Embodied Energy for Some Ready-to-Eat Meals

Type of ready-to-eat meal	Embodied energy (MJ kg ⁻¹)
Cooled, plastic packaging	45–70
Ambient, canned	75–110
Frozen, cardboard packaging	80–115

TABLE VIII

Energy Requirements for Cooled and Frozen Storage

Location	Energy requirement (MJ kg ⁻¹ week ⁻¹)
Warehouse	0.01
At home ^a	2–5
Shop ^a	1–10

Source. 1996–1997 data from Unilever analyses.

^aThe wide ranges are due to the variety of cooling and freezing cabinets, ranging from well-isolated facilities to small open boxes.

does not occur with factory production. In restaurants, the energy requirement and product losses usually are higher than for the two other routes. But in general, the total energy requirements (including the embodied energy of losses in the chain) for food prepared at home and ready-to-eat meals do not differ a great deal.

Finished products are finally stored and distributed to a point of sale. This requires energy for transportation (Table IV). In addition, some food products have to be kept cool or frozen. Average energy requirements for cooling and freezing vary substantially (Table VIII). The most efficient is storage in a well-isolated warehouse, whereas storage in a small open freezing cabinet is very energy demanding. The latter option is often used in shops to stimulate consumers to buy such products. In those cases, priority is given to easy access and not to efficient energy use.

There seems to be a direct relationship between embodied energy and product cost, particularly for commodity food products, as indicated in Fig. 1. Only a few products are situated above the line, representing those products produced in greenhouses with subsidized energy. There are many products below the line, mainly branded and seasonal food products. In those cases, the price that is being paid in the market is higher than the energy equivalent because there also is an emotional value to the product.

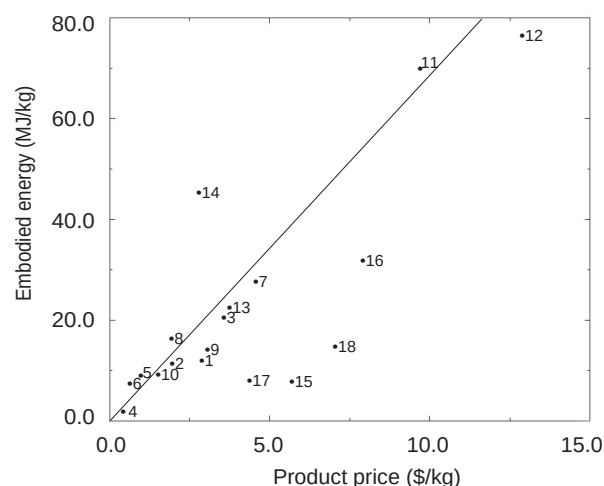


FIGURE 1 Relationship between embodied energy and product price for the following food products: (1) whole-meal bread, (2) wheat flour, (3) rice, (4) potatoes, (5) carrots, (6) onions, (7) jam, (8) sugar, (9) margarine, (10) milk, (11) butter, (12) cheese, (13) eggs, (14) greenhouse tomatoes, (15) beer, (16) wine, (17) fresh vegetables, (18) strawberries.

5. CONSUMPTION

After purchasing the food, a consumer will add to the overall embodied energy of the product by using a car to take the product home, by storing it in a refrigerator, by cooking it, by washing the dishes, and (ultimately) by disposing of the nonconsumed waste. In most countries in the Western world, on average, shopping adds 0.5 to 3.0 MJ kg⁻¹ because a substantial part of all shopping is done by car. Nearly 50% of all Dutch consumers use a car to do their supermarket shopping at a distance of approximately 3 km from their homes, buying on average approximately 20 kg of products each time. Dutch consumer habits seem to be comparable to those in other

countries, although the distances to the supermarkets are longer in some countries (e.g., France, the United States) and the proportion of motorized shopping differs to some extent.

In-house, some food products can be kept at ambient temperature, whereas others require energy for cooled or frozen storage (Table VIII). The amount of energy for storage depends on the energy efficiency of the refrigerator and/or freezer, its storage capacity, the temperature installed, the amount of product stored, and the frequency of opening. The storage capacity that is required for a household is related to the frequency of shopping. This relation is demonstrated in Table IX, where the energy requirement for four different situations has been calculated. In each case, the same amount of food is purchased per week and the shopping trips are done by car. Cars use petroleum (46 MJ/kg) at an average rate of 6.5 L (4.8 kg or 215 MJ) per 100 km. The distance to the shop has been varied (2 or 5 km), as has the shopping frequency (two or three times/week). Our calculations demonstrate that for small distances between house and shop, the total energy requirement for shopping and storage is not very much influenced by the shopping frequency. At longer distances, the energy requirement for transporting the goods is so high that it pays to have a large refrigerator to store the goods.

Ready-to-eat meals are prepared in a factory. This preparation requires relatively little energy per portion because the production is large scale. In many cases, the final product needs to be stored and transported at low temperatures; hence, it requires more energy. Preparation of food at home often starts with fresh ingredients. In the chain from farm to home, there usually is a significant amount of product loss due to material deterioration and processing inefficiencies. Those losses are usually substantially lower in the case of factory production

TABLE IX

Energy Requirements for Shopping and Storage of Food at Different Shopping Distances and Frequencies

Shopping distance (km)	Shopping frequency (week ⁻¹)	Energy requirement for shopping (MJ year ⁻¹)	Storage capacity (L)	Energy requirement for storage (MJ year ⁻¹)	Total energy requirement for shopping + storage (MJ year ⁻¹)
2	2	895	235 ^a	1235	2130
2	3	1341	155 ^b	1020	2361
5	2	2236	235 ^a	1235	3471
5	3	3354	155 ^b	1020	4374

^a As 210 L for cooled storage and 25 L for frozen storage; energy requirement 0.4 kWh per day per 100 L.

^b As 140 L for cooled storage and 15 L for frozen storage; energy requirement 0.5 kWh per day per 100 L.

TABLE X
Energy Requirements for Cooking

Situation	Energy requirement (MJ kg ⁻¹)
Home	
Homemade	5–7
Ready-to-eat	1–2
Restaurant	10–15

due to better controlled logistics. In some restaurants, the energy requirement and product losses are higher than those for eating at home, but that depends very much on the number of customers and the amount of food that has to be discarded. In general, the total embodied energies (including the embodied energy of losses in the chain) for food prepared at home and ready-to-eat meals do not differ a great deal (Table X).

The amount of energy for cooking depends on the method of preparation, the energy source (with gas being more efficient than electricity), the energy efficiency of the cooking device, and the type of recipe. A major additional factor is the number of portions being prepared at the time. In general, the amount of energy required per portion diminishes at higher numbers. For very small numbers, cooking by microwave is to be preferred; for larger numbers, cooking with gas turns out to be most energy efficient.

After a meal, the dishes have to be washed. The energy requirements for this activity depend on the energy needed to heat the water and on whether dishwashing is done manually or with a dishwasher. Dishwashers vary in their energy efficiency. Households that use electrically heated water may save on total energy requirements by using an energy-efficient dishwasher (Table XI). This is due to the fact that the production of electricity as an energy source has an efficiency of only 40%. The data in Table XI are based on 17 L as the average amount of water used for manual dishwashing. The energy used to produce the dishwasher itself is negligible in comparison with the amount of energy that is needed for the actual dishwashing during the life span of the machine (10–15 years).

Finally, at the end of the food chain, consumers dispose of inorganic waste such as packaging materials and organic waste (e.g., leftovers, parts of the foods that were removed during the cleaning and

TABLE XI
Energy Requirements for Dishwashing in an Average Dutch Household (2–3 Persons)

Method	Energy requirement per event (MJ)	Daily energy requirement (MJ)
Manual dishwashing		
Water heated by gas	1.0	1.0
Water heated by electricity	2.5–3.0	2.5–3.0
Dishwasher	4.0–5.5	2.5–3.5

cutting of the ingredients of the meal). On average, some 10 to 15% of the total energy makeup of the foods that are purchased by consumers are not used for consumption but rather discarded. This loss equals an amount of approximately 1000 MJ/year/person.

6. CONCLUSIONS

For all food categories, there is a wide variety of energy requirements over the production chain that are due predominantly to the following:

- *Time and place of consumption in relation to the period and region of production.* Foods that need to be stored for a long period and/or transported over long distances inevitably have higher energy requirements than do foods that are consumed during the season of production close to the area where they originate. However, in some (warm) regions, food can be cultivated much more efficiently than in others, so production of food away from home does not necessarily always lead to increased energy requirements.
- *Production method.* Cultivation under controlled environmental conditions in glasshouses requires additional energy for heating.
- *Degree of processing.* All preservation steps inevitably add to the energy requirements.
- *Consumer food preference.* Consumers can opt for protein sources with high overall energy requirements, such as meat and meat products, or protein from plants with much lower energy requirements.
- *Consumer behavior.* Inadequate shopping, storage, and preparation of food products may add substantially to the energy use of the food system.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Conversion of Energy: People and Animals • Diet, Energy, and Greenhouse Gas Emissions • Fisheries and Energy Use • Food Capture, Energy Costs of • Human Energetics • Industrial Agriculture, Energy Flows in • Lifestyles and Energy • Livestock Production and Energy Use

Further Reading

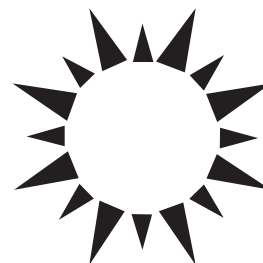
- Bouwman, M. (2000). "Tracking Transport Systems: An Environmental Perspective on Passenger Transport Modes." Ph.D. thesis, University of Groningen, The Netherlands.
- Dutilh, C. E., and Kramer, K. J. (2000). Energy consumption in the food chain. *Ambio* 29(2), 98–101.
- European Commission. (1997). "Harmonization of Environmental Life Cycle Assessment for Agriculture," final report, concerted action AIR3-CT94-2028, appendix 3. European Commission, Brussels, Belgium.
- Hendrickson, J. A. (1997.). Energy use in the U.S. food system: A summary of existing research and analysis. *Sustainable Farming* 7(4).
- Kramer, K. J. (2000). "Food Matters: On Reducing Energy Use and Greenhouse Gas Emissions from Household Food Consumption." Ph.D. thesis, University of Groningen, The Netherlands.
- Smil, V. (2002). Worldwide transformation of diets, burdens of meat production, and opportunities for novel food proteins. *Enzyme Microbial Technol.* 30, 305–311.

Forest Products and Energy

BRYNHILDUR DAVIDSDOTTIR

Boston University

Boston, Massachusetts, United States



-
1. The Forest Products Industry: Organization and Main Processes
 2. Forest Products and Energy Consumption
 3. Forest Products and Energy Generation
 4. The Future: Energy Conservation and Self-Generation Potentials
 5. Conclusion

Glossary

black liquor Spent pulping liquor.

charcoal A black porous solid wood fuel that consists mainly of elemental carbon; produced from heating wood in a vessel devoid of oxygen.

cogeneration A process that produces both thermal and electrical energy.

disembodied technological change Technological change that does not require substantial capital investments; influences only the input efficiency of already installed vintages.

embodied technological change Technological change that requires substantial capital investment; influences only the input efficiency of the most recent vintage of the capital stock.

hogged fuels Forest residues from extracting industries such as woodchips that are used to produce energy.

pulpwood Forest products that are used to produce pulp such as woodchips and logs.

wood fuels Forest products that are used for energy generation; include black liquor, fuelwood, charcoal, and residues.

wood pulp Pulp that is produced from virgin forest fibers.

Wood fuels, or energy derived from forest products, are an important contributor to renewable energy supply worldwide. In developing countries, the household sector is the largest consumer of wood fuels. In many developing countries, total consumption of wood fuels is increasing but with a declining fractional share of wood fuels in total national

energy supply. Of course, substantial regional and subregional differences exist. In the developed world, industrial consumption of wood fuels has mostly replaced traditional household wood fuel use. The pulp and paper industry is largely responsible for this development through its use of chemical pulping processes. The pulp and paper industry is one of the most energy-intensive manufacturing industries. Over the past half-century, it has gone from primarily purchasing the process energy required to produce pulp and paper toward self-generation from wood fuels often more than half of its energy needs. Now in the beginning of the 21st century, some paper companies are even net energy exporters, primarily due to cogeneration capabilities. Continued transition to self-generation of energy contains important implications for material use in the industry and is influenced by economic, organizational, and technological issues. Consequently, an in-depth analysis of the pulp and paper industry is warranted to understand the prospects for a continued increase in the share of energy supplied from forest products in developed countries.

1. THE FOREST PRODUCTS INDUSTRY: ORGANIZATION AND MAIN PROCESSES

The forest products industry produces two major classes of products: (1) lumber and wood products and (2) paper and allied products. The lumber and wood products sector produces whole logs, lumber, woodchips, hardwood flooring, furniture parts, and the like. It operates primarily on purchased energy, and total energy consumption in this sector is relatively low. In the United States, it accounts for 16% of the total energy consumed by the forest products industry as a whole. Thus, this article

focuses on the other half of the forest products industry—the paper and allied products industry.

The paper and allied products industry, or the pulp and paper industry, produces products such as pulp, newsprint, printing and writing papers, and paperboard. The United States currently produces 38% of all pulp produced in the world, followed by Canada, which produces 16%. Correspondingly, the United States produces 36% of all paper and paperboard produced in the world, followed by Japan, which produces 10%. The industry accounts for 0.7% of gross domestic product (GDP) in the United States, but in Finland and Sweden the industry contributes approximately 3.4 and 4.5%, respectively.

The organizational structure of the industry can be divided into three broad sectors: pulp mills, paper and paperboard mills, and finishing. The first two sectors account for more than 97% of all energy use and 100% of all self-generation of energy. In the United States, the industry is mostly competitive and is vertically integrated both upstream and downstream. For instance, it is common for paper companies to own the timber resources they use as a source of raw materials; for example, the three largest paper companies own more than 47% of all privately held timberland. In the United States, it is also common for paper and pulp mills to be located at the same site—what is called an “integrated mill.” The capacity of the mills that are integrated tends to be much larger than that of the mills that are not. On average, an integrated mill has a capacity of 1000 tons/day, whereas a paper mill that is not integrated has a capacity of 200 tons/day.

Internationally, capacity utilization rates are extremely high and have been over 90% in the United States for the past quarter-century. Capacity utilization rates are defined as total annual production divided by total annual production capacity. The reasons for those historically high capacity utilization rates are the immense capital intensity of the industry and low profit margins. To remain profitable, the industry must run its machinery very close to full capacity. Because it is too expensive to shut down the equipment or let it lay idle, it is difficult to retrofit already installed machinery, for example, to improve energy efficiency. Consequently, most capital that is installed is used for a long time. This manifests itself in a slow rate of capital turnover and implies that technological improvements occur only incrementally. New improvements tend to be those that can be retrofitted without major fixed-capital expenditures and extensive downtime or those that

are considered cobenefits to process or product improvements. This has significant implications for energy use and energy conservation in the industry. The production of paper and paperboard closely follows the organizational structure of the industry and is discussed next.

1.1 Pulp and Paper Manufacture

The two main processes in pulp and paper manufacturing are pulp and paper/paperboard production.

1.1.1 Pulping

Pulping is the process where fibers are separated and treated to produce pulp. The wet pulp is then converted into paper at an integrated pulp and paper mill or is dried and transported from the pulp mill to a paper mill. Different pulping processes are used depending on the fiber material and the desired end product. Fiber materials consist mostly of virgin or waste fibers. Virgin fibers are derived from pulpwood such as from woodchips or logs. Virgin fibers are used to produce a variety of pulps, with the main categories broadly characterized as chemical pulp (where the main process is called kraft pulping), semi-chemical pulp, and mechanical pulp. Pulp mills may also produce pulp from plant materials other than wood, collectively known as nonwoods. Nonwood fiber sources include agricultural residues and industrial fiber crops such as hemp. Wastepaper pulping took off during the past decade. For instance, in Denmark and the United States, waste fibers account for close to 100% and 38% of total fiber use, respectively (Fig. 1).

Pulp mills can use only virgin fibers, waste fibers, or a mix of both. When producing pulp from virgin

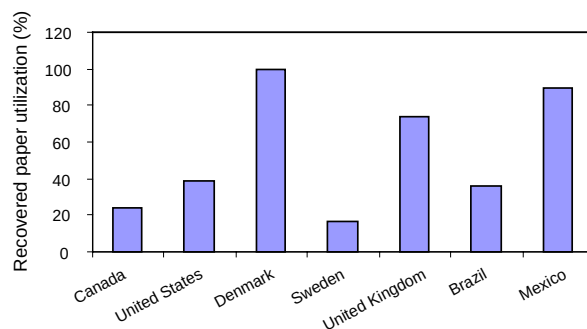


FIGURE 1 Use of wastepaper in selected countries in 2001. Recovered paper utilization rate is defined as the amount of wastepaper used in the production of paper and paperboard divided by the total amount of paper and paperboard produced. From the Food and Agricultural Organization, www.fao.org/forestry.

fibers, the raw materials are first prepared, for example, by debarking and chipping. The next step is the actual pulping process. The primary purpose of pulping is to free the fibers from the lignin that binds the fibers together and/or to separate the fibers in water. Pulp with longer fibers and less lignin is considered to be of highest quality and produces the strongest paper. Mechanical pulping is used for lower grade paper such as newsprint and is the process used for recycled fibers. The advantage of this process is that it gives much higher pulp yields (90–95%) than do other processes but at the same time produces weaker paper because the method yields shorter fibers. Also, if virgin fibers are mechanically pulped, the paper ages faster due to the presence of lignin. Mechanical pulping of virgin fibers is energy intensive and requires 15 million to 16 million BTUs per ton of pulp produced and does not provide the pulping liquors necessary for the self-generation of energy from black liquor. However, the pulping of waste fibers is much less energy intensive and requires approximately 4 million BTUs per ton of pulp.

Chemical pulping is by far the most common method for pulping wood both in Europe and in the United States. In the United States, it accounts for more than 80% of all pulp produced. Chemical pulps have a low yield (40–55%), but the pulp produced is of very high quality. The most common process, the kraft process, accounts for more than 95% of all chemical pulp produced both in the United States and in Europe. In this process, the virgin materials are mixed with a chemical solution, called white liquor, in a digester, where they are pressurized and heated to dissolve the lignin from the wood. After being dissolved, the virgin material is separated into individual fibers and the chemical mixture, now called black liquor, is concentrated in an evaporator. Then, it is burned in a recovery boiler, where the energy is recovered, often cogenerating both steam and electricity, which is then used as process energy. Kraft pulping requires 5 to 7 million BTUs per ton of pulp, but chemical recovery requires an additional 4 to 7 million BTUs per ton of pulp.

Semi-chemical pulping is a combination of chemical and mechanical pulping, where the fibers are chemically treated before they are mechanically pulped. Following any of the pulping processes, the fibers either move on to the bleaching phase to reach a certain level of brightness or move on directly to the papermaking phase. The bleaching phase requires 5.0 to 7.5 million BTUs per ton of pulp.

1.1.2 Papermaking

Different types of pulp are used for the production of various paper products. Given the different energy requirements and opportunities for energy self-generation, the output mix, and thus demand for paper, has major implications for energy use in the industry.

Mechanical pulp is most often used to produce newsprint. Sometimes, chemical pulps are added, for example, to improve strength. Bleached kraft pulp is used mainly in quality printing and writing papers, and unbleached kraft is used to produce linerboard and sack paper. Semi-chemical and chemimechanical pulps typically are used to produce boxboard and sanitary products.

The papermaking stage is made up of paper and paperboard mills and consists of three main processes: stock preparation, paper production, and finishing. First, the pulp is dispersed in water and combined with additives. Then, this slush of pulp, water, and various additives is fed to a paper machine, and eventually dry bulk rolls or sheets of paper or paperboard are produced. The drying process is the most energy-intensive part of papermaking, but the total specific energy use ranges from 3.5 to 10.0 million BTUs per ton of paper depending on the paper type being produced.

2. FOREST PRODUCTS AND ENERGY CONSUMPTION

2.1 Total Energy Use and End Use Efficiency

Internationally, the pulp and paper industry is a major energy consumer and one of the most energy-intensive manufacturing industries. In the United States, only petroleum and coal products production rival its energy intensity (Fig. 2).

The American Paper Institute (API) and later the American Forest and Paper Association (AF&PA) have gathered and analyzed energy consumption in the industry since 1972. According to their data, the industry consumes approximately 12% of total manufacturing energy use. As the industry has grown, total energy consumption has increased on average 1.2% per year. At the same time, the intensity of energy use has declined considerably, from 39.3 million BTUs per ton of paper and paperboard in 1972 to 29.32 million BTUs per ton in 1998. This decline amounts to on average 0.9% per year decline in energy intensity (Fig. 3).

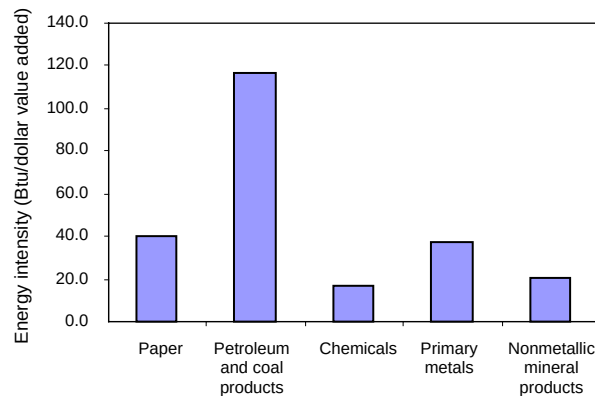


FIGURE 2 Total energy consumption (thousands of BTUs) per dollar of value added by industry. From the U.S. Energy Information Administration, 2002 Energy Consumption Survey, 1998 Annual Survey of Manufacturers.

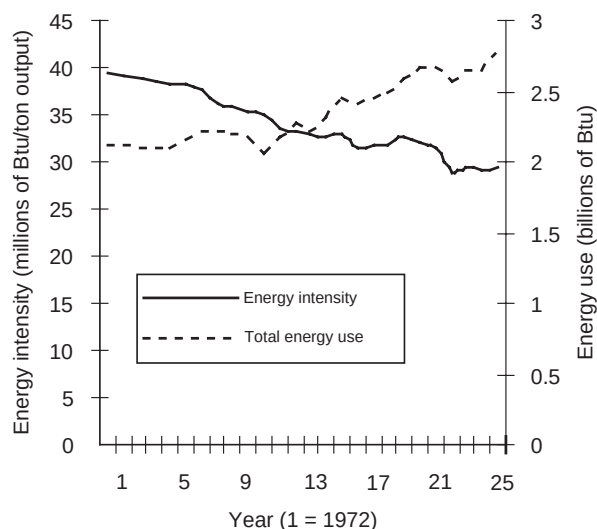


FIGURE 3 Total energy use (billions of BTUs) and energy use per output (millions of BTUs/ton output) in the U.S. paper and allied products industry. From the American Forest and Paper Association, various years.

Despite substantial improvements in energy intensity, total energy use has continued to increase because production growth on average has increased faster than have improvements in energy intensity. As depicted in Table I, which demonstrates the decline in energy intensity in selected countries, energy intensity has declined the most in the United Kingdom and the least in Sweden.

The data in Fig. 2 and Table I depict national annual averages. However, major geographical and temporal differences exist, both in terms of absolute quantities and in terms of changes in total fuel use and energy intensity. Geographical differences rest on

TABLE I

Relative Changes in Energy Intensity Expressed as Annual Changes from 1973 to 1991 (Percentages)

Country	Annual change
United States	-0.9
Japan	-1.9
Germany	-1.4
France	-1.5
United Kingdom	-2.6
The Netherlands	-2.0
Sweden	-0.8
Australia	-1.4
Average	-1.6

Source. Farla *et al.* (1997). Adjusted with permission from Elsevier.

the variation in processes that are used and, thus, on the products produced. They also are influenced by regionally heterogeneous input prices, input availability, public policies, and the like. Studies have shown that the factors having the largest impact on changes in energy intensity and total energy use are changes in output levels, changes in output (and thus process) mix, and changes in production efficiency.

By definition, an increase in output levels, without a change in output mix, increases the use of energy. But based on the relative energy intensities of the processes in use, the total amount of energy used by the industry can either increase or decrease due to a change in the mix of output given that each output category can somewhat be associated with a particular pulping and papermaking process.

The difference in energy intensities of various products rests mainly on three factors: the pulping process that is used (and thus whether chemical recovery is required), the question of whether to bleach the paper or not, and the papermaking process (Table II).

Table II illustrates the relative energy intensities of various paper and pulp manufacturing processes that are in widespread use relative to the least intensive pulping process—wastepaper pulping. The data demonstrate that the energy intensity of different pulping processes can vary by more than 400%.

In addition to the varying energy requirements based on output and process mix, variations in total energy use and energy intensity originate in the amount of energy required to heat air and ground water that is used in the production process. For instance, according to the AF&PA, total energy requirements per ton of output decrease significantly

TABLE II
Relative Energy Intensities of Major Processes Using Current Technologies versus Wastepaper Pulping

Process	Process/wastepaper pulping
Wood preparation	0.12
Pulping	
Kraft pulping	1.92
Sulfite pulping	1.99
Groundwood and thermomechanical	4.09
Dissolving and special alpha	3.71
Semichemical pulping	2.28
Wastepaper pulping	1
Market pulp drying	0.97
Bleaching	1.74
Chemical recovery	
Kraft recovery	2.33
Sulfite recovery	1.75
Semicheical recovery	1.40
Papermaking	
Newsprint	1.90
Printing and writing paper	3.21
Packaging and industrial paper	3.28
Tissue paper	2.60
Paperboard	3.16
Auxiliary	0.35

Source. Energetics (1990). Adjusted with permission from Energetics.

during the summer months in the northern United States by a difference of 15 to 20%.

2.2 Energy-Saving Technological Change

Energy efficiency also has changed due to embodied and disembodied technological change. Energy-saving disembodied technological change is defined as low- or no-cost operational changes in the energy efficiency of the existing capital stock. Embodied technological change requires substantial capital investments through which new technologies diffuse into the capital stock. Such diffusion requires either expansion or replacement investment. Expansion investment, by definition, expands the existing capital stock, enabling increased production levels and increased total energy use, albeit usually at a slightly lower level of energy intensity, assuming new capital is more efficient than existing capital. Replacement investment directly replaces retired capital and includes the investment spent on retrofitted capital. Consequently, replacement investment

does not expand the capital stock. By definition, it does not increase total energy use; thus, it enables a faster decrease in energy intensity than does expansion investment, assuming new capital is more efficient than old capital.

Embodied technological change is considered a more permanent change in energy efficiency, even though disembodied change accounts for a substantial part of energy savings, as is the case in the pulp and paper industry. This is because embodied changes are directly imbedded into the capital stock, whereas disembodied gains are more easily reversible. Disembodied changes occur, for example, due to the accumulated experience of employees (learning by doing). Consequently, if the rate of employee turnover increases, the rate of disembodied change slows down.

Historically, energy-saving technological change in the pulp and paper industry has been incremental and a combination of embodied and disembodied changes, an expected feature of a mature industry. This incremental change in energy intensity has provided a continued counter to the increase in output levels and has resulted primarily from gradual process improvements, a shift to wastepaper pulping, and better housekeeping practices. Again, the historical low rate of capital turnover is one of the main culprits in keeping the rate of change incremental. Other reasons include, for instance, that research and development in the industry is very much geared toward improving existing production technologies, as is the case in most mature industries. In the United States, private research and development (R&D) expenditures in the pulp and paper industry are the lowest among all manufacturing industries—and are spent by companies that specialize in consumer products. This suggests that much of R&D expenditures are directed toward incremental product development rather than toward energy intensity improvements. What this indicates is that the incremental development of the industry's knowledge base over time favors already existing processes and that this acts as a barrier to the adoption of new technologies. This is called technology lock-in. Thus, as the industry becomes more mature, it becomes less inclined to update its technologies and, as a result, closely follows a specific predetermined technology trajectory. Second, the total cost of energy is on average lower than that of other inputs such as fibers, and fossil fuel prices in the United States are low and fluctuating. Thus, other investments besides energy-saving investments may seem more profitable and returns less uncertain. Third, the required rate of capital recovery for an investment in energy-saving

equipment—the payback period—is approximately 2 years. Consequently, if an energy conservation project does not offer benefits other than energy cost savings, the payback period must be less than 2 years.

Combining those factors, and taking into account the low profit margins and the necessity to maintain a high capacity utilization rate, makes it hard to justify and execute substantial investments in energy-saving equipment leading to embodied changes in energy efficiency. As a result, the norm in the pulp and paper industry is a gradual decline in energy intensity through efficiency improvements that generally result from improved housekeeping practices or that are side effects of other process- or capacity-related changes.

2.3 Energy Mix

The pulp and paper industry uses two types of end use energy: thermal and electrical. Electrical energy is used primarily to drive pumps and fans. Thermal energy, which is in the form of steam, is used, for example, in digesters, in evaporators, and in the drying of pulp. The industry is in the enviable position of being able to generate itself a large proportion of both the required thermal energy and the required electrical energy. For instance, in the United States, it generates more than 55% of the sum of required electrical and thermal energy, and more than 80% of that energy is generated directly as a by-product of chemical pulping. As depicted in Fig. 4, the remaining energy consumption (collectively called purchased energy) is divided mostly among purchased electricity, coal, natural gas, and residual fuel oil, with a small fraction being allocated to distillate fuel oil and liquid petroleum gas.

The fuel mix in the U.S. pulp and paper industry has changed substantially since 1972. The industry has slowly increased the use of self-generated energy, with an annual average increase of 3.27% for the past 30 years or so. At the same time, the fractional share of purchased fuels has declined. Of purchased fuels, residual fuel oil is slowly being phased out, and the fractional share of natural gas is increasing. The fractional share of coal has remained relatively constant since the early 1980s.

Between 1972 and 1983, the U.S. pulp and paper industry rapidly switched from purchased fuels to self-generated energy, as indicated by the decline in purchased energy intensity and the increase in the intensity of self-generated energy use (Fig. 5).

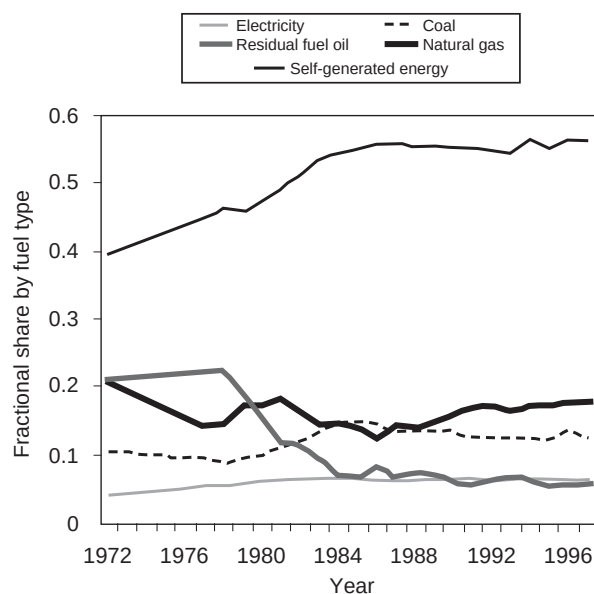


FIGURE 4 Fractional share of fuels by type in the United States pulp and paper industry. From the American Forest and Paper Association, various years.

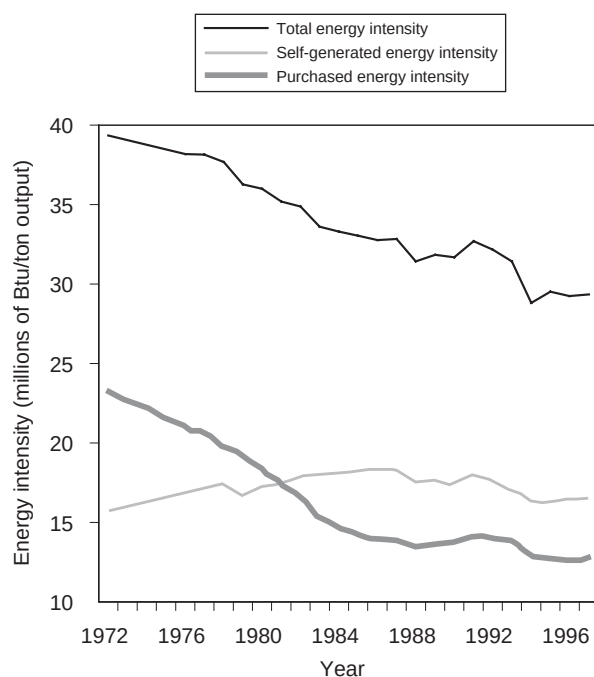


FIGURE 5 Total energy intensity, intensity of purchased energy, and intensity of self-generated energy in the U.S. pulp and paper industry. From the American Forest and Paper Association, various years.

This expansion occurred in part as a result of higher energy prices during the 1970s and a change in the Public Utility Regulatory and Policies Act of 1978 (PURPA). Those two events made cogeneration

of steam and electricity economically efficient. Lower energy prices and the close link between material and energy use reduced the rate of cogeneration after 1983. As previously discussed, more than 80% of the energy that is self-generated is a direct by-product of chemical pulping of virgin fibers. A continued increase in the fractional share of self-generated energy either requires an increase in the share of the chemical pulping of virgin fibers (which has important implications for the use of recycled fibers and the output mix) or requires technological change that will increase the energy and economic efficiency of cogeneration. The former requirement is a major obstacle to an increase in the share of self-generated energy and vividly illustrates the close interconnection between materials and energy cycles in the industry. The latter requirement is a smaller obstacle given that such technology already exists; however, the character of the industry presents a formidable challenge to the ability of this new technology to diffuse into the capital stock. This issue is discussed later in section IV, where the potential for energy conservation and continued expansion into self-generation is explored.

3. FOREST PRODUCTS AND ENERGY GENERATION

As illustrated in section II, and in Figs. 4 and 5, the pulp and paper industry enjoys substantial flexibility in its choice of energy inputs. This historical fuel flexibility of the industry results from its direct access to wood fuels and its heavy use of steam for process heat. Because steam systems can use a wide variety of fuels, the industry has been able to shift with relative ease from residual fuel oil to gas and to substitute bark and forest residues for fossil fuels. In addition, the industry contributes to energy supply indirectly by supplying some of the materials that contribute to methane generation in landfills that can be used for electric power generation and the like.

3.1 Direct Use of Forest Products as Energy

3.1.1 Wood Fuels: Categories and Importance

Wood fuels are a renewable and climate-friendly energy source because they are considered carbon neutral if used and produced in a sustainable and efficient manner. Wood fuels consist of four main

commodities: fuelwood (also called firewood), charcoal, black liquor, and wood waste.

Fuelwood is used extensively in most developing countries and is collected from the forest, often as fallen branches or twigs. In domestic applications, it is converted into heat in simple ovens, often for cooking, baking, or drying, and/or is used for space and water heating. The fuel gases are also used for the smoking of food products. Contrary to popular belief, the gathering of fuelwood is not responsible for large-scale deforestation; however, it can contribute to localized forest and soil degradation. Larger issues associated with the traditional use of fuelwood are the health effects of using inefficient stoves in poorly ventilated areas. This leads to emissions of pollutants such as carbon monoxide, methane, and nitrogen oxides that carry serious health risks. For instance, those pollutants are known to cause respiratory infections, adverse pregnancy outcomes, and tuberculosis. In addition, the collection of fuelwood is hard physical labor. Because women are mostly responsible for collecting fuelwood and cooking, women and small children are most adversely affected. Commercial fuelwood is also used as an industrial source of energy in many developing countries. Typical industrial uses include curing and drying, food processing, and lime and brick burning. Although the total amount of wood consumed by the industrial sector is much smaller than that consumed by households, it is nonetheless significant. For example, in Kenya and Sri Lanka, 26 and 18% of fuelwood is used by industry, accounting for more than 70 and 65% of all industrial energy use, respectively.

Charcoal is produced from wood through carbonization or pyrolysis. The wood is heated in a closed container that is devoid of oxygen. The absence of oxygen prevents the wood from burning to ashes; thus, the wood decomposes to charcoal, among other residues. Charcoal plays an important but a much smaller role than does fuelwood in both industrial and household sectors of most developing countries. In households, it is used for tasks similar to those of fuelwood, but a switch from fuelwood to charcoal provides numerous benefits. Those include the fact that charcoal generally does not decompose in storage, it burns without smoke for the most part, it does not create dangerous flames while cooking, and it requires a simple stove whose heat output is relatively easy to control. As a result, a switch to charcoal often represents an increase in the quality of life. The disadvantages of charcoal use are related primarily to the inefficiencies inherent in its produc-

tion and use. Production efficiency in most developing countries ranges between 8 and 28% and is a function of moisture content and type of equipment. Charcoal is also inefficient in use. Although charcoal stoves are more efficient than fuelwood stoves (20–35% vs 10–25%), charcoal stoves are much less efficient than modern fuel stoves such as kerosene (35–50%) and electric stoves (75–85%). The combined production and use inefficiencies have important consequences for the abundance of local wood resources given that cooking with charcoal often uses three to four times more wood than does cooking with fuelwood. Charcoal is also used in industrial settings in both developing and developed countries, for instance, in the production of iron and steel, in metallurgy, and in the chemicals industry.

The other two categories of wood fuels are derived from secondary sources and consist of black liquors from pulp production and wood waste from forest products industries and other products from society such as wastepaper.

Wood fuels account for approximately 7% of the world's total energy supply and are an extremely important source of energy in many developing countries. According to the Food and Agricultural Organization (FAO), developing countries consume approximately 77% of the world's supply of wood fuels, and this in turn accounts for 15% of total energy consumption in the developing world. Consumption in developed countries accounts for the remaining 23%, but this share amounts to 2% of total energy consumption in the developed world. Of course, substantial regional and national differences exist. Developing countries rely mostly on fuelwood and charcoal, but in most developed countries the pulp and paper industry is instrumental in supplying and using such energy. For instance, in 1997 in New Zealand, black liquor accounted for nearly 100% of all energy derived from forest products, and in Japan, Portugal, Sweden, Norway, and Finland, it accounted for 95, 87, 81, 79, and 80%, respectively (Table III). Forest products provided approximately 17% of total national energy supply in Finland and Sweden.

In the United States, black liquor alone contributes 63% of all energy derived from forest products, accounting for slightly over 2% of total energy supply.

3.1.2 Wood Fuel Use in the Pulp and Paper Industry

The pulp and paper industry uses not only black liquors but also solid biomass residues such as

TABLE III

Wood Fuel Consumption in 1997 by Selected Countries and Types (Petajoules)

Country	Fuelwood	Charcoal	Black liquor	Total
Bangladesh	307.39	—	0.95	318.34
Burkina Faso	110.26	1.07	—	111.33
Belgium	5.55	0.51	4.75	10.81
Cameroon	135.67	2.68	—	138.35
Finland	40.80	0.12	162.36	203.28
Kenya	193.89	42.17	1.53	237.59
Mexico	233.39	3.74	7.01	244.14
New Zealand	—	—	15.44	15.44
Sweden	38.81	0.49	171.20	210.50
United States	698.64	24.68	1225.50	1948.82

Source. Food and Agricultural Organization (www.fao.org/forestry). Adjusted with permission from the Food and Agricultural Organization of the United Nations.

woodchips, forest residues, wastepaper, and bark. When using solid biomass residues, the industry uses so-called “wood-fired” power boilers (which are similar to coal-fired boilers), where the forest residues are burned to generate steam and/or electricity. Also, it is common for boilers to have multifuel capabilities, enabling the aforementioned fuel flexibility in the industry. Wood has an advantage over natural gas, coal, and oil in that the industry has a cheap supply in the form of wood wastes, bark, and hogged fuels, giving a substantially higher thermal energy content per dollar spent when compared with both oil and coal. Furthermore, wood waste can be burned alongside coal. Disadvantages in use of solid biomass residues are the often high moisture content in the wood waste and low volumetric energy density. Despite those disadvantages, the use of hogged fuels, bark, and wood waste accounts for approximately 20% of self-generated energy in the U.S. industry.

The process of using black liquor to produce energy begins with a process called chemical pulping. In that process, sodium compounds called white liquor are used to cook the virgin wood to release wood fibers from the lignin in the wood. After the woodchips are cooked, they are separated from the lignin/sodium solution, washed, and screened after which they proceed on to the papermaking process. However, the lignin/sodium solution, which now is called black liquor, is transported to a condenser, where it is evaporated to create concentrated strong black liquor. Then, the strong black liquor is transported to a recovery boiler. The most

common model is the Tomlinson recovery boiler. In the recovery boiler, the black liquor is combusted to yield recoverable thermal energy and an inorganic solution of sodium compounds. The inorganic chemicals are extracted as smelt at the bottom of the boiler, giving green liquor. Then, the green liquor is converted to white liquor and reused. The thermal steam energy from the boiler is used directly as process energy, such as for the evaporation of black liquor, and also in driving steam turbines to produce electrical energy, which can be either used on site or sold. This double use of waste heat as steam in the production process and the simultaneous production of electricity is called cogeneration. The industry is the leading manufacturing industry in the United States in the cogeneration of steam and electricity, having produced in 1998 39.4% of all cogenerated electricity.

3.2 Indirect Use of Forest Products as Energy

Indirect generation of energy also occurs as a result of paper waste management practices. In 1995, the U.S. Environmental Protection Agency (EPA) estimated that 49.8 million tons of paper and paperboard were discarded as municipal solid waste in the United States, representing a tremendous amount of usable matter that is buried every year. There are at least two means by which this energy can be harnessed: through the direct incineration of wastepaper (preventing the wastepaper from entering the landfill) or through the capture of methane (which forms in landfills as a result of paper decomposition).

Decomposition of paper and wood products in landfills occurs in a series of stages. The first stage of decomposition is characterized by the removal of oxygen from the waste by aerobic bacteria. Then, the environment becomes oxygen limiting and the bacteria shift to fermentative metabolism. During the second stage, fermentative bacteria hydrolyze polymers into sugars, and those are then fermented further into short-chain carboxylic acids, carbon dioxide, and hydrogen. As acids are produced, the acidity of the refuse increases but then stabilizes due to the removal of acetate and the release of metallic cations that dilute the hydrogen ions. Eventually, this leads to the development of a pH level that is close to neutral. A neutral pH is a necessary prerequisite for the landfill to enter the third stage of decomposition. During this final stage, acetogens further oxidize the remaining short-chain carboxylic acids into acetate, carbon dioxide, and hydrogen. Because the acetogens cannot perform this final stage of decomposition

TABLE IV

Methane Generation Potential of Various Paper Products (Gram Methane/Gram Refuse)

Material	Methane potential
Printing and writing	0.098
Newsprint	0.054
Tissue	0.131
"Other"	0.108

Source. Micales and Skog (1997). Adjusted with permission from Elsevier.

unless the hydrogen ions are removed, this process must occur parallel to the growth of hydrogen-consuming bacteria such as those that produce methane, hence the production of methane.

Because lignin inhibits decomposition, the methane generation capacity of paper products varies as a function of the amount of lignin that remains present (Table IV).

Newsprint is produced using mechanical pulp that retains most of the initial lignin; thus, the methane generation capacity is lower for newsprint than for other paper categories. Because tissue is produced using chemical processes, all of the lignin has been removed, hence the highest methane generation potential. Currently in the United States, approximately 10% of the methane generated in landfills is captured for energy purposes, indicating a substantial future potential.

4. THE FUTURE: ENERGY CONSERVATION AND SELF-GENERATION POTENTIALS

The question remains, will the pulp and paper industry be able to become even more self-sufficient in terms of energy generation and even become a net energy exporter and, thereby, increase the role of wood fuels as a source of energy? Will the industry simultaneously be able to reduce total and purchased energy intensity? This section addresses those issues by highlighting technological and economic potentials and the looming barriers to change. A short conclusion follows.

4.1 Technical Potentials

This subsection examines whether technical potentials exist to (1) increase the share of self-generation

of energy and (2) reduce energy intensity. The discussion is based on a study performed at the Lawrence Berkeley National Laboratory by Ernst Worrell, Nathan Martin, and colleagues on U.S. industry potential to increase its energy efficiency and energy self-generation.

4.1.1 Increase in Energy Self-Generation

Black liquor gasification is arguably the most promising new technology to increase self-generation capabilities in the industry. As discussed previously, the cogeneration of electricity and steam usually is performed by recovery boiler/steam-driven turbines. However, the new biomass cogeneration technologies use gas turbines rather than steam turbines. Compared with conventional steam turbines, gas turbine combined cycle cogeneration offers high power to steam ratios and high thermal cycle efficiencies. For instance, power output is three times higher than what is possible with the conventional Tomlinson recovery boiler. Thus, black liquor gasification can potentially have a revolutionary impact on the industry by allowing kraft mills to become self-sufficient in both thermal and electrical energy as well as to become net electricity exporters. Despite high initial capital costs, it is believed that in the long run black liquor gasification systems will prove to be economically efficient.

4.1.2 Improvements in Energy Efficiency

Improvements in efficiency are technologically feasible at every level of pulp and paper production. The discussion in this subsection is a nonexhaustive account of some of the potential improvements in energy efficiency and is intended to illustrate the potential for change. Keep in mind that the average energy intensity of the U.S. pulp and paper industry is close to 29 million BTUs per ton of output. When preparing raw materials, the already available energy-saving technologies, such as ring-style debarkers and cradle debarkers, would save approximately 0.03 million to 0.05 million BTUs per ton of output, and improved screening processes could save up to 0.5 million BTUs per ton of output. Those three investments have a payback period of less than 1 year given U.S. energy prices. In mechanical pulping, biopulping could reduce energy intensity by 3.4 million BTUs per ton of output and could reduce various refiner improvements by 1.6 million BTUs per ton of output. However, the payback period is estimated to be 30 years for biopulping, whereas it is only 3.4 years for refiner improvements. In thermo-mechanical pulping, potential improvements range

from 1 million to 7.5 million BTUs per ton of output via improved heat recovery with a payback period of 4.7 years. Efficiency of chemical pulping can increase by 8.4 million BTUs per ton of output by using continuous digesters with a payback period of 7.7 years. The efficiency of papermaking can increase the most by improving drying technologies such as through the use of Condabelt drying. Energy savings are estimated to be up to 2.40 million BTUs per ton of output with a payback period of 3.8 years. Dry sheet forming, or production of paper without the addition of water, will save up to 5.5 million BTUs of energy per ton of output with a payback period of 88 years. Numerous general measures can also be applied, including installing energy efficient lighting, which will save up to 0.1 million BTUs per ton of paper with a payback period of 3.7 years, and pinch analysis, which optimizes thermal energy flows throughout the plant. This measure is likely to save up to 2.5 million BTUs per ton of paper with a payback period of 1 year. Substantial improvements can also be made to make the production and distribution of steam more efficient, including improved boiler maintenance, which will save approximately 1.7 million BTUs per ton of output with a payback period of 0 years. Other measures include leak repair and steam trap maintenance, with energy savings of up to 0.7 and 2.5 million BTUs and with payback periods of 0.1 and 0.3 years, respectively. Finally, the industry can reduce total primary energy consumption by increasing the use of recycled fibers. This switch might not reduce purchased energy requirements unless the waste pulp is replacing mechanical or thermomechanical pulp.

Clearly, significant economically efficient engineering potentials exist for the pulp and paper industry to increase process energy efficiency. An investment is deemed economically efficient if the payback period is 0 to 3 years. According to the aforementioned study from the Lawrence Berkeley National Laboratory, such cost-effective energy savings amounted to 16% of total primary energy use in the industry in the United States in 1994, and including the expansion of recycling increased that potential to 22%.

4.2 The Barriers to Energy-Saving Potentials

The first culprit, as normally is the case with very capital-intensive mature industries, is the low rate of capital turnover. Because it is extremely expensive to

replace old capital, not just due to capital costs but also due to the required downtime of equipment, capital is kept as part of the capital structure for up to 100 years. Second, energy-related investments compete with other investments for capital, and the relatively low fuel and electricity prices that have prevailed since the mid-1980s have reduced the profitability of energy-saving investments. Combining that with fluctuating energy prices, uncertain returns, and a very short payback time (less than 2 years) explains much of the unused potential for saving energy. Third, the close link among the input materials, output types, and amount and type of energy that the industry can use acts as a limit to the extent to which the industry can switch over to self-generated energy. Fourth, energy savings often seem far from the radar screen of decision makers, who focus more often on improving and extending existing product lines.

5. CONCLUSION

Wood fuels that are derived from forest products remain an important source of renewable, potentially carbon-neutral energy. As countries developed, the role of wood fuels declined and shifted from primarily household use to industrial use. In developed countries, the pulp and paper industry, through its use of chemical pulping, is the main user and provider of wood fuels, thereby keeping them on the horizon as a significant source of energy. The pulp and paper industry is one of the most energy-intensive manufacturing industries, but in the United States the industry self-generates from wood fuels (mostly black liquor) more than 55% of its energy needs. Further extension of energy self-generation partially depends on the rate of diffusion of new technologies, yet the character of the industry poses formidable challenges to technological change, as do economic factors such as low and fluctuating energy prices.

An increase in the share of self-generated energy is also closely tied to the use of specific input materials and output demand because black liquor is a by-product of chemical pulping of virgin fibers. If paper companies increase the use of pulping processes that rely on virgin fibers, they continue to be able to cogenerate steam and electricity from black liquor. An increase in the use of recycled fibers conserves forest resources. This may, on the one hand, be more energy intensive in terms of purchased energy if the waste pulp replaces chemically created pulp. On the other hand, the replacement of recycled pulp for

mechanically produced pulp conserves a significant amount of purchased energy.

Because methane has 21 times the global warming potential of carbon dioxide, important trade-offs also exist regarding greenhouse gas emissions. Those trade-offs are linked directly to waste management. If wastepaper is put into landfills, it will generate methane and carbon dioxide. If it is recycled and the waste pulp replaces nonbleached chemical pulps, it may increase purchased energy consumption and, thus, increase carbon dioxide emissions. If it is incinerated, it will not contribute significantly to methane emissions but will not play a role in the conservation of forest resources. It seems that in this particular industrial system, it is challenging to conserve one resource without using more of another or to reduce emissions of one greenhouse gas without emitting more of another.

Despite formidable barriers to change, the future is not bleak for the pulp and paper industry. Major potentials exist for improving energy efficiency, and new cogeneration technology can transform a substantial part of the industry into a net energy exporter—completely reliant on renewable self-generated energy. In the future, the industry could derive large profits from selling energy or, alternatively, could let the utilities sector invest in and operate the new cogeneration technology. Either way, the potential exists for an increase in the energy efficiency of the pulp and paper industry as well as for an increase in the role of forest products, and thus wood fuels, in the total supply of energy in the developed world.

SEE ALSO THE FOLLOWING ARTICLES

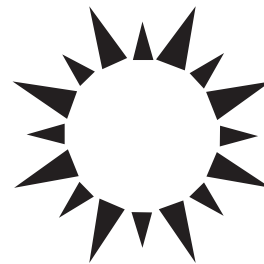
Biomass, Chemicals from • Biomass for Renewable Energy and Fuels • Biomass Resource Assessment • Coal, Fuel and Non-Fuel Uses • Cogeneration • Fire: A Socioecological and Historical Survey • Wood Energy, History of • Wood in Household Energy Use

Further Reading

- American Forest and Paper Association. (various years). "Statistics of Paper, Paperboard, and Woodpulp." AF&PA, Washington, DC.
- Biermann, C. J. (1996). "Handbook of Pulp and Papermaking." Academic Press, San Diego.
- Energetics. (1990). "Industrial Technologies: Industry Profiles." Energetics, Columbia, MD.

- Farla, J., Blok, K., and Schipper, L. (1997). Energy efficiency developments in the pulp and paper industry. *Energy Policy* 25, 745–758.
- Food and Agricultural Organization of the United Nations. (2001). Recovered paper data 2001. Rome, Italy.
- Food and Agricultural Organization of the United Nations. (2002). Woodfuel consumption 1997. Rome, Italy.
- Gilbreath, K. R., Gupta, A., Larson, E., and Nilson, L. J. (1995). Background paper on energy efficiency and the pulp and paper industry. Prepared for the ACEEE 1995 Summer Study on Energy Efficiency in Industry, Grand Island, NY.
- Herzog, H., and Tester, J. W. (1991). Energy management and conservation in the pulp and paper industry. In “Energy and the Environment in the 21st Century” (J. Tester, D. O. Wood, and N. Ferrari, Eds.), pp. 437–447. MIT Press, Cambridge, MA.
- Martin, N., Anglani, N., Einstein, D., Khrushch, M., Worrell, E., and Price, L. K. (2000). “Opportunities to Improve Energy Efficiency and Reduce Greenhouse Gas Emissions in the U.S. Pulp and Paper Industry,” LBNL-46141. Lawrence Berkeley National Laboratory, Berkeley, CA.
- Micales, J. A., and Skog, K. E. (1997). The decomposition of forest products in landfills. *Intl. Biodeterioration Biodegradation* 39(2–3), 145–158.
- Smith, M. (1996). “The U.S. Paper Industry and Sustainable Production: An Argument for Restructuring.” MIT Press, Cambridge, MA.

Forms and Measurement of Energy



TREVOR M. LETCHER

Natal University
Durban, South Africa

ARTHUR WILLIAMSON

University of Canterbury
Christchurch, Canterbury, New Zealand

-
1. Introduction
 2. The History of the Concept of Energy
 3. Kinetic and Potential Energy
 4. Heat and Work
 5. Measurement of Energy in Terms of the SI System of Units
 6. Forms of Energy
 7. Conclusion

Glossary

energy The capacity to do work.

heat The transfer of energy as a result of a temperature difference.

kinetic energy Energy by virtue of movement.

potential energy Energy by virtue of position.

radiant energy Energy due to electromagnetic vibrations.

work The transfer of energy into a useful form.

An understanding of the physical world requires the concept of energy, or the capacity to do work. Our modern society depends on energy for heating, cooling, transport of all kinds, lighting, and communications, to name but a few vital needs. Thousands of years ago, human energy requirements were limited to the body's own energy and that was fueled by food—the carbohydrates and fats of a hunter-gatherer's diet. Today, the total energy required to ensure the survival and well-being of a person in a modern society is many hundreds of times greater than that of a person in a primitive society.

1. INTRODUCTION

Energy is defined as “the capacity to do work.” It can take many forms. We talk of kinetic energy (obtained

from ocean waves and from wind), gravitational energy (obtained from falling water), electrical energy (obtained from batteries and turbines), chemical energy (obtained from exothermic reactions such as combustion of diesel and gasoline), thermal energy (obtained from burning coal or wood), radiant energy (from sunlight), and nuclear energy (obtained from fission or uranium atoms or fusion of hydrogen nuclei). Some of these energy forms are more useful than others and many can be transformed from one form to another. For example, the energy from a nuclear reaction can be used to heat water, which in turn can be used to do work by turning a turbine to make electrical energy to heat our houses.

A fundamental law of thermodynamics (the study of the transformations of energy) is that energy is conserved in any process. This implies that when energy transfer does take place, no energy is lost in the process. Usually in the transfer process, a significant fraction of the energy is lost to useless heating of the environment. The efficiencies of conversions can be as low as 10 or 20%. This law of conservation of energy was a major breakthrough in the understanding of energy and has never been violated. One of its implications is that energy cannot be created. In other words, perpetual motion is impossible. This, however, has not stopped hundreds of would-be millionaires from sending their plans to patent offices around the world.

2. THE HISTORY OF THE CONCEPT OF ENERGY

The concept of energy took a long time to develop and was hindered by incorrect and unsubstantiated

ideas of Aristotle and many others. The turning point came when Galileo and then Newton, in the 17th century, properly defined the force, f , acting on a body of mass, m , moving with an acceleration, a , as

$$f = ma. \quad (1)$$

It took another two hundred years before energy was correctly formulated by William Thomson in 1851. It was the steam engine that catalyzed the concept. Energy was defined as the work done in moving a body a distance $\times$ against a force, f :

$$E = w = fx. \quad (2)$$

But it was not until the middle of the 20th century that the separate units for heat (the calorie) and work (the erg) were dropped in favor of the joule and the link between energy, heat and work was finally made.

During the early part of the twentieth century, it was shown that energy can only occur in discrete packets or quanta, and the concept of the quantization of energy was born. For most forms of energy, these packets are so very close to each other that they effectively form a continuum and the sum of all the quantized energies approximates to the energy as defined by Newtonian physics. The quantization of energy becomes important when dealing with small energies related to small particles such as electrons and atoms. In particular, vibration, rotation, and electronic energies are affected and quantum effects must be considered.

3. KINETIC AND POTENTIAL ENERGY

We tend to think of energy in its many manifestations, some of which are given in the introduction. However, in reality, there are only three types of energy: kinetic, potential and radiant energy. The kinetic energy of the body of mass, m , moving with a velocity, v , is defined as

$$E_K = 1/2mv^2. \quad (3)$$

It is the energy a body has as a result of its movement.

The potential energy of a body, on the other hand, is the capacity of the body to do work by virtue of its position. A body of mass, m , dragged up a height, h , against gravity, has the potential energy

$$E_{PE} = mgh, \quad (4)$$

where g is the acceleration due to gravity, which on Earth is about 9.8 m s^{-2} . This body has potential

energy as it rests in position. From that position it can be made to do useful work, such as if it is allowed to fall under gravity while linked to a pulley which lifts a weight.

The link between kinetic and potential energy and other forms of energy is not always clear. Electrical energy, for example, is really the potential energy of electrons in the presence of positive charges. Chemical energy is a sum of the potential energy of electrons in chemical bonds and the kinetic energy of the bonding electrons as they move around an atom or a molecule. Nuclear energy is due to the potential energy and kinetic energy of the subatomic particles within an atomic nuclei.

4. HEAT AND WORK

The terms heat and work are synonymous with energy, and both are simply ways of transferring energy. Heat refers to the transfers of energy as a result of a temperature difference (ΔT). The energy transferred from a hot body of mass, m (in contact with a cold body), as the body's temperature drops from T_1 to T_2 is given by

$$\Delta E_{\text{HEAT}} = mS\Delta T, \quad (5)$$

where $(T_1 - T_2) = \Delta T$ and S is the specific heat of the body. The specific heat is defined in terms of Eq. (5) and is the energy required to heat 1 g of the material of the body, through 1 K. For pure water it is $4.18 \text{ J g}^{-1} \text{ K}^{-1}$. (This is equivalent 1 calorie per gram per Kelvin.)

Work, on the other hand, refers to the transfer of energy into a useful form (i.e. lifting a weight). For example, the work done, w , in lifting a mass, m , through a height, h , against gravity is

$$w = mgh. \quad (6)$$

Work is an agent of transfer and can take many forms. Types of work include expanding a surface, expanding a volume of a gas, pushing electrons along a wire (electricity) and lifting weights.

Heat and work are best understood in terms of the movement of molecules. The molecules of a stationary body at a temperature greater than 0 K (-273.15 K) move in a chaotic manner. If the molecules all moved in the same direction, the body would move! These molecules may rotate, vibrate, or translate. The latter refers to kinetic motion. These movements are collectively known as thermal motion. Molecules in a hot body have a greater thermal motion than the molecules in a cold body. The

transfer of energy from a hot body to a cold body is a result of the stimulation of molecules in the cold body to move more vigorously by the molecules in the hot body. Heat can thus be considered as the transfer of energy that makes use of the chaotic motion of the molecules (see Fig. 1A).

When work is done, the molecules or electrons all move in the same direction in an organized way. When a weight is lifted against gravity, all the molecules of the weight move in the same direction—upward against gravity. When electrical work is done, the working electrons all move in the same direction. Work must therefore be considered as the transfer of energy that makes use of organized motion (Fig. 1B).

Looking back at the historical use of energy, it was relatively easy for early humans to use heat to transfer some forms of energy to thermal motion. For many thousands of years, humans used burning logs of wood and presumably sunlight (radiant energy) on sun-facing caves to warm their caves. The transfer of energy so that work is done is a very much more difficult process and was not really done by machines until the advent of steam engines, which heralded the Industrial Revolution less than 200 years ago. Work, other than that done by human muscles in lifting, pushing, and pulling, is a recent phenomenon.

5. MEASUREMENT OF ENERGY IN TERMS OF THE SI SYSTEM OF UNITS

The measurement of energy involves no less than six of the seven fundamental quantities and units that make up the *Système International d'Unités*. The quantities and units together with their symbol are given in Table I. All other quantities can be derived

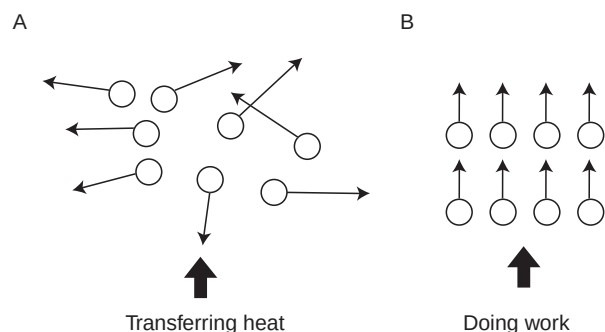


FIGURE 1 (A) Heat is the transfer of energy that makes use of the chaotic motion of the molecules. (B) Work is the transfer of energy that makes use of the organized motion of molecules.

from these fundamental ones. For example, acceleration can be written in terms of distance and time (L/T^2), force is the product of mass, and acceleration and energy can be written as the product of force and distance (ML^2/T^2); this combination is often called the dimensionality of the quantity. All energy forms have the same dimensionality, as do all units of pressure and all units of force and so on. The unit size of a quantity in the SI is the product of the fundamental units of which the quantity is composed (Table II).

The Newton meter⁻² (Pascal) is a rather small unit of pressure for many everyday purposes and one more commonly uses the bar (100 000 Pa), which is very close to atmospheric pressure (1 standard atmosphere = 101 325 Pa).

The SI has a set of conventional multipliers, which are used for quantities larger and smaller than the basic units (Table III). The rate of energy flow is the Watt (W)

$$1 \text{ Watt} = 1 \text{ J s}^{-1}.$$

Electrical energy is often metered and measured in kilowatt hours (kWh):

$$1 \text{ kWh} = 3\,600\,000 \text{ J}.$$

The distinction between power (rate of flow of energy) and energy is often badly made especially in

TABLE I

The Seven SI Units

Quantity	Name	Symbol
Length	Meter	m
Mass	Kilogram	kg
Time	Second	s
Temperature	Kelvin	K
Amount of substance	Mole	mol
Electric current	Ampere	A
Luminous Intensity	Candela	cd

TABLE II

Some Secondary SI Units

Quantity	Definition	Name	Symbol
Energy	$\text{kg m}^2 \text{s}^{-2}$	Joule	J
Force	kg m s^{-2}	Newton	N
Pressure	$\text{kg m}^{-1} \text{s}^{-2}$		
	N m^{-2}	Pascal	Pa
Power	$\text{kg m}^2 \text{s}^{-3}$	Watt	W

TABLE III
Prefixes Used in the SI System of Units

Multiplication	Multiple or submultiple	Symbol
10^{12}	tera	T
10^9	giga	G
10^6	mega	M
10^3	kilo	k
10^2	hecta	h
10^1	deca	da
Base unit		
10^{-1}	deci	d
10^{-2}	centi	c
10^{-3}	milli	m
10^{-6}	micro	μ
10^{-9}	nano	n
10^{-12}	pico	p
10^{-15}	femto	f
10^{-18}	atto	a

reporting electrical power/energy. This confusion frequently leads to the use of the term “kilowatts per hour,” which is neither an energy unit nor a unit of power, and kilowatts when kilowatt hours is meant.

Extensive properties, those that are proportional to the amount of substance such as heat capacities are in SI units expressed in terms of energy per mole. Except in chemical practice the mole is little used and extensive properties are frequently expressed as specific properties—that is, properties per unit mass. For example, the molar heat capacity of copper is $24 \text{ J mole}^{-1} \text{ K}^{-1}$ or in specific terms $378 \text{ J kg}^{-1} \text{ K}^{-1}$.

It is to be noted that if all quantities in a calculation are expressed in SI units then the result will also be in SI units. This coherence is of considerable practical use in ensuring that the results of calculations do not come out in strange units.

6. FORMS OF ENERGY

6.1 Mechanical and Gravitational Energy

Mechanical forms of energy can be considered as those that relate to movement and position. The most important type is translation or kinetic energy and it can be calculated from Eq. (3). Kinetic energy can be found in the energy of the wind and ocean waves.

Wind energy and its transformation into electricity have the advantage of being environmentally friendly. There are no emission of toxic substances, and it is a renewable resource. It is however, not a steady source and furthermore, it is a secondary energy form as it depends on the radiant energy from the sun. All this means that wind energy is a dilute and weak form of energy. In spite of this, a large number of wind turbines are being erected in most continents, and wind energy could be an important renewable energy source in the future.

The amount of energy reaching a wind machine can be calculated from Eq. (3), given the wind speed, the wind collecting area of the vanes, and the density of the air. For example, if the wind velocity is 10 m s^{-1} , the wind vane collecting area is 10 m^2 , and the density of air is 1.2 kg m^{-3} , then the kinetic energy, E_K , per second is

$$E_K = 1/2 mv^2 = (0.5 \times 10 \times 10 \times 1.2 \times 10^2) \text{ J}$$

$$\text{Power} = 6.5 \times 10^3 \text{ J s}^{-1}.$$

The harnessing of ocean waves is severely hampered by the corrosive nature of the sea, and very few such projects remain in the world after much enthusiasm in the 1980s and 1990s. The ocean waves are dependent on the moon-earth gravitational attraction and also on wind direction. Wave energy is thus also a secondary energy source and hence is rather dilute compared to primary sources.

The gravitation potential energy between two bodies of mass m_1 and m_2 , separated by a distance r , depends on the force between the bodies, f_G :

$$f_G = \frac{Gm_1m_2}{r^2}, \quad (7)$$

where G is the universal gravitational constant and has a value of $6.673 \times 10^{-11} \text{ Newton meter kg}^{-1}$. The potential energy E_G is thus equal to

$$E_G = \frac{Gm_1m_2}{r}. \quad (8)$$

The acceleration, g , due to gravity on any mass, due to m_2 , is given by

$$g = \frac{Gm_2}{r^2}, \quad (9)$$

and on earth, g has a value of about 9.8 m s^{-2} . (The radius of the earth is about $6.37 \times 10^6 \text{ m}$ and the mass of the earth is about $6.0 \times 10^{24} \text{ kg}$.)

The gravitational potential energy of falling water can be considerable if the flow of water and the fall is large. There are many very large power stations

around the world that utilize hydroelectric power. It is a clean energy source, free of chemical pollutants. Unfortunately, there is a limit to the number of places on earth where large volumes of falling water can be found. Hydroelectric power contributes about 2% of the total amount of energy used in the world (a 1994 figure).

There are other types of mechanical energy forms that are scientifically interesting and important but are not on a scale large enough for exploitation as an energy source. These include vibrational, rotational, and electronic energies. The potential energy of a spring is given by

$$E_{\text{VIB}} = 1/2 kx^2, \quad (10)$$

where x is the displacement of the spring from its rest position and k is the spring constant. Such energies are important in understanding the bonding that takes place between atoms. These vibrational energies are quantized and the available energies are small and given by

$$E_{\text{VIB}} = (v + 1/2)h\nu \quad v = 0, 1, 2, 3, \dots, \quad (11)$$

where h is Planck's constant (6.626×10^{-34} J s) and ν is the frequency of the vibration. This frequency can be calculated from the spring constant and the masses of the two atoms involved in the bond.

Rotational energy is found in molecules and in planets. The rotational energy of molecules is even smaller than the vibration energy of the bonds between atoms. Rotational energy is also quantized and the energies confined to

$$E_{\text{ROT}} = \frac{J(J+1)h}{4\pi I} \quad J = 0, 1, 2, \dots, \quad (12)$$

where I is the moment of inertia about an axis.

Electronic energy is the extra energy an outer electron needs to move to a higher electronic orbit. It is also the energy given out when an excited outer electron moves back to its ground state. This energy is radiant energy and sometimes the emission is in the visible region of the electromagnetic spectrum, and we can see it as a color. When sodium atoms are excited (NaCl in a bunsen-burner flame or electrically excited sodium vapor in a street lamp), they emit an intense yellow flame. Electronic energy is quantized, and the energies are related to the frequency, ν , of the emitted electromagnetic energy by

$$E_{\text{ELEC}} = h\nu. \quad (13)$$

The energies of translating atoms and molecules are also quantized but the allowed energy levels are so close together that the energies can be considered

as a continuum, and the equations of Newtonian physics apply to these particles. The allowed energy levels for a particle of mass, m , in a one-dimensional box of length L is given by

$$E_{\text{TRANS}} = \frac{n^2 h^2}{8mL^2} \quad n = 1, 2, 3, \dots \quad (14)$$

There is another type of mechanical energy that is related to steam engines and to expanding gases or vapors. The work done by a gas or vapor when it expands (with a volume change of ΔV), against a pressure p , is

$$w = p\Delta V. \quad (15)$$

As an example, let us consider the work done by a mole of water at its boiling point, as it turns into a vapor (with a volume change of $\Delta V = 30 \times 10^{-3}$ m³), against atmospheric pressure (101325 Pa). The work done is

$$w = (101325 \times 30 \times 10^{-3}) \text{ J} = 3040 \text{ J}.$$

This is the energy that drives steam engines.

To put the other larger mechanical energies into perspective, let us consider the kinetic energy of a moving tennis ball and the potential energy of water about to fall.

The energy of a tennis ball ($m = 60$ g) hit by Serena Williams, traveling at a velocity of 100 km s^{-1} , has a kinetic energy of

$$E_{\text{KE}} = 1/2 mv^2 = (1/2 \times 0.060 \times 100^2) \text{ J} = 300 \text{ J}.$$

The potential energy of 1.0 kg of water, about to fall 100 m under gravity ($g = 9.8 \text{ m s}^{-2}$) is

$$E_{\text{PE}} = mgh = (1 \times 9.8 \times 100) \text{ J} = 980 \text{ J}.$$

Both energies are rather small.

6.2 Chemical Energy

Chemical energy is stored in molecules through the bonds that bind atoms, and some of this energy can be released when chemical reactions take place. In this process, some bonds are broken (involving an input of energy) and some bonds are made (involving a release of energy). The overall effect is either endothermic or exothermic. However, for most reactions, the process is exothermic and energy is released. This release is usually in the form of thermal energy.

The thermal energy associated with a chemical reaction is always measured at constant pressure and is known as the enthalpy of the reaction, $\Delta_r H^\circ$, and relates to conditions of 1 bar pressure (10^5 Pa)

and 298.15 K. As an example, let us consider the reaction between hydrogen and oxygen:



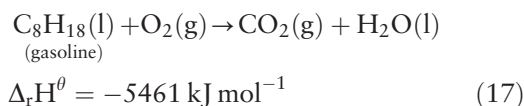
This reaction involves breaking two moles (or molecules) of H–H bonds, one mole (or molecule) of O–O bonds and making four moles of O–H bonds. As the values of the bond energies are known ($\Delta H_{\text{H-H}}^\theta = 436 \text{ kJ mol}^{-1}$, $\Delta H_{\text{O-O}}^\theta = 497 \text{ kJ mol}^{-1}$ and $\Delta H_{\text{O-H}}^\theta = 463 \text{ kJ mol}^{-1}$), it is possible to determine the enthalpy of the reaction, $\Delta_r H^\theta$:

$$\begin{aligned} (\Delta_r H^\theta) &= (-4 \times 463 + 497 + 2 \times 436) \text{ kJ} \\ &= -483 \text{ kJ} \text{ (the negative sign implies an} \\ &\quad \text{exothermic process).} \end{aligned}$$

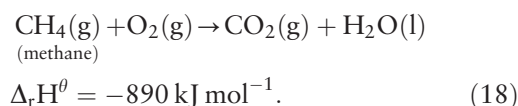
The experimental enthalpy value for the formation of two moles of $\text{H}_2\text{O}(\text{g})$ from two moles of $\text{H}_2(\text{g})$ and a mole of $\text{O}_2(\text{g})$ is indeed -483 kJ .

It is important to know that the sum of the bond energies in a molecule do not make up the total energy of the molecule. The bond energies relate only to chemical reactions. The total energy involves all types of energies, including the energy of binding protons to neutrons and inner electrons and so on. The total energy of a molecule is difficult, if not impossible, to determine.

Chemical energy is the most important type of energy on Earth. It accounts for 96% of all energy sources. These include oil (37%), coal (24%), natural gas (21%), and biomass (14%). These are 1994 figures. Chemical energy supplies the energy for all modes of transport, whether it be ships, cars, trucks, buses, or planes, through the combustion of hydrocarbons such as gasoline and diesel. Chemical energy also supplies energy to most of the electricity power stations in the world through the combustion of natural gas (CH_4), oil, or coal. Typical combustion reactions are



and

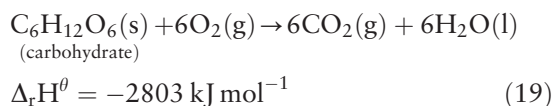


These reactions involve a relatively large amount of energy for a small mass of fuel. The gasoline reaction (Eq. 17) produces 47900 J per gram of n-octane and the natural gas reaction (Eq. 18) produces

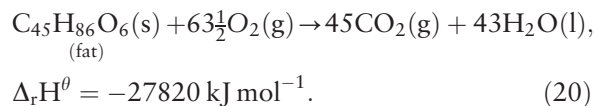
55600 J per gram of methane. The steam engine, by comparison, produces 169 J per gram of H_2O .

Biomass energy refers to the use of plant material as an energy resource. It includes the growing of sugar cane and corn for alcohol production (a viable exercise in Brazil and also in the mid-west United States) and the burning of wood for heat and energy. Much of the energy used in third world countries is of the latter type. Another interesting and useful biomass process is the exploitation of methane gas from landfills. The decomposition of organic material, under anaerobic conditions, produces an equal volume of methane and carbon dioxide gases. The methane gas may be used for heating or to generate electricity. Energy from landfill projects are now commonplace in many countries of the world, notably in the United Kingdom and the United States. The energy is a renewable resource and exploitation does reduce the amount of methane gas (a greenhouse gas, which is 27 times more effective than carbon dioxide) that enters the atmosphere.

The energy used by humans for mechanical activities such as running, lifting weights, pulling, and so forth comes from the combustion of food (carbohydrates, fats, etc.) in much the same way as does the energy of an internal combustion engine. Not a very nice thought. The combustion process in the human body is done through a series of complex reactions, and the overall reaction looks similar to that of burning wood or diesel! Typical carbohydrate and fat reactions are



and



The carbohydrate (glucose) and the fat produce 15.6 kJ g^{-1} and 38.5 kJ g^{-1} , respectively. The fat produces more than twice the energy per gram that does the carbohydrate. You might have seen these energy values in health food literature as 3.7 and 9.2 kcal per gram, respectively. The calorie is an old-fashioned unit for energy and $4.18 \text{ J} = 1 \text{ calorie}$. You might also have seen these energy values written as 3.7 and 9.2 Calories per gram, respectively. The Calorie is a health food unit for a kilocalorie.

Chemical energy is measured in calorimeters and the simplest type is known as a “bomb” calorimeter. The bomb is made of thick steel, and is of the order of 100 mm high with an internal volume of about 300 cm³. The only type of reactions that can be investigated in these calorimeters are combustion reactions. The inside of the bomb contains a sample holder and contacts to attach a high-resistance fuse wire. The contacts are connected to a cable through gas-tight seals, to a battery and a switch, outside the bomb. The bomb is also fitted with an oxygen inlet pipe and a pressure valve, rated at 30 atmospheres. The sample (e.g., a few grams of glucose) is weighed out (m_1) and placed in the sample holder, and the fuse wire is connected to the contacts and bent so that it passes through the sample. Oxygen is pumped into the bomb until the pressure is about 20 atmospheres. The bomb is then placed in a tank of water, which has been weighed (about 3 kg) (m_2) and a thermometer and stirrer placed in the tank. Once the bomb and water has reached equilibrium (30 minutes), the temperature of the water is noted (T_1), the fuse is fired through the external switch and battery, and the combustion reaction (Eq. 19) takes place. The heat of combustion heats the bomb, which in turn warms the water. When an equilibrium has been reached (another 30 minutes), the temperature of the water is noted (T_2). The bomb and the water should have the same change of temperature i.e., $(T_2 - T_1) = \Delta T$.

The energy given out by the glucose sample is balanced by the energy absorbed by the water and the bomb. If the energy per mole of glucose combustion reaction is $\Delta_r U^\theta$, then

$$[(m_g/MM)\Delta_r U^\theta]_{\text{glucose}} = [m_2 s_2 \Delta T]_{\text{water}} + [m_1 s_1 \Delta T]_{\text{bomb}}, \quad (21)$$

where m_g is the mass of glucose, MM is the molar mass of glucose, s_2 is the specific heat of the water (4.18 J g⁻¹ K⁻¹), and m_1 and s_1 are the mass and specific heat of the calorimeter. The product ($m_1 s_1$) is unique for each calorimeter (it is called the bomb's heat capacity) and can be determined from calibration experiments using a standard substance such as benzoic acid. The value of $\Delta_r U^\theta$ is calculated from Eq. (21). This energy refers to a constant volume experiment. The value of $\Delta_r H^\theta$ is usually very close to $\Delta_r U^\theta$. The exact value can be calculated from

$$\Delta_r H^\theta = \Delta_r U^\theta + \Delta n RT, \quad (22)$$

where Δn is the change in the amount of gas in the reaction. For this reaction (see Eq. 19), there are six

moles of O₂(g) on the reactant side and six moles of CO₂(g) on the products side. For this combustion reaction $\Delta n = 0$ and $\Delta_r H^\theta = \Delta_r U^\theta$.

There is another type of chemical energy that should be considered and that is the energy of a phase change. By phase change we mean a solid changing to a liquid or a liquid changing to a vapor. A common example is the change of liquid water to steam. This type of energy has been called the latent heat and is due to intermolecular interactions which bind molecules together as a result of the attractions of dipoles and fluctuating dipoles that are created by the electrons wizzing around molecules. The larger the molecule, the more electrons in the molecule and the greater will be this interaction and the greater will be its latent heat or its enthalpy of phase change. For water, the enthalpies of fusion (melting) at 273.15 K and vaporization at 373.15 K are 6.0 and 40.7 kJ mol⁻¹, respectively. It obviously takes a lot more energy to separate the liquid molecules to form a vapor than it does to separate the molecules in solid water to form a liquid. Much of the energy (coal) used to fire up a steam engine is used to overcome the enthalpy of vaporization of water and create steam from liquid water.

It is possible to store energy using the reversible property of phase equilibria. For example, solar energy can readily be used to melt Glauber salts (Na₂SO₄·10H₂O) because of its low melting temperature (306 K). This molten material will revert to a solid when the temperature drops below the melting point (at night) and will release the stored energy, which is the enthalpy of fusion.

6.3 Electrical Energy

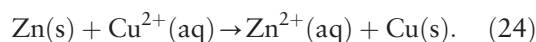
Electrical energy is the most useful of energies as it can be readily transported along copper or aluminum cables for long distances. It can also be conveniently switched on and off. Electrical energy is not unlike the gravitational energy of falling water with the electrons moving under a potential difference replacing water falling under the influence of gravity.

The electrical energy of electrons moving under a potential difference of V with a current i for a time t is

$$E_E = Vit. \quad (23)$$

Electrical energy is readily converted from chemical energy, creating a flow of electrons. Such devices are batteries and one such battery is the Daniel cell. The natural reaction between zinc and a solution of copper sulfate to form zinc sulfate and metallic

copper can be demonstrated by placing a piece of metallic zinc in a solution of copper sulfate whereupon the zinc will begin to dissolve and red metallic copper will be precipitated. The overall reaction can be represented as the displacement of the copper ions from solution as metallic copper by the formation of zinc ions from the zinc in the process of which electrons are transferred from the zinc to the copper ions:



This reaction can be made to take place in a more controlled way in an electrochemical cell (the Daniel cell), in which zinc and copper electrodes are each immersed in a solution of its sulfate and are separated by a membrane which allows the diffusion of ions. If an external metallic connection (a piece of wire) is made between copper and the zinc rods, an electric current (electrons) can flow between the two rods; zinc metal will dissolve to form zinc ions and the copper ions will deposit as metal on the copper rod. The completion of the flow of electric charge will be achieved by the movement of zinc ions through the membrane and the movement of copper ions toward the copper rod. This electrochemical cell is a prototype of the many different kinds of battery that are used to power various devices from hearing aids to video cameras and flashlights in daily use.

6.4 Radiation of Electromagnetic Energy

Radiation energy is a unique form of energy. It is characterized by a wave with a wavelength and a frequency:

$$c = \lambda\nu. \quad (25)$$

where c is the speed of light ($3 \times 10^8 \text{ m s}^{-1}$). The magnitude of radiant energy is given by

$$E_R = h\nu. \quad (26)$$

Radiation energy is emitted from an object in the form of electromagnetic waves, which consist of

electric and magnetic fields whose amplitudes vary with time. These waves propagate out from the object at the speed of light. Familiar examples of electromagnetic radiation include gamma rays, X-rays, ultraviolet waves, visible waves, infrared waves, microwaves, TV waves, and radio waves. The full range of the electromagnetic spectrum spans cosmic rays at the short wavelength and radio waves at the long wavelength end (see Fig. 2). It is the wavelength that categorizes the different types of radiation. Only a small region of the spectrum is visible to the human eye. That part is called the visible spectrum. The radiant energy that causes sunburn and possible cancer is the ultraviolet rays. These are of shorter wavelength (hence greater energy) than the visible rays and the infrared rays. It is the latter that gives us the sensation of warmth. These rays have similar energies to the vibration energies of the molecules in our bodies and hence are readily absorbed.

The main source of electromagnetic radiation on Earth is the sun. This radiation involves only the UV, visible and IR parts of the electromagnetic spectrum. Sunlight is the mainstay and support for all plant and animal life on earth. The total amount of sunlight reaching the earth each day is $3.11 \times 10^{23} \text{ J}$. This is an enormous amount of energy and 11 days of sunshine over the whole earth is equivalent to all the fossil fuel (gas, oil, coal) in the earth. However, the earth's surface is large and the amount of sunshine reaching the earth peaks at about 1.2 kW m^{-2} on the equator at midday. This is really a small amount of energy. A solar energy collector of almost $8 \text{ km} \times 8 \text{ km}$ is needed to create a reasonably large electricity generating power station of 1.0 GW (assuming a conversion factor of 10%). Even this is unrealistic as it assumes all the solar collecting panels are dust free and clean and that the sun shines brightly 8 hours every day. Solar energy is a dilute form of energy, but it is a renewable and clean energy source and will certainly be an important source in the future as the earth's fossil fuel resources become exhausted.

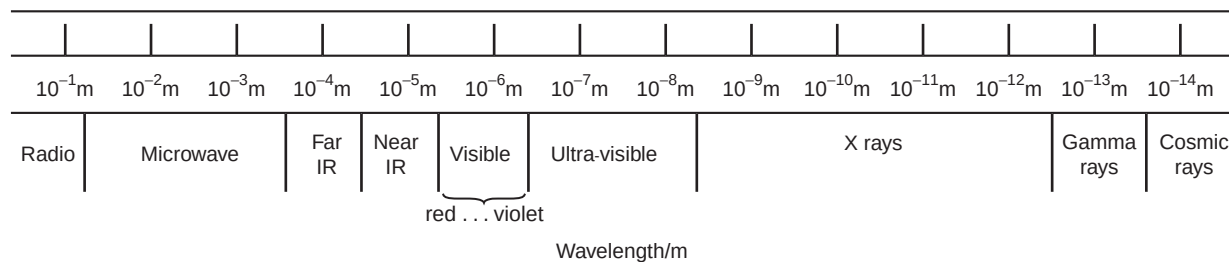
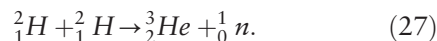


FIGURE 2 The electromagnetic spectrum.

6.5 Nuclear Energy

Nuclear energy comes from reactions involving the nuclei of atoms. In these reactions, the subatomic particles, neutrons and protons of the reacting species, find a more stable arrangement in the products. For nuclear reactions to take place, the reacting species must penetrate the negative electron shield surrounding the nucleus and then penetrate the positively charged nucleus. Thus, the reacting species must have large kinetic energies and preferably should have zero charge. That is why neutrons (zero charge) are often used. In chemical reactions, by contrast, the nucleus is never part of the process and the rearrangement that takes place relate only to the outer electrons of the atoms.

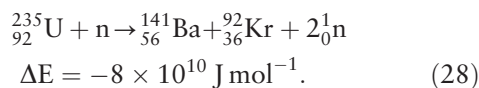
A simplified version of a typical nuclear reaction taking place in our sun is the reaction between two hydrogen isotopes (deuterium) to form an isotope of helium:



Reactions of this type produce large amounts of energy. This particular reaction produces 3×10^8 kJ per gram of deuterium. This is significantly greater (by about six orders of magnitude) than the magnitude of the chemical reactions we met earlier (Eqs. 16 through 20). The deuterium reaction in Eq. (27), involving a fusion of nuclei, is called a fusion reaction. The energy that is released in this reaction and in other fusion reactions comes in many forms: radiation energy (δ -radiation, U-V, IR, and visible light), kinetic energy, and thermal energy.

Another type of nuclear reaction involves the splitting of a nuclei by high-energy neutrons. Such reactions are referred to as fission reactions and the material that undergoes the splitting is called fissile material. Uranium-235, uranium-233, and plutonium-239 are examples of fissile material and all of these are used in nuclear power plants. To ensure fission, the kinetic energy of the neutrons must be high enough to enter the nucleus and split it. For many isotopes of many elements, the arrangement of neutrons and protons in their nuclei is sufficiently stable to ensure that the bombardment by high-energy neutrons has no effect.

A typical fission reaction is



The amount of energy is again more than six orders of magnitude greater than typical chemical reactions used in power generation.

If the number of neutrons and the concentration of uranium-235 nuclei is high enough, the nuclear reaction becomes self-sustaining. The minimum mass is known as the critical mass. If the reaction is too vigorous, an explosion will occur. This is what happened in Nagasaki and Hiroshima in 1945 and in Chernobyl in 1986. Nuclear reactors for power generation have strict mechanical control over the uranium-235 enriched rods and moderators to reduce the possibility of runaway nuclear reactions (chain reactions).

The release of energy in nuclear reactions always involves a loss of mass. The relationship between the energy output, ΔE , and the loss of mass, Δm , is given by Einstein's famous equation:

$$\Delta E = \Delta mc^2, \quad (29)$$

where c is the speed of light, $3.00 \times 10^8 \text{ m s}^{-1}$.

In a nuclear reactor, the energy heats a heat exchanger (e.g., water or sodium), which transfers heat to water to make steam, which is then used to turn a turbine and create electricity. Nuclear energy makes up 2% of all the energy used in the world. France has taken the lead here, and 80% of its electricity is produced from nuclear energy. In 1995 there were 400 nuclear powerstations in more than 30 countries. This represented about 10% of the world's electricity.

There are two major problems with fission nuclear power: (1) the products are radioactive and need to be contained for hundred of years because of their long half-lives and (2) there is the possible danger of a runaway chain reaction resulting in an explosion and the subsequent contamination of the environment with radioactive fallout. Solutions to these problems will bring about a resurgence of nuclear power production worldwide, especially as the world's fossil fuel reserves are beginning to run down and the temptation to extract large amounts of nuclear energy from small masses of fuel becomes irresistible.

Fusion reactions would, in theory, be the saving grace of the nuclear industry because the products are essentially free of radioactivity and, moreover, the fuel (deuterium) is available from seawater. Unfortunately, the fusion process is difficult to achieve in practice because of the difficulty in supplying the deuterium molecules with enough energy for the nuclei to strike each other with enough force to allow them to interact and for the neutrons and protons to rearrange. Much work is currently being done in a number of laboratories

around the world, but the process is costly and technically challenging.

Geothermal energy (the energy that heats natural hot-water springs) is due to energy from radioactive sources deep inside the earth. This energy is being tapped in Iceland and in Italy.

The measurement of the energy from nuclear reactions can be determined from a special type of calorimeter, but it can also be calculated from a detailed analysis of the relative atomic masses of the products and reactants, followed by the application of Einstein's equation (Eq. 29).

7. CONCLUSION

A knowledge of the forms of energy available on our earth (and a knowledge of how to measure and calculate the magnitudes of each type of energy) gives us the necessary data to begin to make the important decisions as to which type of energy can be exploited in the future. This is particularly pertinent as the world's fossil fuel reserves begin to dwindle.

SEE ALSO THE FOLLOWING ARTICLES

Combustion and Thermochemistry • Conservation of Energy Concept, History of • Electrical Energy and Power • Electromagnetism • Energy in the History and Philosophy of Science • Entropy • Heat Transfer • Mechanical Energy • Nuclear Power, History of • Temperature and Its Measurement • Thermodynamic Sciences, History of • Work, Power, and Energy

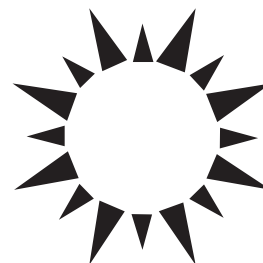
Further Reading

- Atkins, P. (2003). "Galileo's Finger—The Ten Great Ideas of Science." Oxford University Press, Oxford.
- Atkins, P., and de Paula, J. (2002). "Physical Chemistry." Oxford University Press, New York.
- Atkins, P., and Friedman, R. S. (1997). "Molecular Quantum Mechanics." Oxford University Press, Oxford.
- Boyle, G. (ed.). (1996). "Renewable Energy: Power for a Sustainable Future." Oxford University Press, Oxford.
- Ramage, J. (1997). "Energy, a Guide Book." Oxford University Press, Oxford.

Fuel Cells

NIGEL BRANDON

Imperial College
London, United Kingdom



-
1. What Is a Fuel Cell?
 2. Fuel Cell Types
 3. Fuel Selection
 4. Fuel Cell Applications
 5. Fuel Cell Costs
 6. Environmental Benefits

Glossary

- anode** The fuel electrode, where electrons are donated as the fuel is oxidised. The negative electrode in a fuel cell.
- balance of plant** All the components that make up a power system, other than the fuel cell stack, for example, the control system, reformer, power conditioner, or compressor.
- bipolar plate** A dense, electronic (but not ionic) conductor that electrically connects the anode of one cell to the cathode of another. It also distributes fuel or air to the electrodes.
- cathode** The air electrode, where electrons are accepted and oxygen is reduced. The positive electrode in a fuel cell.
- electrocatalyst** A catalyst incorporated into both the anode and the cathode to promote the electrode reactions.
- electrolyte** A dense ionic (but not electronic) conductor. Each fuel cell type is distinguished by the nature of the electrolyte used within it.
- interconnect** Another term for the bipolar plate (see above).
- membrane electrode assembly (MEA)** An assembled anode, cathode, and electrolyte.
- open circuit voltage** The voltage from the fuel cell when no current is being drawn from it.
- overpotential** The voltage loss within an operating fuel cell due to electrode kinetics, mass transport limitations, and component resistance.
- positive electrolyte negative (PEN)** The assembled cathode, electrolyte, and anode of a solid oxide fuel cell.
- reformer** The fuel processor that converts the fuel to a hydrogen-rich gas suitable for the fuel cell.
- stack** An assembly of many individual fuel cells, complete with gas manifolds and electrical outputs.

A fuel cell is a device for directly converting the chemical energy of a fuel into electrical energy in a constant temperature process. Fuel cells operate on a wide range of fuels, including hydrogen, and are seen as a clean, high-efficiency power source and an enabling technology for the hydrogen economy.

1. WHAT IS A FUEL CELL?

The fuel cell can trace its roots back to the 1800s when a Welsh-born, Oxford-educated barrister, Sir William Robert Grove, realized that if electrolysis, using electricity, could split water into hydrogen and oxygen, then the opposite would also be true. Grove subsequently built a device that would combine hydrogen and oxygen to produce electricity—the world's first gas battery, later renamed the fuel cell.

It was another British scientist, Francis Thomas Bacon, who was the first to develop a technologically useful fuel cell device. Bacon began experimenting with alkali electrolytes in the late 1930s, using potassium hydroxide instead of the acid electrolytes used by Grove. Bacon's cell also used porous "gas-diffusion electrodes" rather than Grove's solid electrodes. These increased the surface area over which the electrochemical reactions occurred, improving power output. In the early 1960s, Pratt and Whitney licensed the Bacon patents and won the National Aeronautics and Space Administration contract for the Apollo spacecraft utilizing onboard liquid hydrogen and oxygen to provide both power and water. Innovation and development has continued since that time with pioneering work by, for example, Westinghouse, Union Carbide, and General Electric, among others.

Today, potential applications for fuel cells range from battery replacement in consumer goods and portable computers, through residential scale combined heat and power (CHP), to distributed energy generation. However, the key problem limiting the

significant commercial uptake of fuel cells is cost, and it is cost reduction, together with the need to demonstrate high levels of reliability and durability, that are the primary concerns of fuel cell developers today.

As demonstrated by Grove's gas battery, a fuel cell is analogous to a battery, but one that is constantly being recharged with fresh reactants. In a similar manner to a battery, each fuel cell comprises an electrolyte, which is an ionic conductor, and two electrodes (the negative anode and positive cathode), which are essentially electronic conductors.

The nature of the ion transfer varies between the different types of cell, but the principle shown in Fig. 1 for a polymer electrolyte membrane fuel cell (PEMFC) is representative. In this case, hydrogen is fed to the anode of the cell where it splits into a proton and electron, the former passing through the electrolyte and the latter forced around an external circuit where it drives a load. The proton and electron combine with oxygen from the air at the cathode, producing pure water and a small amount of heat. The overall reaction is thus



The anode is then the negative electrode of the device, and the cathode the positive.

The fuel cell differs from a conventional heat engine (such as the internal combustion engine or the gas turbine), in that it does not rely on raising the temperature of a working fluid such as air in a combustion process. The maximum efficiency of a heat engine is subject to the Carnot efficiency limitation, which defines the maximum efficiency that any heat engine can have if its temperature extremes are known:

$$\text{Carnot efficiency} = (T_H - T_L)/T_H, \quad (2)$$

where T_H is the absolute high temperature and T_L is the absolute low temperature. In contrast, the theoretical efficiency of a fuel cell is related to the ratio of two thermodynamic properties, namely the chemical energy or Gibbs free energy (ΔG^0) and the enthalpy (ΔH^0) of the fuel oxidation reaction:

$$\text{Maximum fuel cell efficiency} = \Delta G^0/\Delta H^0. \quad (3)$$

Figure 2 provides an illustration of the theoretical efficiency possible from a fuel cell running on hydrogen and air as a function of temperature and compares this to the Carnot efficiency of a heat engine at the same temperature, assuming a low temperature of 25°C. As the Gibbs free energy for reaction (1) falls with increasing temperature, while the enthalpy remains largely unchanged, the

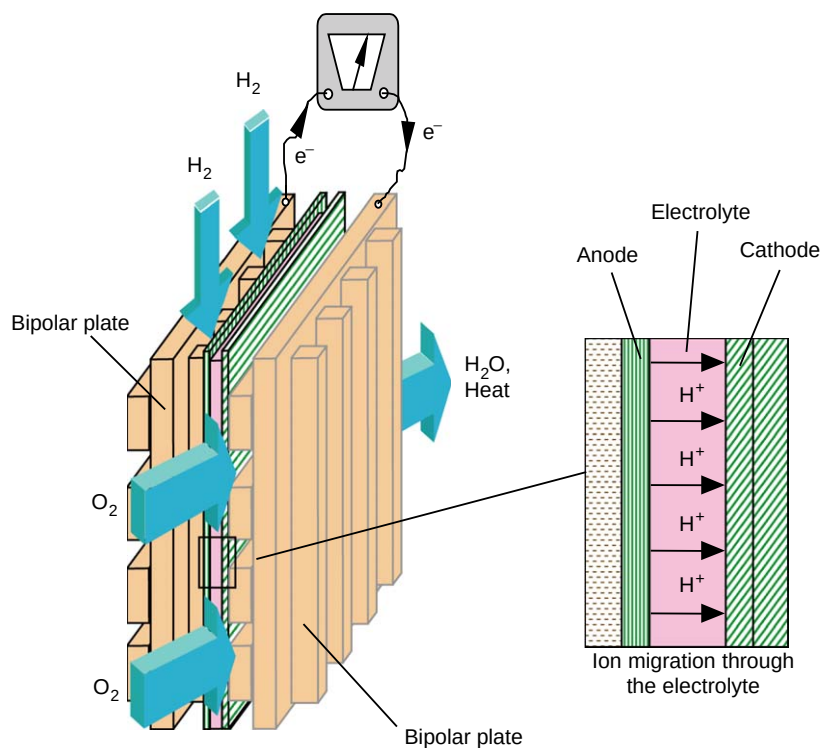


FIGURE 1 Schematic illustration of a polymer electrolyte fuel cell.

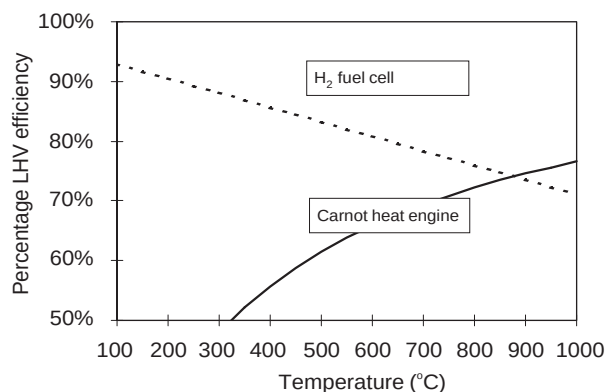


FIGURE 2 Maximum efficiency (on a lower heating value basis) of a hydrogen/air fuel cell, and a heat engine limited by the Carnot cycle, as a function of temperature.

theoretical efficiency of the fuel cell falls with increasing temperature. Indeed, at high temperatures the theoretical efficiency of a heat engine is higher than that of a hydrogen driven fuel cell. However, because of the need for motion in a heat engine, either rotary or linear, significant materials issues are associated with operating them at high temperatures, from the perspective of both durability and cost. Fuel cells do not have moving parts operating at high temperatures and thus are less susceptible to this problem.

However, other factors play a role in determining the actual efficiency of an operating fuel cell, in particular operating temperature, fuel type, and materials selection. For example, losses associated with the kinetics of the fuel cell reactions fall with increasing temperature, and it is often possible to use a wider range of fuels at higher temperatures. Equally, if a fuel cell is to be combined with a heat engine, for example, in a fuel cell/gas turbine combined cycle, then high fuel cell operating temperatures are required to maximize system efficiency. All these factors mean that there is considerable interest in both low-temperature and high-temperature fuel cells, depending on the application.

Figure 3 illustrates the shape of the current-voltage characteristics that would be expected from a typical fuel cell. When no current is being drawn from the fuel cell, the cell voltage is at a maximum, termed the open circuit voltage (E), which increases with the partial pressures of the fuel and air gases and decreases with increasing temperature, according to the Nernst equation:

$$E = E^0 - \frac{RT}{nF} \ln\left(\frac{p_{H_2O}}{p_{H_2} p_{O_2}^{0.5}}\right). \quad (4)$$

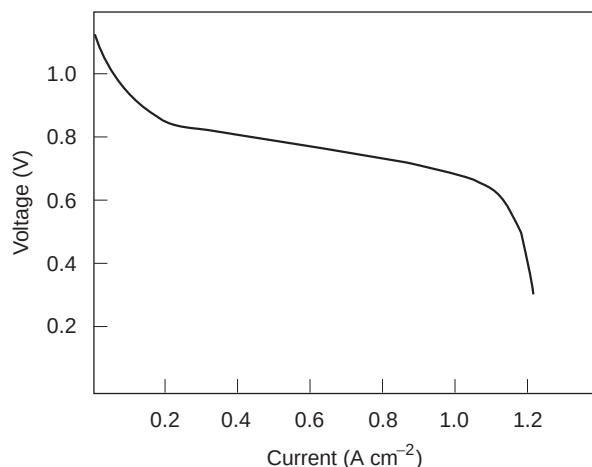


FIGURE 3 Schematic illustration of a typical fuel cell current-voltage curve.

where E^0 is related to the Gibbs energy for the reaction via:

$$\Delta G_{298}^0 = -nE^0F, \quad (5)$$

where n is the number of electrons involved, 2 for reaction (1), and F is the Faraday constant (96495 C mol^{-1}). As the value of ΔG_{298}^0 for the reaction of hydrogen with oxygen to form water is -229 kJ mol^{-1} , then an open circuit voltage of around 1.2 V would be expected from a hydrogen/air fuel cell operating at near ambient temperatures under standard conditions.

As current is drawn from the cell, additional irreversible losses result in a decrease in the cell voltage (E_{cell}), according to:

$$E_{\text{cell}} = E - iR - \eta_a - \eta_c, \quad (6)$$

where iR refers to ohmic losses within electrodes, interconnects, and current take-off's due to the finite resistance of the materials used, η_a refers to the overpotential at the anode, reflecting losses due to both electrode kinetics and mass transport limitations, and η_c refers to the equivalent overpotential at the cathode. Different loss terms are reflected in different regions of the current-voltage curve, with the initial fall in voltage reflecting electrode kinetics, the central linear region being dominated by iR losses, before mass transport limitations dominate at high current densities.

Hence, under load a single cell produces a reduced cell voltage due to the losses highlighted here. It is the task of the fuel cell designer to minimize these losses by judicious selection of materials and cell geometry. Electrodes, for example,

are required to be porous to enable gas transport to and from the active catalyst region adjacent to the electrolyte, yet also to be good electronic conductors. As such, there is considerable research into electrode structures and materials, so as to minimize overpotentials. Another choice facing the fuel cell system designer is the operating cell voltage.

At any point other than very close to the open circuit voltage (where parasitic loads become significant), then, the higher the cell voltage, the higher the cell efficiency—but the lower the power density. In general, a cell voltage of about 0.6 to 0.7 V is taken as the operating point. This is a compromise between efficiency (a high-efficiency reducing fuel consumption) and power density (a low-power density increasing stack size and weight, and hence capital cost). To achieve a useful output power, individual cells are connected together in a “stack”. This is achieved using an interconnect or bipolar plate, which joins the anode of one cell to the cathode of the next cell. The interconnect also separates and often distributes the fuel and oxidant. An example of a PEMFC stack is shown in Fig. 4. In this case, the interconnect is known as a flow field plate, and the combination of electrolyte and electrodes as the membrane electrode assembly (MEA). Other fuel cell types use different terminology (positive electrolyte negative or PEN for an SOFC, for example) for these components, but their function remains the same.

The fuel cell stack is then incorporated into a system, which meets the demand of a specific application. Fuel cell systems can include a range of

balance of plant (BoP), depending on the fuel used, the application, and the fuel cell type. BoP can include a fuel processor, compressor, power conditioner, and control system. Together, these can add significant cost to the overall system. As well as high efficiency and low emissions levels (discussed further in Section 6), a number of further advantages are cited for fuel cells:

1. They are quiet and involve few moving parts, other than some fans or a compressor to blow air into the device, and hence do not require much maintenance.
2. They are modular, such that several can be coupled together to increase the capacity of a system, but can be mass-manufactured to reduce cost.
3. They exhibit an increase in efficiency at low loads, unlike a heat engine, which normally only exhibits maximum efficiency around the design point for the device.

2. FUEL CELL TYPES

There are five main classes of fuel cell, each with differing characteristics, and differing advantages and disadvantages. The five types are summarized in Table I, with each taking the name of the electrolyte used in its fabrication. These five classes of fuel cell can essentially be further grouped into one of two classes, distinguished as either low-temperature fuel cells (AFC, PEMFC, PAFC), or high-temperature fuel

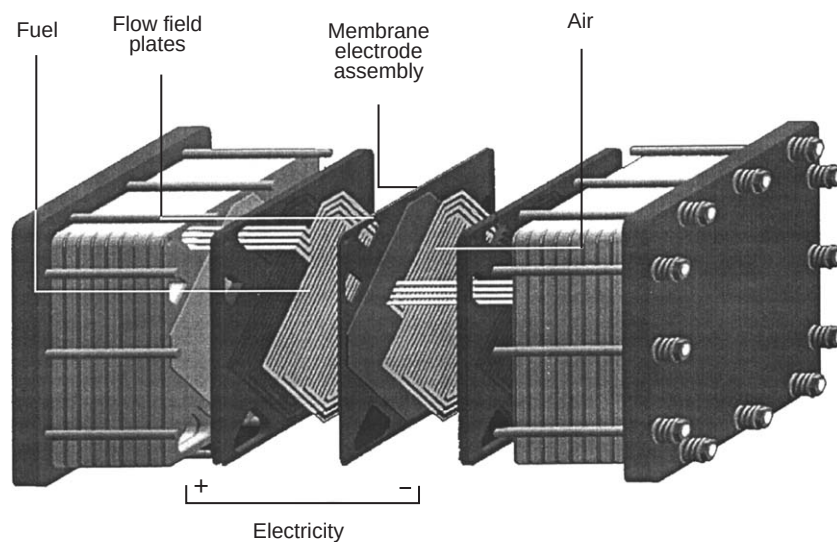


FIGURE 4 Example of a polymer electrolyte fuel cell (PEMFC) stack.

TABLE I
Summary of the Five Main Fuel Cell Types

Fuel cell type	Alkaline	Polymer	Phosphoric acid	Molten carbonate	Solid oxide
Acronym	AFC	PEMFC	PAFC	MCFC	SOFC
Operating temperature	60–250°C	80–120°C	150–220°C	600–700°C	600–1000°C

cells (MCFC, SOFC). The low-temperature fuel cells can be distinguished by the following common characteristics:

- They require a relatively pure supply of hydrogen as a fuel (e.g., AFCs are sensitive to carbon dioxide, PEMFCs to carbon monoxide). This usually means that a fuel processor and some form of gas cleanup is required, adding cost and reducing system efficiency.
- They incorporate precious metal electrocatalysts to improve performance.
- They exhibit fast startup times.
- They are available commercially (AFC, PAFC) or are approaching commercialization (PEMFC).

In contrast, high-temperature fuel cells can be classed as having the following general features:

- They have fuel flexibility; they can be operated on a range of hydrocarbon fuels at high efficiency.
- Their increased operating temperature reduces the need for expensive electrocatalysts.
- They can generate useful “waste” heat and are therefore well suited to cogeneration applications.
- They exhibit slow startup times.
- They can require expensive construction materials to withstand the operating temperature, particularly in the balance of plant.
- Reliability and durability are concerns, again due to the operating temperature.
- They are suitable for integration with a gas turbine, offering very high efficiency combined cycles.
- They are further from commercialization, though a significant number of demonstrators are in operation.

The previous summary is, of course, a generalization. For example, a significant amount of work is ongoing to develop intermediate temperature SOFCs (IT-SOFCs), which operate at temperatures toward 600°C and which aim to overcome a number of the disadvantages of high-temperature fuel cells cited here. Similarly, work also is being carried out to develop low-temperature fuel cells that operate directly on methanol rather than clean hydrogen,

eliminating the need for a fuel processor. Nonetheless, it is reasonable to distinguish between the low-temperature and high-temperature variants as being broadly best suited to transportation and stationary power applications, respectively, applications that place differing requirements on the fuel cell stack and system. Each of the main fuel cell types is discussed in more detail in the following sections.

2.1 Alkaline Fuel Cells (AFCs)

The alkaline fuel cell has a long history in the space program. It is still used in the space shuttle in an expensive guise, producing power for the onboard systems by combining the pure hydrogen and oxygen stored in the rocket-fuelling system.

The electrolyte is concentrated (85 wt%) potassium hydroxide (KOH) in AFCs operated at 250°C, or less concentrated (35 to 50 wt%) KOH for operation below 120°C. The electrolyte is retained in a porous matrix (typically asbestos), and electrocatalysts include nickel and noble metals. The main difficulties with this fuel cell type are (1) carbon monoxide, always found in hydrogen produced by reforming hydrocarbon or alcohol fuels (see Section III), is a poison to the precious metal electrocatalysts, and (2) carbon dioxide (in either fuel or air) will react with the KOH electrolyte to form potassium carbonate. As such, applications are essentially limited to those where either pure oxygen and hydrogen can be used.

2.2 Polymer Electrolyte Fuel Cells (PEMFCs)

PEMFCs have high-power density, rapid startup, and low-temperature operation (around 80 to 120°C), and so are ideal for use in applications such as transport and battery replacement. The electrolyte used is a proton conducting polymer. This is typically a perfluorinated polymer, though other hydrocarbon-based membranes are under development in an attempt to reduce cost or to enable

operation at temperatures approaching 200°C. The catalytically active layer sits adjacent to the membrane, supported on a PTFE treated carbon paper, which acts as current collector and gas diffusion layer. For operation on pure hydrogen, platinum is the most active catalyst, but alloys of platinum and ruthenium are used when higher levels of carbon monoxide are present (CO is a poison in all low-temperature fuel cells). Water management in the membrane and electrodes is critical for efficient performance, because the membrane must be hydrated, while avoiding flooding of the electrode pores with water.

The present cost of PEMFC systems is high. While mass-manufacturing techniques will help bring down those costs, further technical innovation is also needed. The potential for replacing diesel standby generators that are often noisy, expensive, and unreliable is engendering great interest. As many countries move toward liberalization of their energy systems, the opportunities for decentralized generation are also growing. Demonstrator plants of 250 kWe are also being produced and operated. These run on natural gas using a reformer, for the main part, and offer very low emissions but high efficiency as their benefit.

2.3 Phosphoric Acid Fuel Cells (PAFCs)

Phosphoric acid fuel cells have been very successful in fuel cell terms over the past 5 years as stationary cogeneration plants, with more than 220 commercial power plants delivered. The PAFC fleet has demonstrated over 95%+ availability, and several units have passed 40,000 hours operation. PAFCs operate at 150 to 220°C, using a 100% phosphoric acid electrolyte retained in a silicon carbide matrix. Generally platinum electrocatalysts are used in both anode and cathode.

Typical applications lie in hospitals, where the waste heat can be used in laundry and other areas and where consistent and reliable power is required; in computer equipment power provision, where the absence of power surges and spikes from the fuel cell enables systems to be kept running; and in army facilities and leisure centers that have a suitable heat and power requirement. While the PAFC is by far the most commercial of the fuel cells to date, it may well be superseded in the longer term by PEMFC plants that can potentially be produced more cheaply and by SOFC or MCFC plants that have more useful heat output and that operate at higher efficiencies on hydrocarbon fuels such as natural gas.

2.4 Molten Carbonate Fuel Cells (MCFCs)

The electrolyte in the MCFC is usually a combination of alkali carbonates, retained a ceramic LiAlO_2 matrix. The temperature of operation is 600 to 700°C, where the alkali carbonates form a highly conducting molten salt, with carbonate ions providing the means for ionic conduction. The increased temperature of operation means that precious metal electrocatalysts are not needed, and generally nickel anodes and nickel oxide cathodes are used. A design constraint with the MCFC is the need for CO_2 recirculation, meaning that it is difficult to operate below the 100 kWe's scale, removing the market in micro combined heat and power (micro-CHP).

2.5 Solid Oxide Fuel Cells (SOFCs)

Solid oxide fuel cells operate at elevated temperatures, generally above 800°C for the all ceramic high-temperature variant and in the range 600 to 800°C for metal-ceramic intermediate-temperature solid oxide fuel cells. The electrolyte is a dense ceramic, usually yttria stabilized zirconia (YSZ), which is an oxide ion conductor at elevated temperatures. The cathode is typically a perovskite material such as strontium doped lanthanum manganite, often mixed with the YSZ in the form of a composite. The anode is a cermet of nickel and YSZ. The main difference between the high temperature SOFC and the IT-SOFC lies in (1) the thickness of the electrolyte, which tends toward 20 μm thick films for IT-SOFCs to reduce ionic resistance, and (2) the interconnect material, with stainless steel being used at the lower temperatures of the IT-SOFC, whereas more expensive high chrome alloys, or oxides such as lanthanum chromite, are needed at higher temperatures.

SOFCs lend themselves to applications in which their high-temperature heat can be used. This heat can be used in two basic ways: for heating processes such as those in industry or in homes or for integration with turbines in hybrid cycles for very high efficiency electricity production. Recent advances in microturbines have led to the concept of a combined power plant of 250 kWe with an efficiency approaching 70%. Specific applications in which SOFCs may be used are in decentralized electricity generation of 250 kWe to 30 MWe, industrial cogeneration of 1–30 MWe, or domestic applications of 1–5 kWe. Intermediate temperature SOFCs are also of interest for vehicle auxiliary power unit (APU) applications, operating on diesel or gasoline. Carbon

monoxide is not a poison for SOFCs, meaning that a wide range of fuels can be used, together with a simpler, and therefore cheaper, fuel processor. It is also possible to recuperate heat from the fuel cell within the fuel reformer, improving system efficiency when compared to low temperature fuel cells when operating on hydrocarbon fuels.

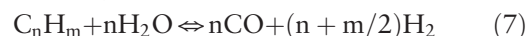
3. FUEL SELECTION

All fuel cells can run on hydrogen as a fuel. This is combined with oxygen, normally fed into the fuel cell as air, to form water, as shown in reaction (1). However, high-temperature fuel cells can also run on other fuels, especially hydrogen-rich gases, and some low-temperature systems are able to run on specific liquid fuels. The application of the fuel cell system often determines on which fuel it will run. Although fuel cells require comparatively clean fuels, they are also flexible in which ones they will accept. High-temperature fuel cells will operate directly on hydrogen-rich gases such as methane and are being successfully tested on forms of biogas. They, and their lower temperature cousins, will also accept any form of gas produced from liquid or solid fuels, provided that it contains a high percentage of hydrogen. This allows for great flexibility according to location and enables the efficient conversion of many indigenous or waste resources. This is discussed in more detail in the following for two applications, stationary power generation and transportation.

3.1 Stationary Power Generation

The common fuel for those commercial stationary systems that exist (mainly PAFC based) is natural gas, which is reformed by a separate steam reformer before the hydrogen is fed into the fuel cell stack. Demonstration PEMFC systems also use this method. In contrast, high-temperature fuel cells are able to operate directly on some gas streams, though they may need to be cleaned up by the removal of sulfur and higher hydrocarbons. The advantage is that high-temperature cells are able to reform the fuel (methane, for example) directly on the anode of the fuel cell because they have sufficient thermal energy and catalyst present to do so. Both steam reforming and partial oxidation reforming can be used. The former is preferred on the grounds of efficiency, though partial oxidation may be needed in an external reformer or for startup:

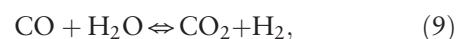
Steam reforming:



Partial oxidation:



Good thermal integration by incorporating the reformer within a high-temperature fuel cell stack offers significant efficiency gains over an external reformer. However, the internal reforming process can cause severe temperature gradients across the stack, and many designers prefer to use a separate catalyst bed, distributed within the fuel cell stack. These processes are called direct and indirect internal reforming, respectively. As well as the reformer, low temperature fuel cells also require a high- and low temperature shift reactor to carry out the water gas shift reaction of CO to CO₂:



together with selective oxidation, to reduce the CO content to parts per million levels.

Fuel cells will also run on reformat from other fuel sources, and there are successful demonstration PAFC systems using gas from landfill sites or sewage farms.

3.2 Transportation

The fuelling problem in transport is more complex than in stationary systems, because there are far more severe size, weight, cost, and performance constraints. Vehicles could be fuelled using pure hydrogen, stored onboard the vehicle. Indeed, this is how most demonstration buses are supplied, making them true zero-emission vehicles. However, there is rarely provision for hydrogen fuelling, so if fuel cells are to be used in cars there may have to be another solution.

Like stationary systems, transport fuel cells (usually PEMFC) can operate on a hydrogen-rich reformat. This is less efficient than using pure hydrogen but may be a pragmatic approach to the fuelling issue. Methanol has been suggested as a good compromise fuel. Although it also requires a supply infrastructure to be developed, it is a liquid and can be handled in a similar way to gasoline, though concerns have been raised regarding its toxicity level. It is also comparatively easy to process into hydrogen. Gasoline can also be reformed using a sequence of autothermal, partial oxidation, and water gas shift reformers, though the resulting reformat stream contains less hydrogen than from methanol and reduces the system performance somewhat more.

The fuelling question remains significant. If methanol turns out to be the fuel of choice, then not only will methanol reformers with sufficiently high performance and low cost have to be integrated into all fuel cell cars, but the problem of supplying methanol to the consumer will also have to be addressed. However, the consensus of opinion in this sector seems to be moving toward hydrogen as the fuel of choice. For this to become a reality, hydrogen storage systems need to be improved, though it is possible to use compressed hydrogen in cylinders designed for compressed natural gas. Equally important are the standards and regulatory decisions that have to be made to allow hydrogen vehicles on the road.

In some cases, fuels other than hydrogen can be used within the fuel cell without reforming into hydrogen. Significant effort, for example, is being applied to the development of a direct methanol fuel cell that can use methanol as a fuel directly without any preprocessing. While this decreases system complexity, it makes the fuel cell more difficult to design and could result in lower efficiency. However, this route is particularly attractive for microfuel cells, which aim to replace battery technology.

4. FUEL CELL APPLICATIONS

There are many areas in which fuel cells could potentially be used to replace conventional power equipment, discussed under the broad headings of stationary power generation, transport, and battery replacement.

4.1 Stationary Power Generation

Even within stationary applications there are a number of distinct divisions, though the most important ones have to do with temperature and the amount of waste heat that can be used. In general, plant sizes of several hundred kWe's (enough for a leisure center or small office block) down to 1 to 5 kWe (a single house or a portable power unit) are under development.

High-temperature systems can be used in more demanding applications where larger systems are required or additional heat is useful. Such systems are also more efficient when operating on fuels such as natural gas, as they enable internal reforming. While heat of 80°C or more can be used for space heating, hot water and possibly absorption chillers to allow for cooling, industrial steam raising, or gas

turbine bottoming cycles require temperatures of at least 500°C. This not only allows for the potential of generating extra electrical power and thus improving the overall system efficiency to nearly 70%, but also the possibility of using cogenerated heat and increasing total energy efficiency to 90%. Either of these options brings down the cost per unit of energy even if the capital cost of the system is high, but can only be effectively based around a high-temperature fuel cell variant.

4.2 Transport

Transport applications tend to demand rapid startup and instant dynamic response from fuel cell systems, so a high-temperature fuel cell is unlikely to be competitive as the main engine in applications such as cars and buses. The prime candidate for these vehicle propulsion systems is the PEMFC, which exhibits both of the above characteristics while also having very high power density. This is important as it must also occupy a similar amount of space to an internal combustion engine. Of recent interest has been the development of auxiliary power units for vehicles, in which the fuel cell meets the onboard electric load of the vehicle. Both PEMFCs and ITSOFCs are under development for this application.

AFC systems have traditionally been used in space applications by NASA—in the *Gemini*, *Apollo*, and space shuttle programs—but are also being investigated for certain transport applications, mainly in vehicles with limited duty cycles such as delivery vehicles and fork lifts.

Of course, transport is not confined to the car and aerospace markets—locomotives, ships, scooters, and a whole host of other applications offer potential for a variety of fuel cell systems. For ships and trains, where the application is almost akin to having a stationary power plant running all the time, startup and system dynamics are less important than noise, emissions, fuel consumption, and vibration. Serious investigations are under way as to the benefits of installing SOFCs, for example, on ships to remove the marine diesel—traditionally a source of heavy pollution, very high noise levels, and damaging vibration. However, in these cases the preferred fuel would be heavy fuel oil, one of the most difficult to process for a fuel cell application.

PEMFC systems have also been commissioned for a number of the world's navies—Ballard and Siemens have each been active in their respective countries putting PEMFCs into submarines.

4.3 Battery Replacement

An area in which it is possible that fuel cells may break through to commercialization in the very near future is in the replacement of conventional batteries. Battery power for laptop computers, mobile phones, and many other devices is expensive and often inconvenient if recharging is required every few hours. A small fuel cell with an equally small fuel source could potentially operate for longer than a battery, but with refueling only taking a few minutes instead of many hours. Batteries cost much more for the amount of power they can supply than almost any other application, and this gives the fuel cell a good chance of entering the market, although it is still expensive. The ideal fuel cell for these applications is the PEMFC, which not only has the high power density required for miniaturization, but also has the least challenging electrolyte management issues of the low-temperature fuel cells.

4.4 The Fuel Cell Industry

The recent upsurge of interest in fuel cell systems has led to rapid growth in industries developing fuel cell technology around the world. As well as fuel cell technology developers, there are an even larger number of component suppliers, which are not dedicated to the fuel cell industry but have a significant role to play, for example, in the supply of compressors, pumps, valves, and so forth. Indeed, it is estimated that typically only around one-third of the cost of a fuel cell system is the actual fuel cell stack, with around one-third being balance of plant, and one-third integration, installation, and commissioning costs. Cost issues are discussed in more detail in the following section.

5. FUEL CELL COSTS

The true costs associated with fuel cells are not yet clear—either from a capital or operating perspective. Present costs are well above conventional technologies in most areas, though this depends slightly on the type of fuel cell and the market area in which it may play a part.

The economics of fuel cell systems are also different in different market niches. The fuel cell has the potential to usurp many traditional technologies in a variety of markets, from very small batteries and sensors to multimegawatt power plants. Each system has very different characteristics

and will accept very different prices. For example, a laptop battery substitute that could run for 20 h instead of 2 h could command a high price, especially if it could be refueled in seconds from a canister rather than recharged over several hours. At the other end of the scale, the potential for building modular power plants in which maintenance can be carried out on each module without shutting down the system is worth a significant amount of money to the owner.

Traditional economic calculations have suggested that the fuel cell system for large-scale power generation needs to be less than \$1500/kWe before it will be competitive, while the fuel cell system for automobiles and mass production must be competitive with the internal combustion engine at \$50/kWe or below. That said, it must be remembered other drivers exist for the technology, including environmental benefits, and an issue of increasing strategic importance for many countries, namely a reduced reliance on oil. Some fuel cell systems will sell themselves at \$10,000/kWe, however, if they can be installed where there is currently no available technology capable of meeting requirements.

However, it is clear that all fuel cell costs at present—and these are estimated at anything between 500 and 10,000 dollars per kilowatt (a mature technology such as a gas turbine costs about \$400/kWe)—are high because they are representative of an emerging technology. Once in mass production, recent estimates predict costs of \$40 to \$300/kWe for PEMFCs for transport applications, depending on assumptions regarding technology development. It is clear that both further technical innovation as well as mass manufacture will be needed to compete solely on a cost basis with the internal combustion engine.

High-temperature systems tend to be more expensive as they require significant investment in associated balance of plant, but should still be able to be manufactured for close to \$600 per kilowatt, not far from the current price for a gas turbine or gas engine.

6. ENVIRONMENTAL BENEFITS

One of the areas in which fuel cells should demonstrate significant advantages is in their potential for minimal environmental impact. As well as offering a high theoretical efficiency, all fuel cells emit low levels of pollutants such as the oxides of sulfur (SO_x) and nitrogen (NO_x). SO_x emissions are

low because low sulfur fuels such as methanol or desulfurized natural gas are used. NO_x emissions are low because even the high-temperature fuel cells such as the molten carbonate fuel cell and the solid oxide fuel cell operate at temperatures well below those needed to form NO_x by the thermal combination of nitrogen and oxygen. The formation of NO_x at high temperatures is a problem for those trying to push up the efficiency of heat engines by increasing their maximum operating temperature. Fuel cell systems are already exempt from permitting requirements in some U.S. states, including New York, meaning that they can bypass one of the planning stages (in itself a useful bonus in reduced cost and time).

The high efficiency of the electrochemical process is an added advantage, since less fuel is required to produce a given amount of power than would otherwise be needed. This means that less fuel is used and, of increasing importance, that less CO₂ is released. Estimates suggest that a fuel cell power plant running on a traditional cycle could produce 20 to 30% less CO₂ than traditional power plants, and that vehicles powered by fuel cells would provide similar benefits.

In the case of vehicles, it is particularly important to examine the fuel on which they run. While fuel cell vehicles powered on gasoline reformer technology will have little or no benefit in reduced greenhouse gas emissions, methanol-powered cars may produce 25% less CO₂. Using hydrogen produced from natural gas will result in cuts of up to 40%, while hydrogen from renewably generated electricity, such as wind or solar, emits no greenhouse gas at any stage of the process.

In the ultimate renewable/zero-emission scenario, fuel cells can be run on pure hydrogen produced from renewable energy, using electrolysis of water.

This ensures that there are no polluting emissions at any stage of the fuelling process (though there will inevitably be some from the manufacture and construction of the plant). This scenario is applicable both to stationary energy and to transport, and would enable countries to reduce their dependence on imported energy in various forms, in addition to being very clean. Hydrogen can also be produced from a whole variety of primary resources, including biomass and fossil fuels.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Batteries, Overview • Batteries, Transportation Applications • Cogeneration • Fuel Cell Vehicles • Fuel Economy Initiatives: International Comparisons • Hydrogen, End Uses and Economics • Hydrogen, History of • Hydrogen Production • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

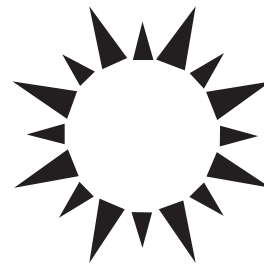
- Fuel Cell Today. www.fuelcelltoday.com. Accessed on November, 2003.
- Kordesch, K., and Simader, G. (1996). "Fuel Cells and Their Applications." VCH, Weinheim, Germany.
- Larminie, J., and Dicks, A. (2000). "Fuel Cell Systems Explained." Wiley, Chichester, United Kingdom.
- Minh, N. Q., and Takahashi, T. (1995). "Science and Technology of Ceramic Fuel Cells." Elsevier, Amsterdam.
- Steele, B. C. H., and Heinzel, A. (2001). Materials for fuel cell technologies. *Nature* **414**, 345–352.
- U.S. Department of Energy (2000). "Fuel Cell Handbook." 5th Ed. Produced under contract DE-AM26-99FT40575.

Fuel Cell Vehicles

JOHN M. DECICCO

Environmental Defense

Dexter, Michigan, United States



-
1. Introduction
 2. Factors Shaping Automobile Propulsion Technology
 3. Fuel Cell Technologies
 4. Vehicular Fuel Cell Systems
 5. Fuel Supply, Storage, and Processing
 6. Performance and Cost
 7. Commercialization Issues

Glossary

- autothermal** A characteristic of fuel processor design that combines exothermic and endothermic reactions so that the resulting system is close to thermally neutral.
- criteria air pollutant** An airborne substance regulated according to public health-based standards; key criteria pollutants pertinent to motor vehicles include carbon monoxide, various toxic substances, ozone and fine particulate matter, and their precursors including reactive hydrocarbons and nitrogen oxides.
- electrolyte** A substance that dissociates into ions, so that it can conduct electricity by means of ion transport (as opposed to electron conduction as in a metallic conductor, for example).
- energy density** Energy content of a fuel or energy storage device per unit mass.
- fuel processor** A device that converts a fuel into hydrogen-rich gas and removes impurities so that the resulting gas can supply the anode of a fuel cell; also termed a *reformer*.
- platform** The primary load-bearing structural assembly that determines the basic size of a motor vehicle and supports its powertrain, body panels, and suspension components.
- powertrain** The interconnected set of components that drives the wheels to propel a motor vehicle, commonly including engine, transmission, driveshaft and differential, or transaxles.
- POX** Partial oxidation, a combustion process in which the reactants are only partially oxidized, allowing hydrogen to be dissociated from a fuel without being completely consumed.

PrOx Preferential oxidation, a catalyzed chemical reaction in which species other than hydrogen are oxidized so as to yield a product stream containing maximum amounts of hydrogen with minimum amounts of undesirable constituents (such as CO).

reformat The gaseous output stream from a reformer or fuel processor, containing hydrogen for supplying a fuel cell anode.

reformer A device that converts a fuel into hydrogen-rich gas and removes impurities so that the resulting gas can supply the anode of a fuel cell; also termed a *fuel processor*.

regenerative braking The process of slowing a vehicle by drawing kinetic energy into a generator (typically an electric motor operating “in reverse”), thereby exerting a rotational drag on the wheels.

specific energy Energy content of a fuel or energy storage device per unit volume.

stack A set of individual fuel cells arranged next to each other and connected in series.

ultracapacitor An electrochemical device that stores electricity through capacitive charging and is capable of providing high power in a compact size and mass.

1. INTRODUCTION

A fuel cell is an electrochemical device that directly converts a fuel to electricity by means of reactions on the surfaces of electrodes and transport of ions through an electrolyte. A fuel cell can be thought of as a chemical battery whose reactants are fed from external sources rather than packaged as part of the battery. A key feature of a fuel cell is transformation of the chemical potential energy of a fuel directly into electricity, a high-value form of energy that can be put to many uses. The fuel cell's direct energy conversion process occurs without an intermediate step of heat generation, as involved in combustion engines. Fuel cells designed to operate at relatively low temperatures, as suitable for automotive propul-

sion, need catalysts such as platinum that enable reaction rates sufficient to provide adequate power.

A fuel cell engine couples a fuel cell to an electric motor to provide mechanical power. A fuel cell vehicle is a motor vehicle having a fuel cell engine as its prime mover. Thus, a fuel cell vehicle has the advantages of electrical propulsion while retaining the advantages of carrying its own chemical fuel. Electrical propulsion is quiet and offers high efficiency, through avoidance of many mechanical frictional losses. Another advantage is its controllability, through fully programmable “tuning” of the vehicle’s driving characteristics for responsiveness and smoothness. An electric drive system can further increase a vehicle’s efficiency by recovering energy during braking, energy that otherwise is lost as low-grade heat in friction brakes. Advantages of an onboard chemical fuel include the relatively low mass occupied by the stored fuel, good driving range between refueling, and ease of refueling. Such advantages are related to a chemical fuel’s energy density, or amount of energy per unit mass. Liquid fuels (such as hydrocarbons and alcohols) have higher energy densities than gaseous fuels (such as natural gas and hydrogen), and chemical fuels typically have higher energy densities than rechargeable batteries.

In addition to the fuel cell itself, a fuel cell vehicle requires a variety of additional components. Some are the auxiliary parts needed for a complete fuel cell power system. Others include fuel storage and processing systems, an electric motor, battery, associated electronics, and control systems. The major functional components are illustrated schematically in Fig. 1. The stack consists of a series of individual fuel cells, each consisting of electrodes, electrolyte, plates with channels for supplying the reactants and removing the exhaust products, and connections.

Auxiliaries include blowers, pumps, heat exchangers, and other items needed to regulate the heat and mass flows from the stack. Fuel is supplied from the onboard tank, holding a liquid fuel or compressed hydrogen. Some designs entail a fuel processor, or reformer, to convert the fuel to a hydrogen-rich gas for feeding the fuel cell; that process may create pollutants requiring emissions controls for cleanup. The stack produces direct current that is regulated by a power electronics package, which in turn supplies electricity to one or more electric traction motors for driving the vehicle. The motor may recapture braking energy that is fed back through the power electronics to recharge the battery. At minimum, a fuel cell vehicle has a battery for startup; a more powerful battery or other high-power storage device is used to provide supplemental traction power in a hybrid fuel cell vehicle. Thus, a fuel cell vehicle is a complex machine, requiring numerous, high-technology components that must be integrated together and manufactured at the low cost and high reliability expected of mass-produced automobiles.

2. FACTORS SHAPING AUTOMOBILE PROPULSION TECHNOLOGY

Fuel cell vehicles are being developed because they promise to meet the requirements expected of automobiles in a market increasingly constrained by environmental and resource limitations. Air pollution and oil dependence have been persistent challenges for vehicles powered by petroleum fuels (gasoline and diesel). Global warming presents a new challenge in the need to limit carbon dioxide (CO₂) emissions from fossil fuel combustion. It is unclear whether automobiles powered by combustion engines running

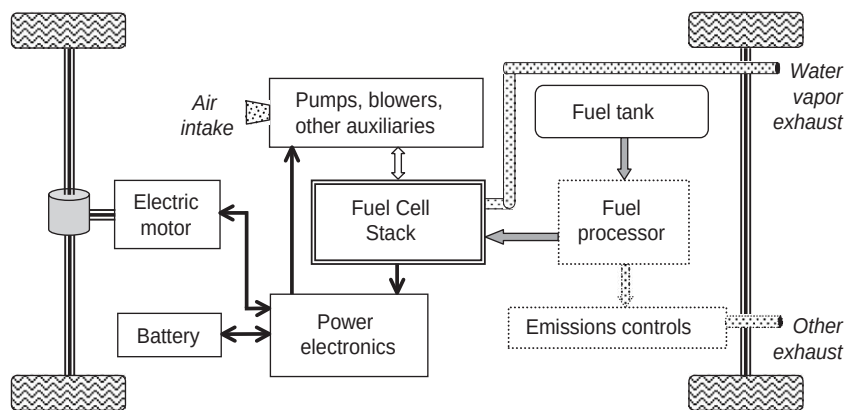


FIGURE 1 Major components of a fuel cell vehicle.

on conventional fossil-derived fuels will be able to meet these challenges. Moreover, electric propulsion has its attractive features of quiet, precisely controllable operation and high efficiency for both powertrain and accessories. The growing number of electrical amenities in automobiles further raises the value of a propulsion system that provides ample electric power capacity. In short, eliminating the internal combustion engine has great appeal and motivates the development of fuel cell vehicles.

Fuel cell vehicles are not the only option for addressing future automotive expectations. Improvements to internal-combustion engines can partially address the challenges, and further refinements can provide additional progress. Hybrid combustion-electric vehicles extend the efficiency horizon of petroleum-fueled designs while boosting onboard electric power capacity. Advanced strategies for controlling criteria air pollutants generated by gasoline vehicles (including sophisticated onboard combustion control and cleanup systems as well as reformulated fuels) make air quality concerns less a driver for technology change than the global warming issue, which presents a more fundamental challenge. Effective control of criteria air pollution is a greater hurdle for diesel engines, which otherwise have an energy-efficiency advantage. Conventional alternative fuels (such as alcohols, ethers, plant oils, and natural gas) also offer partial solutions, but their benefits have not been great enough to overcome the high costs of transition. Using low-CO₂ producing fuels in hybrid vehicles may enhance the prospects for some alternatives. Rechargeable battery electric vehicles have cost and range limitations that inhibit extensive use with existing transportation infrastructure. What makes fuel cell vehicles attractive is their promise for affordably addressing the multiple design requirements likely to be faced by future automobile propulsion systems.

3. FUEL CELL TECHNOLOGIES

Fuel cells are typically classified according to type of electrolyte. While many varieties of fuel cells have been demonstrated in the laboratory, five major types are seeing development for commercial applications:

- Polymer electrolyte membrane (PEM) cells use a plastic (polymer) membrane that becomes electrically conducting when hydrated (saturated with water); they operate near 100°C.
- Alkaline fuel cells use a caustic electrolyte such as potassium hydroxide (KOH); they also operate near 100°C.
- Phosphoric acid fuel cells use that acid as the electrolyte and operate around 200°C.
- Molten carbonate fuel cells use electrolytic mixtures of metal (Li, Na, K) carbonates and operate at 600°C or higher; they are highly efficient and can reform a fuel within the cell.
- Solid oxide fuel cells use zirconium oxide as a solid-state electrolyte and operate at 600°C or higher, and also have advantages of high efficiency and in-cell fuel reforming.

As fuel cell technologies mature, they may see a variety of uses, including battery replacement for portable electronics, large-scale electric power generation, distributed generation, combined heat and power, and transportation. Both PEM and alkaline fuel cells have been used on spacecraft. The first fuel cell vehicle was a farm tractor demonstrated in 1959 using an alkaline cell. Phosphoric acid fuel cells are in limited commercial use for specialized power generation but are too expensive and difficult to manage for vehicle use. Molten carbonate and solid oxide fuel cell technologies are being developed for stationary electric power generation. The need to bring the relatively large quantities of these electrolytes up to a high operating temperature makes these fuel cells unsuitable as a prime mover for motor vehicles, although small solid oxide fuel cells are being investigated as auxiliary power units.

The PEM fuel cell is the focus of automotive development efforts because of its low operating temperature and potential for low-cost mass production. Only PEM cells running on hydrogen have been developed enough to power practical motor vehicles. However, PEM cells that run on methanol are being investigated. It is also premature to dismiss alkaline fuel cells as a long-run option. Figure 2 illustrates the functional elements of a hydrogen PEM fuel cell. Fuel (hydrogen or a hydrogen-rich gaseous mixture) is supplied to the anode (negative electrode) where it diffuses to a catalyst layer. There it dissociates into electrons, supplied to the external electrical load, and hydrogen ions (protons), which are conducted across the polymer electrolyte membrane (also known as a proton exchange membrane) to the cathode. At the cathode, electrons from the returning current and oxygen from the air complete the reaction at a second catalyst layer, forming water vapor that is exhausted from the fuel cell.

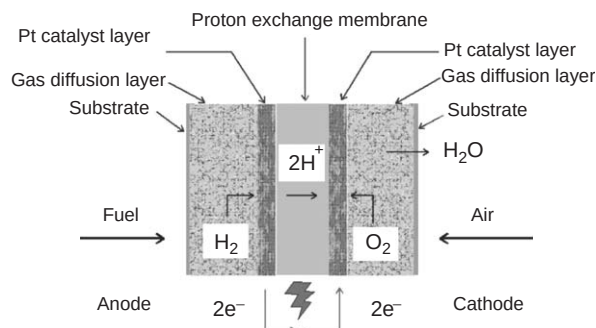


FIGURE 2 Schematic diagram of a polymer electrolyte membrane (PEM) fuel cell. Courtesy of Johnson Matthey Fuel Cells Today Web site (www.fuelcelltoday.com), with permission.

The ideal, open-circuit voltage of a hydrogen-air cell at typical operating conditions is approximately 1.2 volts. Under practical load conditions, a single hydrogen PEM cell might produce 0.7 volts. The voltage produced in part determines the cell's energy efficiency; at 0.7 volts, a hydrogen PEM cell is roughly 60% efficient in converting the chemical potential energy of hydrogen into electricity. Research and development on fuel cell design seeks to obtain a maximum amount of current from a given cell surface area while maintaining voltage, which determines the efficient power output of the fuel cell. A stack of cells connected in series is used to produce the power levels needed for vehicle propulsion.

4. VEHICULAR FUEL CELL SYSTEMS

The components of a fuel cell vehicle, as shown in Fig. 1, draw on a variety of technologies, not all of which are at the same state of development. The particular requirements for fuel cell vehicle systems depend on the configuration. Three categories of components are common to all fuel cell vehicle systems: the fuel cell stack itself, auxiliary equipment that supports the operation of the stack, and the electric drive system. Fuel storage and processing components depend on the choice of onboard fuel; these systems are discussed in Section V of this article.

4.1 Fuel Cell Stacks

It is progress in the development of hydrogen-air PEM stacks that has made fuel cells a contender for powering automobiles of the future. For many years, the energy and power densities of PEM cells were so

low and the amount of platinum catalyst required was so high that most commercial applications seemed out of the question. For example, the platinum requirements for the PEM cells used on Gemini space missions of the 1960s were on the order of 100 g/kW, for a cost factor of \$1500/kW (assuming a platinum cost of \$15/g). A typical automotive fuel cell stack would be 80 kW, implying a cost of \$120,000 for the catalyst material alone. By comparison, current automotive catalytic converters require roughly 0.05 g/kW of platinum-group metals, costing on the order of \$100 for an average car. More stringent emissions standards are pushing precious metal requirements higher, so that future gasoline vehicles may need 0.1 to 0.2 g/kW of platinum group metals.

During the 1980s, research on electrode catalyst deposition led to large reductions in the amount of platinum needed. By the early 1990s, PEM fuel cell platinum requirements had dropped below 10 g/kW; as of 2002, state-of-the-art designs needed less than 2 g/kW, with further reductions still in sight. Such breakthroughs have motivated engineering development work on the other critical elements of a fuel cell stack, such as the membrane, electrode assemblies, and flow fields that supply hydrogen and air and exhaust the reaction products from the cells. Platinum needs remain a cost challenge for PEM fuel cells, however, and ongoing research targets another order-of-magnitude reduction in platinum catalyst requirements as needed to place the technology in the range of affordability and precious metal availability that would make widespread use feasible.

Fuel cell stacks can be designed as compact modules, with connections for their electrical output as well as for feeding fuel and air into the stack and removing exhaust, cooling the stack, and monitoring its operating state. Since PEM cell stacks are built in modular form, a great deal of flexibility will be available to designers in terms of how stacks are placed and arranged in the vehicle. Using current designs, a typical fuel cell stack for automotive applications, delivering 80 kW of peak power, occupies a volume of about 80 liters and weighs roughly 90 kg. However, even more compact designs are likely to evolve as experience is gained. Stacks could be built in modules of specified sizes, so that multiple modules would be arranged and connected to supply the differing power requirements of different sized vehicles while permitting enhanced economies of scale in mass production. A remaining engineering challenge for fuel cells is achieving sufficient membrane durability. Other developmental issues include

freeze protection and further cost reduction for electrodes, membranes, and separator plates.

The latter serve the multiple functions of separating adjoining cells, providing the flow fields that transport fuel-air reactants and exhaust products to and from the cells, and removing waste heat. Early PEM cells have used graphite plates, but these are costly to produce. Coated metallic and composite materials are being developed for separator plates, with goals of meeting low-cost mass production targets while achieving performance and durability requirements in the highly reactive fuel cell environment.

4.2 Fuel Cell Engine Auxiliaries

Numerous other parts are needed for a complete fuel cell engine suitable for motor vehicles. These auxiliary components include many familiar pieces of equipment, such as blowers, pumps, heat exchangers, sensors, and controls. However, fuel cells have specialized performance requirements, so substantial engineering effort is needed to develop auxiliary systems suitable for fuel cell vehicles.

A blower or compressor is needed to supply air to a fuel cell stack, and air compression represents the largest parasitic load in the system. The pressure of the air supplied to the cells influences the stack efficiency, but a trade-off is involved. The higher output that comes with higher pressure entails an energy penalty for running the compressor. Designers are experimenting with various pressure regimes, and although pressure may vary depending on operating conditions, most designs operate at no more than three atmospheres. Compression power requirements approach 10% of stack output at peak power, while other loads (pumps, cooling, controls, etc.) are generally in the 1 to 2% range. Optimized designs can hold the average air compression penalty to 3 to 5% of stack output. Energy in the fuel cell exhaust stream is typically captured in a turbine expander to help power the compressor.

Another critical auxiliary component is an integrated system for moisture control and thermal management. This system maintains membrane moisture, exhausts water vapor, removes heat from the stack, and manages the flow of heat among various components for optimum efficiency. Fuel cell membranes need to be precisely hydrated to function properly. Too much water accumulation blocks the reactions, but the membrane can fail or degrade if it becomes too dry. Fuel cells need to start in very cold weather; motor vehicles are typically designed to be able to start at -40°C . Therefore, special “anti-

freeze” hydrating solutions are being developed that do not poison the stack; mechanisms are needed to address the water vapor that is present in the exhaust stream and other fuel cell subsystems as well as quickly bring a stack up to operating temperature.

Heat removal presents its own special challenges, since PEM cells operate at low temperatures and so need to efficiently reject low-trade heat. Ambient pressure designs operate near 80°C . This level is much lower than the over 2000°C flame temperature and roughly 500°C exhaust gas temperature of a gasoline combustion engine. Pressurized PEM fuel cells use only relatively low pressures, so temperatures are under 200°C in any case, offering but a small difference to ambient temperature for driving heat transfer. Even at its high, roughly 50% efficiency, a 80 kW stack must dissipate 5 to 10 kW of waste heat under average loads and as much as 40 kW without risking membrane degradation under peak loads. Some of this heat is carried away in the exhaust stream, but much must still be handled by directly cooling the stack and transferring heat to the ambient air. Thus, sizable heat exchangers and fans are needed.

4.3 Electric Drive Systems

The primary electric drivetrain components for fuel cell vehicles are the same as those for any electric vehicle: traction motors, power electronics, and batteries. Electric drive components require their own sets of auxiliaries and management systems, for control and cooling of the equipment. A fuel cell vehicle may have a hybrid powertrain, in which the fuel cell is sized at less than the vehicle’s peak power requirement and additional power is supplied from an electricity storage device, such as a high-power battery or ultracapacitor. Although a hybrid powertrain is more complex than one using a fuel cell alone, it offers advantages of regenerative braking, which helps vehicle efficiency, and the availability of greater power for startup and auxiliary loads. Hybridization also would have an economic advantage if the cost of peak power from the electricity storage device is less than that of meeting the peak power requirement with a fuel cell alone.

The electric drive system for a fuel cell vehicle is comparable to that for a battery electric vehicle or series hybrid, with the traction motor and power electronics scaled to meet the full power needs of the vehicle. This situation contrasts with electric drive system for a parallel hybrid as used for contemporary models (such as the Toyota Prius and Honda Civic Hybrid), in which the electric drive components

supply only a portion of the tractive power with the rest mechanically provided from the combustion engine. The specific components used in a fuel cell vehicle, such as type of motor, depend on the particular electric drive system design, which is itself a new and evolving area of engineering.

Although more developed than fuel cells themselves, electric drive technologies are not mature by automotive standards. As of 2002, hybrid electric vehicle production volumes were still an order of magnitude lower than those of typical mainstream vehicle models, and costs remain high. For fuel cell vehicles, electric drive system costs compound those of the core fuel cell system and its fuel storage or processing components. The cost of electric drive components is expected to fall substantially, with some estimates as low as \$30/kW (retail-equivalent, excluding batteries). Time and production experience will be needed for costs to fall to such levels, which are likely to be essential for widespread commercialization of any type of electric vehicle, including fuel cell vehicles. As experience is gained and research progresses, the types of electric drive components used in fuel cell vehicles will also evolve. For example, instead of traditional electric traction motors, compact wheel motors might be used as suggested by GM's Autonomy fuel cell vehicle concept.

5. FUEL SUPPLY, STORAGE, AND PROCESSING

Depending on the fuel they carry, fuel cell vehicles entail various requirements for fuel supply (including production and distribution), fuel storage on the vehicle, and on-vehicle fuel processing that might be needed.

A fuel cell vehicle operated on gasoline may require minimal changes in existing fuel supply infrastructure, but faces special processing requirements onboard the vehicle. A fuel cell vehicle operated directly on hydrogen, on the other hand, does not require a fuel processor, but needs a new fuel supply infrastructure and workable technology for onboard fuel storage. Methanol is also being considered for fuel cell vehicles, since it has the convenience of liquid while being easier to process into hydrogen than gasoline. Table I summarizes the issues and technical status of these three key fuel options for fuel cell vehicles. Ethanol has processing requirements similar to those for gasoline and a long-term potential for production from renewable resources. Natural gas has processing requirements similar to those for methanol and is already widely distributed, but faces the challenges associated with gaseous storage.

The ultimate fuel supply, storage, and processing configuration of fuel cell vehicles will be shaped by how well various fuel-related technologies progress through research and development. The configuration will also be influenced by the broader changes in transportation infrastructure (such as increasing use of intelligent systems) that will occur over the time frame needed to bring these technologies to states of workability and affordability that would enable widespread commercialization.

5.1 Fuel Supply

The auto and oil industries are among the world's largest, and the extensive infrastructure that supplies gasoline and diesel motor fuel is built of years of investment in both equipment and know-how. If practical ways can be found to power fuel cells with

TABLE I
Fuel Options for Fuel Cell Vehicles

	Hydrogen	Methanol	Gasoline
Fuel economy gain	High	Modest	Small/nil
CO ₂ reduction (near-term) ^a	Modest	Low	Small/nil
Emissions control need	Zero	Low	Low
Fuel supply	Needs major development	Very limited	Already exists
On-board storage	Needs major development	Readily feasible	Already exists
Fuel processing	Not needed	Under development	Under development
Added car cost/weight	Modest	High	High

^aNear-term CO₂ reduction assumes hydrogen and methanol derived from natural gas; the long-term potential is greater for all options as renewable resources or carbon sequestration systems for fossil fuels become available.

existing hydrocarbon fuels, or variants of them, adoption of fuel cell vehicles could proceed more quickly than if it were constrained by the rate at which infrastructure can be changed. On the other hand, fuel cells powered directly by hydrogen are already technically workable. Thus, another commercialization path could be niche vehicles that can accept the range restrictions of compressed hydrogen storage using new infrastructure that is constructed incrementally.

Gasoline-like hydrocarbon fuels for powering fuel cell vehicles would have properties tailored to the new technology, just as today's motor fuels are formulated for optimum performance in combustion engines. Impurities such as sulfur would have to be essentially eliminated; this change is already under way as sulfur levels in conventional fuels are being lowered for pollution control reasons. The mix of particular hydrocarbons would also change. But the key characteristics of being an energy-dense, convenient-to-distribute liquid would remain, and that is the attraction. Hydrocarbon fuel supply is also quite energy efficient. The "well-to-tank" efficiency—that is, the amount of fuel energy delivered to a vehicle per unit of primary energy—is typically over 80% for gasoline and over 90% for diesel.

The major problems for use of hydrocarbons in fuel cell vehicles are (1) the fact that hydrocarbons still carry carbon to the vehicle and (2) the difficulties of fuel processing (discussed later). All known hydrocarbon fuel technologies for motor vehicles involve releasing the carbon as CO_2 . Practical ways to "scrub" or capture carbon from vehicles do not exist. Thus, even with the fuel cell vehicle's potentially higher efficiency and the possibility of offsets through carbon sequestration, there will be limits to the effectiveness of fossil-derived fuels for addressing global warming. Thus, hydrocarbons supplied by an evolved version of the current motor fuel supply infrastructure are viewed as a possible transitional pathway rather than the ultimate fuel. Hydrocarbons derived from renewable resources could offer a zero or very low net release of CO_2 to the atmosphere (and of course, such fuels could also be used in combustion engine vehicles).

Hydrogen is the ideal fuel for PEM cells and any other low-temperature fuel cell technology. Currently and for the near-term, natural gas is the most economical source of hydrogen. It can be produced by steam reforming at central plants and then distributed to filling stations, or natural gas could be locally reformed at filling stations to provide the hydrogen. The former approach is how most hydro-

gen is now made. The primary energy efficiency for supplying compressed hydrogen in this manner is typically 70% (that is, the energy content of the compressed hydrogen at its point of distribution is about 70% of the energy content of the natural gas at the well). Local reforming would take advantage of existing natural gas distribution infrastructure and is generally viewed as the most economical way to begin building a hydrogen fuel distribution system. Hydrogen can also be made by reacting coal with water; this approach holds economic promise because coal is so plentiful. Research is under way to find ways to sequester the resulting CO_2 and such a system could enable the use of coal to generate hydrogen for a variety of applications, including vehicles, with lower CO_2 emissions than existing direct use of fossil fuels.

The ultimate attraction of hydrogen is that it is a carbon-free energy carrier which, if produced from renewable resources, offers the potential of an energy system with zero greenhouse gas emissions. If vehicular hydrogen supply systems were initially developed using natural gas, for example, they ultimately might be converted to hydrogen supplied from renewable resources.

Other fuel options have their own supply characteristics, which differ in the near and long term. Methanol is now made from natural gas and that would be the most economical source for the foreseeable future. In the long run it could be produced from coal or biomass, with the latter resource option offering very low greenhouse gas emissions. But methanol also has toxicity issues that are difficult to manage. Ethanol is now made from sugar or starch-bearing crops in an energy- and resource-intensive agricultural system, making it costly and of dubious net environmental benefit even though it is in some sense renewable. However, ethanol can be made through efficient, enzymatic conversion of cellulose from organic wastes and specialized crops, thereby offering near-zero greenhouse gas emissions. Alcohols still carry carbon to the vehicle, so achieving their potential for zero or very low greenhouse gas emissions would require both truly renewable, zero-fossil-fuel input energy crops plus essentially zero-emissions processing, refining, and distribution systems.

Finally, from a long-term fuel supply perspective, it cannot be discounted that the most effective way to use fuel cells to power vehicles might be indirectly, at least for many vehicles in urban and near-urban areas. In this case, the fuel cells are kept off the vehicle, which would be powered by advanced

batteries charged through an intelligent, electrified infrastructure to which electricity is supplied by stationary fuel cells. This approach would represent an even closer coupling of motor vehicles to a power grid than other options for grid-connected vehicles, including vehicles with onboard fuel cells, that could supply power to the grid but would still need a separate fuel infrastructure.

5.2 Fuel Storage

Fuel storage onboard an automobile is straightforward and well established for liquids, but presents cost, range, and packaging challenges for any gaseous fuel. The challenge is particularly acute for hydrogen, which is the least dense of substances. Thus, a fundamental breakthrough in onboard hydrogen storage technology is needed for widespread use of direct hydrogen fuel cell vehicles to become feasible. Even if fuel cells and their auxiliary systems achieve low costs, and hydrogen supply and infrastructure systems become competitive, all known options for storing hydrogen have limitations that would inhibit commercialization of fuel cell vehicles for personal automobiles as used today.

Hydrogen is the least dense of all fuels in spite of its high specific energy of 142 MJ/kg, compared to 46 MJ/kg for gasoline. All vehicular storage options for hydrogen present difficulties, rooted in the specific energy and energy density values shown in Table II. Options include compressed gas, liquid hydrogen and metal or chemical hydrides (which involve chemical reactions to store and release the hydrogen). Devices that react iron oxide with water to produce hydrogen have been explored. Adsorption in carbon nanotubes and high-pressure confinement in glass microspheres have been proposed but remain largely speculative.

Compressed gas storage is well known through experience with natural gas vehicles; most fuel cell demonstration vehicles use compressed hydrogen. Problems with gaseous storage include onboard space requirements and costs inherently higher than those of liquid fuel tanks. Packaging challenges exist even for natural gas, which has an energy density nearly four times higher than that of hydrogen. Development is under way on very-high pressure tanks made of carbon-fiber composites that offer greater shape flexibility than metal cylinders. Automakers have successfully tested tanks rated at 5000 psi, which is higher than the 3000 psi range now typical for compressed gas storage, and designs in the 10,000 psi range are being developed.

TABLE II
Onboard Fuel Storage System Characteristics

Fuel storage option	Specific energy (MJ/kg)	Energy density (MJ/L)
Gasoline (liquid in steel tank)	36	32
Methanol (liquid in steel tank)	18	18
Hydrogen		
Compressed gas (5000 psi/340 atm)	7	2
Cryogenic liquid	15	5
Metal hydride	3	12
Organic hydride	9	7
Sponge iron	5	6
Nanotubes	50	100

Note: Comparisons are made on a higher heating value (HHV) basis: 46 MJ/kg for gasoline, 23 MJ/kg for methanol, and 142 MJ/kg for hydrogen, counting the mass and size of fuel system components but not counting powertrain energy conversion losses. Values for hydride and sponge iron options are fairly uncertain but representative of what is expected through research; the nanotube option is still considered speculative.

Storing hydrogen as a liquid requires very cold, cryogenic temperatures. At atmospheric pressure, hydrogen liquefies at 20 K (−253°C or −423°F); the saturation temperature rises by only 10 K going to 10 atmospheres. Special systems are needed to maintain such low temperatures on a vehicle and safety manage the hydrogen boil-off that occurs while the vehicle is not in use. Liquid hydrogen storage systems have been demonstrated in compact packages suitable for vehicles, along with automated refueling systems that approach gasoline in terms of speed and convenience. A serious handicap is high energy cost of liquefying hydrogen, amounting to roughly one-third of the fuel's energy content. Because it involves unrecoverable low-grade heat rejection, this energy loss is fundamental and degrades the environmental benefits and economics no matter what resource is used for hydrogen production. Further losses occur during distribution, refueling, and storage (due to boil-off), so that the net supply efficiency of liquid hydrogen may be no more than 50%.

Hydrides of various forms are being investigated for storing hydrogen; they are relatively compact but have a high weight penalty. Metal hydrides store roughly 7% hydrogen by weight, not counting

containment, but not all of that is recoverable in practical storage systems (by comparison, gasoline is about 16% hydrogen by weight). Energy penalties are involved in charging (filling) and discharging hydride systems. Some metallic systems (such as magnesium alloys) are costly but compact; other materials (such as sodium borohydride) are less costly but face other technical challenges. Another system under investigation is lithium hydride in an organic mineral oil slurry, and other organo-metallic hydrides may hold promise, although the charging and discharging reactions involve elevated temperatures and catalysts to effectively release the hydrogen. Some systems would have to be refilled by simultaneous pumping of fresh, “charged” hydride into the vehicle while pumping out the hydrogen-depleted chemical for reprocessing. In general, hydride systems are still in early stages of development and have not yet been shown to be workable for automobiles.

5.3 Fuel Processing

The challenges of supplying hydrogen and storing it onboard a vehicle and have prompted efforts to develop onboard fuel processors (reformers) for converting liquid hydrocarbon or alcohol fuels into a hydrogen-rich anode feed gas. In general, such fuel processing involves two major steps: (1) dissociating the hydrogen from the fuel and (2) purifying the resulting gaseous stream to have maximum hydrogen levels and minimal constituents such as CO that can poison the stack. Each step involves multiple reactions and processes and a preliminary cleanup stage is needed if sulfur or other detrimental substances are present in the fuel.

Methanol can be processed by steam reforming, a catalytic process that is endothermic (requiring heat input). Steam reforming is a standard process for large-scale production of hydrogen from natural gas. Low-cost catalysts, such as copper/zinc oxides on alumina for methanol and nickel on alumina for natural gas, can be used for steam reforming. Two reactions are involved: (1) the fuel and water are converted to a synthesis gas containing CO and H₂, and (2) a water-gas shift reaction converts H₂O plus CO into CO₂ plus additional H₂. The resulting reformat is rich in H₂ and low in CO. Since air input is not used in steam reforming, the reformat is not diluted with nitrogen. But because it is not possible to run a complete water-gas shift reaction, further CO cleanup is needed using a catalyzed preferential oxidation (“PrOx”) processor (that is, a device that

preferentially oxidizes CO rather than H₂). A precious metal catalyst is required for the PrOx processors shown to be effective to date.

Because methanol is a liquid that can be steam-reformed at a relatively low temperature of about 250°C, some researchers consider methanol-based systems more feasible than systems based on hydrocarbons that require reforming temperatures in excess of 650°C. However, steam reforming involves a trade-off between fast, responsive reaction rates needed for driving and the slower reaction rates that are limited by the heat transfer characteristics of the reformers. Slow startup and high cost also remain difficult issues for methanol processing.

Any hydrocarbon can be converted into a H₂ reformat by using a partial oxidation (“POX”) process followed by a water-gas shift reaction and then PrOx cleanup of the residual CO. This approach is being developed for gasoline as well as flexible-fuel reformers that could handle alcohols as well as other fuels. Partial oxidation is exothermic, since it involves burning some of the fuel, and the POX process does not itself require a catalyst. However, most hydrocarbon fuel processors use a catalyst to lower the reaction temperature, improve hydrogen yields, and avoid soot formation. Many are designed to be autothermal, using the heat released from the POX as input for the steam reforming process, so that the resulting system is close to thermally neutral. This approach helps the fuel processor achieve a high efficiency.

Startup and transient response are less problematic for POX and autothermal reformers than for steam reformers. However, they face a greater CO cleanup challenge, since their raw reformat can have a CO concentration as high as 10% (from gasoline, for example). Also, nitrogen in the input air dilutes the H₂ in the reformat, thereby lowering the efficiency of the fuel cell, and can produce ammonia, adding to the cleanup requirements. Providing a very clean H₂ reformat is critical, because impurities in the reformat not only poison the fuel cell catalyst, but also cause the fuel cell membrane to degrade. To achieve high effectiveness and durability, most fuel processing catalysts contain noble metals, leading to high catalyst costs.

All approaches to onboard liquid fuel processing have problems of cost, size and weight, startup and dynamic response, and overall workability that need to be resolved before practical designs can be developed. Although the basic processes involved are well known, combining them into an affordable device that would fit into a car is a formidable

challenge. Although size and weight are a lesser concern in nonautomotive applications, many similar technical issues must be resolved even for fuel cells in stationary power generation.

6. PERFORMANCE AND COST

The ultimate appeal of fuel cell vehicles rests on whether they can provide superior environmental performance (low emissions of air pollutants and greenhouse gases) while meeting or exceeding the driving performance and other consumer expectations of competing technologies. At the same time, fuel cell vehicles will need to be affordable in terms of providing such features with customer value sufficient to be marketed successfully.

A fuel cell vehicle's environmental performance advantage is determined largely by the operating efficiency of the stack and associated components. Part of the fuel cell's attraction is its inherently high energy conversion efficiency, exceeding 50% over much of the load range, compared to the 35% or 45% efficiency achieved by gasoline or diesel engines, respectively, only in their ideal operating range. However, net fuel cell system efficiency is strongly influenced by auxiliary systems, particularly air compression. For liquid fueled systems, fuel processing penalties also come into play. Overall vehicle efficiency also depends the balance-of-vehicle parameters that determine mass, aerodynamic drag, tire rolling resistance, and accessory loads. The level of greenhouse gas emissions also depends on the source of the fuel, as it does for all vehicle propulsion options.

The performance and cost of fuel cell vehicles and key competing advanced vehicles is illustrated in relative terms in Fig. 3. While numerous other alternative fuels (such as natural gas, alcohols, synthetic fuels) could also enter into such a comparison, much of what is known can be inferred by comparing direct hydrogen and gasoline-based fuel cell vehicles to evolved versions of gasoline and diesel internal combustion engine vehicles, including hybrid-electric vehicles that retain internal combustion engines. The comparisons of Fig. 3 are normalized, showing both retail price increases and greenhouse gas emissions reductions in percentage terms. The estimates are referenced to a contemporary gasoline car, represented by "Today's Vehicles" at origin of the axes, rated at 28 miles per gallon (mpg) on an average of U.S. EPA regulatory city and highway driving tests. Cost and performance depend on numerous assumptions regarding

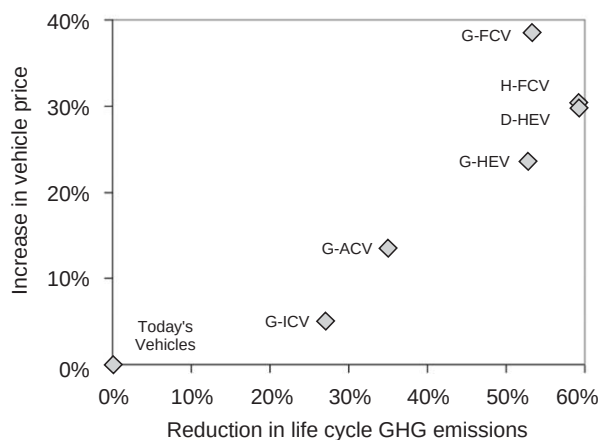


FIGURE 3 Greenhouse gas emissions reductions and retail prices increases of fuel cell vehicles relative to other advanced vehicles. All fuels assume oil or natural gas feedstocks and projections are given relative to today's vehicles as represented as a midsize car rated at 28 miles per gallon on the U.S. EPA composite test cycle. Key: G, gasoline; D, diesel; H, hydrogen, ICV/ACV, improved/advanced conventional vehicle; HEV, hybrid electric vehicle (on advanced platform); FCV, fuel cell vehicle (on advanced platform).

future technologies; the values could change depending on R&D progress. Nevertheless, the ordinal ranking reflects current knowledge of the engineering potential of fuel cell cars compared to other future technologies.

Even using fossil fuel feedstocks (oil for the gasoline and diesel vehicles, natural gas for methanol and hydrogen), fuel cell vehicles and other options all show lower greenhouse gas emissions than today's cars. A good part of this reduction is associated with an advanced vehicle platform having reduced mass, better aerodynamics, and other reductions in the "load," or amount of energy that needs to be delivered by the vehicle's power plant. The comparisons for the fuel cell vehicles, hybrid-electric vehicles, and advanced conventional vehicles assume a uniform advanced platform. This platform has substantial load reduction compared to today's cars, reflected in how the gasoline-powered advanced conventional vehicle (G-ACV) in Fig. 3 achieves a greater GHG emissions reduction than the improved conventional vehicle shown for reference (G-ICV).

Among fuel cell vehicles, the direct hydrogen design (H-FCV) provides the greatest greenhouse gas reduction and has the lowest expected cost, since it avoids the cost and other penalties associated with onboard fuel processing. The H-FCV illustrated in Fig. 3 assumes compressed gas hydrogen storage

which, while technically feasible, may not be practical for mass-market vehicles. The gasoline fuel cell vehicle (G-FCV) represents the highest cost option. Although it may not reduce greenhouse gas emissions more than a hybrid-electric combustion engine vehicle, the G-FCV is being researched because it offers a transition path to hydrogen fuel cell vehicles that make use of existing infrastructure.

The comparisons shown here assume hydrogen produced from natural gas. Substantially lower or even zero greenhouse gas emissions are hypothetically possible with hydrogen produced from solar energy, other renewable resources, or nuclear power. Very low greenhouse gas emissions could be obtained with fossil-derived hydrogen if ways were found to sequester the carbon, since direct use of hydrogen avoids carrying carbon atoms to the vehicle. Similarly, very low greenhouse gas emissions would be feasible with methanol or other fuels that might be processed into onboard hydrogen if they were derived from renewable resources or other feedstocks enabling offset of the carbon.

Because of diesel fuel's high efficiency in piston engines and low energy requirements for production from crude oil, diesel vehicles have lower greenhouse gas emissions than similar gasoline vehicles. A diesel hybrid-electric vehicle (D-HEV) offers lifecycle greenhouse gas emissions reductions as great as those of a H-FCV using hydrogen derived from natural gas. However, it is not yet known whether diesel engines can attain adequately low levels of criteria air pollutants, as already demonstrated for gasoline internal combustion engines and likely to be feasible for any fuel cell vehicle. Costs of an advanced D-HEV are expected to be similar to those of a H-FCV, due to the significant costs premium of diesel engines and their emissions controls compared to gasoline engines.

To put costs into perspective, the average real (inflation-adjusted) price of new cars in the United States has been rising linearly for many years, as a result of consumer features as well as regulatory changes to address safety and environmental hazards. Between 1980 and 2000, average new car real prices rose 34%. Continuing a linear trend over the next 20 years—a plausible early horizon for mainstream commercialization of fuel cell vehicles—implies a further 21% price increase. Current expectations, as reflected in Fig. 3, suggest fuel cell vehicle costs exceeding such a trend. Nevertheless, a combination of research progress (lowering costs) and greater value placed on the environmental or other attributes offered by the technology could yield

a situation where fuel cell vehicles are competitive in future market and regulatory contexts. Indeed, the prospects for such a business case underpin the research and development investments now being made by automakers in fuel cell technologies.

7. COMMERCIALIZATION ISSUES

The commercialization prospects for fuel cell vehicles depend not only on their performance and cost, but also on how well they can compete with other technology options that address similar market and policy needs. While market forces have not traditionally motivated design change for reasons of environmental performance, customer values and expectations can evolve and such characteristics could grow in importance. However, inherent market conservatism will favor less disruptive ways to address evolving needs, which might be met by improved gasoline and diesel vehicles, including hybrid-electric versions. Yet looking over the long-run, particularly the need to substantially reducing greenhouse gas emissions, hydrogen fuel cells may well provide a solution that is superior to other alternatives.

While the use in mass-market private vehicles is the prospect that motivates extensive fuel cell R&D, cars and light trucks may not be the initial market for fuel cell vehicles. Fuel cells seem likely to become competitive sooner in urban buses and delivery vehicles. Centrally managed fleets of such vehicles could be serviced by a limited hydrogen refueling infrastructure; onboard hydrogen storage (e.g., with compressed gas) is also more feasible with such vehicles. Engines for buses and other heavy vehicles are produced at lower volumes with higher unit costs than those for light-duty vehicles, enabling fuel cells to compete sooner in this more limited market. Urban operation is well suited for electric drive and causes higher population exposures to exhaust pollution, providing a higher value on the very low or zero emissions offered by fuel cells. The higher cost tolerance, lower infrastructure barriers, and higher value of a very clean technology suggest that buses and other urban vehicles offer greater opportunities for initial fuel cell vehicle commercialization than light duty vehicles. Nevertheless, improvements in diesel emissions control and the emergence of diesel electric hybrid designs will still provide notable competition for fuel cell vehicles.

For light vehicles, the most difficult barriers faced by fuel cell vehicles are those associated with the fuel,

namely, the dual challenges of onboard hydrogen storage and liquid fuel reforming. Both of these technical hurdles present fundamental R&D challenges. Until such challenges are overcome, it will not be possible to confidently project when commercialization of fuel cell cars might proceed. Moreover, electric drivetrain components (motors, controllers, high-power batteries) are just now beginning to see mass production for hybrid vehicles. It is likely to take over a decade to gain experience and reduce costs for such components sufficiently to enable fuel cell vehicles to be affordable to the mass market. Thus, although automakers began introducing fuel cell demonstration vehicles in 2002, widespread commercialization will not be possible for at least another 15 or more years. Even over a 15- to 20-year time frame, building a business case for fuel cell vehicles will require the convergence of three pathways: (1) engineering development and maturation of fuel cell stacks and auxiliary components, (2) cost reduction for electric drive components, and (3) resolution of the barriers related to the right fuel for fuel cell vehicles.

Given such the long time horizon, non-energy-related needs could begin to reshape the personal transportation system. These other needs include issues such as congestion, safety, personal security, and convenience, and the opportunities associated with telecommunications and computer technology. It is not obvious that any "drop in" new fuel and technology, including fuel cells, that does little to address such other needs will provide benefits compelling enough to displace evolved versions of the internal combustion engine running on hydrocarbon fuels. Nevertheless, given the fuel cell's promise for meeting the long-term challenges for motor vehicle energy use, research on the technology has captivated the interests of industry and governments around the world, so that fuel cells may well

become the prime mover of choice in a future transportation system whose details are yet to be imagined.

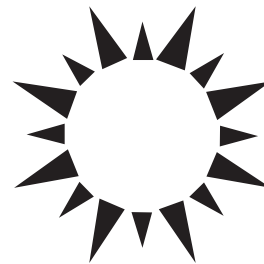
SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Batteries, Transportation Applications • Biodiesel Fuels • Ethanol Fuel • Fuel Cells • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Fuel Economy Initiatives: International Comparisons • Hybrid Electric Vehicles • Hydrogen, End Uses and Economics • Transportation Fuel Alternatives for Highway Vehicles

Further Reading

- Angrist, S. W. (1976). "Direct Energy Conversion." 3rd Ed. Allyn & Bacon, Boston.
- Appleby, A. J. (1999, July). The electrochemical engine for vehicles. *Sci. Am.* 28, 74–79.
- DeCicco, John, M. (2001). "Fuel Cell Vehicles: Technology, Market and Policy Issues." Society of Automotive Engineers, Warrendale, PA.
- Borroni-Bird, C. E. (1996). Fuel cell commercialization issues for light-duty vehicle applications. *J. Power Sources* 61, 33–48.
- Koppel, T. (1999). "Powering the Future." Wiley, Toronto.
- Oei, D-G. (1995). Direct hydrogen-fueled proton exchange membrane (PEM) fuel cell for transportation. SAE Paper No. 952763. Society of Automotive Engineers, Warrendale, PA.
- Stobart, R. K. (1999). Fuel cell power for passenger cars—what barriers remain? SAE Paper No. 1999-01-0321. Society of Automotive Engineers, Warrendale, PA.
- Thomas, C. E., James, B. D., Lomax, F. D., and Kuhn, I. F. (2000). Fuel options for the fuel cell vehicle: Hydrogen, methanol, or gasoline? *Int. J. Hydrogen Energy* 25, 551–567.
- Weiss, M. A., Haywood, J. B., Drake, E. M., Schafer, A., and AuYeung, F. F. (2000, October). On the road in 2020: A life-cycle analysis of new automobile technologies. Energy Laboratory Report No. MIT EL00-003. Massachusetts Institute of Technology, Cambridge, MA.

Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles



MICHAEL WANG

Argonne National Laboratory
Argonne, Illinois, United States

-
1. Introduction
 2. Alternative Transportation Fuels and Alternative Fuel Vehicles
 3. Fuel Cycle Analyses of Vehicles and Transportation Fuels
 4. Fuel Pathways and Vehicle/Fuel Combinations
 5. Energy and Emissions: Items of Interest
 6. Issues Involved in Transportation Fuel Cycle Analyses
 7. Completed Studies
 8. Fuel Cycle Energy and Emission Results of Vehicle/Fuel Systems
 9. Conclusions

Glossary

advanced vehicle technologies The engineering and design processes that lead to vehicles with high energy efficiencies and low emissions; include direct-injection, hybrid electric, fuel cell, and battery-powered electric vehicle systems.

alternative transportation fuels Energy sources other than petroleum gasoline and diesel for motor vehicle applications. These fuels can be produced from nonpetroleum sources and may have energy and emission benefits.

feedstock stage The first stage of a transportation fuel cycle; the feedstock stage includes fuel feedstock recovery, storage, and transportation.

fuel cycle analysis A procedure used to analyze energy use and emissions during the cycle of energy feedstock production and transportation; transportation fuel production, transportation, and distribution; and transportation fuel use in vehicles (also called well-to-wheels analysis).

fuel stage The second stage of a transportation fuel cycle; the fuel stage includes fuel production, storage, transportation, and distribution, in addition to activities such

as fuel liquefaction and compression for long-distance transportation and onboard vehicle storage.

greenhouse gases The by-products of energy use that are believed to have the potential to cause global warming. Although there are many greenhouse gases, the three major greenhouse gases that result from motor vehicle fuel use are carbon dioxide, methane, and nitrous oxide.

life cycle analysis The process of evaluating energy use and emissions of a product from “cradle to grave” (including raw material recovery, material fabrication, product manufacture, product use, product recycling, and disposal). This process is typically used for consumer products; fuel cycle analysis (well-to-wheels analysis) is usually used for transportation vehicles/fuels.

pump-to-wheels stage The second stage of a well-to-wheels analysis (also called vehicle operation stage). It begins with refueling of vehicle fuel tanks. Emissions at this stage may come from fuel combustion or conversion, fuel evaporation, and brake and tire wear.

vehicle operation stage The last stage of a transportation fuel cycle; this stage begins with refueling of vehicle fuel tanks. Emissions at this stage may come from fuel combustion or conversion, fuel evaporation, and brake and tire wear.

well-to-pump stage The first stage of the well-to-wheels analysis (also called the upstream stage), it covers all activities from the wells where fuel feedstock is recovered, to fuels available at the pumps of vehicle refueling stations. In particular, it includes activities associated with the feedstock stage and the fuel stage of a transportation fuel cycle analysis.

well-to-wheels analysis Also called a fuel cycle analysis, this process assesses all activities from the wells where fuel feedstock is recovered, to vehicle wheels where useful energy is generated and used to move vehicles. The analysis focuses in particular on energy use and emissions of energy feedstock production and transportation; transportation fuel production, transportation, and distribution; and transportation fuel use in vehicles.

The spark-ignition and compression-ignition engine technologies that are currently employed in motor vehicles were developed more than 100 years ago. These conventional vehicle technologies are fueled by petroleum-derived gasoline and diesel fuels (the so-called conventional fuels). Over the past 100 years, the conventional technologies have been dramatically improved, reducing cost and increasing performance. Humanity enjoys the mobility provided by these technologies more than ever. However, their use causes negative effects, such as the exclusive dependence on petroleum by the transportation sector, air pollution (especially urban air pollution), and large amounts of greenhouse gas (GHG) emissions. Alternative transportation fuels to replace conventional fuels and advanced vehicle technologies (such as electric drive technologies) to increase vehicle energy efficiency have long been studied and promoted worldwide. Alternative transportation fuels and advanced vehicle technologies could help solve urban air pollution problems, reduce GHG emissions, and reduce U.S. dependence on imported oil. To evaluate the energy and emission effects of alternative fuels and vehicle technologies accurately and adequately, researchers must consider emissions and energy use from fuel production processes as well as from vehicle operations (so-called fuel cycle analysis). This is especially important for technologies that employ fuels with distinctly different primary energy sources and fuel production processes because the emissions and energy use associated with fuel production and distribution can be significantly different.

1. INTRODUCTION

Fuel cycle analysis is requisite when comparing different vehicle technologies powered by different fuels. Figure 1 shows the significantly different greenhouse gas (GHG) emission results obtained by considering only vehicle tailpipe emissions (pump to wheels) versus considering both fuel production (well to pump) and tailpipe emissions (fuel cycle). A number of studies have been conducted to estimate fuel cycle emissions and energy use associated with various transportation fuels and vehicle technologies. The results of those studies were influenced by the assumptions made by individual researchers regarding technology development, emission controls, primary fuel sources, fuel production processes, and many other factors. Because different methodologies and parametric assumptions were used by different

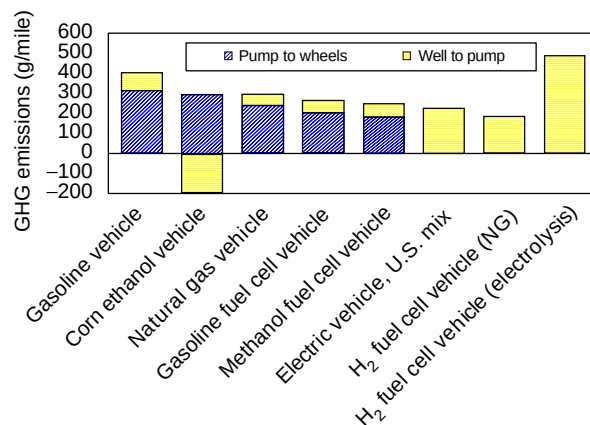


FIGURE 1 Tailpipe vs fuel cycle greenhouse gas emissions of selected vehicle/fuel systems. NG, natural gas.

researchers, it is difficult to compare and reconcile the results of different studies and to conduct a comprehensive evaluation of fuel cycle emissions and energy use. Computer models for calculating emissions and energy use are needed to allow analysts and researchers to test their own methodologies and assumptions and make accurate comparisons of different technologies.

The effects on energy use and emissions of using advanced vehicle/fuel systems can differ considerably because (1) the production technologies and distribution pathways for the fuels are significantly different and (2) the fuel economy and emissions performance associated with the vehicles powered by these fuels vary. If the energy and environmental effects of vehicle/fuel systems are to be thoroughly examined, the fuels must be evaluated on the basis of a full fuel cycle, including both fuel production and in-vehicle usage.

2. ALTERNATIVE TRANSPORTATION FUELS AND ALTERNATIVE FUEL VEHICLES

2.1 Alternative Transportation Fuels

At present, in the United States and worldwide, motor vehicles are fueled almost exclusively by petroleum-based gasoline and diesel fuels. Since the first oil price shock in 1973, efforts have been made to seek alternative fuels to displace gasoline and diesel fuels and achieve energy and environmental benefits. Some of the alternative fuels that have been researched and used are liquefied petroleum gas (LPG), compressed

natural gas (CNG), liquefied natural gas (LNG), methanol (MeOH), dimethyl ether (DME), Fischer-Tropsch diesel (FTD), hydrogen (H_2), ethanol (EtOH), biodiesel, and electricity. Production processes associated with gasoline, diesel, and each of these alternative fuels differ.

For fuel cycle analyses, gasoline often serves as the baseline fuel to which other fuels are compared. Because of emission concerns, requirements for gasoline quality are in place in many countries to help reduce motor vehicle emissions. In the United States, so-called reformulated gasoline, compared to the gasoline used in the 1970s, has much better quality in terms of emission reduction potential. Oxygen additives such as methyl *tert*-butyl ether (MTBE) or ethanol are often mixed with gasoline. To control emissions of nitrogen oxides (NO_x) and particulate matter of size smaller than $10\text{ }\mu\text{m}$ (PM_{10}) with after-combustion treatments for diesel vehicles, regulations are in place or are being developed to maintain the sulfur content of diesel fuels at a very low level. The demand for both low-sulfur gasoline and low-sulfur diesel requires significant desulfurization efforts in petroleum refineries.

Liquefied petroleum gas, primarily propane, is produced in either petroleum refineries or natural gas (NG) processing plants. In the United States, about 40–50% of LPG is produced in petroleum refineries and the remaining 50–60% is produced in NG processing plants. Natural gas can be delivered to vehicle refueling stations, where it is compressed to above 3600 psi for storage on CNG vehicles, or it can be liquefied in central liquefaction plants and transported to refueling stations for LNG vehicle applications. Compared with compressed natural gas vehicles (CNGVs), liquefied natural gas vehicles (LNGVs) have one distinct advantage, a longer driving range per refueling. However, cryogenic storage of LNG on-board a vehicle presents technical and cost challenges.

Methanol is produced through synthesis of a gaseous mixture of H_2 , carbon monoxide (CO), and carbon dioxide (CO_2). Although methanol can be produced from biomass, coal, heavy oil, naphtha, and other feedstocks, the availability of reasonably cheap NG feedstock makes steam methane reforming (SMR) technology an economical way to produce methanol. SMR is a mature technology and is widely used in existing methanol plants. Methanol plants may be located near remote gas fields to take advantage of cheap, remote NG supplies.

Dimethyl ether, which has physical properties similar to those of LPG, has been proposed and tested as an alternative to diesel fuel in compression-

ignition (CI) engines. Use of DME in diesel engines offers NO_x and PM emissions reduction benefits. DME is now used predominantly as an aerosol propellant and is produced from methanol through a dehydration process. Research and development efforts have focused on technologies to produce DME directly from coal and NG.

The Fischer-Tropsch (FT) process produces middle distillates containing no sulfur and virtually no aromatics; the process also produces naphtha and wax. Using FT diesel in CI engines helps reduce NO_x and PM emissions. The FT reaction process was used by Germany during World War II to produce diesel fuel from coal and by South Africa during the oil embargo against that country's apartheid. Currently, several major companies are actively pursuing the production of FTD from remotely located ("stranded") natural gas.

Although H_2 can be produced from many feedstocks, most H_2 is currently produced from NG via SMR technologies. Several alternative schemes, such as partial oxidation (POX), autothermal reforming (ATR), and plasma reforming, have been developed and used commercially to produce H_2 . Hydrogen plants can generate a significant amount of steam; some of the steam produced is used as process steam within the plant, and the remainder can be exported to nearby chemical plants. The gaseous H_2 ($G.H_2$) produced can be compressed to above 5000 psi for storing on board hydrogen-powered vehicles, or the $G.H_2$ can be liquefied to produce liquid H_2 ($L.H_2$) and used in hydrogen-powered vehicles. Besides production in central plants, H_2 can be produced in refueling stations to avoid construction of expensive hydrogen distribution infrastructure. In addition, H_2 can be produced via electrolysis of water with electricity that is generated by using different energy sources.

Biomass can be used to produce ethanol and biodiesel. Ethanol fuel is currently produced from grains such as corn and sugarcane. Research and development efforts to develop and commercialize technologies to produce ethanol from cellulosic biomass are ongoing. Methyl or ethyl esters that are produced from vegetable oils and animal fats are commonly called biodiesel (BD). Biodiesel, which can be used in diesel engine vehicles, is produced via a transesterification process from natural vegetable oils (soybean oil, cotton oil, and rapeseed oil). Outside the United States production is mainly from rapeseed oil; in the United States, biodiesel is produced primarily from soybeans.

Electricity can be used directly as a transportation fuel in battery-powered electric vehicles (EVs); in

combination with an onboard engine, electricity can supply the motive power for hybrid electric vehicles (HEVs). Also, electricity can be used to produce H_2 via electrolysis of water. The energy use and emission effects of electricity as a transportation fuel are primarily determined by the type of fuels (e.g., coal, NG, or oil) used and the type of fuel combustion technologies (e.g., steam boilers, simple-cycle gas turbines, or combined-cycle gas turbines for NG-fired power plants). Because the electric generation mix is so different from region to region, the energy and emission effects of electricity also vary greatly.

2.2 Alternative Fuel Vehicles

Figure 2 shows the numbers and types of alternative fuel vehicles on the road in the United States. Besides these conventional vehicle technologies, advanced vehicle technologies such as HEVs and fuel cell vehicles (FCVs) are also under development. Past fuel cycle analyses have covered both conventional and advanced vehicle technologies. Some of the alternative fuels (such as methanol, ethanol, and biodiesel) can be blended with gasoline or diesel for vehicle applications.

3. FUEL CYCLE ANALYSES OF VEHICLES AND TRANSPORTATION FUELS

The fuel cycle for a given transportation fuel includes the following processes: energy feedstock (or primary

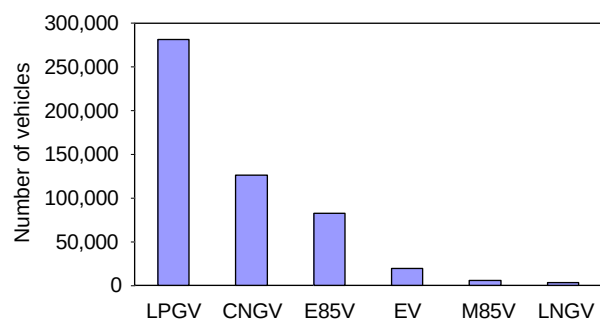


FIGURE 2 Stocks of alternative fuel vehicles in the United States in 2002. LPGV, Liquefied petroleum gas vehicles; CNGV, compressed natural gas vehicles; E85V, vehicles fueled by 85% ethanol/15% gasoline (by volume); EV, electric vehicles (battery powered); M85V, vehicles fueled by 85% methanol/15% gasoline (by volume); LNGV, liquefied natural gas vehicles. Data from the Energy Information Administration (U.S. Department of Energy) Web site at <http://www.eia.doe.gov>, accessed on Nov. 12, 2002.

energy) production; feedstock transportation and storage; fuel production; fuel transportation, storage, and distribution; and vehicle operations that involve fuel combustion or other chemical conversions (Fig. 3). The processes that precede vehicle operations are often referred to as the well-to-pump (WTP) stage, the vehicle operations are referred to as the pump-to-wheels (PTW) stage, and the entire fuel cycle is referred to as the well-to-wheels (WTW) cycle. Various models have been developed that allow researchers to conduct fuel cycle analyses of vehicle/fuel systems. The GREET (greenhouse gases, regulated emissions, and energy use in transportation) model, developed by Argonne National Laboratory, is in the public domain and, so far, more than 1000 users worldwide have used the model to predict the energy use and emissions associated with alternative fuels and vehicles. The GREET model and its documents are posted on the Internet at <http://greet.anl.gov>.

Mark Delucchi of the University of California at Davis has developed another transportation fuel cycle models. The model has been used by Delucchi for applications in the United States, Canada, and Europe. Another analytic tool, the TEAM model, developed by Ecobalance, is a commercial model used primarily for life cycle analysis of consumer products. In certain configurations, the model can be used to analyze vehicle/fuel systems. The National Renewable Energy Laboratory in the United States has applied the model to analyze bioethanol and biodiesel. Shell (Royal Dutch Petroleum Company) recently applied the model to analyze FT diesel.

For fuel cycle analysis of vehicle/fuel systems, the results need to be per distance traveled. If the analysis is conducted for each unit of energy produced and used, the energy efficiency and emission differences during the pump-to-wheels (vehicle operation) stage may be excluded, resulting in incomplete analysis of

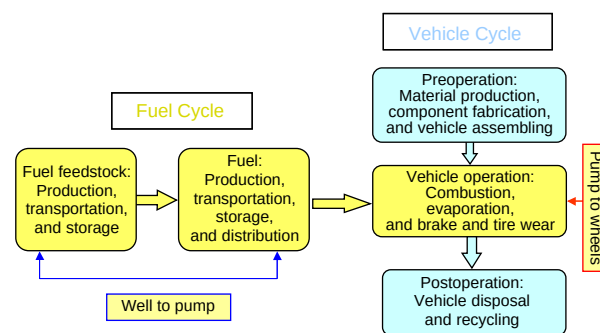


FIGURE 3 Coverage of a fuel cycle analysis.

the systems. To estimate well-to-wheels energy use and emissions, the well-to-pump energy use and emissions per unit of fuel available at refueling pumps are estimated first. The WTP estimates take into account energy uses (including use of process fuels and energy losses during WTP activities) and emissions associated with the activities, as discussed in reports prepared by Delucchi and Wang about their models. Estimates of WTW emissions usually include emissions of greenhouse gases [primarily CO₂, methane (CH₄), and nitrous oxide (N₂O)] and criteria pollutants such as volatile organic compounds, CO, NO_x, PM₁₀, and sulfur oxides (SO_x). Emissions occurring during WTP activities include those resulting from the combustion of process fuels and from noncombustion processes such as chemical reactions, fuel leakage, and evaporation. Of these pollutants, CO₂ emissions are relatively easy to estimate because they depend almost exclusively on the carbon content of the fuels used. Emissions of other pollutants are affected by combustion technologies, as well as the amounts and the types of fuels used. To a large extent, emissions estimates for these other pollutants rely heavily on emission factors (in grams per unit of fuel used) for different combustion technologies fueled by different process fuels. These emission factors can be obtained from databases such as the AP-42 document created by the U.S. Environmental Protection Agency (EPA).

Sulfur oxide emissions are calculated by assuming that all sulfur contained in process fuels is converted into sulfur dioxide (SO₂) for combustion technologies fueled with process fuels except for coal and residual oil. The SO_x emissions for these two process fuels are usually subject to after-combustion control devices. PTW energy use and emissions are calculated on a per-mile basis. Usually, energy use per distance traveled is calculated from vehicle fuel economy values, which are available from measurements or vehicle simulations. Vehicle emissions values are available from emission certifications, estimations of on-road vehicle operations, and laboratory tests.

Although fuel economy and emissions for benchmark gasoline vehicles and diesel vehicles are presented as actual values (e.g., miles per gallon for fuel economy and grams per mile for emissions), fuel economy and emissions for alternative fuel vehicles (AFVs) are usually presented as ratios to benchmark vehicles. In this way, differences in vehicle operating conditions can be reduced to some extent, as long as the fuel economy and emission ratios are for similar vehicle types and under similar operating conditions. For criteria pollutants, location (e.g., urban, rural) is

also important, and fuel cycle analyses need to be conducted with a special focus on where emissions occur. In the GREET model, criteria pollutant emissions are separated into “total emissions” (those that occur everywhere) and “urban emissions” (those that occur within metropolitan areas, where large populations are exposed to the pollutants).

4. FUEL PATHWAYS AND VEHICLE/FUEL COMBINATIONS

Table I shows different types of vehicle propulsion systems and the transportation fuels that have been studied for their potential to power the vehicles. Gasoline, CNG, LNG, LPG, methanol, ethanol, and hydrogen can be used in vehicles equipped with conventional spark-ignition (SI) engines. Interest in developing efficient, low-emission, spark-ignition direct-injection (SIDI) engine technologies has heightened in recent years. Although vehicles with SIDI engines can be fueled efficiently with liquid fuels (gasoline, methanol, ethanol, and liquid hydrogen), gaseous fuels do not appear to offer inherent fuel economy benefits when used in these engines. Compression-ignition direct-injection (CIDI) engines, which can be fueled by diesel, DME, FT diesel, and biodiesel, help increase vehicle fuel

TABLE I
Potential Combinations of Vehicle Technologies and Transportation Fuels

Vehicle technology ^a	Potential fuels for use ^b
Conventional SI engine	Gasoline, CNG, LNG, LPG, MeOH, EtOH, H ₂
SIDI engine	Gasoline, MeOH, EtOH, L.H ₂
CIDI engine	Diesel, dimethyl ether, Fischer–Tropsch diesel, biodiesel
SIDI engine HEV	Gasoline, CNG, LNG, LPG, MeOH, EtOH, L.H ₂
CIDI engine HEV	Diesel, dimethyl ether, Fischer–Tropsch diesel, biodiesel
Battery-powered EV	Electricity
FCV	Hydrogen, MeOH, gasoline, EtOH, diesel, CNG, LNG, LPG, naphtha

^a Abbreviations: SI, spark ignition; CIDI, compression ignition, direct injection; HEV, hybrid electric vehicle; EV, electric vehicle; FCV, fuel cell vehicle.

^b Abbreviations: CNG, compressed natural gas; LNG, liquefied natural gas; LPG, liquefied petroleum gas; MeOH, methanol; EtOH, ethanol; L.H₂, liquid hydrogen.

economy substantially, although they face the challenge of reducing NO_x and PM emissions.

Hybrid electric vehicles achieve significant fuel economy gains compared with conventional vehicle powertrain configurations. HEVs can be equipped with conventional SI, SIDI, or CIDI engines. HEVs can also be designed to obtain electricity from the electric grid and combine it with power from onboard engines. Battery-powered EVs, charged with grid electricity, offer large fuel economy gains and zero tailpipe emissions. The limited driving range, limited performance characteristics, and high battery costs of EVs limit their market penetration, but EVs continue to be promoted.

Major automakers and the government have programs aimed at development of fuel cell vehicles. Although the ultimate fuel for FCVs is hydrogen, hydrocarbon fuels are being considered, at least for a transition period, before expensive hydrogen production and distribution infrastructure is in place. In theory, any hydrocarbon fuel (e.g., methanol, gasoline, ethanol, diesel, CNG, LNG, LPG, and naphtha) can be used to produce hydrogen on board FCVs. Use of hydrocarbon fuels in FCVs requires onboard fuel processors to produce H_2 .

5. ENERGY AND EMISSIONS: ITEMS OF INTEREST

A typical life cycle analysis includes analysis of the inventory of environmental releases, evaluation of potential environmental impacts, and interpretation of the inventory analysis and impact evaluation. Environmental releases include emissions to air, water, and solid waste emissions; these are defined in document 14040 of the International Standards Organization (ISO), *Environmental Management, Life-Cycle Assessment—Principles and Framework*. Prior fuel cycle analyses of vehicle/fuel systems have focused primarily on air emissions and energy use. For air emissions, studies often include the following criteria:

1. Greenhouse gas emissions: carbon dioxide (CO_2), methane (CH_4), nitrous oxide (N_2O), and several other gases, including hydrofluorocarbons, perfluorocarbons, and sulfur hexafluorides.
2. Criteria pollutant emissions: volatile organic compounds (VOCs), carbon monoxide (CO), nitrogen oxides (NO_x), particulate matter [usually smaller than $10\text{ }\mu\text{m}$ (PM_{10})], and sulfur oxides (SO_x).
3. Emissions of air toxics: acetaldehyde, benzene, 1,3-butadiene, and formaldehyde.
4. Energy use parameters: total energy use, fossil energy use, and petroleum use.

To estimate CO_2 -equivalent GHG emissions, the GHGs are often combined with their global warming potentials (GWPs) as described by the Intergovernmental Panel on Climate Change (IPCC) in its 2001 report, *Climate Change 2001: The Scientific Basis, Technical Summary of the Group I Report*. The three major GHGs for evaluating vehicle/fuel systems are CO_2 , CH_4 , and N_2O . The GWPs of the greenhouse gases vary with the time horizon used to calculate them. With a time horizon of 100 years, GWPs are 1, 23, and 296 for CO_2 , CH_4 , and N_2O , respectively.

Emissions of criteria pollutants are regulated in the United States by the EPA (the term “criteria pollutants” comes from the EPA’s establishment of criteria for ambient concentrations and emissions of these pollutants). These pollutants primarily cause adverse health effects. Thus, the locations in which they are emitted as well as their amounts are important. Total emissions are emissions that occur everywhere, including unpopulated or sparsely populated regions, and urban emissions are those that occur in urban areas with large populations.

Air toxics are hydrocarbons identified by EPA as potential cancer-causing species. Acetaldehyde, benzene, 1,3-butadiene, and formaldehyde are the major air toxics emitted by motor vehicles. Air toxic emissions of various vehicle/fuel combinations have been assessed in a study at Argonne National Laboratory by Winebrake and colleagues.

All complete fuel cycle analyses of energy impacts include total energy use; which means that all energy sources, such as renewable energy and nonrenewable energy (i.e., fossil energy), are taken into account. Some analyses further analyze fossil energy impacts; in these studies, assessments take into account fossil energy, including petroleum, natural gas, and coal. Because fossil energy resources are finite, analyses of fossil energy use reveal the impacts of particular vehicle/fuel systems on energy resource depletion. This aspect of life cycle analysis is especially important for biofuels such as ethanol and biodiesel, for which analysis of total energy impacts may not provide adequate information because the energy in ethanol and biodiesel comes from solar energy during crop growth, and use of solar energy per se is not a concern. Some studies also include analysis of petroleum displacement effects by vehicle/fuel

systems. This is an important issue for countries that rely on imported oil and seek to reduce their oil consumption.

6. ISSUES INVOLVED IN TRANSPORTATION FUEL CYCLE ANALYSES

There are several key issues that transportation researchers need to keep in mind when conducting and using the data that result from life cycle analysis studies.

6.1 Usefulness of Fuel Cycle Analysis Results

ISO document 14040 identifies four areas for using life cycle analysis (LCA) results: (1) identifying opportunities to improve the environmental aspects of products, (2) helping industry, governments, and nongovernmental organizations make sound environmental decisions, (3) selecting relevant environmental indicators for measuring the environmental performance of products, and (4) marketing claims of the environmental performance of products. For transportation fuel cycle analyses (FCAs), the first two areas are important. The results of FCAs of vehicle/fuel systems are helpful in comparing the energy and emissions performance of different systems on a complete basis. Such comparisons help society identify and pursue the technologies that offer the greatest potential energy and emissions benefits; for example, they can be used to rank technologies in relative terms rather than estimating the absolute amount of energy use and emissions for given technologies. FCA results may be less useful for estimating total energy use and total emissions for all the activities associated with a vehicle/fuel system.

As ISO 14040 states, the results of LCAs in general, and transportation FCAs in particular, can be very helpful in identifying opportunities to improve energy performance and environmental conditions. That is, FCA results should be used in a dynamic way to identify the subsystems that can contribute to improved energy and environmental performance of a vehicle/fuel system and to identify potential system improvements, and not in a static way to show simply which systems are better than others.

6.2 Allocation Methods

For processes that produce multiple products, energy and emission burdens have to be allocated to individual products. ISO 14040 advocates reducing the need to allocate by expanding the system boundary. That is, when comparing two systems, the system boundaries for both can be expanded so that all products from both systems are within a newly expanded single system. In transportation FCAs, the system boundary expansion is also called the displacement method (for a full discussion, see the 1999 report by Wang, *Transportation Fuel-Cycle Model, Volume 1: Methodology, Development, Use, and Results*). In practice, however, the system expansion approach can become time and resource intensive. Also, in some cases, systems may be expanded subjectively without considering marketplace effects. Thus, allocation of energy and environmental burdens is still widely used in transportation FCAs. The burdens can be allocated based on shares of output products in terms of energy content, market value, or mass. Of these three options, the energy content basis is appropriate if all products are energy products. The market value basis could be used if the market values of products will drive market behaviors, which can, in turn, affect technologies. The mass basis usually is the least favorable method for transportation FCAs. In any event, all products need to be taken into account in evaluating multiple-product processes and systems. Ignoring coproducts, as some past studies have done, results in unfair treatment of the primary product of a process or a system. Also, the fact that the coproducts indeed have market values is ignored.

6.3 Technology Maturity Levels

In comparative transportation FCAs, different fuels and vehicle technologies are often analyzed together. Comparisons of technologies should take into consideration different commercial maturity levels; some fuels and technologies are already mature and in use, whereas others may still be in the research and development stage. Because FCAs generally do not address technology maturity and commercial readiness, it is important to remember that FCAs and their results cannot be used to assess the commercial viability of fuels and vehicle technologies. FCAs simply provide information on energy and emission effects that would occur if the evaluated fuels and technologies were to be introduced into the marketplace.

The level of technology maturity also affects FCA data quality; For technologies at the research and

development stage, few or no actual operational data are available. Sources of data for these technologies are often limited to engineering analyses, predictions, and simulations. There is typically an abundance of reliable performance data for fuels and vehicle technologies already in the marketplace.

6.4 Uncertainties in Input Parameters and Output Results

Because FCAs cover many activities from wells to wheels, there is a demand for data. Data associated with different activities can come from a wide variety of sources of differing quality, and different engineering designs and facility operation can vary significantly. To address the reality that the data obtained are subject to a large degree of uncertainty, many FCAs have been conducted using sensitivity analyses that change key input parameters. Many researchers now incorporate Monte Carlo simulations in their studies to address uncertainties on a systematic basis. With Monte Carlo simulations, inputs parameters and output results are associated with probability distribution functions.

6.5 Transparency

Transportation FCA results are determined by assumptions regarding input parameters, simulation methodologies, and system boundaries. These assumptions need to be presented explicitly in transparent ways to allow studies to be compared and to make clear whether general conclusions can be drawn from individual studies. Sensitivity analyses can be included in FCA studies to demonstrate the importance of some of the assumptions in determining results. Sometimes, seemingly similar studies are combined to generalize conclusions.

7. COMPLETED STUDIES

Some of the results of past studies are summarized in the following sections. Although many of the studies estimated both (1) absolute energy use and emissions of vehicle/fuel combinations and (2) energy use and changes in emissions of alternative vehicle/fuel systems relative to benchmark technologies (usually gasoline-fueled light-duty vehicles), the summary here presents relative changes only, because they are more comparable than absolute amounts across studies.

Of the vehicle types that use alternative transportation fuels, most completed studies evaluated only

passenger car applications. However, many alternative transportation fuels and advanced vehicle technologies can also be applied to heavy-duty trucks. The benchmark vehicles for heavy-duty vehicle comparisons are usually diesel trucks. Because heavy-duty vehicle technologies can be considerably different from light-duty vehicle technologies and because heavy-duty vehicle driving patterns are very different from those for passenger cars, the relative FCA results summarized here are generally not applicable to heavy-duty vehicles.

7.1 Kreucher, Ford Motor Company (1998)

The study by W. M. Kreucher included evaluations of well-to-wheels emissions, energy use, and costs. Vehicle fuel economy for a Ford Escort was simulated by using Ford's proprietary model on the federal test procedure (FTP) cycle. Well-to-pump energy and emissions were simulated with a model developed by Arthur D. Little for the U.S. Department of Energy (DOE). Emissions analysis included CO₂ and criteria pollutants.

7.2 Weiss *et al.*, Massachusetts Institute of Technology (2000)

Weiss and others at the Energy Laboratory at the Massachusetts Institute of Technology (MIT) conducted a study to compare several vehicle propulsion systems fueled by alternative transportation fuels. In conducting the comparison, the researchers projected vehicle and fuel production technologies from 1996 to 2020. Their analysis was primarily for North American applications. Greenhouse gas emissions in that study include CO₂ and CH₄. A midsize passenger car, such as a Toyota Camry, was used for vehicle simulations. Well-to-pump data were based on a critical review of published data. Vehicle fuel economy results were simulated by using an MIT vehicle computer simulation model. One of the study's strengths is vehicle characterization and simulations. The study also presents some detailed information on vehicle performance and technology advancement.

7.3 (S&T)² Consultants, Inc., for Methanex (2000)

The (S&T)² study was conducted primarily for Canadian well-to-pump conditions, but the results

are somewhat applicable to U.S. WTP conditions. The study included GHG emissions only (CO_2 , CH_4 , and N_2O) and employed Canadian oil recovery efficiencies and carbon emissions, which are high because a significant amount of oil is recovered from oil sands in Canada. Also, the study assumed low refinery efficiency and high GHG emissions for petroleum refineries. Emissions from vehicle materials cycles were also included. A derivative of the model developed by Delucchi was used for simulating well-to-wheels GHG emissions. The study is unique in its collection and use of Canadian WTP data on GHG emissions and fuel production efficiencies.

7.4 Ishitani, University of Tokyo (2000)

The study by H. Ishitani was conducted assuming Japanese conditions for both WTP activities and vehicle technologies. Vehicle fuel economy was simulated with a model that the author developed. The study included well-to-wheels energy efficiencies only. It appears that well-to-pump and pump-to-wheels efficiency assumptions were more optimistic in this study than in the other studies reviewed here.

7.5 Louis, Shell Global Solutions (2001)

The Shell study estimated well-to-wheels GHG emissions and energy use for 24 vehicle/fuel systems. Greenhouse gas emissions included CO_2 and CH_4 . The vehicle technologies that were evaluated were internal combustion engine vehicles (ICEVs), HEVs, and FCVs. The analysis for WTP activities was conducted for the United Kingdom primarily with Shell data. A Mercedes A-class passenger car was chosen for comparison. The GREET model developed by Argonne National Laboratory was used in the study to conduct WTW simulations.

7.6 Delucchi, University of California at Davis (2001)

The study by M. A. Delucchi evaluated motor vehicles, forklifts, and space and water heaters. It included emissions of GHGs and criteria pollutants, but not energy use, although the Delucchi model includes estimation of energy use. The CO_2 -equivalent GHG emissions addressed in the study were CO_2 , CH_4 , N_2O , nonmethane organic compounds (NMOCs), CO , NO_x , SO_x , PM, and hydrofluorocarbon (HFC)-134a and chlorofluorocarbon (CFC)-

12. Delucchi included both fuel and vehicle cycles in his analysis. The results summarized in Fig. 3 are fuel cycle results. Delucchi showed that vehicle cycle GHG emissions could be 16% of vehicle tailpipe GHG emissions.

7.7 Arthur D. Little (2002)

Arthur D. Little (ADL) has conducted studies for the DOE and the state of California since the early 1990s. The report cited here was prepared for the DOE to evaluate fuel choices for FCVs. The study considered energy efficiencies, GHG emissions, costs, and safety for use of the following fuels in FCVs, ICEVs, and battery-powered EVs: hydrogen, methanol, ethanol, gasoline, and diesel. The analysis was conducted for year 2010. ADL developed a vehicle model to simulate vehicle fuel economy under different driving cycles. A modified GREET version was used for WTW simulations.

7.8 Summary

Tables II and III summarize changes in WTW energy use (Table II) and GHG emissions (Table III) by vehicle/fuel system relative to baseline gasoline vehicle data from the studies described in the preceding sections. Energy use results from five of the seven studies (the Methanex and Delucchi studies did not address energy use) indicate that, of the nine SI engine vehicle/fuel systems, use of cellulosic ethanol, corn ethanol, methanol, and H_2 in SI engines could lead to increased WTW energy uses, whereas use of LPG and CNG results in reduced energy use. Of the three CIDI engine vehicle/fuel systems, use of FT diesel could result in increased WTW energy use, because of large energy losses during FT diesel production. Use of all of the eight HEVs results in reductions in energy use. Among the FCVs, use of electrolysis H_2 , cellulosic ethanol, and corn ethanol could result in increased energy use, because of very large energy losses during WTP activities for these fuels. Of the four studies analyzing gasoline FCVs, only the MIT study concluded that gasoline FCVs could have higher WTW energy use compared to baseline gasoline vehicles. Of the five studies that addressed EVs, two showed increases in energy use by EVs, but three showed reductions.

The WTW GHG emission changes presented in Table III for six of the reviewed studies (the study by the University of Tokyo did not present GHG emission results) reveal that, of the nine fuels for SI engine vehicle applications, NG-based LH_2 could

TABLE II

Changes in Well-to-Wheels Total Energy Use by Vehicle/Fuel Systems Relative to Gasoline Vehicles^a

Study ^b	Research group ^c				
	Ford	MIT	Univ. Tokyo	Shell	ADL
SI engine vehicles		1–			
Advanced gasoline ICEV			3–	1–	
LPG ICEV	1–				
CNG ICEV	–				
Corn EtOH ICEV	4+				
Cellulosic EtOH ICEV	3+				
Methanol ICEV	2+				
NG G.H ₂ ICEV				1+	1+
NG L.H ₂ ICEV				2+	
NG MH H ₂ ICEV				1–	
CIDI engine vehicles					
Diesel ICEV	2–	2–		2–	2–
FTD ICEV		2+		1+	
DME ICEV				1–	
HEVs					
Gasoline HEV		2–		2–	2–
CNG HEV		3–		2–	
NG G.H ₂ HEV				2–	
NG L.H ₂ HEV				1–	
NG MH H ₂ HEV				2–	
Diesel HEV		3–		3–	2–
FTD HEV		1–		1–	
DME HEV				2–	
FCVs and EVs					
NG G.H ₂ FCV		2–	3–	2–	3–
kWh G.H ₂ FCV					2+
NG L.H ₂ FCV				1–	2–
NG MH H ₂ FCV				2–	3–
Gasoline FCV		1+	3–	2–	3–
Methanol FCV		1–	2–	2–	2–
Corn EtOH FCV					1+
Cellulosic EtOH FCV					1+
CNG FCV				3–	
FTD FCV				1–	
DME FCV				2–	
EV	4+	2–	3–	2+	2–

^a Change in use: 4+, increase of 60–80%; 3+, increase of 40–60%; 2+, increase of 20–40%; 1+, increase of 0–20%; 1–, decrease of 0–20%; 2–, decrease of 20–40%; 3–, decrease of 40–60%; 4–, decrease of 60–80%; 5–, decrease of greater than 80%.

^b Abbreviations: ICEV, internal combustion engine vehicle; LPG, liquefied petroleum gas; CNG, compressed natural gas; EtOH, ethanol; NG, natural gas; G.H₂, L.H₂, MH H₂, gaseous, liquid, and metal hydride hydrogen; CIDI, compression ignition, direct injection; FTD, Fischer–Tropsch diesel; DME, dimethyl ether; HEV, hybrid electric vehicle; FCV, fuel cell vehicle; EV, electric vehicle.

^c MIT, Massachusetts Institute of Technology; ADL, Arthur D. Little.

result in increased WTW GHG emissions. For methanol, one study showed increased GHG emissions and another study showed reductions. Use of

all of the other seven fuels results in GHG emissions reductions. Of the three fuels for CIDI engine vehicle applications and eight fuels for HEV applications, all

TABLE III

Changes in Well-to-Wheels Greenhouse Gas Emissions by Vehicle/Fuel Systems Relative to Gasoline Vehicles^a

Study	Research group					
	Ford	MIT	Methanex	Shell	Delucchi	ADL
SI engine vehicles		1–				
Advanced gasoline ICEV						
LPG ICEV	1–				2–	
CNG ICEV	1–			2–	2–	
Corn EtOH ICEV	2–				1–	
Cellulosic EtOH ICEV	3–				3–	
Methanol ICEV	1+				1–	
NG G.H ₂ ICEV				1–	2–	1–
NG L.H ₂ ICEV				1+		
NG MH H ₂ ICEV				2–		
CIDI engine vehicles						
Diesel ICEV	1–	2–		2–	2–	2–
FTD ICEV		1–		1–		
DME ICEV				2–		
HEVs						
Gasoline HEV		2–		2–		2–
CNG HEV		3–		3–		
NG G.H ₂ HEV				3–		
NG L.H ₂ HEV				2–		
NG MH H ₂ HEV				3–		
Diesel HEV		3–		3–		2–
FTD HEV		2–		2–		
DME HEV				3–		
FCVs and EVs						
NG G.H ₂ FCV		2–	3–	3–		3–
kWh G.H ₂ FCV			1–			1+
NG L.H ₂ FCV			3–	2–		3–
NG MH H ₂ FCV	3–			3–		4–
Gasoline FCV		1+	2–	2–		2–
Methanol FCV		2–	3–	3–		2–
Corn EtOH FCV						3–
Cellulosic EtOH FCV						5–
CNG FCV				3–		
FTD FCV			2–	2–		
DME FCV				2–		
EV	2–	2–		2–	2–	2–

^a Change in use: 3+, increase of 40–60%; 2+, increase of 20–40%; +, increase of 0–20%; 1–, decrease of 0–20%; 2–, decrease of 20–40%; 3–, decrease of 40–60%; 4–, decrease of 60–80%; 5–, decrease of greater than 80%.

result in GHG emissions reductions. All of the 12 FCV and EV types are shown to have GHG emissions reduction benefits, except for electrolysis H₂ and gasoline for FCV applications, for which two studies showed increases (the four studies showed decreases for these two options).

8. FUEL CYCLE ENERGY AND EMISSION RESULTS OF VEHICLE/FUEL SYSTEMS

Because different studies have different system boundaries and parametric assumptions, the studies

described in Section 7 resulted in different magnitudes of changes in energy use and GHG emissions. The GREET model was used to develop the results presented in this section. In conducting the analyses with the GREET model, key assumptions were specified by taking into account the results from various completed studies. Because some of the vehicle/fuel systems are emerging technologies and are not in the marketplace yet, all technologies were analyzed for the time frame around 2010, so that both emerging technologies and technologies already in the marketplace can be compared within the same period.

8.1 Specifications of and Key Assumptions for Fuel Production Pathways

Table IV presents specifications of and key assumptions of fuel production pathways for North American applications. For each fuel pathway, the table presents the fuel pathway (including feedstocks), key parametric assumptions, and main pathway characteristics. Note that the petroleum gasoline pathway is the baseline pathway to which other fuel production pathways are compared.

8.2 Vehicle Fuel Economy Assumptions

Table V lists the fuel economy of baseline gasoline vehicles (GVs) and the fuel economy ratios of alternative vehicle/fuel systems relative to the baseline GV. The values are based on research at Argonne National Laboratory of various vehicle technologies and fuel economy simulations from studies completed at Argonne and elsewhere. The baseline GV fuel economy [27.5 miles per gallon (mpg)] represents the on-road fuel economy value for a typical midsize passenger car produced in 2010 (there is a considerable penalty from laboratory-tested mpg to on-road actual mpg). The mpg ratios for alternative vehicle/fuel systems are based on test results and vehicle simulations. These values were used in the GREET model to conduct WTW simulations.

8.3 Well-to-Pump Energy Efficiencies

Figure 4 shows the energy efficiencies of various fuel production pathways from well to pump. The efficiencies shown are defined as the energy in a given fuel (available at pumps in vehicle refueling stations) divided by total energy inputs during all

WTP activities, including the energy content of the fuel. One way to interpret the WTP efficiencies in the figure is as the difference between 100% and the energy efficiencies, which roughly represent energy losses during WTP stages for making a given fuel available at the pump. As stated in Section 3, WTP activities include feedstock production; feedstock transportation and storage; fuel production; and fuel transportation, storage, and distribution.

The reformulated gasoline (RFG) pathway is considered to be the baseline pathway to which other pathways are compared. Figure 4 shows that the baseline RFG pathway has a WTP efficiency of about 80%. There are four fuel pathways, LPG, CNG, LNG, and low-sulfur (LS) diesel, that have WTP efficiencies higher than that of the baseline RFG pathway. These pathways have fuel production efficiencies higher than the gasoline refining efficiency (see Table IV).

Fuel pathways in the next group have energy efficiencies of 50–70%. These pathways include biodiesel from soybeans, DME from NG, methanol from NG, FT diesel from NG, G.H₂ from NG, and ethanol from corn. For renewable biodiesel and ethanol, energy losses from photosynthesis during soybean and corn plant growth are not included because solar energy is abundant. However, energy is consumed for soybean and corn farming and production of biodiesel and ethanol. These effects result in lower WTP efficiencies for these two renewable fuel pathways, compared to the baseline RFG pathway. The NG-based pathways for DME, methanol, FT diesel, and G.H₂ have lower WTP efficiencies because they are subject to considerable energy losses during fuel production.

The next five fuel pathways have WTP energy efficiencies of 30–50%. They include European Union (EU) electricity generation, ethanol from cellulosic ethanol, L.H₂ from NG, U.S. electricity generation, and G.H₂ from EU electricity via electrolysis. Electricity generation with the EU/U.S. generation mix has low WTP efficiencies because electricity generation efficiencies are usually low (see Table IV). Note that for nuclear power generation, an energy efficiency of 100% is assumed here, even though nuclear power plant efficiencies are usually in the range of 35–40%. This is because the energy content in uranium used for power generation is not a concern, because a unit of mass of uranium can generate a large amount of electricity. In a way, nuclear energy could be treated in a manner similar to renewable energy. Because of the way nuclear power and renewable power (such as hydropower)

TABLE IV

Specifications and Key Assumptions for Fuel Production Pathways^a

Fuel pathway	Key assumptions	Main pathway characteristics
Petroleum to gasoline	Petroleum refining efficiency of 85.5%	For U.S. refineries; 60% crude is imported and 40% domestic; gasoline with sulfur content of 30 ppm and with EtOH as the oxygenate.
NG to CNG	NG compression efficiency of 93%	NG is recovered in North America
NG to LNG	LNG production efficiency of 90.5%	NG is recovered outside of North America; LNG is produced outside of North America and transported to North America
Petroleum and NG to LPG	NG to LPG production efficiency of 96.5%; petroleum to LPG production efficiency of 93.5%	Of the total LPG produced for North American use, 60% is from NG and 40% from petroleum
NG to MeOH	MeOH production efficiency of 68%	Methanol is produced outside of North America and transported to North America
Corn to EtOH	Energy use for ethanol production of 45,500 Btu/gal; ethanol yield of 2.6 gal per bushel of corn	Both dry and wet milling EtOH plants are included; coproducts (primarily animal feeds) are considered
Herbaceous biomass to EtOH	EtOH yield of 92 gal per dry ton of biomass	Cogenerated electricity in cellulosic ethanol plants is considered
Petroleum to LSD	Petroleum refining efficiency of 87%	For U.S. refineries; 60% crude is imported and 40% domestic; diesel with sulfur content of 15 ppm
NG to DME	DME production efficiency of 70%	DME is produced outside of North America and transported to North America
NG to FTD	FTD production energy efficiency of 63% and carbon efficiency of 80%	FTD is produced outside of North America and transported to North America
Soybeans to biodiesel	Biodiesel yield of 1.36 gal per bushel of soybeans	Coproducts (animal feeds and glycerine) are considered
U.S. average electricity	Coal-fired power plant efficiency of 36.5%; NG-fired power plant efficiency of 41%; oil-fired power plant efficiency of 35%	The U.S. electricity generation mix: 54% coal; 15% NG; 1% oil; 18% nuclear; 12% others
European union average electricity	Coal-fired power plant efficiency of 36.5%; NG-fired power plant efficiency of 41%; oil-fired power plant efficiency of 35%	The European Union electricity generation mix: 25% coal; 18% NG; 7% oil; 34% nuclear; 16% others
NG to G.H ₂	G.H ₂ production efficiency of 71.5%; gaseous H ₂ compression efficiency of 85%	G.H ₂ is produced in North America with North American NG
NG to L.H ₂	G.H ₂ production efficiency of 71.5%; H ₂ liquefaction efficiency of 71%	L.H ₂ is produced outside of North America and transported to North America
Electricity to G.H ₂ via electrolysis	G.H ₂ production efficiency of 71.5%; G.H ₂ compression efficiency of 86%	G.H ₂ is produced at refueling stations
Electricity to L.H ₂ via electrolysis	G.H ₂ production efficiency of 71.5%; H ₂ liquefaction efficiency of 66%	L.H ₂ is produced at refueling stations

^aFuels are for use in North America. Abbreviations: EtOH, ethanol; NG, natural gas; CNG, compressed natural gas; LNG, liquefied natural gas; LPG, liquefied petroleum gas; MeOH, methanol; LSD, low-sulfur diesel; DME, dimethyl ether; FTD, Fischer–Tropsch diesel; G.H₂, L.H₂, gaseous and liquid hydrogen.

are treated here, the EU generation mix has a higher WTP efficiency compared to the U.S. generation mix (the EU generation mix has higher shares of nuclear power and hydropower). L.H₂ from NG has an energy efficiency of about 42% because this pathway is subject to large energy losses during H₂ production and liquefaction. The G.H₂ from EU electricity via electrolysis has an energy efficiency of 32% because

of the energy losses during electricity generation and H₂ production.

The last three pathways are electrolysis-based H₂ production pathways (G.H₂ from U.S. electricity, L.H₂ from EU electricity, and L.H₂ from U.S. electricity). They have WTP efficiencies of 23–30%. These pathways are subject to large energy losses during electricity generation, H₂ production, H₂

TABLE V

Gasoline Vehicle Fuel Economy and Fuel Economy Ratios of Alternative Vehicle/Fuel Systems

Vehicle ^a	Fuel economy ratio (relative to baseline GVs)
Baseline GVs (mpg)	27.5
CNGV	1.05
LNGV	1.05
LPGV	1.08
M90 vehicle	1.08
E90 vehicle	1.08
LS diesel CIDI vehicle	1.21
DME CIDI vehicle	1.21
FTD CIDI vehicle	1.21
BD20 CIDI vehicle	1.21
Gasoline HEV	1.43
LS diesel HEV	1.63
EV	3.50
H ₂ FCV	2.36
Methanol FCV	1.60
Gasoline FCV	1.50

^aAbbreviations: GV, gasoline vehicle; CNGV, compressed natural gas vehicle; LNGV, liquefied natural gas vehicle; LPGV, liquefied petroleum gas vehicle; E90, blend of 90% ethanol/10% gasoline (by volume); LS, low sulfur; CIDI, compression ignition, direct injection; DME, dimethyl ether; BD20, blend of 20% biodiesel/80% petroleum diesel (by volume); EV, electric vehicle.

compression (for G.H₂), and H₂ liquefaction (for L.H₂). The WTP energy efficiency of 23% for L.H₂ from U.S. electricity implies that, if FCVs are to be powered with L.H₂ produced this way, they need to achieve vehicle efficiencies at least four times as high as conventional GVs in order to have equal energy effects—a formidable challenge.

8.4 Well-to-Wheels Total Energy and Fossil Energy Results

The combination of well-to-pump results and the energy use and emissions associated with vehicle operation (also called the pump-to-wheels stage) results in full well-to-wheels energy and emission results for vehicle/fuel systems. The energy efficiencies of 19 fuel pathways were presented in the preceding section. The well-to-wheels energy use and emissions results for 23 representative vehicle/fuel systems, combined with some of the vehicle technology options listed in Table I, are now considered.

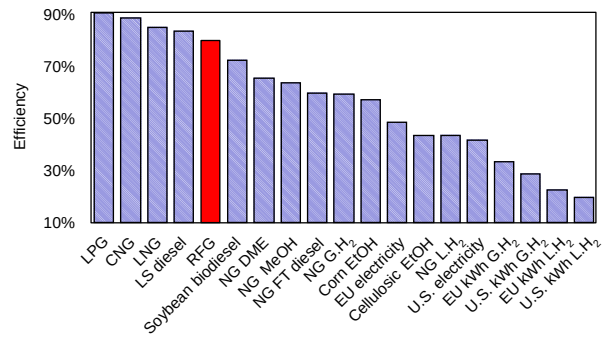


FIGURE 4 Well-to-pump energy efficiencies of fuel production pathways. LPG, Liquefied petroleum gas; CNG, compressed natural gas; LNG, liquefied natural gas; LS diesel, low-sulfur diesel; RFG, reformulated gasoline; NG DME, natural gas (dimethyl ether); NG MeOH, natural gas (methanol); NG FT diesel, natural gas (Fischer–Tropsch diesel); NG G.H₂, natural gas (gaseous hydrogen); corn EtOH, corn-derived ethanol; EU, European Union; cellulosic EtOH, cellulosic-derived ethanol; NG L.H₂, natural gas (liquid hydrogen).

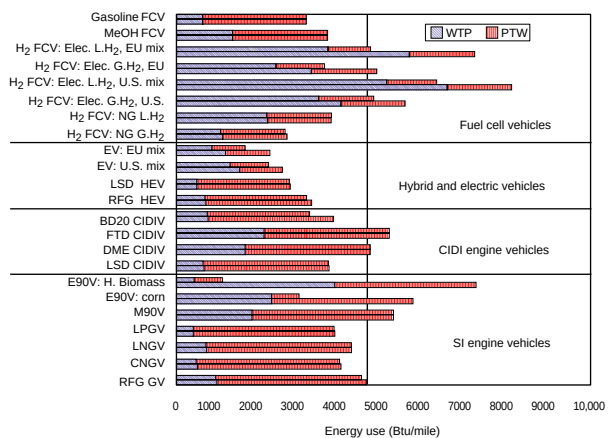


FIGURE 5 Well-to-wheels total energy use and fossil energy use of vehicle/fuel systems. WTP, Well to pump; PTW, pump to wheels; FCV, fuel cell vehicle; LSD, low-sulfur diesel; HEV, hybrid electric vehicle; BD20, blend of 20% biodiesel/80% petroleum (by volume); CIDI, compression-ignition, direct-injection vehicle; SI, spark ignition; FTD, Fischer–Tropsch diesel; E90 V, vehicles fueled by 90% ethanol/10% gasoline (by volume); M90 V, vehicles fueled by 90% methanol/10% gasoline (by volume); GV, gasoline vehicle; see legend to Fig. 4 for other abbreviations.

Figure 5 shows WTW results for total energy use and fossil energy use. Total energy use includes both fossil and renewable energy sources. Fossil energy use includes three fossil energy sources, petroleum, NG, and coal. In this and the following six figures, the 23 vehicle/fuel systems are separated into four groups by vehicle technologies: SI engine vehicles, CIDI engine vehicles, HEVs and battery-powered EVs, and FCVs. For each system, the bottom bar represents total energy use and the top bar represents

fossil energy use. Also, for each bar, the results are separated into well-to-pump and pump-to-wheels components.

For total energy use, of the seven SI engine vehicle options, use of methanol and ethanol could increase total energy use relative to baseline RFG vehicles. The increases are caused by high amounts of energy use during the WTP stage. On the other hand, use of CNG, LNG, and LPG results in reduced total energy use, because of lower WTP energy use. Among the four CIDI vehicle options, use of DME, and especially FT diesel, could lead to increases in total energy use. But, because of high CIDI engine efficiencies, use of LS diesel and a blend of LS diesel and biodiesel (BD20; a blend of 20% biodiesel and 80% LS diesel) results in reduced total energy use. All four HEV and EV options reduce total energy use, because HEVs and EVs have much higher vehicle efficiencies compared to vehicles with SI engines (see Table V). The eight FCV options have diversified WTW energy results. The four electrolysis-based H_2 pathways actually increase total energy use, even though FCVs are much more efficient compared to SI engine vehicles. This is because electrolysis-based H_2 pathways have much lower efficiencies than the baseline RFG fuel pathway (Fig. 4) has. The two NG-based H_2 pathways and the two hydrocarbon fuels (gasoline and methanol) achieve energy benefits.

For fossil-energy-based fuels (i.e., fuels produced from petroleum and gasoline), WTW fossil energy results are similar to total energy results. However, renewable fuels (ethanol and biodiesel) achieve great fossil fuel benefits, even though they have potential total energy use penalties. Also, H_2 produced from both EU and U.S. electricity has much smaller increases in fossil fuel use because a significant portion of electricity is generated from nonfossil energy sources.

8.5 Well-to-Wheels GHG Emissions

Figure 6 shows WTW GHG emissions of the 23 vehicle/fuel systems. For each system, the bottom bar represents CO_2 -equivalent emissions of CO_2 , CH_4 , and N_2O , combined with their global warming potentials (GWPs). The top bar represents CO_2 emissions only. For the two ethanol pathways (corn and cellulosic ethanol), there are some negative emissions. They represent carbon uptake during biomass growth. The carbon is eventually emitted to the air during ethanol combustion. Thus, the net results from the two ethanol pathways are the difference between the bars on the right of the axis

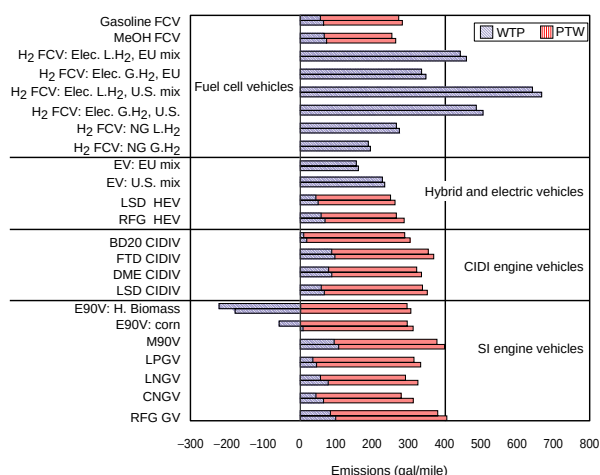


FIGURE 6 Well-to-wheels greenhouse gases and CO_2 emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

and the bars on the left. SI engine vehicles fueled with cellulosic ethanol have the lowest GHG emissions because a large amount of carbon is sequestered during biomass growth. Furthermore, biomass, not fossil fuels, is used in cellulosic ethanol plants to provide heat and steam for plant operations.

The group with the next lowest GHG emissions includes battery-powered EVs and FCVs fueled with G.H₂ made from NG. In the EV case, the very high efficiency overwhelmingly outweighs the low efficiency of electricity generation. In addition, some of the electricity used by EVs is generated from zero-emission sources such as nuclear power and hydropower. In the FCV case, the vehicle efficiency outweighs the inefficiency of G.H₂ production and compression. The third group with low GHG emissions includes HEVs, FCVs fueled with L.H₂ from NG, FCVs fueled with methanol and gasoline, and CIDI vehicles fueled with BD20. These vehicle/fuel systems achieve moderate GHG emissions reductions either because of efficiency gains of the vehicle technologies or because of reduced WTP GHG emissions for fuel production and distribution. The fourth group achieves some GHG emission reductions. This group includes SI engine vehicles fueled with CNG, LNG, LPG, M90 (a blend of 90% methanol and 10% gasoline, by volume), and E90 (a blend of 90% ethanol and 10% gasoline, by volume) with ethanol from corn; CIDI vehicles fueled with LS diesel, DME, and FT diesel; and FCVs fueled with G.H₂ from EU electricity. The last group includes FCVs fueled with L.H₂ from EU and U.S. electricity and G.H₂ from U.S. electricity. These vehicle/fuel systems actually increase GHG emissions because of

the high energy losses during electricity generation, H₂ production, H₂ compression (for gaseous H₂), and H₂ liquefaction (for liquid H₂).

For most vehicle/fuel systems, CO₂ emissions are slightly lower than GHG emissions, because CH₄ and N₂O emissions are excluded in the former. In a few cases, such as with ethanol and NG-based fuels, the differences between CO₂ and GHG emissions could be significant. In these cases, exclusion of CH₄ and N₂O emissions could give the evaluated fuel options unwarranted benefits.

8.6 Well-to-Wheels VOC Emissions

Figure 7 shows total VOC emissions of the evaluated vehicle/fuel systems. For each technology option, the bottom bar represents total VOC emissions (emissions that occur everywhere), whereas the top bar represents urban emissions (emissions that occur within urban areas). Urban emissions, a subset of the total emissions, are determined by information on locations of facilities within the United States. For vehicle operation emissions, urban emissions are estimated by using the vehicle distance traveled within U.S. urban areas. The separation of criteria pollutant emissions is intended to provide information for potential human exposure to the air pollution caused by these pollutants. Because the major concern with criteria pollutants is human health effects, the results of urban emissions are discussed here, even though there are large differences between total and urban emissions for many technology options.

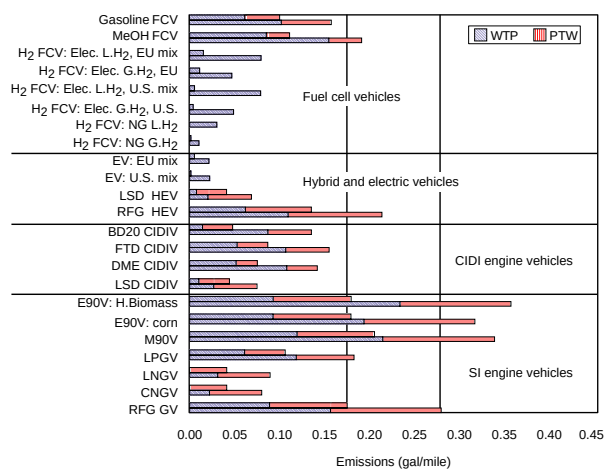


FIGURE 7 Well-to-wheels total and urban volatile organic compound emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

Of the seven SI engine technology options, CNGVs, LNGVs, and LPGVs achieve large urban VOC emissions reductions. These fuels generally have lower vehicle emissions. The E90 vehicles have VOC emissions equivalent to those of baseline GVs; M90 vehicles have a little higher VOC emissions, mainly because of high emissions from WTP stages. The four CIDI engine vehicle options have much lower urban VOC emissions compared to baseline GVs. Of the two HEV options, RFG HEVs have somewhat lower VOC emissions. Low-sulfur diesel HEVs (equipped with CIDI engines) have much lower VOC emissions. EVs and FCVs fueled by H₂ have the lowest urban VOC emissions: these vehicles are zero-emission vehicles. WTP urban VOC emissions from these options are usually small. FCVs fueled with methanol and gasoline do not achieve VOC emissions reductions as high as those offered by EVs and H₂-powered FCVs, although they offer considerable emissions reductions. This is because WTP urban VOC emissions for methanol and gasoline are relatively large, and production of H₂ from methanol and gasoline used to fuel FCVs generates VOC emissions.

8.6 Well-to-Wheels CO Emissions

Figure 8 presents the WTW CO emissions of the evaluated systems. Internal combustion engines produce a large quantity of CO emissions. Consequently, the technologies powered by these engines have high CO emissions. Also, the two FCV options with onboard fuel processors produce a significant amount of CO emissions. EVs and direct H₂ FCVs

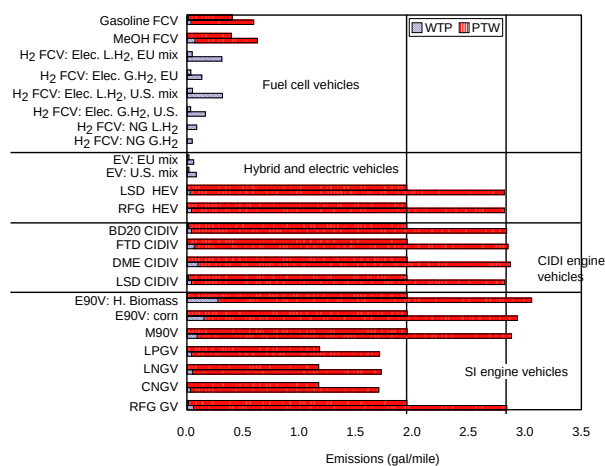


FIGURE 8 Well-to-wheels total and urban CO emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

have much lower CO emissions than the baseline GVs do.

8.7 Well-to-Wheels NO_x Emissions

Figure 9 shows NO_x emissions for the 23 vehicle/fuel systems. The differences between total and urban NO_x emissions for all technology options are large. Of the seven SI engine vehicle options, all options fueled by alternative fuels achieve significant urban NO_x emissions reductions, primarily because of low WTP urban NO_x emissions for these fuels. This is due to the plan to locate production facilities for alternative fuels outside of U.S. urban areas, in contrast to the siting of a large proportion of U.S. petroleum refineries within U.S. urban areas.

The five CIDI engine technology options and the two EV options (including LS diesel HEVs) offer small or no urban NO_x emissions reduction benefits. These technology options have high vehicle emissions. Their lower WTP NO_x emissions are not enough to offset high tailpipe NO_x emissions to achieve large WTW NO_x emissions reduction benefits. In the two EV cases, because some electric power plants are located in urban areas, NO_x emissions from these power plants cause urban NO_x emissions similar to those of baseline GVs.

FCVs fueled with H₂ from NG and with methanol and gasoline have lower urban NO_x emissions as compared to baseline GVs. The nonzero urban NO_x emissions for these vehicle options are caused primarily by their WTP urban NO_x emissions. The four electrolysis H₂ production options with EU and U.S. electricity actually increase urban NO_x emissions, even though FCVs do not generate any NO_x emissions.

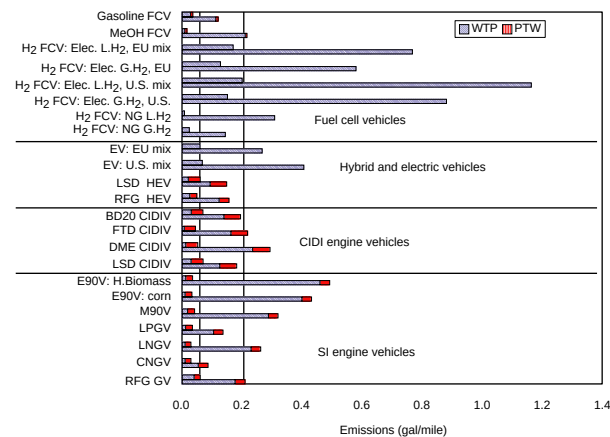


FIGURE 9 Well-to-wheels total and urban NO_x emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

8.8 Well-to-Wheels PM₁₀ Emissions

Figure 10 shows PM₁₀ emissions for the evaluated vehicle/fuel systems. The majority of urban PM₁₀ emissions are from vehicle operations. Of these, a large proportion is from brake and tire wear. This will be especially true when future stricter tailpipe PM₁₀ emission standards take effect. On the other hand, for total PM₁₀ emissions, the technologies relying on electricity as the ultimate fuel have much higher PM₁₀ emissions because of the high PM₁₀ emissions from coal-fired power plants.

8.9 Well-to-Wheels SO_x Emissions

Figure 11 presents SO_x emissions of the evaluated systems. Because SO_x emissions are a regional as well as local air pollutant, both total and urban SO_x

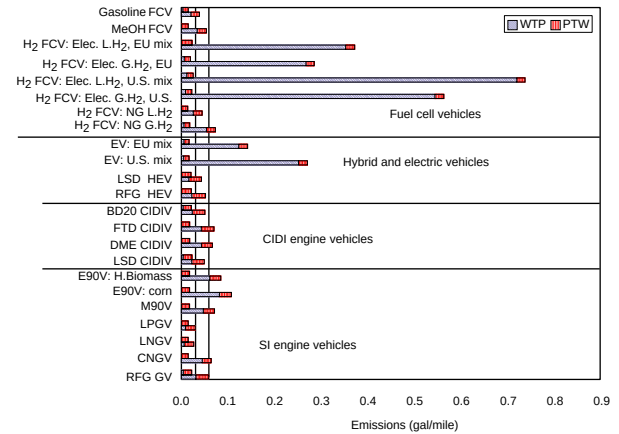


FIGURE 10 Well-to-wheels total and urban PM₁₀ emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

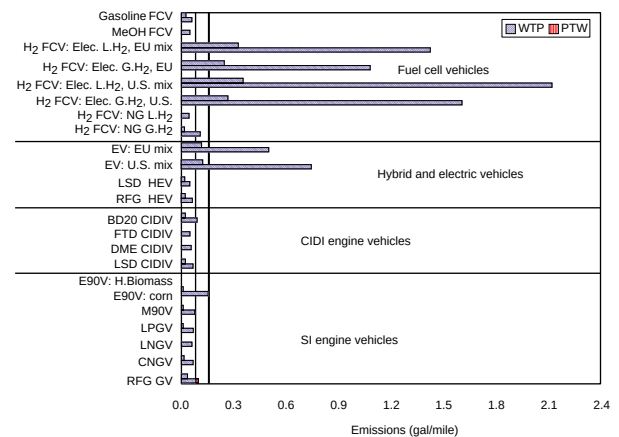


FIGURE 11 Well-to-wheels total and urban SO_x emissions of vehicle/fuel systems; see Figs. 4 and 5 for abbreviations.

emissions could be important. All the alternative vehicle/fuel systems achieve SO_x emissions reduction benefits because RFG (used in the baseline vehicle) has a higher sulfur content than do other fuels, resulting in high tailpipe SO_x emissions. Also, petroleum refineries generate considerable SO_x emissions.

All the options that rely on electricity generate very high total SO_x emissions. Corn-based ethanol and FCVs fueled with G.H_2 from NG generate somewhat high SO_x emissions. In the case of corn ethanol, this is because a significant amount of steam in ethanol plants is generated from coal combustion.

9. CONCLUSIONS

To evaluate the energy and emissions impacts of advanced vehicle technologies and new transportation fuels fully and accurately, researchers must conduct well-to-wheels analyses for two reasons: (1) fuel production pathways can be significantly different for new transportation fuels, resulting in different WTP energy and emission effects, and (2) vehicle tailpipe emissions, especially emissions of criteria pollutants, continue to be controlled and could become a smaller portion of the WTW emissions associated with vehicle/fuel systems.

In general, advanced vehicle/fuel systems achieve energy and emissions benefits relative to the benchmark gasoline vehicle. However, certain fuel pathways must be avoided to achieve the energy and emission benefits of some vehicle/fuel systems. Emission controls for criteria pollutants need to be considered for well-to-pump activities, as well as pump-to-wheels stages, so that advanced vehicle/fuel systems can help solve urban air pollution problems.

SEE ALSO THE FOLLOWING ARTICLES

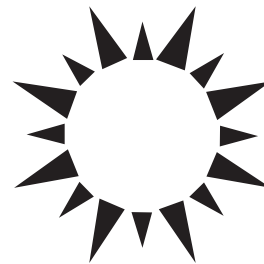
Alternative Transportation Fuels: Contemporary Case Studies • Biodiesel Fuels • Ethanol Fuel • Fuel Cell Vehicles • Fuel Economy Initiatives: International Comparisons • Hybrid Electric Vehicles • Internal Combustion Engine Vehicles • Transportation Fuel Alternatives for Highway Vehicles • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

- Arthur D. Little. (2002). "Guidance for Transportation Technologies: Fuel Choice for Fuel Cell Vehicles." Final report, Cambridge, Massachusetts, prepared for the U.S. Department of Energy, Feb. 6. Arthur D. Little, Inc., Cambridge, Massachusetts.
- Delucchi, M. A. (2001). A life-cycle emissions analysis: Urban air pollutants and greenhouse gases from petroleum, natural gas, LPG, and other fuels for highway vehicles, forklifts, and household heating in the U.S. *World Resource Rev.* 13(1), 25–51.
- Energy Information Administration (EIA). (1998). "International Energy Annual 1996." DOE/EIA-0219(96), Energy Information Administration, Office of Integrated Analysis and Forecasting, U.S. Department of Energy, Washington, D.C.
- Intergovernmental Panel on Climate Change (IPCC). (2001). "Climate Change 2001: The Scientific Basis, Technical Summary of the Group I Report." Cambridge University Press, Cambridge, U.K.
- International Standard Organization (ISO). (1997). "ISO 14040, Environmental Management, Life Cycle Assessment—Principles and Framework." ISO, Barrow, U.K.
- Ishitani, H., Baba, Y., and Kobayashi, O. (2000). Well-to-wheel energy efficiency of fuel cell electric vehicles by various fuels. [Translated from the original paper published in Japanese in *Journal of Energy and Resources*.]
- Kreucher, W. M. (1998). "Economic, Environmental and Energy Life-Cycle Inventory of Automotive Fuels." SAE technical paper 982218. Society of Automotive Engineers, Warrendale, Pennsylvania.
- Louis, J. J. J. (2001). "Well-to-Wheel Energy Use and Greenhouse Gas Emissions for Various Vehicle Technologies." SAE technical paper 2001-01-1343. Society of Automotive Engineers, Warrendale, Pennsylvania.
- (S&T)² Consultants, Inc. (2000). "Assessment of Emissions of Greenhouse Gases from Fuel Cell Vehicles." Prepared for Methanex Corporation, June 5. (S&T)² Consultants, Inc., Delta, British Columbia, Canada.
- U.S. Environmental Protection Agency (EPA). (2002). "Inventory of U.S. Greenhouse Gas Emissions and Sinks: 1990–2000." EPA 430-R-02-003, U.S. Environmental Protection Agency, Office of Atmospheric Programs, EPA, Washington, D.C.
- Wang, M. (1999). "GREET 1.5—Transportation Fuel-Cycle Model, Volume 1: Methodology, Development, Use, and Results" ANL/ESD-39, Vol. 1, Center for Transportation Research, Argonne National Laboratory, Argonne, Illinois.
- Wang, M., and Huang, H. (1999). "A Full Fuel-Cycle Analysis of Energy and Emissions Impacts of Transportation Fuels Produced from Natural Gas." Center for Transportation Research, Argonne National Laboratory, Argonne, Illinois.
- Wang, M. Q., Saricks, C. L., and Wu, M. (1997). "Fuel-Cycle Fossil Energy Use and Greenhouse Gas Emissions of Fuel Ethanol Produced from U.S. Midwest Corn." Prepared for the Illinois Department of Commerce and Community Affairs by the Center for Transportation Research, Argonne National Laboratory, Argonne, Illinois.
- Wang, M., Saricks, C. L., and Santini, D. (1999). "Effects of Fuel Ethanol Use on Fuel-Cycle Energy and Greenhouse Gas Emissions." ANL/ESD-38. Center for Transportation Research, Argonne National Laboratory, Argonne, Illinois.
- Wang, M. Q., Saricks, C. L., and Wu, M. (1999). Fuel ethanol produced from U.S. Midwest corn: Help or hindrance to the vision of Kyoto? *J. Air Waste Manage. Assoc.* 49, 756–772.

- Wang, M., Saricks, C. L., and Santini, D. (1999). Greenhouse gas emissions of fuel ethanol produced from corn and cellulosic biomass. *Environ. Mgr.*, October, pp. 17–25.
- Weiss, M. A., Heywood, J. B., Drake, E. M., Schafer, A., and AnYeung, F. F. (2000). “On the Road in 2020, a Life-Cycle Analysis of New Automobile Technologies.” Energy Laboratory Report # MIT EL 00-003. Massachusetts Institute of Technology, Cambridge, Massachusetts.
- Winebrake, J., He, D., and Wang, M. (2000). “Fuel-Cycle Emissions for Conventional and Alternative Fuel Vehicles: An Assessment of Air Toxics.” ANL/ESD-44, Center for Transportation Research. Argonne National Laboratory, Argonne, Illinois.

Fuel Economy Initiatives: International Comparisons



STEVEN PLOTKIN

Center for Transportation Research, Argonne National Laboratory
Washington, D.C., United States

-
1. Overview
 2. Fuel Economy Standards
 3. Fuel Taxes
 4. Vehicle Tax Incentives Promoting Higher Fuel Economy
 5. Information Programs to Promote Higher Fuel Economy
 6. Government Sponsorship of Research and Development

Glossary

driving cycle A carefully measured trip that is defined by the vehicle speed during every second of the trip over a specified distance; using a driving cycle allows different vehicles to be tested and compared in a consistent manner.

fuel economy A measure of vehicle fuel efficiency expressed as distance traveled divided by fuel consumed (e.g., miles per gallon, kilometers per liter).

hybrid drivetrains (hybrid vehicles) Generally refers to vehicles that combine a conventional fuel-driven engine with an electric drivetrain, that is, a battery (or other means of storing electricity) and a motor/controller.

light-duty vehicles Automobiles, compact passenger vans, sport utility vehicles, and pickup trucks; in the United States, statutes regulating light-duty vehicles generally restrict these to vehicles less than 8500 pounds “gross vehicle weight” (i.e., curb weight plus carrying capacity of passengers plus cargo).

top runner A means of setting vehicle efficiency, emissions, or other standards by identifying representative vehicles that excel in the characteristic being targeted and using their performance as the basis for setting standards for an entire class of vehicles.

This article describes initiatives aimed at improving the fuel economy of light-duty passenger vehicles—automobiles and light trucks (sport utility vehicles,

pickup trucks, and small passenger vans). The article’s primary focus is on direct measures that create fuel economy targets for national and/or corporate fleets of new vehicles.

1. OVERVIEW

Reducing the growth of fuel consumption and greenhouse emissions from light-duty vehicles (LDVs) is a critical component of attempts to prevent global warming and to reduce the vulnerability of national economies to the adverse effects of oil price volatility. In the United States, where travel by LDVs is most intensive, LDVs consume nearly 8 million barrels of oil per day, or 40% of U.S. oil consumption, and account for nearly 17% of U.S. emissions of carbon dioxide. According to the U.S. Department of Energy’s Energy Information Administration, energy use and greenhouse emissions from LDVs in the United States are expected to grow by nearly 2% per year over the next few decades. Although the rest of the world is, on average, far behind the United States in vehicle ownership and use (nearly 800 vehicles/1000 population in the United States vs ~120 vehicles/1000 population for the world average and ~30 vehicles/1000 population for the developing world), growth rates of both vehicle ownership and fuel use are much higher outside the United States and especially in the developing world. It is unlikely that U.S. levels of vehicle ownership and per capita personal travel will become the global norm, but it appears inevitable that vehicle ownership and personal travel will continue to grow rapidly throughout the world.

Restraining the growth of greenhouse emissions and oil use of LDVs can be accomplished by improving vehicle efficiency (fuel economy), substituting nonpetroleum fuels (with lower lifecycle greenhouse emissions), reducing travel demand

(e.g., by raising the cost of travel or promoting compact development), and promoting shifts in travel to other modes that are more efficient and/or use nonpetroleum fuels (e.g., rapid transit). All of these approaches are worth pursuing.

In addition to reducing oil consumption and greenhouse emissions, improvements in fuel economy will affect emissions that directly affect human and ecosystem health. The primary positive effect is through the reduced need to produce, refine, and distribute vehicle fuels, yielding a proportional reduction in so-called “upstream” emissions and fewer hydrocarbon emissions from vehicle refueling. However, the effect on vehicle tailpipe emissions is decidedly mixed and may well be negative. On the one hand, reductions in fuel consumption generally yield a proportional reduction in emissions from the engine. On the other hand, this may make little practical difference in actual tailpipe emissions if emissions regulations are written in terms of maximum allowed weight of emissions per unit of distance driven (as they are in the United States, Europe, and Japan). In this case, manufacturers are likely to use fewer emissions controls on those vehicles with inherently lower emissions, yielding the same emissions outcome—with the important caveat that, if vehicle emissions controls malfunction, the more efficient vehicles will then tend to have lower tailpipe emissions than their less efficient counterparts. A decidedly negative impact is associated with any increase in travel caused by the lower per mile fuel costs associated with improved efficiency (the so-called “rebound effect”); this will increase emissions. The overall net impact on health-related emissions of these several effects is not clear.

The three major markets that currently have active fuel economy initiatives—the United States, Japan, and the European Union (EU)—have approached the pursuit of higher levels of fuel economy in different ways. The United States has pursued what some consider to be a contradictory policy, combining an economic policy that maintains low fuel prices with a regulatory policy that demands attainment of fixed fuel economy levels. U.S. standards have remained at 27.5 mpg for passenger cars and between 20 and 21 mpg for light trucks since 1985 (although there is a pending proposal to increase the light truck standard to 22.2 mpg by model year 2007). At the same time, U.S. gasoline prices have been half or less than half the prices prevailing in Europe and Japan (due to much lower tax rates in the United States), and this has provided minimum economic incentive for U.S. motorists to

seek out vehicles with high fuel economy. On the other hand, until recently neither Japan nor the EU had used formal standards to stimulate higher levels of fuel economy, apparently relying on high fuel prices (generally \$3.00/gallon and up) to maintain efficient fleets. However, Japan recently introduced a set of “weight class standards” that require attainment of fuel economy targets in each of nine weight categories, with an estimated fleet fuel economy improvement (from 1995 levels) of about 23%—to approximately 35 mpg—by 2008. Also, the EU accepted a voluntary commitment from its vehicle manufacturers associations for a reduction in “per vehicle” CO₂ emissions of 25% percent—to 140 g CO₂/km—by 2010 (2009 for the Japanese and Korean associations), also from a 1995 baseline. If fuel composition does not change, this corresponds to a fuel economy improvement of 33%—to about 41 mpg.

In addition to fuel taxes and fuel economy standards, approaches to improving LDV fuel economy include

- vehicle purchase and ownership taxes, tax credits, and direct purchase subsidies rewarding high efficiency;
- government requirements for public access to information about fuel economy and fuel costs, allowing vehicle purchasers to appropriately incorporate these factors into their purchase decisions;
- government sponsorship of research and development (R&D) on higher efficiency vehicles;
- government requirements for fleet purchase of high-efficiency vehicles.

Although this article does not address regulations controlling health-related vehicle emissions, many developing nations are beginning to adopt emissions standards based on those of the industrialized nations, especially recent EU standards. Complying with these standards may require adoption of improved engine controls and fuel injection, both of which should yield improved fuel efficiency.

2. FUEL ECONOMY STANDARDS

2.1 Overview

The most direct way in which to improve LDV fuel economy is to establish clear targets for new vehicle fuel economy or CO₂ emissions (although the latter

can theoretically be attained partly, or even solely, by switching to lower carbon fuels). In 1974, U.S. automakers voluntarily agreed to achieve a 40% improvement in fuel economy by 1980. This voluntary agreement was superseded by mandatory standards promulgated in 1975, aiming for still higher gains by 1985. Canada followed the U.S. example with a set of voluntary fuel economy targets mimicking the United States' mandatory ones. Within the past few years, both the EU and Japan have established new targets, the EU through a voluntary commitment with its automobile manufacturers associations and Japan through a set of mandatory fuel economy targets based on vehicle weight classes.

Setting fuel economy or CO₂ emissions targets for future vehicles or fleets raises a number of important concerns. Some of the key concerns include the following:

- *Economic efficiency and impacts.* Many in the environmental community have championed new fuel economy standards because they appear to guarantee results and perhaps because the U.S. standards appear to have been quite successful. However, economists have generally not supported fuel economy standards because they fear that such standards may force manufacturers to produce vehicles that consumers will not want, they do not address the demand for vehicle travel (and actually will tend to boost travel by reducing fuel costs, thereby offsetting some of the fuel savings associated with the increased efficiency—the rebound effect), and they inevitably create strong incentives for manufacturers to “game play” to take advantage of quirks in the regulations. A number of studies have compared fuel economy standards with other means of reducing oil consumption and greenhouse gas emissions, and others have examined the economic impacts of new standards, particularly the impact on employment. A thorough examination of these studies would require a separate discussion, but they can be summarized with extreme brevity as follows. First, standards generally are less efficient than direct economic incentives, in particular fuel taxes, which will suppress travel as well as stimulate higher fuel economy. However, there is little agreement about the magnitude of the difference in economic efficiency among the different approaches due to the wide range in estimates of the fuel price sensitivity of travel and fuel economy and similar disagreements about the actual costs of new standards. Second, employment impact studies sponsored by the vehicle manufacturers generally conclude that new standards

will reduce employment, whereas studies sponsored by the environmental community conclude that new standards will increase employment. These differences are driven primarily by very different underlying assumptions about the success and costs of new standards as well as about the approach that manufacturers will take to comply with the standards.

- *Selection of targets and assumptions about future technological and market development.* Unless fuel economy targets are quite modest and short term, the process of developing reasonable standards requires a set of assumptions about how automotive technology will change within the time frame of the standards, how consumer tastes will change (or not change), and how manufacturer market behavior will change (e.g., whether companies will go after new markets, whether companies will produce new classes of vehicles). Analyses leading to new standards have tended to avoid these questions by assuming little or no change, for example, no change in vehicle attributes such as weight and power (except possibly to address existing emissions and safety standards). All aspects of the vehicle marketplace have changed substantially over time, with strong impacts on fuel economy improvement potential.

Those who might wish to apply any of the existing standards to their own markets, or who simply wish to understand how ambitious the targets are, should pay careful attention to a variety of factors that affect the fuel economy potential of a particular market. The more important factors include the following:

- *Fuel taxes.* All else equal, higher fuel taxes and prices will mean that vehicle purchasers will have greater incentive to buy more fuel-efficient vehicles and manufacturers will have greater incentive to make them. Although raising fuel prices may be a desirable strategy for increasing fleet fuel efficiency (and reducing vehicle miles driven), higher fuel taxes might not be politically viable. In any case, attaining a given fuel economy target should generally be easier in markets with high fuel prices.

- *Driving cycles and conditions.* The driving cycle over which fuel economy is measured can strongly influence the performance of fuel economy technologies such as improved engines and use of lightweight materials. In general, regulators try to measure fuel economy on a driving cycle that corresponds well to actual driving conditions in their market. For example, Japanese driving conditions are best represented by congested city driving, and

the official Japanese “10.15” driving cycle used to measure fuel economy for regulatory purposes is a slow, stop-and-go cycle. (10.15 refers to the two sets of speed profiles in the cycle: one with 10 and the other with 15 separate driving modes—acceleration, deceleration, or constant speed driving. The 10-mode portion of the cycle is driven three times, followed by one 15-mode portion.) The U.S. regulatory cycle uses much faster speeds, with the European cycle somewhat in between, and these cycles represent average driving conditions in their markets reasonably well. One effect of these differences is to reward certain key technologies that focus on improving “low-load” efficiency—variable valve timing, hybrid drivetrains, idle-off, and the like—with significantly higher fuel economy “boosts” on the Japanese cycle than on the U.S. cycle. In contrast, improvements in aerodynamics are well rewarded on the U.S. cycle, and are nearly irrelevant on the Japanese cycle, because aerodynamic drag forces are much more important at high speeds than at low speeds. A critical effect of differences in driving cycles among the different markets is that the fleet fuel economy value and targets in each market are not directly comparable to those in the other markets. In general, if the current fuel economy levels in each market were to be measured on the U.S. driving cycle, the European fleet would have its measured fuel economy levels raised a bit (from their value on the European cycle) and the Japanese fleet would be raised even more. However, the percentage changes needed to “normalize” the fleets to a single driving cycle will change over time as technology changes.

- *Emissions regulations.* Emissions regulations can affect the performance and cost of key engine technologies and in extreme circumstances can prevent certain technologies from being used. Currently, emissions standards in the three markets with fuel economy standards are strictest in the United States and most lenient in Japan. Unless excellent progress is made in lean burn NO_x control technology, strict U.S. NO_x standards might serve to ban or significantly limit diesel use in the LDV fleet and reduce the fuel economy performance of direct injection gasoline engines. In contrast, European and Japanese standards allow for such “lean technologies” and, moreover, do not hold diesel-fueled vehicles to NO_x emissions levels as stringent as those for gasoline-fueled vehicles.

- *Vehicle taxes.* High vehicle taxes can serve to make new fuel economy technologies more expensive because they apply to the price of the total vehicle. For example, EU countries have “value-added” taxes

that apply to the total price of the vehicle. On the other hand, countries may impose taxes on vehicle sales and/or ownership that vary with fuel efficiency or with factors that are likely to affect fuel efficiency, especially engine power or displacement, and these taxes would provide a positive incentive for high-efficiency vehicles. Several EU countries and Japan have such taxes. Also, some countries may give tax credits or deductions for certain kinds of high-technology vehicles, increasing the cost-effectiveness of these vehicles and making it somewhat easier for fleets to achieve high fuel economy. For example, the United States gives tax deductions to purchasers of hybrid electric vehicles.

- *Vehicle miles driven.* More driving means higher annual fuel savings from new technologies, thereby improving the cost-effectiveness of these technologies. For example, LDVs are driven farther in the United States ($\sim 13,000$ miles/year for the first 5 years) than in Europe (8000–9000 miles/year), and much farther than in Japan (~ 6000 miles). This tends to raise the cost-effectiveness of new technologies in the United States versus that in the other markets.

- *Average vehicle efficiency.* In considering how high a percentage increase in current fuel economy might be sought, a high current fuel economy will mean that a particular percentage increase will save less fuel than would be saved if current fuel economy was low. In other words, a 20% increase would save more fuel in the United States than in Europe or Japan because the current U.S. fuel economy level is lower than in the other markets.

2.2 U.S. Corporate Average Fuel Economy Standards

In the United States, all major automakers are subject to Corporate Average Fuel Economy (CAFE) standards requiring their new vehicle fleets to meet specific fuel economy targets. The standards were established by the Energy Policy and Conservation Act of 1975 in response to the 1973 oil embargo and subsequent oil price shock. Under the act, each manufacturer’s fleet of new vehicles is subdivided into passenger cars and light trucks (including pickup trucks, vans, and sport utility vehicles) and the passenger car fleet is further subdivided into domestic and import vehicles, where vehicles with less than 75% U.S./Canadian content are considered to be imports. Each individual fleet must then meet a separate target. In 2002, each manufacturer’s domestic and import passenger car fleet had to average at

least 27.5 mpg and each manufacturer's light truck fleet had to average at least 20.7 mpg. Manufacturers that do not comply with these standards are liable for a civil penalty of \$5.50 for each tenth of 1 mpg its fleet falls below the standard, multiplied by the number of vehicles in the noncomplying fleet. The U.S. government has enforced the standards; for example, in calendar year 2001, manufacturers paid civil penalties totaling more than \$33 million for failing to comply with the standards in model years 1999 and 2000.

Additional important characteristics of the standards are as follows:

- *Low-volume manufacturers.* The standards can be waived for low-volume manufacturers, that is, those producing fewer than 10,000 passenger cars annually worldwide.
- *Vehicle weight limits.* The standards apply only to vehicles whose gross vehicle weight rating (GVWR) is 8500 pounds or less. A number of large sport utility vehicles have GVWRs above this level and are not subject to the standards.
- *Setting and amending standards.* The U.S. Congress set the original schedule for passenger car standards and authorized the National Highway Traffic Safety Administration (NHTSA), in the Department of Transportation, to set the light truck standards. Manufacturers were required by Congress to comply with the 27.5-mpg standard for passenger cars by 1985, and the standard has remained unchanged since (except for a brief reduction from 1986 to 1989). NHTSA is authorized to amend the passenger car and light truck standards.
- *Testing responsibility.* Manufacturers are responsible for measuring their own fleet fuel economy. They test preproduction prototypes of their new vehicle models and submit the results to the U.S. Environmental Protection Agency (EPA). EPA retests approximately 10% of the tested vehicles to confirm the manufacturer's tests.
- *Test procedure.* The procedure for testing vehicle fuel economy consists of having a professional driver drive the vehicle on a dynamometer through two driving cycles designed to simulate city and highway driving. Each vehicle is driven in the same manner on the same type of equipment, so that results can be compared with confidence. The city test begins with a cold start (as if the vehicle had been parked overnight) and simulates an 11-mile, stop-and-go trip at mostly low speeds (although the maximum speed is 56 mph), averaging 20 mph. The trip takes 31 min with 23 stops and a total of about

5.5 min of idle time. The highway test begins with the engine warmed up and simulates a 10-mile combined highway/rural road trip at an average speed of 48 mph and a maximum of 60 mph. The resulting two fuel economy values are averaged by assuming that 55% of driving is city driving and 45% is highway driving, with the average calculated by the following formula:

$$\begin{aligned} 1/\text{vehicle mpg} = & 0.55/(\text{city mpg}) \\ & + 0.45/(\text{highway mpg}). \end{aligned}$$

Calculating CAFE. Each manufacturer's CAFE is the sales-weighted average of the tested fuel economy levels of its vehicles. This average is calculated in a manner similar to the preceding formula, that is, the inverse of the fuel economy level (the fuel consumption level) of every vehicle in the manufacturer's fleet is added, with the sum divided by the total number of vehicles. To clarify, if a manufacturer sold two types of vehicles with annual sales of 100,000 and 300,000, its fleet average fuel economy (CAFE) would be calculated using the following formula:

$$\begin{aligned} 1/\text{CAFE} = & [100,000/(\text{mpg of vehicle 1}) \\ & + 300,000/(\text{mpg of vehicle 2})] \\ & \div 400,000. \end{aligned}$$

Credits. Manufacturers can earn credits if they exceed the standards, and these credits can be applied to any of the subsequent or preceding 3 years to offset shortfalls. Also, a manufacturer failing to meet the standard for a given model year can avoid paying the civil penalties by submitting a compliance plan to NHTSA demonstrating that it will earn enough credits within the next 3 model years to offset the penalties. Manufacturers also earn credits by producing and selling alternative fuel vehicles, with greater credits awarded for dedicated vehicles than for dual-fuel vehicles (which can burn either gasoline or an alternative fuel). Although each alternative-fuel vehicle obtains a high fuel economy value based on the theoretical amount of oil-based fuel it consumes (theoretical because dual-fuel vehicles may burn gasoline exclusively), total credits awarded to a manufacturer cannot increase its official CAFE value by more than 1.6 mpg.

• *Original schedule.* Although the passenger car CAFE targets have remained constant since 1990 and the light truck targets have remained constant since 1996, the original legislation was crafted to allow companies to raise their fleet fuel economy levels gradually, taking into account the impossibility of redesigning entire corporate fleets simultaneously as

well as the time needed to develop improved technology. Thus, for passenger cars, the 1978 standard was 18.0 mpg, with yearly increases thereafter (to 19.0 mpg in 1979, 20.0 in 1980, 22.0 in 1981, 24.0 in 1982, 26.0 in 1983, and 27.0 in 1984), leading to the 27.5-mpg target in 1985.

Although there have been arguments about whether increases in U.S. LDV fuel economy after 1975 were caused mostly by the CAFE regulations or were instead caused largely by market forces, the U.S. fleet has generally complied with the regulations and has maintained fuel economy values above the required levels despite low gasoline prices and shifts in consumer tastes toward larger, more powerful vehicles. However, since attaining the standards, U.S. fleet fuel economy has essentially stopped improving, and the combined car and truck fuel economy has actually declined due to increasing market shares of (less efficient) light trucks. Over the past two decades, the U.S. fleet has used a continuing influx of new technology to allow fleet fuel economy levels to be maintained while other factors—weight, acceleration time, luxury features, and the like—moved in directions that would otherwise have reduced fuel economy. EPA has estimated that the tested fuel economy (not adjusted downward for onroad conditions) of the model year 2001 fleet of cars and trucks, which was 23.9 mpg, would have instead been 30.3 mpg had the weight mix and 0-to-60 mph acceleration times of the fleet been the same as that of the 1981 model year. In other words, the fleet traded away between 6 and 7 mpg for higher weight and better acceleration during that 20-year period.

A number of controversies have arisen over the CAFE standards. Probably the most important is the claim that size and weight reductions demanded by the standards have had an adverse impact on overall vehicle safety. This claim has tended to dominate discussions about raising future CAFE standards. The National Research Council's (NRC) Committee on the Effectiveness and Impact of Corporate Average Fuel Economy Standards concluded in 2002 that the change in size and weight of the fleet between 1976 and 1993 caused about 2000 additional traffic fatalities in 1993. There was a strong dissent to the committee's conclusions from within the committee itself as well as from other analysts. Furthermore, an earlier national academy panel concluded that the available analyses showing vehicle weight/traffic fatality correlations should not be used to project fatalities caused by fleetwide weight changes because the analyses could not

adequately control for various confounding factors such as driver age. The issue is complicated by the fact that fuel economy can be improved without weight reduction and that the distribution of weight within the fleet is likely to be more important to safety than the average weight. A reanalysis of traffic fatality data sponsored by Honda, using more recent data but the same analytical method cited by the NRC committee, found essentially no impact on traffic fatalities from a 100-pound reduction in the average weight of the fleet.

Manufacturer responses to some provisions of the standards have caused some controversy. For example, the division into domestic and import fleets has created incentives for manufacturers to alter the domestic "content" of individual models to move them into one fleet or the other depending on circumstances. The original purpose of the division was to create an incentive for manufacturers to produce small cars within the United States so as to boost domestic employment (manufacturers needed the smaller cars to provide a counterweight to larger, less efficient models that previously had been their major market focus for domestic production). The division succeeded in stimulating domestic small car production, but it also created an incentive to add import content to larger domestic models so as to get them out of the domestic fleet and into a fleet of more efficient import models that could more easily offset their low fuel economy. For example, Ford temporarily made its full-size (and relatively low-fuel economy) Crown Victoria model into an "import" to get it out of the domestic fleet, thereby aiding that fleet in complying with its CAFE standard.

In 2002, the State of California passed legislation (Assembly Bill 1493) that requires the California Air Resources Board to establish greenhouse gas standards for vehicles in model year 2009 and beyond, applying to corporate fleet averages. Although automakers will be able to partially comply with the standards by reducing greenhouse gas emissions from nonvehicle sources and by shifts in fuels, the standards will essentially have an effect similar to that of fuel economy standards. The bill has been challenged in the courts by the vehicle manufacturing industry on the basis that the standards will be fuel economy standards and that only the federal government can set such standards.

2.3 Japanese Weight Class Standards

The Japanese government has established a set of fuel economy standards for gasoline- and diesel-

powered light-duty passenger and commercial vehicles, with fuel economy targets based on vehicle weight classes. Table I shows the vehicle targets for gasoline- and diesel-powered passenger vehicles, as measured on the Japanese 10.15 driving cycle. The targets for gasoline-powered vehicles are to be met by 2010, and 2005 is the target year for diesel-powered vehicles. Figure 1 illustrates the percentage improvements required for each weight class of gasoline-powered vehicles as well as the market share for each class at the time the regulations wave created. The original regulations demanded that the targets be met by each automaker

for each weight class; that is, automakers could not average across weight classes by balancing a less-than-target vehicle in one class with a better-than-target vehicle in another class. Revisions to the regulations in January 2001 allow automakers to accumulate credits in one weight class and use them in another class, with important limitations.

Assuming no change in vehicle mix, these targets imply a 22.8% improvement in gasoline passenger vehicle fuel economy (15.1 km/L in 2010 vs 12.3 km/L in 1995) and a 14.0% improvement in diesel passenger vehicle fuel economy (11.6 vs 10 km/L) compared with the 1995 fleet. In other words, compliance with

TABLE I

Japanese Weight Class Fuel Economy Standards for Passenger Vehicles

	Vehicle weight class (kg)								
	750	875	1000	1250	1500	1760	2000	2250	2500
A. Gasoline-Fueled Passenger Vehicles for 2010									
Vehicle curb weight (kg)	–702	703–827	828–1015	1016–1265	1266–1515	1516–1765	1766–2015	2016–2265	2266–
Vehicle curb weight (pounds)	1548	1550–1824	1826–2238	2240–2789	2791–3341	3343–3892	3894–4443	4445–4994	4997–
Fuel economy target (km/L)	21.2	18.8	17.9	16.0	13.0	10.5	8.9	7.8	6.4
Fuel economy target (mpg)	49.9	44.2	42.1	37.6	30.6	24.7	20.9	18.3	15.0
B. Diesel-Powered Passenger Vehicles for 2005									
Vehicle weight (kg)	–1015	1016–1265	1266–1515	1516–1765	1766–2015	2016–2265	2266–		
Vehicle weight (pounds)	–2238	2240–2789	2792–3341	3343–3892	3894–4443	4445–4994	4997		
Target standard (km/L)	18.9	16.2	13.2	11.9	10.8	9.8	8.7		
Target standard (mpg)	44.5	38.1	31.0	28.0	25.4	23.0	20.5		

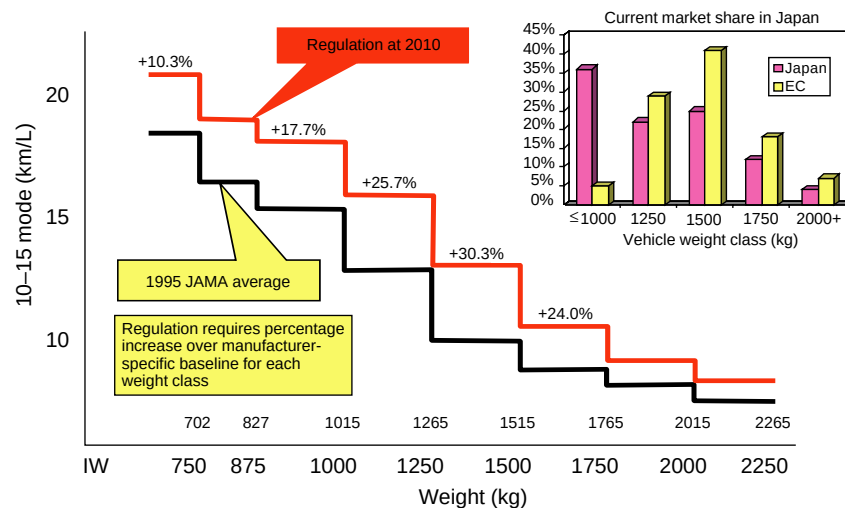


FIGURE 1 New Japanese fuel economy regulations. (Source: Japan Ministry of Transport.)

these standards will produce a Japanese gasoline-fueled, light-duty passenger vehicle fleet of 35.5 mpg and a diesel fleet of 27.3 mpg by 2010 and 2005, respectively, as measured using the Japanese 10.15 driving cycle (and assuming no changes in the market shares of the various weight classes). The Japanese 10.15 driving cycle is substantially slower than the combined U.S. city/highway cycle, and the U.S. equivalent miles per gallon for this fleet would be significantly higher.

The regulations call for civil penalties if the targets are not met, but these penalties are very small. The formal enforcement procedure includes a government assessment of whether the automakers have taken the necessary steps to comply with the targets, recommendations by the regulating agencies about steps they might take, and another assessment by the agencies of the adequacy of the automakers' compliance with the recommendations, with the procedures taking more than 1 year. Realistically, enforcement will come from pressure from the government and the auto companies' desire to avoid public embarrassment, not from the financial penalties.

The fuel economy targets were selected by identifying representative "best in class" fuel economies in each weight class and demanding that the average new vehicle meet that level in the target year. The Japanese call this the "top-runner" method of selecting fuel economy targets. The actual top-runner fuel economy target might not be the highest in the weight class. As described in the next subsection, some vehicles will not be candidates for top-runner status so as to avoid saddling the automakers with impractical targets based on exotic technology or nonrepresentative characteristics.

As noted previously, manufacturers may accumulate credits in a weight class by achieving levels of fuel economy higher than the target levels. These credits, measured as (difference in fuel consumption between the vehicle and the target fuel consumption rate) * (number of vehicles sold), can be applied to vehicles in another weight class at a 50% rate; that is, two credits are needed to offset one unit of underperformance. However, the credits can be

applied only if the automakers' performance in the failed weight class meets the threshold values of fuel efficiency shown in Table II (for gasoline vehicles).

The fuel economy regulations have additional requirements over and above the actual fuel economy targets. These are as follows:

1. For new vehicles, fuel economy and major efficiency technologies onboard must be recorded in catalogs and displayed at exhibits.
2. Government is charged with providing education and other incentives for vehicle users and manufacturers, making sure that fuel economy regulation proceeds in harmony with other regulations (especially new emissions and fuel quality standards), reviewing manufacturers' efforts to improve fuel economy, and trying to harmonize this regulation with similar efforts in Europe and the United States.
3. Manufacturers are expected to develop new efficiency technologies, design vehicles of outstanding efficiency, and help to educate the public.
4. The public is charged with selecting efficient vehicles and using them in an efficient manner.

Although the Japanese standards do not take effect until 2010, some concerns have already arisen. First, changes in the weight composition of the Japanese fleet may change the degree of improvement in fuel economy demanded by the standards. Figure 2 shows how the weight distribution of the fleet has changed between 1995 and 2000. Primary changes include a large reduction in the lightest (750-kg) market segment and substantial increases in the heavier segments. If the year 2000 weight distribution is applied to the fuel economy standards, the expected improvement in fleet fuel economy over the 1995 baseline would be 18.0% versus the 22.8% percent projected by the Japanese government based on the 1995 weight distribution. Conceivably, future changes in the fleet's weight distribution might drive down the expected improvement even further.

TABLE II

Fuel Economy Thresholds for Use of Credits (Japanese Weight Class Standards)

Weight class (kg)	750	875	1000	1250	1500	1760	2000	2250	2500
Fuel economy threshold (km/L)	20.2	18.5	17.1	14.1	12.6	9.8	8	6.8	6.1
Target fuel economy (km/L)	21.2	18.8	17.9	16	13	10.5	8.9	7.8	6.4

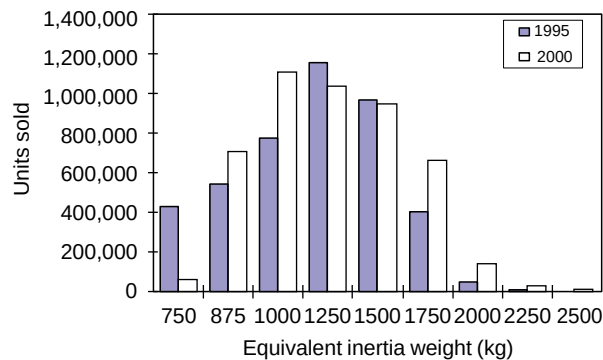


FIGURE 2 Weight distribution of the Japanese new LDV fleet, 1995–2000. (Source: Kiyoyuki Minato, Japan Automobile Research Institute.)

Another concern is the limitation on averaging fuel economy across weight classes. This allowed “50% averaging” still leaves manufacturers with far less flexibility than if full averaging were allowed. This lack of flexibility particularly affects European and U.S. manufacturers that export into the Japanese market and that are likely to have limited numbers of models, and possibly only one model, in a weight class.

2.3.1 The Japanese Top-Runner Selection Process

The new Japanese weight class standards for LDV fuel economy are based on identifying top runners—vehicles with exemplary fuel economy—in each weight class and requiring future vehicles in that class to attain similarly high fuel economy. According to Japanese industry sources, the top-runner selection process is more nuanced than simply selecting the highest fuel economy vehicle in each class and establishing that as the standard. Instead, the regulators—the Ministry of International Trade and Industry (MITI) and the Ministry of Transportation (MOT)—with the consultation of the Japanese Auto Manufacturers Association (JAMA), select top-runner targets based on their judgment of the relevance of class-leading vehicles and their technology to the class as a whole. For example, because the large majority of personal vehicles use automatic transmissions, only vehicles with automatic transmissions are considered as top-runner candidates for personal vehicles. For commercial vehicles, separate goals are established for vehicles with automatic and manual transmissions. Top-runner candidates that have very small sales volumes and/or very expensive technology are not likely to be considered because the regulators are interested in

technologies that have been shown to be practical for the class as a whole. For example, the hybrid electric vehicles currently in the Japanese marketplace were not used as top runners.

2.4 ACEA Voluntary Commitment to the EC on Vehicle Carbon Emissions

The European Automobile Manufacturers Association (ACEA) has offered, and the European Commission (EC) has accepted, a voluntary commitment to reduce the CO₂ emissions from new light-duty passenger vehicles, with firm fleetwide targets of 140 g CO₂/km (~41 mpg for gasoline) by 2008, measured under the new European test cycle (Directives 93/116/EC and 99/100/EC). This represents approximately a 25% reduction from the 1995 average of 187 g/km (~30 mpg) on this cycle. The European cycle is likely to produce lower fuel economy ratings than the U.S. combined urban/highway cycle, so the “U.S. equivalent” miles per gallon ratings of the year 2008 European fleet will likely be higher than 41 mpg if the targets are met.

Note that the 140-g CO₂/km goal is a collective target, not a target for each individual company. The ACEA participants in the voluntary commitment—BMW, Fiat, Ford of Europe, GM Europe, Daimler-Chrysler, Porsche, PSA Peugeot Citroen, Renault, Rolls-Royce, Volkswagen, and Volvo—have not publicly committed to individual objectives.

The voluntary commitment applies to light passenger vehicles classified as M1 in European Council Directive 93/116/EC, which includes vehicles with no more than eight seats in addition to the driver. The commitment includes a promise to introduce some models emitting 120 g/km (~48 mpg) or less by 2000 and a nonbinding 2003 target range of 165 to 170 g/km (~34–35 mpg). In addition, the commitment was to be reviewed in 2003 with the aim of moving toward a fleet goal of 120 g/km by 2012. Finally, ACEA agrees to monitor compliance with the commitment jointly with the EC.

In exchange for their commitment to meet the 2008 CO₂ emissions goal, industry members insisted that the voluntary commitment be contingent on a number of conditions, including the following:

- *Clean fuels availability.* Because the industry believes that direct injection engines will play a key role in achieving the targets, the commitment requires “full-market availability” of clean fuels needed for this technology by 2005—gasoline with 30 ppm sulfur content and less than 30%

aromatics, diesel fuel at 30 ppm sulfur and cetane number greater than or equal to 58. The low sulfur level is needed to avoid poisoning the catalysts needed for NO_x control.

- *Protection against unfair competition.* Non-ACEA members must commit to similar goals, and the European Community will agree to try to persuade other car manufacturing countries to embrace equivalent efforts. The latter is designed to protect ACEA members from suffering in world market competition for their European efforts.
- *Regulatory ceasefire.* There are no new regulatory measures to limit fuel consumption or CO₂ emissions.
- *Unhampered diffusion of technologies.* The companies assume that the EC will take no action that would hamper the diffusion of efficiency technologies, particularly direct injection gasoline and diesel engines. Presumably, this could include tighter emissions standards on NO_x and particulates. ACEA assumes that up to 90% of the new vehicle fleet will use direct injection engines by the 2008 compliance date, and tighter standards for these pollutants might interfere with the diffusion of this technology.

The voluntary commitment also incorporates a statement that might be seen as an escape hatch for the companies:

On the basis of the outcome of the monitoring, or if the impacts of this commitment on the European automotive industry, particularly its employment situation and its global competitive environment, are detrimental, ACEA and the [European] Commission will review the situation and make any necessary adjustments in good faith.

Both JAMA and the Korean Automobile Manufacturers Association (KAMA) have agreed to a revised version of the ACEA targets—achievement of the 2008 target levels in 2009.

A concern with the voluntary commitment is that it does not specify individual corporate targets, nor does it define mechanisms to guarantee attainment of the fleet target. The flexibility afforded by the lack of specified individual targets theoretically could lead to an efficient market solution if companies could agree among themselves to attain levels of fuel economy improvement that represent approximately equal marginal net value (value of fuel savings minus technology cost). However, it is not clear that there is any agreement among the companies defining their individual targets, and there is no guarantee that all companies will make a fair effort toward meeting the fleet target. The lack of specified corporate targets

might tempt companies to provide less than their full effort and might make it difficult for the EU to assess responsibility if the industry fails to meet the target.

3. FUEL TAXES

3.1 Overview

Economists have overwhelmingly favored fuel taxes over fuel economy standards as a means to reduce fuel consumption because taxes give maximum flexibility to both vehicle manufacturers and purchasers and because they influence fuel consumption through both fuel demand and vehicle supply by making travel more expensive (thereby reducing vehicle miles traveled) and by creating an economic incentive for manufacturers to build efficient vehicles and for consumers to purchase them. However, it is important to note that the presumed superior effectiveness of fuel taxes over CAFE standards depends on the assumption that markets for fuel economy and travel are very efficient. This assumption about an efficient market, especially as it pertains to the market for fuel economy, is debatable, and the interested reader is urged to examine the considerable literature on this subject.

Because analyses of future light-duty transportation energy use require estimates of the impact of fuel prices on travel and fleet fuel economy, estimates of the fuel price elasticity of travel, fuel economy, and fuel consumption are ubiquitous in the literature. However, there is substantial disagreement about the magnitude of these elasticities because travel volumes, fuel economy, and fuel consumption are dependent on several variables other than fuel price and because fuel prices have tended to be volatile during the past few decades, thereby complicating attempts to estimate long-run elasticities. Thus, the magnitude of the effect of changes in fuel taxes on fleet fuel economy and on travel volumes and fuel consumption is also subject to considerable disagreement. For the United States, recent studies by D. L. Greene and others imply that a 10% gasoline price increase would cause at most approximately a 2% increase in average fuel economy, a 2 or 3% decrease in travel, and a 4% decline in gasoline consumption (compared with a baseline, and increasing, level) over the long term. However, others conclude that the decline in gasoline consumption would be higher—as much as 6 or 8%. The effect elsewhere will be different. For example, in Europe and Japan, fuel prices make up a larger share of the total cost of

travel, implying that fuel price elasticities there will likely be higher than those in the United States.

3.2 United States

In the United States, motor vehicle fuel taxes have generally been viewed as “user taxes” at the federal and state levels. In other words, the federal government and most state governments set their per gallon tax rate at a level that will pay for a substantial share of road-building requirements. In addition, states and localities add sales taxes as a percentage of fuel prices. Federal taxes in 2000 were \$0.184/gallon for gasoline and \$0.244/gallon for diesel fuel. Gasohol, a blend of 90% gasoline and 10% ethanol by volume, was taxed at only \$0.131/gallon. State taxes averaged about \$0.20/gallon plus sales taxes, the latter of which vary between 1 and 7%. In addition, 12 states allow localities to add their own taxes.

The result is that taxes on gasoline and diesel fuel in the United States have been quite low in comparison with those in Europe and Japan. In 2000, the average U.S. gasoline price (less local taxes) was \$1.48/gallon, of which 28% was federal and state taxes. The average total gasoline price in 2001 was \$1.63/gallon, and the average diesel price was \$1.52/gallon.

3.3 Japan and Europe

Europe and Japan have traditionally taxed gasoline at high rates, so that gasoline prices have been much higher than those in the United States. These higher prices, along with differences in land use patterns and other factors, play an important role in the strong differences in average per vehicle miles traveled among Europe, Japan, and the United States; over their first 5 years of life, LDVs are driven approximately 13,000 miles in the United States, 8000 miles in Europe, and 6000 miles in Japan.

In 2001, Japanese gasoline prices averaged \$3.80/gallon as compared with \$3.68/gallon in France, \$3.74/gallon in Germany, and \$4.63 in the United Kingdom. With the exception of the United Kingdom, diesel prices were substantially lower—\$2.70/gallon in Japan, \$2.80/gallon in France, and \$2.91/gallon in Germany. U.K. diesel prices were virtually the same as gasoline prices—\$4.42/gallon.

In Europe, the price differences among the various countries and the gasoline/diesel price difference are due nearly entirely to differences in taxation. Table III displays the gasoline and diesel taxes in the several European countries. As shown by the table, every

TABLE III

European Taxes on Gasoline and Diesel Fuel as of 2001 [(Euros/Liter)/(Euros/Gallon)]

Country	Gasoline	Diesel	Value-added tax (percentage)
Austria	0.408/1.55	0.283/1.07	20.0
Belgium	0.507/1.92	0.290/1.10	21.0
Denmark	0.524/1.98	0.367/1.39	25.0
Finland	0.552/2.09	0.300/1.14	22.0
France	0.571/2.16	0.374/1.42	19.6
Germany	0.593/2.25	0.409/1.55	16.0
Greece	0.297/1.13	0.245/0.93	18.0
Ireland	0.348/1.32	0.249/0.94	20.0
Italy	0.520/1.97	0.382/1.45	20.0
Luxembourg	0.372/1.41	0.253/0.96	15.0
The Netherlands	0.402/1.52	0.339/1.28	19.0
Portugal	0.289/1.09	0.246/0.93	17.0
Spain	0.372/1.41	0.270/1.02	16.0
Sweden	0.528/2.00	0.356/1.35	25.0
United Kingdom	0.765/2.90	0.765/2.90	17.5

European country but the United Kingdom taxes diesel fuel substantially less than it does gasoline on a volumetric basis despite diesel fuel's higher volumetric energy content. This differential has stimulated sales of diesel-powered vehicles, which have soared in Western Europe—to nearly 40% of new vehicle sales during the first 6 months of 2002.

Figure 3 shows the average gasoline prices in Canada, France, Germany, Japan, the United Kingdom, and the United States for the years 1990 and 2001, illustrating the importance of taxes in the overall prices.

4. VEHICLE TAX INCENTIVES PROMOTING HIGHER FUEL ECONOMY

4.1 Overview

A number of countries have tax programs designed to reward purchasers and owners of fuel-efficient vehicles (or to penalize purchasers and owners of fuel-inefficient vehicles) by tying vehicle purchase and/or ownership taxes either directly to fuel economy or to vehicle features associated with fuel economy (e.g., engine displacement). Aside from taxes aimed specifically at fuel economy or associated vehicle characteristics, all countries in the major auto markets have vehicle taxes that are based on the total purchase

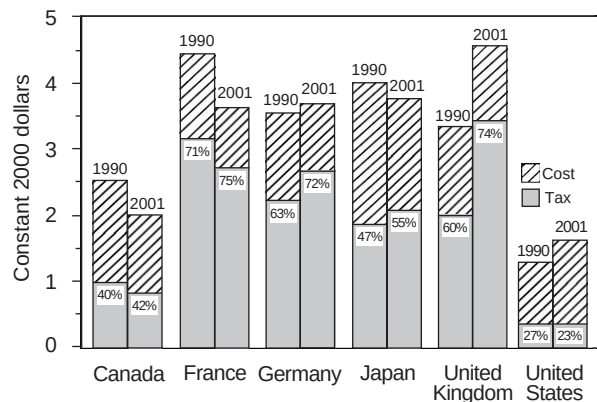


FIGURE 3 Gasoline prices for selected countries, 1990 and 2001, showing effect of taxes. From Oak Ridge National Laboratory. (2002). "Transportation Energy Data Book," 22nd ed.

price of the vehicle (e.g., sales taxes in the United States, value-added taxes in Europe). This latter form of taxes can actually serve as an impediment to higher fuel economy to the extent that increased efficiency is attained with high-cost technology, such as continuously variable automatic transmissions, hybrid electric drivetrains, and variable valve controls, and such taxes add to the cost of these technologies. Of course, such taxes, by making vehicles more expensive, reduce sales and so limit fuel consumption in this manner.

4.2 United States

The United States has two distinct tax-based programs aimed at improving vehicle efficiency. The first is the well-known "gas guzzler tax," established by the Energy Tax Act of 1978, which adds a substantial purchase tax to the price of passenger cars (light-duty trucks are not subject to the gas guzzler tax) whose fuel economy falls below an EPA test value of 22.5 mpg. The current tax rate begins at \$1000 U.S. for cars with fuel economy between 21.5 and (less than) 22.5 mpg, and it escalates to \$7700 U.S. for cars with fuel economy of less than 12.5 mpg. For the 2001 model year, there were 35 models subject to the tax, and all of these were expensive luxury or performance cars such as the Dodge Viper, Mercedes SL500, Rolls-Royce Corniche, and BMW 750IL. Most analysts believe that mass market manufacturers have tried hard to keep their larger models above the tax's minimum fuel economy, not only to avoid the payment but also to avoid having these vehicles subject to the stigma attached to being identified as "guzzlers."

The second U.S. program is a substantial tax deduction from federal income taxes of up to \$2000 U.S. (i.e., \$2000 U.S. may be deducted from reported net income before calculating taxes) for hybrid vehicles. The tax credit was originally established as a "Clean Fuel" program designed to promote the use of alternative nonpetroleum fuels, but the U.S. government's Internal Revenue Service determined that the program applies to hybrids. This program is scheduled to phase out between 2004 and 2006, although Congress might extend the deadline.

4.3 Japan

Japan has instituted a number of tax incentives for higher fuel economy or for characteristics that are associated with higher fuel economy. They are as follows (with U.S. dollar values calculated at 120 yen/dollar):

- Regarding taxes directly promoting hybrid vehicles, the vehicle acquisition tax is reduced to 2.2% for hybrid vehicles (the national acquisition tax for conventional vehicles is 5%). For example, the purchaser of a Toyota Prius would save about \$400. Second, there is better tax treatment of corporate income tax for company purchases (this applies to all vehicles meeting 2010 fuel economy targets and generating 25% less emissions than current standards). Third, there are purchase incentives for private individuals and special government loans for company purchases. The private incentive is half the price difference between the hybrid and equivalent conventional vehicles. For a Prius, the incentive is worth about \$1800.
- The annual tax of vehicles meeting both low emissions standards and fuel economy standards is reduced. A Toyota Prius owner would save approximately \$120/year for the first 2 years of ownership.
- The vehicle tax is based on weight and assessed at inspection (6300 yen or ~\$53/half ton).

4.4 Europe

As in Japan, many members of the EU have passenger car taxes that either directly reward high fuel efficiency or reward vehicle characteristics that are associated with high fuel efficiency. Table IV shows the variety of taxes on both vehicle purchase and ownership in the 15 European countries that have such taxes. Of particular interest are yearly taxes in Denmark that vary inversely with fuel

economy—at approximately \$30/vehicle/year for gasoline vehicles with fuel economy above 20 km/L (47 mpg), approximately \$330/vehicle/year for vehicles at about 23 mpg, and more than \$1100/vehicle/year for vehicles below 10.6 mpg.

The United Kingdom has introduced a number of economic incentives for higher fuel economy, including a vehicle purchase tax based on CO₂ emissions ranging from 60 to 160 pounds (~\$95 to \$250 U.S. at 2002 exchange rates), tax credits for companies purchasing vehicles with 120 g/km or less CO₂ emissions, and 1000-pound (~\$1600 U.S.) grants to consumers purchasing hybrid vehicles.

Another example of a tax program based directly on fuel economy is the tax rebate program adopted by the Netherlands in January 2002 that awarded a purchase tax rebate of 1000 euros for cars with an “A” fuel economy rating (at least 20% more efficient than other cars of similar size) and 500 euros for cars with a “B” rating (10–20% higher fuel economy than cars of similar size). However, this program ended after its first year in existence due to a change in governments.

TABLE IV
Taxes on Vehicle Purchase and Ownership Based on Fuel Economy-Related Variables: European Union and United Kingdom

Country	Sales tax	Ownership tax
Belgium	Based on cc	Based on cc
Denmark		Fuel economy, weight
Germany		cc
Spain	< 1.6 L: 7% > 1.6 L: 12%	Horsepower
Finland		
France		
Greece		cc
Ireland	< 1.4 L: 22.5% 1.4–2 L: 25% > 2 L: 30%	cc
Italy		Horsepower
Luxembourg		cc
The Netherlands		Weight
Austria	Fuel economy Maximum 16%	Horsepower
Portugal	cc (e.g., 1.6 L): 5032 euros	cc
Sweden		Weight
United Kingdom		CO ₂ emissions

Note: cc, cubic centimeters of engine displacement.

4.5 Measures in Other Countries

We know of no studies that have examined vehicle taxation programs elsewhere in the world, but it seems likely that there are a number of tax regimes that affect vehicle fuel economy, albeit indirectly. For example, Brazil reduces its 25% tax on industrialized products to 10% for automobiles with engines less than 1 L in size, greatly encouraging sales of low-powered vehicles that are likely to be fuel efficient.

5. INFORMATION PROGRAMS TO PROMOTE HIGHER FUEL ECONOMY

5.1 Overview

A variety of information programs exist in the major automobile markets to assist potential purchasers of new cars in comparing fuel economy characteristics (and other attributes) among competing vehicles. However, there is no evidence that the existence of such systems plays a significant role in consumer choice of vehicles. For example, in 1995, the light-duty fleets in the two European countries with fuel economy labeling systems (Sweden and the United Kingdom) were the least efficient in Europe (according to International Energy Agency data).

5.2 United States

The fuel economy information program in the United States is the oldest and most highly developed of the major vehicle markets. The current system includes a label on the window of all new cars and light trucks available for sale that lists the approximate fuel economy the average driver can expect to obtain in city and highway driving (obtained by adjusting the fuel economy test values obtained on the official EPA city and highway test cycles downward by 10% for the city cycle and 22% for the highway cycle) and other relevant information. More specifically, the label includes the following:

- a description of the vehicle sufficient for the customer to distinguish it from other similar models, including the identical model with a different engine or transmission;
- the (adjusted) city and highway fuel economy values;
- the statement of the likely range of fuel economy values obtained by the majority of drivers,

estimated at 15% below to 15% above the city and highway fuel economy values;

- a statement of the range of fuel economy levels obtained by comparable vehicles;
- an estimate of the annual fuel cost.

The second component of the system is a fuel economy guide, available at automobile dealers and on a Web site (www.fueleconomy.gov), that offers an extensive menu of information about fuel economy and other vehicle attributes. Main features include the following:

- for every vehicle marketed, identified by manufacturer, model, engine option, and transmission type: city and highway fuel economy, annual fuel cost (assuming 15,000 miles/year [45% highway driving and 55% city driving]), annual greenhouse gas emissions (with the same assumptions), and an EPA air pollution score;
- on the Web site, a search engine that allows selection by market or size class, by manufacturer, or by fuel economy;
- on the Web site, capability to display competing vehicles side by side;
- identification of “best” and “worst” vehicles, on the basis of fuel economy, for each size class;
- compilation of advice on driving more efficiently;
- simple information about advanced technology vehicles (e.g., hybrid electric vehicles).

EPA also offers a “Green Vehicle Guide” at its Web site (www.epa.gov/autoemissions) that lists, for each new vehicle, its fuel economy value and a ranking from 0 to 10 based on its emissions of air pollutants.

In addition to the consumer information available in the Fuel Economy Guide and on the Web site, EPA publishes a yearly compilation of fuel economy trends that provides guidance to policymakers and analysts about national progress in fuel efficiency. The 2001 version of the report, *Light-Duty Automotive Technology and Fuel Economy Trends 1975 through 2001*, is available on the Web (www.epa.gov/oms/cert/mpg/fetrends/rdrtrnd01.htm). The report highlights key fuel economy and technology trends in the U.S. LDV fleet, including trends in vehicle attributes, such as horsepower and weight, that affect fuel economy. Fuel economy information is broken down in a variety of ways (e.g., total fleet, all cars, all light trucks) by size and weight class, by manufacturer, by marketing group (e.g., Ford Group—Ford, Mazda, Volvo, Jaguar, Land Rover),

and the like. Also, the report shows the difference in fleet fuel economy using the standard EPA assumption of 55% city driving and 45% highway driving and shows actual measured values of city and highway driving (the “actual” fuel economy values are about 2% lower than the estimated values using the 55/45 split). Trends in technology (e.g., four-wheel drive, four valves per cylinder) and vehicle characteristics (e.g., horsepower, weight) are shown in tables and figures. One interesting feature is a series of analyses that show how the fleet fuel economy would change if all of the vehicles in each size class or weight class attained the fuel economy of the best vehicles in that class. Finally, the report includes an analysis of what fleet fuel economy would have been had there been no change in average vehicle weight and acceleration capabilities over the past 10 or 20 years. This analysis shows the impact on fuel economy of changes in vehicle weight and acceleration capability. The analysis is then repeated with size class substituted for weight.

5.3 Japan

The government of Japan does not require fuel economy labels on new cars, although it has started to consider introducing such labeling. However, the Ministry of Land, Infrastructure, and Transport does publish a fuel economy guide (and maintains the guide on its Japanese-language Web site) that includes information for each vehicle on fuel economy and CO₂ emissions and highlights vehicles that already meet 2010 fuel economy standards. Also, manufacturers must list fuel economy values in sales brochures and vehicle specification sheets at car dealerships.

5.4 Europe

Two European countries, Sweden and the United Kingdom, have had vehicle fuel economy labeling systems in place for the past two decades or so. The United Kingdom’s system has been mandatory, with fines set for noncompliance, but the Swedish system generally has not been enforced.

The EC has established (in Directive 1999/94/EC) consumer information requirements for new passenger cars that demand the following:

- fuel economy labels on all new cars that include both fuel economy and CO₂ emissions rates;
- ready availability of a fuel economy guide containing fuel economy ratings of all new vehicles sold in that market;

- posters in new vehicle showrooms listing the fuel economy ratings of all vehicles for sale at those locations;
- incorporation of fuel economy/fuel consumption values in advertising material.

The official labeling program in the Netherlands complies with the EC directive in that labels contain fuel consumption and CO₂ emissions values, and it goes a step further in giving each vehicle a CO₂ emissions rating, from “A” (>20% more efficient) to “G” (>30% less efficient), that compares the car with others of similar size (defined as car width × car length) as well as with the average of all cars. The rating is 75% dependent on the comparison with cars of similar size and 25% dependent on the comparison with the average of all cars. This feature was adopted as a compromise, on the one hand, to ensure that small cars that are rather inefficient for their size are not always labeled as if to appear superior to larger cars that are quite efficient for their size and, on the other, to avoid giving large cars the same rating as small cars of similar technical efficiency even though the large cars have much higher CO₂ emissions.

6. GOVERNMENT SPONSORSHIP OF RESEARCH AND DEVELOPMENT

6.1 Overview

Although the automotive industry is a vigorous sponsor of research and development—worldwide, nearly \$20 billion U.S. worth in 1997—governments throughout the world sponsor additional automotive R&D, both separately from and in partnership with the industry. This work focuses primarily on four areas: emissions, safety, fuels, and fuel economy. Within the past few years, government sponsorship of automotive R&D has moved sharply in the direction of attempting to advance the performance and cost-effectiveness of automotive fuel cells, which address three of the four areas: emissions, fuels, and fuel economy. Aside from large programs in the three major market areas—the United States, Japan, and Western Europe—there is a significant fuel cell program in Canada and there are small efforts in South Korea and elsewhere.

Aside from direct sponsorship, governments promote R&D in the automotive industry (and elsewhere) through generous tax credits for R&D. Although the actual amount of tax credits taken by

automakers in the various markets is proprietary information, Japanese vehicle companies (excluding parts producers) had approximately \$190 million U.S. available to them in 1999 and \$122 million in 2000. U.S. companies (including parts producers) had approximately \$540 million available to them in 1997 and \$617 million available in 1998 in U.S. R&D tax incentives.

6.2 United States

The U.S. government has sponsored R&D in vehicle efficiency technology for decades, primarily through the Department of Energy, whose Electric and Hybrid Vehicle Program, begun in 1976, spent about \$500 million from its inception to the beginning, in 1992, of the Partnership for a New Generation of Vehicles (PNGV). PNGV was a joint R&D program between the federal government and the U.S. Council for Automotive Research (USCAR), an R&D partnership of Ford, General Motors, and DaimlerChrysler. For approximately a decade, PNGV was the centerpiece of the government's vehicle research efforts, with one of its key goals being the demonstration of a family car that could obtain 80 mpg while meeting all safety, emissions, performance, and affordability requirements. Research areas included engine powertrain systems (including lean-burn engines and emissions controls), advanced materials, electric energy storage and electric motors and controllers (for electric hybrids), and vehicle manufacturing. Federal research dollars were spread among the USCAR companies and their suppliers, the national laboratories, and other R&D performers such as universities.

In January 2002, the FreedomCAR Partnership superseded the PNGV, with the research focus shifting to the development of hydrogen as a primary fuel for cars and trucks. In fiscal year 2002, the Department of Energy spent approximately \$150 million on FreedomCAR-related research, with \$60 million focused on fuel cell and hydrogen infrastructure R&D. Despite the shift in primary focus, FreedomCAR's research goals retain some of PNGV's focus on materials and engine powertrain systems.

6.3 Japan

The Japanese government is sponsoring extensive research activity in high-efficiency vehicle technology for both conventional and fuel cell drivetrains. The 2001 Prime Minister's Action Plan calls for a \$155

million R&D program (among other initiatives such as vehicle purchase subsidies, tax incentives, and low-interest loans) aimed at putting 10 million low-emissions vehicles (including fuel cell, hybrid, and alternative-fuel vehicles) on the road by 2011, with 50,000 of these being fuel cell vehicles. The plan further calls for 5 million fuel cell vehicles to be on the road by 2020. More recent recommendations from senior ministers of the key transportation and environment ministries call for increased fuel cell R&D funding, more rapid vehicle deployment than is called for in the action plan, and additional measures to achieve global fuel cell leadership.

Prior to the new focus on fuel cells, the major government/industry R&D project for high-efficiency vehicles was the Advanced Clean Energy Vehicle (ACE) Project. The ACE Project's primary goal was to develop vehicles using clean fuels that would double energy efficiency with ultra-low emissions vehicle (ULEV) emissions. The total government plus private funding over the project's lifetime was approximately \$175 million, but the LDV part of this was substantially smaller (the project included development of technologies for buses and cargo trucks).

6.4 Europe

Both European auto manufacturers and the EU responded vigorously to the U.S. PNGV program, the manufacturers by forming the European Council for Automotive R&D (EUCAR) to coordinate R&D among the companies and the EU by initiating the Car of Tomorrow program. EU funding for EUCAR starting with the initiative was \$2.3 billion for 5 years.

Fuel cell funding by European governments was approximately \$60 million in 2001, with half from the EU and half from individual countries. However, the majority of these funds is aimed at transit projects. Key research areas are component cost reduction, development of a cost-effective hydrogen supply infrastructure, and development of vehicle systems. At the national government level, France and Germany have the most robust programs at approximately \$11.5 million and \$8 million, respectively (for fiscal year 2000). The Netherlands, Spain, Denmark, Italy, and the United Kingdom have programs at or above \$2 million per year, with Switzerland, Sweden, and Norway having smaller programs. In addition, individual governments have initiated more generalized efficiency R&D programs. For example, the United Kingdom's Foresight Vehicle Initiative, with nearly \$20 million of government money invested,

has focused on hybrid, electric, and alternative-fuel vehicles; advanced materials and structures; and powertrain technology—a program quite similar to that of the PNGV. However, even this program's focus is shifting more to fuel cell technology.

6.5 Canada

The Canadian government's R&D support has focused strongly on fuel cell research, in which it is a global leader. Canadian fuel cell research began in 1983 at Ballard Power Systems through a defense contract to further develop General Electric/National Aeronautics and Space Administration technology developed for the U.S. space program. Canadian government fuel cell support was approximately \$40 million in 2001, about one-third of the U.S. government's effort.

Acknowledgments

This article was produced by the University of Chicago as operator of Argonne National Laboratory under Contract W-31-109-ENG-38 with the U.S. Department of Energy. The U.S. government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in the article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the government.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • European Union Energy Policy • Fuel Cycle Analysis of Conventional and Alternative Fuel Vehicles • Hybrid Electric Vehicles • National Energy Policy: Japan • National Energy Policy: United States • Taxation of Energy • Transportation Fuel Alternatives for Highway Vehicles • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

- Greene, D. L. (1996). "Transportation and Energy." Eno Transportation Foundation, Landsdowne, VA.
- International Energy Agency. (2001). "Saving Oil and Reducing CO₂ Emissions from Transport: Options and Strategies." OECD/IEA, Paris.
- National Research Council. (2002). "Effectiveness and Impact of Corporate Average Fuel Economy (CAFE) Standards." National Academy Press, Washington, DC.
- Plotkin, S. (1999). Technologies and policies for controlling greenhouse gas emissions from the U.S. automobile and light truck fleet. *Transport. Q.* 53(2), 7–30.

Fuzzy Logic Modeling of Energy Systems

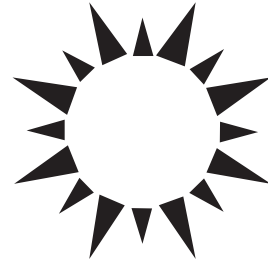
BJØRN LUDWIG

Technowledge Consulting

Göttingen, Germany

Technical University of Clausthal

Clausthal, Germany



-
1. Introduction
 2. Methodical Aspects
 3. Modeling Approach with Soft Computing
 4. Application to the Energy Supply
 5. Results and Discussion

Glossary

fitness function A mathematical expression transforming a certain combination of several input values into one output value, representing the evaluation of the input, often in relation to the optimum.

fuzzy logic A soft computing method to carry out logic operations; uses verbal expressions described by fuzzy sets generated from membership functions.

genetic algorithms A soft computing method that uses biological processes of evolution for optimization.

linguistic variable A variable for which the values are linguistic terms (for instance, “small”) instead of numbers. Fuzzy sets generated by membership functions build the connection between the linguistic expression and the numerical information.

membership function A mathematical expression that represents to what degree an element belongs to a mathematical set, considering that these memberships do not have to be total memberships.

operator A mathematical manipulation on numbers or terms. Classical examples are addition, multiplication, differentiation, etc. In soft computing, other operators are needed, e.g., fuzzy or genetic operators such as aggregation, accumulation, crossover, mutation, and reproduction.

soft computing Mathematical methods of calculating, developed mainly within the past 30 years, and not necessarily operating on classical functions and numbers, but taking into account the human kind of information processing and principles of biological operating.

sustainable development A goal for the future development of mankind under intragenerational and intergenerational justice. The negotiated road map for future economical, ecological, and societal sound development is the Agenda 21, from the United Nations Conference for Environment and Development held in Rio de Janeiro in 1992.

technology assessment An interdisciplinary scientific and societal approach to the questions of what impacts technology application and the technological progress have on mankind, and what technologies are needed and preferred.

The energy supply system is crucial to the further development of humans. Evaluations concerning the impact of energy conversion technologies on this development are difficult due to the complexity of the problem. This leads to different conclusions about how to proceed. Using soft computing techniques, both technically measurable quantities and nontechnical evaluation criteria can be taken into account to obtain an objective evaluation, to discover hidden evaluation criteria, and to formulate a goal for a future sustainable development.

1. INTRODUCTION

Current global development increasingly calls attention to the realities of population growth, extended mobility, and increasing energy demands and their contribution to environmental problems. This collision between civilization and the ecosystem illuminates two dilemmas: (1) The goals of maintaining society's prosperity and protecting the environment are not independent. Rather, they are competing goals, because technological applications increasingly

require extensive use of nonrenewable resources. (2) The globally increasing demand for nonrenewable resources necessitates even more technology application and development. Thus, guided by the vision of sustainability, the concept of technology assessment (TA) must deal with the question of technology's ability to enable sustainable development. To avoid increasing both resource use and risk, new technologies should be appropriate to address developmental needs, yet ecologically "reasonable." Designing such technologies requires improved information on and knowledge of technical systems. This knowledge encompasses development and application of technical systems and the connections among economic, social, and political systems, as well as impacts on the environment. Such knowledge should be used when evaluating both existing and new technologies within decision support systems. Problems characteristically encountered in making such evaluations involve multicriteria decision making, aggregation of facts, weighting of goals, uncertain information, and integrating people into the decision-making process. The overall problem is not a lack of data, but the difficulty in generating knowledge from a vast amount of partially uncertain, partially fuzzy data and information, often structured in such a way as to be rather useless. The global energy supply system, for example, is structured to consist of different primary energy carriers. This leads inherently to different evaluations of the system (for instance, due to the different ranges or durations of the primary energy carriers, or to the differences in the dangers of operating different conversion technologies), thus the present structure cannot be the most desirable one. An essential question for the future, concerning technology assessment studies, is how to achieve a sustainable energy supply. Under the present energy supply structure, evaluations for each applied energy conversion technology are needed.

Approaches to this entire problem are possible using a hybrid method of the soft computing techniques, fuzzy logic and genetic algorithms, first introduced by the author. This approach of recursive optimization considers different technical and non-technical evaluation criteria to achieve optimum primary energy percentages that form the best energy mix. The aim is to reduce the multidimensional decision-making problem to a comparison of aggregated one-dimensional numbers. This should be achieved by combining the advantages of human thought-process structures and processes of biological development with the logical and analytical accuracy of computers.

2. METHODOLOGICAL ASPECTS

Examining such problems analytically requires cognitive information processing. The desired systematic procedures require application of methods to organize and to extend our knowledge on complex systems. Generally, problem-specific combinations of methods are applied. Methodical problems and deficits usually emerge due to the complexity of the processes or of the impacts. Problems in describing complex dynamical systems are due to nonlinear impacts and boundary conditions, to boundary conditions that are time dependent and/or solution dependent, and to different time constants within the subsystems.

2.1 Deficit Analysis

Knowledge about technical systems is commonly gained by analyzing the underlying interactions. This is mainly characterized by the building of models. These are mainly mathematical models that allow systematic simulations to identify potential reactions of the system.

In the case of a holistic and systemic view, the system border is extended to all participating and impacting areas. Thus, interactions and impacts have to be considered, but often resist analysis using a description with classical mathematics—say, a formalistic handling. These entities have one of the following properties: unique basic variables are not identifiable, unique causalities are not recognizable among them, or measurable indicators cannot be determined in case of known interactions. This plays a considerable role in the discussion of sustainability, because the entities could be impacts such as risk, quality of life, or diversity. These are hardly measurable quantities. Mainly they are considered in models by boundary conditions or verbal discussion. Neglecting these entities or disaggregating them to other available variables leads, in any case, to an information loss, so the complexity of the system is not represented completely by the model.

2.2 Deficit Synthesis

Complex systems in practice are handled, or at least are made more accessible, by composite numbers or indicators. This is familiar, for instance, in process or chemical engineering. In contrast to the information content of single-measurement data, composite indicators generate new knowledge by aggregating single data and information. The procedure of

aggregation varies depending on the indicators. Nevertheless, indicators of different aggregation levels always provide the same information content. Indicators are especially important in environmental discussions. Assessing or evaluating a situation or guiding a decision requires differently aggregated indicators that characterize a process or a state.

The aggregation problem points out the following dilemma: the methodical deficits of assessments are due to problems of comparison because of the transformation of indicators to highly aggregated assessment measures. On the other hand, this procedure is necessary, which becomes obvious, for instance, if entities such as environmental pollution are to be quantified. This property allows less aggregated indicators to be used at the beginning of systems analysis (for instance, within an inventory analysis of modern environmental management approaches) and enhances transparent aggregation procedures.

2.3 Deficit Assessment

A multidimensional assessment problem can be considered as a logical measurement operation. The assessment problem then is the optimal arrangement of a set of possible alternatives, with respect to the relevant aims and to the preferences of the decision makers, considering given restrictions. The aim of an assessment is to determine a particular value for any alternative so that its relative advantages are represented in only a single expression. The solution of this selection problem will be especially difficult if the problem structure is complex (that is, when many objectives are to be considered, different assessment scales emerge, objectives are weighted differently, information is at least partially uncertain, the problem is time-dependent, many participants are to be involved in the decision-making process, or no unique criterion exists for decision making). Contributing variables could be factors such as risk or diversity, which represent heterogeneous fuzzy expressions that cannot be broken down any further, often with many different impacts. This makes comparability of these variables difficult, especially when factors of different aggregation degree are involved, and makes any evaluation subjective. Nevertheless, to describe real systems, these influences are indispensable quantities.

2.4 Goal

Designing innovative technologies, considering widespread aspects such as sustainable development,

requires environmental information for design, assessment, and decision processes. The difficulties mainly emerge during recording and assessing of entities. The problem is caused by aggregated system variables that are incomparable. More adequate methodical approaches maintain the complexity of these variables. Necessary subjective inputs during a system analysis then have to be made absolutely clear to maintain transparency, and they have to be introduced as early as possible in an algorithmic procedure to avoid a conscious manipulation of the solution.

3. MODELING APPROACH WITH SOFT COMPUTING

At a level of thinking that does not have to be conscious, humans use a kind of knowledge that is not sharp and is hard to formulate explicitly. This fuzziness is not a property of reality, but of perception. This kind of knowledge is characterized by the handling of incomplete and not well-structured patterns and information. Here the question emerges as to how and with what language we can model such uncertainties. The methodical deficits mentioned previously can be softened by application of the techniques of soft computing. Apart from artificial neural nets, these are primarily fuzzy logic and genetic algorithms; their usefulness is due to their potential to integrate complexity into calculation.

3.1 Fuzzy Logic

Within the context of this discussion, only fuzzy set theory, one branch of fuzzy logic, is of interest for application to technology assessment, because the important questions of TA do not deal with maintaining states of operation, the subject of control theory, but with fuzzy information processing. Fuzzy logic enables the handling of verbal and indeterminate information using exact mathematics.

Traditional binary logic is part of fuzzy logic, as a special case operating only with two values of interpretation. A set is fuzzy limited if not all the members of the set are assigned a value of 1, or 100%, which means total membership. A fuzzy set is defined by the generalized characteristic function, called the membership function μ . This real function can take on any values, but usually it is normalized into the interval $[0, 1]$.

The key notion when modeling with fuzzy logic is the linguistic variable. In contrast to mathematically precise quantifiable equations, verbal descriptions of

activities contain terms of properties with smooth gradation transitions. Linguistically formulated variables are generally more intelligible, but have a higher aggregated information content. These entities are therefore more difficult to quantify without an information loss. The values of a linguistic variable are verbal expressions, called linguistic terms (for instance, the term “small”). The content of each linguistic term is identified with one fuzzy set and is assigned to the related numerical scale of the basic variable by a particular membership function. Thus, the fuzzy sets build the connection between the linguistic expression and the numerical information.

To process fuzzy formulated knowledge, several linguistic variables must be linked by linguistic operators. The connecting rules represent the knowledge, which is stored in a rule base or knowledge base, similar to expert systems. The procedure consists of the sequential steps fuzzification, inference, and defuzzification.

3.2 An Assessment Method Using Fuzzy Logic

Along with the preceding remarks, a general assessment method using fuzzy logic has been developed. An assessment is a determination of the goodness of a considered issue. This will be made by a function that systematically integrates knowledge into the procedure of modeling and also handles the aggregation procedure. The assessment procedure is based on a multidimensional vector of criteria containing all knowledge input, which characterizes the matter to be evaluated related to the chosen degree of aggregation. The aim is to get statements of the form “sth is good, fair, satisfying, to introduce, desirable, etc.” This kind of statement can be understood as a result of logic operations.

The concept of linguistic variables enables the formulation of scales of values with linguistic terms as gradations. The essential and critical process is the formalistic assignment of human knowledge to fuzzy sets. However, through this procedure, existing subjective preferences are integrated transparently at each stage, so a conscious manipulation has only a limited impact on the final result.

The general five-step top-down procedure is as follows:

1. Definition of a scale for evaluation.
2. Definition of subordinate criteria for ranking onto this scale. The question concerning which of the possibly independent criteria and impacts will influence the superior scale from step 1 has to be answered.

3. Stepwise disaggregation of these basic criteria. The subordinate impacts are disaggregated to further arbitrary substages, depending on the given knowledge. The result is an aggregation hierarchy of the different impacts. Generally, steps 2 and 3 answer the question of which of the less aggregated and possibly independent impacts are impacting on the next higher aggregation level. Depending on the character of an impact, it can be described by a measurable quantity on a less aggregated level. Otherwise, it keeps the character of an aggregated entity. Both kinds of characters can be considered simultaneously.

4. Fixing the specifications of the fuzzy system parameters. This includes fixing the numbers of linguistic terms and the verbal connotations of all attendant variables, the shape of the membership functions (default: triangular), and the logic operators for aggregation (default: minimum) and for accumulation (default: maximum), as well as the defuzzification method (default: center of gravity). It is recommended to choose an odd number of linguistic gradations for each input variable (default: three) to ensure a certain resolution, with the additional possibility of naming a medium term. The number of linguistic terms of all superior aggregated variables then can be taken as determined in order to get an objective rule base (see step 5).

5. Formulation of the rule base. The phenomenological dependencies between the variables are stored in the rule base. The rules state which combination of linguistic input terms is connected with which linguistic output term. To achieve a definite assignment in the rule base, the number of linguistic terms of a superior aggregated variable then is determined by all possible combinations of input linguistic terms of all related subordinate variables.

This is carried out by a numerical codification. The N linguistic terms of a subordinate variable are assigned to numbers in the interval $[-(N-1)/2, \dots, -1, 0, 1, \dots, (N-1)/2]$ ($N=3, 5, 7, \dots$), where the medium term is assigned to zero. For instance, a variable with $N_{In}=3$ linguistic terms (small, medium, large) would be assigned to $(-1, 0, 1)$. Thus, we can determine the number N_{Out} of linguistic terms of an output variable from its M input variables ($M=1, 2, 3, \dots$) with their $N_{In,1}, \dots, N_{In,M}$ linguistic terms as follows [Eq. (1)]:

$$N_{Out} = 1 + 2 \sum_{i=1}^M \frac{N_{In,i}-1}{2} \quad (1)$$

For instance, an output variable aggregated from $M=2$ input variables with $N_{In,1}=3$ and $N_{In,2}=5$

linguistic terms would then have $N_{\text{Out}} = 1 + 2(1 + 2) = 7$ linguistic terms.

The rules can be derived from the following procedure. The conditions of all rules that are the permutations of all linguistic input terms can be written numerically codified. The conclusion is then represented by the linguistic term of the output variable (numerical codification is assigned by the sum of the numerical input codifications). For instance, considering one three-term input variable with the terms (small, medium, large) codified as $(-1, 0, 1)$, and a second input variable with five terms (very small, small, medium, large, very large) codified as $(-2, -1, 0, 1, 2)$, the particular rule may be written as follows:

$$\begin{aligned} &\text{IF Input}_1 = \text{small AND Input}_2 = \text{medium} \\ &\text{THEN Output} = ? \end{aligned} \quad (2)$$

which is codified as

$$\begin{aligned} &\text{IF Input}_1 = -1 \text{ AND Input}_2 = 0 \\ &\text{THEN Output} = -1 + 0 = -1 \end{aligned} \quad (3)$$

According to Eq. (1), the output variable has seven terms codified as $(-3, -2, -1, 0, 1, 2, 3)$. Therefore, the conclusion $\text{Output} = -1$ means that the result is the linguistic output term that is located one term smaller than the medium term, whatever the verbal connotation may be.

3.3 Genetic Algorithms

The idea of technology assessment to point out alternatives to any problem solution can be taken as an optimization problem, where optimization of an application is aimed at, for instance, its energy efficiency, its material intensity, or its sustainability. The application of usual optimization procedures is limited. Restraints can be nonlinear problems or problems with many parameters or many restrictions due to the programming effort and memory demand.

Genetic algorithms, instead, refer to a behavior that would be similar to a biological process of evolution. These techniques do not have requirements for linearity, differentiation, or a certain structure, and often require less programming effort. Problem solutions are kept or changed due to the principle of survival of the fittest. The point is that the search for more capable solutions takes place with a high probability at sites of the solution space, where an enhancement higher than average is required. These features are used in genetic algorithms to evaluate systematically created solutions. A genetic algorithm can be defined as an iterative

procedure maintaining a population of structures that are candidate solutions to specific domain challenges. Each candidate solution is called a chromosome. As in biological systems, a chromosome can make a copy of itself. The copy might be slightly different from its parent. During each temporal increment, called a generation, the structures in the current population are rated for their effectiveness as domain solutions, and on the basis of these evaluations, a new population of candidate solutions is formed using specific “genetic operators,” which are used by most genetic algorithms, such as reproduction, crossover, and mutation:

1. *Reproduction*: Through reproduction, genetic algorithms produce new generations of improved solutions by selecting parents with higher fitness ratings or by giving such parents greater probability to be contributors.

2. *Crossover*: Usually genetic algorithms use binary strings to represent solutions. Crossover means choosing a random position on the binary string (that is, after a certain digit) and exchanging the segments either to the right or to the left of this point with another string partitioned similarly.

3. *Mutation*: This genetic operator is an arbitrary change of a single bit of a chromosome. It is needed to keep the algorithm from getting stuck. This operator changes a 1 to 0 or a 0 to 1.

The natural optimization techniques can be distinguished by their main operators. In case of genetic algorithms, the main operator is crossover, which is chosen with a probability of 80–100%. Mutation, in contrast, is a very subordinate operator. Such a bit change occurs at a low probability of 1–10%. Genetic algorithms use a three-step iterative process: (1) test a solution to see how good it is, (2) select the best parents, and (3) generate offspring. Genetic algorithms provide a set of efficient, domain-independent search heuristics. Information that enables rejection of inferior solutions and accumulation of good ones is received. Only genetic algorithms and their derivatives provide methods powerful enough for optimizing complex problems. Traditional optimizing tools are not applicable.

4. APPLICATION TO THE ENERGY SUPPLY

Decreasing resources of fossil primary energy carriers and globally increasing mean per capita energy demand and population growth bring to view the

question how and by which technologies the future energy supply can be guaranteed. Of primary concern here is the evaluation of the sustainability of the energy supply system. An energy supply system provides energy via different conversion technologies. In the present global energy supply system, 90% of the demand is still provided by fossil fuel-based primary energy carriers. Increasing energy demands and increasingly urgent global impacts require development of a new supply system, keeping in mind future generations and threshold countries with vastly growing economies; this is referred to as intergenerational and intragenerational justice. Due to the requirement for sustainability of a new energy supply system, one of the most important questions to be answered concerns which primary energy carriers can be and should be used, and to what extent in consideration of limited resources and potential risks. This question concerns the optimal structure of percentages of the different primary energy carriers, of the different energy conversion technologies, or the optimal energy mix. The applied recursive optimization is a two-step iterative procedure. First, a given energy supply structure must be evaluated, considering the technologies involved, with regard to certain evaluation criteria. Second, the input structure will be varied to find the best rated energy mix.

4.1 Assessment: The Fuzzy Fitness Function

To evaluate a given primary energy structure, the fuzzy logic assessment method described previously is used. The assessment procedure is based on a multidimensional vector of criteria that characterizes a conversion technology. The method is applied as follows, according to the previous description:

1. Definition of a scale for evaluation. The interesting entity is the sustainability of an energy conversion technology, here reduced to the related primary energy carrier.
2. Definition of subordinate criteria for the ranking onto the scale sustainability. In case of energy conversion technologies, the independent criteria duration of the primary energy carrier and ecological relevance of the technology are chosen.
3. Stepwise disaggregation of these subordinate criteria. In this case (Fig. 1), the ecological relevance could be considered as being made up from the basic criteria climate impact, area demand, and risk. Area demand is composed of the value and the size of the

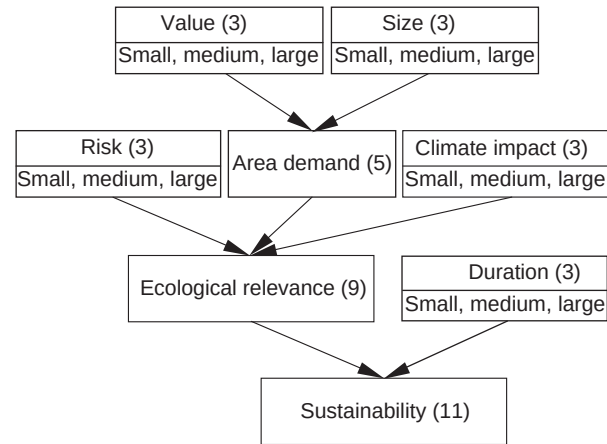


FIGURE 1 Aggregation hierarchy of the assessment function. The number of linguistic terms of each variable determined with Eq. (1) is in parentheses. Reproduced with permission from the Society of Automotive Engineers, SAE 1999-01-2708. Copyright 1999, SAE.

used area. Duration, for instance, may depend on the level of technology, on the reserve of the primary energy carrier considered, and on social factors. The climate impact can be specified with technical quantities such as emissions and global warming potential. However, both are not differentiated further here. Risk in this approach should specify the potential danger in case of an accident.

4. Fixing the specifications of the fuzzy system parameters. Here, for all specifications, the default standard parameters are chosen. Each input variable has been aligned to three linguistic terms (small, medium, large). According to Eq. (1) and the hierarchical aggregation structure (Fig. 1), the numbers of linguistic terms for the superior variables given in Fig. 1 are obtained. All fuzzy sets at the edges of the intervals of the basic variables are performed in such a way that defuzzification of the minimum (maximum) linguistic term would provide the value of 0 (1) of the basic variable. Therefore, the membership functions of these linguistic terms of the input basic variables have to be defined symmetrical to 0 (1).

5. Formulation of the rule base. With the preceding procedure, the objective rule bases for the three aggregation levels (for area demand with 9 rules, for ecological demand with 45 rules, and for sustainability with 27 rules) are obtained. For example, the aggregation levels would have the following formulations:

First aggregation level:

IF Value = medium

AND Size = small

THEN Area Demand = small (4)

Second aggregation level:

IF Risk = small AND

Area Demand = small

AND Climate Impact = small

THEN Ecological Relevance = very very small (5)

Third aggregation level:

IF Ecological Relevance = very very small

AND Duration = large

THEN Sustainability = very very very large. (6)

Each primary energy carrier being considered can now be characterized by the vector of criteria (elements are the values of the inputs, each on a specific aggregation stage). The assessment result can be a verbal or a numerical value of the highest aggregated output variable, "sustainability." In case of an energy mix, different technologies are considered with their relative percentages. An input value for the mix then has to be built by the sum of the values of the single technologies weighted with the percentage within the mix. This assessment procedure operates as the fitness function of the genetic algorithm.

4.2 The Optimization Procedure

The optimization problem consists of determining the percentages of each considered technology that participates in supplying the energy demand. During the optimization procedure, the following fitness function F has to be optimized:

$$F = F(\alpha, \beta, \gamma) \rightarrow \max \quad (7)$$

where α represents the contribution of fossil-fuel-based primary energy carriers, β represents that one of the regenerative (renewable)-energy-based primary energy carriers, and γ represents the contribution of nuclear-energy-based primary energy carriers. The boundary condition, of course, is $\alpha + \beta + \gamma = 1$, or 100%. First, the aggregated inputs to the fuzzy block are determined (Fig. 2). Therefore, each single, specific evaluation of a technology taken from Table I is weighted by its percentage and then summarized. This resulting vector of criteria for a considered energy structure serves as input vector for the fuzzy fitness function that determines the total goodness of the structure. The genetic algorithm now varies the input vector of percentages of a given energy structure until the maximum fitness is achieved (Fig. 2).

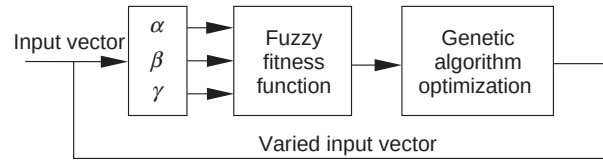


FIGURE 2 Recursive optimization, where the contributions of primary energy carriers are α , fossil fuels; β , regenerative (renewable) energy; and γ , nuclear energy. Reproduced with permission from the Society of Automotive Engineers, SAE 1999-01-2708. Copyright 1999, SAE.

5. RESULTS AND DISCUSSION

Vectors of criteria for different technologies and sustainability results are shown in Table I. The current energy mix has been determined twice, first considering the nuclear percentage based on uranium-235 (mix 1) and second based on the breeder technology (mix 2). A separate control calculation with the analytical hierarchy process method has resulted in the same ranking tendency. Although fossil energy carriers use oil and gas in addition to coal, oil and gas have not been differentiated here, which should be done for more precise results. Also, in reality, hydro energy is the most significant

TABLE I

Vectors of Criteria for Different Primary Energy Carriers

Vector	Criteria ^a					
	CI	Value	Size	Risk	D	S
Test cases ^b						
Max	0	0	0	0	1	1
Min	1	1	1	1	0	0
Technologies ^c						
Fossil (coal)	1.0	1.0	0.01	0.01	0.3	0.439
Nuclear (uranium-235)	0.1	1.0	0.01	0.95	0.1	0.400
Nuclear (breeder)	0.1	1.0	0.01	1.0	0.8	0.510
Regenerative (solar)	0.1	0.01	1.0	0.01	1.0	0.780
Mix 1995 ^d						
(1)	0.906	0.968	0.042	0.079	0.308	0.444
(2)				0.082	0.359	

^aData from SAE 1999-01-2708. CI, climate impact; D, duration; S, sustainability.

^bThe two test cases verified the algorithm.

^cThe single values for the technologies were determined from the literature.

^dEvaluation of the 1995 energy structure in Germany: (1), standard nuclear energy; (2), fast breeder technology.

contributor to regenerative energy supplies, but the percentage in Table I has been taken as consisting entirely of solar energy (photovoltaics). Therefore, the values in Table I for climate impact, value, and size of area are not calculated properly. Nevertheless, the results allow the following statements:

1. The assessment procedure provides correct results for given vectors of criteria.
2. Maximum sustainability cannot be achieved.
3. The traditional nuclear energy conversion technology has been ranked worst because of the high risk and the simultaneous short duration.
4. The current energy mix is debatable, considering this evaluation.
5. The best sustainability is achieved for regenerative energies, and is strongly dependent on the value of the size of area used.

Not surprisingly, the optimization procedure provides 100% regenerative energy carriers as the best solution if risk is considered, and 100% nuclear energy if not. Far more interesting are results evaluated as being a little worse than the pure regenerative solution, but better than today's mix. In this case, the dominant regenerative part decreases with decreasing evaluation values, which also start to oscillate. Simultaneously, the percentages of fossil-fuel-based and nuclear-based primary energy carriers increase alternately, also with increasing oscillations. However, there are obviously several solutions with the same evaluation result.

Using a linguistic notion of arrangement, fuzzy logic allows simultaneous consideration of entities and technically measurable quantities by a verbal description of interactions. The evaluation results for the current energy mix and for certain other possibilities with respect to meeting the worldwide future energy demand show that this kind of consideration is useful

to discover hidden evaluation criteria and to formulate a goal for a future sustainable development.

SEE ALSO THE FOLLOWING ARTICLES

Computer Modeling of Renewable Power Systems • Modeling Energy Supply and Demand: A Comparison of Approaches • Neural Network Modeling of Energy Systems • Population Growth and Energy

Further Reading

- Gore, A. (1993). "Earth in the Balance." Plume, New York.
- Grefenstette, J. (1982). Optimization of control parameters for genetic algorithms. *IEEE Transact. Syst. Mgmt. Cybernet.* 16(1).
- International Standards Organization (ISO). (1997). "Environmental Management—Life Cycle Assessment." ISO 14040. Beuth, Berlin.
- Ludwig, B. (1995). "Methoden zur Modellbildung in der Technikbewertung." Papierflieger, Clausthal-Zellerfeld, Germany.
- Ludwig, B. (1997). On the sustainability of future energy systems. *Energy Convers. Mgmt.* 38(15–17), 1765–1776.
- Ludwig, B. (1997). The concept of technology assessment—An entire process to sustainable development. *Sus. Dev.* 5(3), 111–117.
- Ludwig, B. (1997). "Optimization of Energy Systems Under the Aspect of Sustainability." Proceedings, 32nd Intersociety Energy Conversion Engineering Conference (IECEC), Honolulu, July 27–August 1, 1997, pp. 2070–2075.
- Ludwig, B. (1998). Fuzzy logic applications in technology assessment studies. *J. Intelligent Fuzzy Syst.* 6, 375–388.
- Mitchell, G. (1996). Problems and fundamentals of sustainable development indicators. *Sus. Dev.* 5, 1–11.
- Saaty, T. L. (1980). "The Analytic Hierarchy Process." McGraw-Hill, New York.
- United Nations. (1997). "Yearbook of World Energy Statistics." United Nations, New York.
- Zadeh, L. A. (1965). Fuzzy sets. *Inform. Control* 8, 338–353.
- Zimmermann, H.-J. (1991). "Fuzzy Set Theory and Its Applications." Kluwer Academic Publ., Boston.

Gas Hydrates

TIMOTHY S. COLLETT

United States Geological Survey
Denver, Colorado, United States



-
1. Implications for Gas Hydrates
 2. Gas Hydrate Technical Review
 3. Energy Resource Potential of Gas Hydrates

Glossary

clathrate A class of compounds characterized by cages of molecules that encapsulate other molecules; from the *Latin clathratus*, meaning to encage.

dissociation A chemical process by which temperature and/or pressure changes cause a group of molecules to be separated into simpler groups.

methane A colorless flammable gas consisting of carbon and hydrogen.

methanol Methyl alcohol; a colorless, flammable liquid derived from natural gas.

permafrost Ground (soil or rock) that remains at or below 0°C for at least 2 years.

seismic reflectors An interface between two media with different elastic properties as shown on a seismic profile.

well log A graphic record of the measured physical characteristics of a subsurface rock section as encountered in a well.

Gas hydrates, composed of water and gas, are crystalline substances in which solid lattices of water molecules trap gas molecules in a cagelike structure, or a clathrate. Gas hydrates occur in sedimentary deposits under conditions of pressure and temperature present in permafrost regions and beneath the sea in outer continental margins. Despite the fact that relatively little is known about the ultimate resource potential of natural-gas hydrates, it is certain that they are a vast storehouse of natural gas. There is much to learn about these compounds before this potentially enormous resource can be turned into an economically producible reserve.

1. IMPLICATIONS FOR GAS HYDRATES

The combined information from Arctic gas hydrate studies shows that, in permafrost regions, gas hydrates may exist at subsurface depths ranging from about 130 to 2000 m. Anomalous seismic reflectors, known as bottom-simulating reflectors (BSRs), have been mapped at depths below the seafloor ranging from about 100 to 1100 m. The presence of gas hydrates in offshore continental margins has been inferred mainly from these BSRs. Although methane, propane, and other gases can be included in the clathrate lattice structure of gas hydrates, methane hydrates appear to be the most common in nature. The amount of methane sequestered in gas hydrates is probably enormous, but estimates of the amounts are speculative and range over three orders of magnitude, from about 3114 trillion to 7,634,000 trillion cubic meters, which represents a very large potential energy resource.

Gas hydrates also represent a significant drilling and production hazard. Russian, Canadian, and American drilling engineers have described numerous problems associated with gas hydrates, including blowouts and well-bore casing failures. Seafloor stability is also an important issue related to gas hydrates. Seafloor stability refers to the susceptibility of the seafloor to collapse and slide as the result of gas hydrate disassociation. It is possible that both natural and anthropogenic changes can contribute to *in situ* gas hydrate destabilization, which may convert a hydrate-bearing sediment to a gassy water-rich fluid, triggering seafloor subsidence and catastrophic landslides. Evidence implicating gas hydrates in triggering seafloor landslides has been found along the Atlantic Ocean margin of the United States. However, the mechanisms controlling gas hydrate-induced seafloor subsidence and landslides are not well understood.

Recent studies indicate that atmospheric methane, a greenhouse gas, is increasing at a rate such that the current concentration will probably double in the next 50 years. Because methane is 21 times more radiatively active than carbon dioxide, it is predicted that methane will surpass carbon dioxide as the predominant atmospheric greenhouse gas in the second half of the next century. The source of this atmospheric methane is uncertain; however, numerous researchers have suggested that destabilized natural-gas hydrates may be contributing to the buildup of atmospheric methane.

2. GAS HYDRATE TECHNICAL REVIEW

Under appropriate conditions of temperature and pressure (Fig. 1), gas hydrates usually form one of two basic crystal structures (Fig. 2). Each unit cell of Structure I gas hydrate (Fig. 2A) consists of 46 water molecules that form two small dodecahedral voids and six large tetradecahedral voids. Structure I gas hydrates can hold only small gas molecules such as methane and ethane, with molecular diameters not exceeding 5.2 Å. The chemical composition of a Structure I gas hydrate can be expressed as $8(\text{Ar}, \text{CH}_4, \text{H}_2\text{S}, \text{CO}_2)46\text{H}_2\text{O}$ or $(\text{Ar}, \text{CH}_4, \text{H}_2\text{S}, \text{CO}_2)5.7\text{H}_2\text{O}$. The unit cell of Structure II gas hydrate consists of 16 small dodecahedral and 8 large hexakaidecahedral voids formed by 136 water molecules. Structure II gas hydrates may contain gases with molecular dimensions in the range of 5.9 to 6.9 Å, such as propane and isobutane. The chemical composition of a Structure II gas hydrate can be expressed as $8(\text{C}_3\text{H}_8, \text{C}_4\text{H}_{10}, \text{CH}_2\text{Cl}_2, \text{CHCl}_3)136\text{H}_2\text{O}$ or $(\text{C}_3\text{H}_8, \text{C}_4\text{H}_{10}, \text{CH}_2\text{Cl}_2, \text{CHCl}_3)17\text{H}_2\text{O}$. At conditions of standard temperature and pressure (STP), one volume of saturated methane hydrate (Structure I) will contain as much as 189 volumes of methane gas; because of this large gas storage capacity, gas hydrates are thought to represent an important source of natural gas.

The geologic occurrence of gas hydrates has been known since the mid-1960s, when gas hydrate accumulations were discovered in Russia. Cold surface temperatures at high latitudes on Earth are conducive to the development of onshore permafrost and gas hydrates in the subsurface (Fig. 3). Onshore gas hydrates are known to be present in the West Siberian Basin and are believed to occur in other permafrost areas of northern Russia, including the

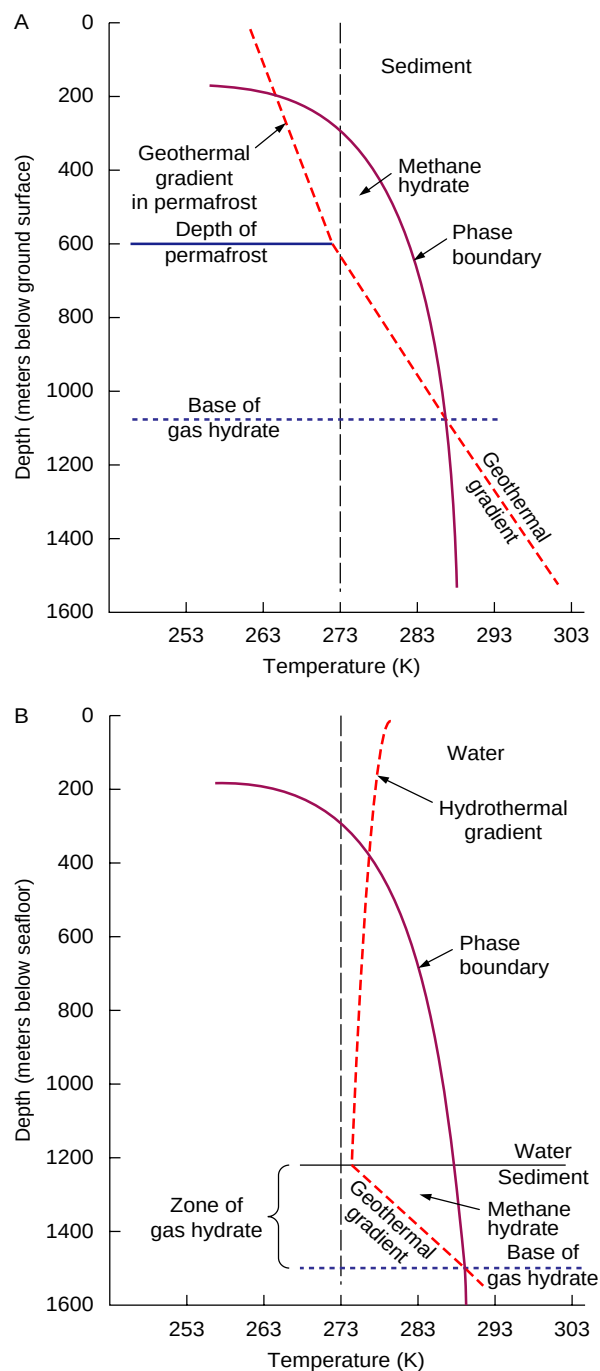


FIGURE 1 Graphs showing the depth-temperature zone in which methane hydrates are stable in (A) a permafrost region and (B) an outer continental margin marine setting.

Timan-Pechora province, the eastern Siberian Craton, and the northeastern Siberia and Kamchatka areas. Permafrost-associated gas hydrates are also present in the North American Arctic. Direct evidence for gas hydrates on the North Slope of Alaska comes from a core test in a gas hydrate exploration well, and

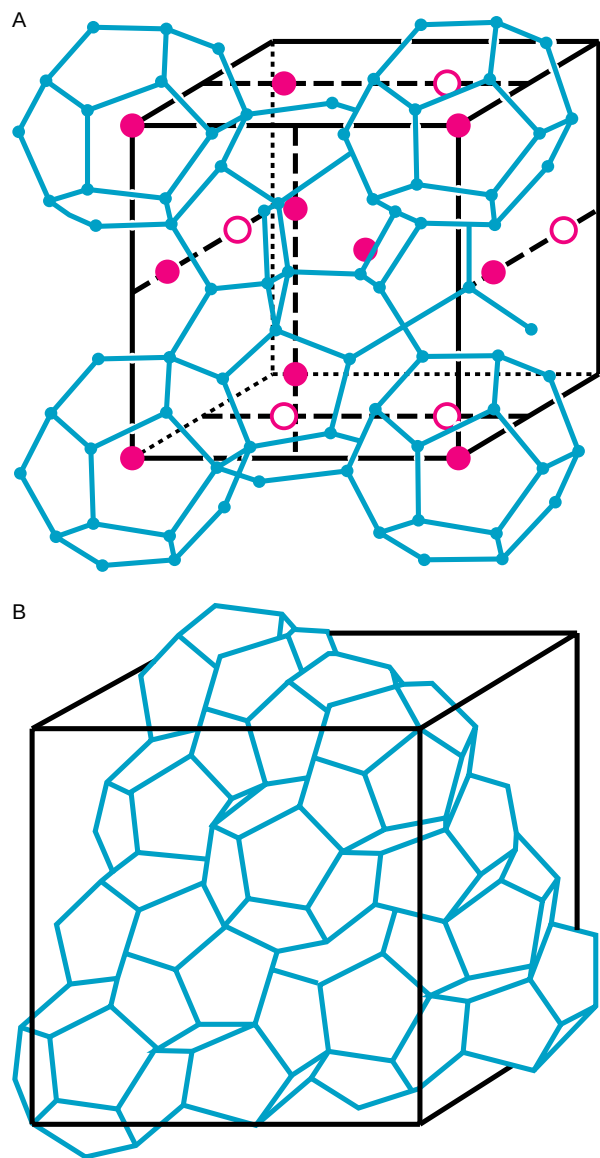


FIGURE 2 Two gas hydrate crystal structures: (A) Structure I and (B) Structure II.

indirect evidence comes from drilling and open-hole industry well logs that suggest the presence of numerous gas hydrate layers in the area of the Prudhoe Bay and Kuparuk River oil fields. In Canada, well-log responses attributed to the presence of gas hydrates have been obtained in about one-fifth of the wells drilled in the Mackenzie Delta, and more than half of the wells in the Arctic Islands are inferred to contain gas hydrates. The recently completed Mallik 2L-38 gas hydrate research well confirmed the presence of a relatively thick, highly concentrated, gas hydrate accumulation on Richards Island in the outer portion of the Mackenzie River Delta.

In marine environments, gas hydrates have been recovered in gravity cores within 10 m of the seafloor in the sediment of the Gulf of Mexico, the offshore portion of the Eel River Basin of California, the Black Sea, the Caspian Sea, and the Sea of Okhotsk. Also, gas hydrates have been recovered at greater subbottom depths during research coring along the southeastern coast of the United States on the Blake Ridge, in the Gulf of Mexico, in the Cascadia Basin near Oregon, in the Middle America Trench, on offshore Peru, and on both the eastern and western margins of Japan.

3. ENERGY RESOURCE POTENTIAL OF GAS HYDRATES

World estimates for the amount of natural gas in gas hydrate deposits range from 14 trillion to 34,000 trillion cubic meters for permafrost areas and from 3100 trillion to 7,600,000 trillion cubic meters for oceanic sediments. Current estimates of the amount of methane in the world's gas hydrate accumulations are in rough accord at about 20,000 trillion cubic meters. If these estimates are valid, the amount of methane in gas hydrates is almost two orders of magnitude larger than the estimated total remaining recoverable conventional methane resources, estimated to be about 250 trillion cubic meters.

The 1995 National Assessment of United States Oil and Gas Resources, conducted by the U.S. Geological Survey, focused on assessing the undiscovered conventional and unconventional resources of crude oil and natural gas in the United States. This

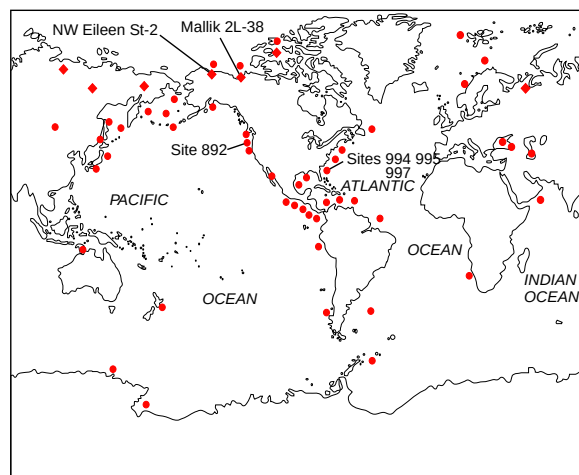


FIGURE 3 Locations of known and inferred gas hydrate occurrences.

assessment included for the first time a systematic resource appraisal of the in-place natural gas hydrate resources of the onshore and offshore regions of the United States. In this assessment, 11 gas hydrate plays were identified within four offshore and one onshore gas hydrate provinces. The offshore gas hydrate provinces assessed lie within the U.S. Exclusive Economic Zone adjacent to the lower 48 states and Alaska. The only onshore province assessed was the North Slope of Alaska. In-place gas resources within the gas hydrate regions of the United States are estimated to range from about 3200 trillion to 19,000 trillion cubic meters of gas, at the 0.95 and 0.05 probability levels, respectively. Although this wide range of values shows a high degree of uncertainty, it does indicate the potential for enormous quantities of gas stored as gas hydrates. The mean in-place value for the entire United States is calculated to be about 9000 trillion cubic meters of gas, which is a very large volume of gas, when it is considered that the United States has only about 7 trillion cubic meters of conventional natural gas reserves.

Even though gas hydrates are known to occur in numerous marine and Arctic settings, little is known about the technology necessary to produce gas hydrates. Most of the existing gas hydrate "resource" assessments do not address the problem of gas hydrate recoverability. Proposed methods of gas recovery from hydrates (Fig. 4) usually deal with dissociating or "melting" gas hydrates *in situ* by (1) heating the reservoir beyond hydrate formation temperatures, (2) decreasing the reservoir pressure below hydrate equilibrium, or (3) injecting an inhibitor, such as methanol or glycol, into the reservoir to decrease hydrate stability conditions. Gas recovery from hydrates is hindered because the gas is in a solid form and because hydrates are usually widely dispersed in hostile Arctic and deep marine environments. First-order thermal stimula-

tion computer models (incorporating heat and mass balance) have been developed to evaluate hydrate gas production from hot water and steam floods; the models have shown that gas can be produced from hydrates at sufficient rates to make gas hydrates a technically recoverable resource. However, the economic cost associated with these types of enhanced gas recovery techniques would be prohibitive. Similarly, the use of gas hydrate inhibitors in the production of gas from hydrates has been shown to be technically feasible, but the use of large volumes of chemicals such as methanol comes with a high economic and environmental cost. Among the various techniques for production of natural gas from *in situ* gas hydrates, the most economically promising method is considered to be the depressurization technique. However, the extraction of gas from a gas hydrate accumulation by depressurization may be hampered by the formation of ice and/or the reformation of gas hydrate due to the endothermic nature of gas hydrate dissociation.

The Messoyakha gas field in the northern part of the Russian West Siberian Basin is often used as an example of a hydrocarbon accumulation from which gas has been produced from *in situ* natural-gas hydrates. Production data and other pertinent geologic information have been used to document the presence of gas hydrates within the upper part of the Messoyakha field. It has also been suggested that the production history of the Messoyakha field demonstrates that gas hydrates are an immediately producible source of natural gas, and that production can be started and maintained by conventional methods. Long-term production from the gas hydrate part of the Messoyakha field is presumed to have been achieved by the simple depressurization scheme. As production began from the lower free-gas portion of the Messoyakha field in 1969, the measured reservoir pressures followed predicted decline relations; however, by 1971, the reservoir pressures

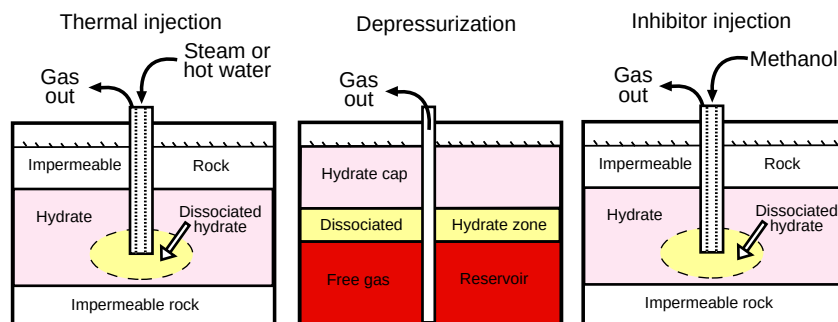


FIGURE 4 Schematic of proposed gas hydrate production methods.

began to deviate from expected values. This deviation has been attributed to the liberation of free gas from dissociating gas hydrates. Throughout the production history of the Messoyakha field, it is estimated that about 36% (about 5 billion cubic meters) of the gas withdrawn from the field has come from the gas hydrates. However, several recent studies suggest that gas hydrates may not be significantly contributing to gas production in the Messoyakha field.

It should be noted that the current assessment of proposed methods for gas hydrate production does not consider some of the more recently developed advanced oil and gas production schemes. For example, the usefulness of downhole heating methods such as *in situ* combustion, electromagnetic heating, or downhole electrical heating have not been evaluated. In addition, advanced drilling techniques and complex downhole completions, including horizontal wells and multiple laterals, have not been considered in any comprehensive gas hydrate production scheme. Gas hydrate provinces with existing conventional oil and gas production may also present the opportunity to test relatively more advanced gas hydrate production methods. For example, in northern Alaska, existing "watered-out" production wells are being evaluated as potential sources for hot geopressured brines that will be used to stimulate gas hydrate production thermally.

It has been suggested that some of the methane liberated from a producing gas hydrate field could be lost to the atmosphere; methane is believed to be a significant greenhouse gas. It is possible that large-scale production of gas from hydrates could contribute methane to the atmosphere; however, the rates of gas release from a hydrate accumulation would probably be similar to those associated with conventional natural gas fields. In addition, any modern gas hydrate field completion plan would attempt to monitor and remediate any potential accidental release of gas.

Because of uncertainties about the geologic settings and feasible production technology, few economic studies have been published on gas hydrates. The National Petroleum Council (NPC), in its major 1992 study of gas, published one of the few available economic assessments of gas hydrate production (Table I). This information assessed the relative economics of gas recovery from hydrates using thermal injection and depressurization. It also benchmarked the cost of gas hydrate production with the costs of conventional gas production on Alaska's North Slope. Since the 1992 NPC report, the cost of

TABLE I

Economic Study of Gas Hydrate Production

Economic factor ^a	Thermal injection ^b	Depressurization	Conventional gas
Investment ($\times 10^6$ \$)	5084	3320	3150
Annual cost ($\times 10^6$ \$)	3200	2510	2000
Total production (MMcf/year) ^c	900	1100	1100
Production cost (\$/Mcf)	3.60	2.28	1.82
Break-even wellhead price (\$/Mcf)	4.50	2.85	2.25

^a Mcf, thousand cubic feet; MMcf, million cubic feet.

^b Assumed process: injection of 30,000 barrels/day of water at 150°C.

^c Assumed reservoir properties: h (reservoir thickness) = 7.3 m, ϕ (reservoir porosity) = 40%, k (reservoir permeability) = 600 md.

conventional gas production has declined. However, within countries with considerable production of cheaper conventional natural gas, hydrates appear not to be an economically viable energy resource in a competitive energy market. However, in certain parts of the world with unique economic and/or political motivations, gas hydrates may become a critical sustainable source of natural gas within the foreseeable future.

SEE ALSO THE FOLLOWING ARTICLES

Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Oil and Natural Gas Exploration • Oil and Natural Gas: Offshore Operations • Oil and Natural Gas Resource Assessment: Geological Methods • Oil Sands and Heavy Oil • Oil Shale • Polar Regions, Impacts of Energy Development • Remote Sensing for Energy Resources

Further Reading

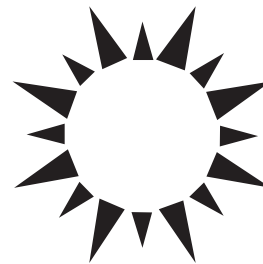
- Cherskiy, N. V., Tsarev, V. P., and Nikitin, S. P. (1985). Investigation and prediction of conditions of accumulation of gas resources in gas-hydrate pools. *Petrol. Geol.* **21**, 65–89.
- Collett, T. S. (1993). Natural gas hydrates of the Prudhoe Bay and Kuparuk River area, North Slope, Alaska. *Am. Assoc. Petrol. Geol. Bull.* **77**, 793–812.
- Collett, T. S. (1995). Gas hydrate resources of the United States. In "1995 National Assessment of United States Oil and Gas

- Resources" (D. L. Gautier, G. L. Dolton, K. I. Takahashi, and K. L. Varnes, Eds.). Available on CD-ROM from the U.S. Geological Survey, Digital Data Series 30.
- Dallimore, S. R., Uchida, T., and Collett, T. S. (1999). Scientific results from JAPEX/JNOC/GSC Mallik 2L-38 gas hydrate research well, Mackenzie Delta, Northwest Territories, Canada. *Geol. Surv. Can. Bull.* **544**, 1–403.
- Holbrook, W. S., Hoskins, H., Wood, W. T., Stephen, R. A., and the ODP Leg 164 Scientific Party. (1996). Methane hydrate, bottom-simulating reflectors, and gas bubbles: Results of vertical seismic profiles on the Blake Ridge. *Science* **273**, 1840–1843.
- Kvenvolden, K. A. (1988). Methane hydrate—A major reservoir of carbon in the shallow geosphere? *Chem. Geol.* **71**, 41–51.
- Kvenvolden, K. A. (1993). Gas hydrates as a potential energy resource—A review of their methane content. In "The Future of Energy Gases" (D. G. Howell, Ed.), pp. 555–561. U.S. Geological Survey Professional Paper 1570. USGS, Reston, Virginia.
- Kvenvolden, K. A., and Barnard, L. A. (1983). Hydrates of natural gas in continental margins. In "Studies in Continental Margin Geology: American Association of Petroleum Geologists, Memoir 34" (J. S. Watkins and C. L. Drake, Eds.), pp. 631–641. AAPG, Tulsa, Oklahoma.
- MacDonald, G. T. (1990). The future of methane as an energy resour. *Annu. Rev. Energy* **15**, 53–83.
- Makogon, Y. F. (1981). "Hydrates of Natural Gas." PennWell Publ. Company, Tulsa, Oklahoma.
- Max, M. D. (2000). "Natural Gas Hydrate in Oceanic and Permafrost Environments." Kluwer Academic Publ., London.
- Sloan, E. D. (1998). "Clathrate Hydrates of Natural Gases." 2nd Ed. Marcel Dekker, New York.

Gasoline Additives and Public Health

SERAP ERDAL

University of Illinois at Chicago
Chicago, Illinois, United States



-
1. Introduction to Gasoline Additives
 2. Human Health Effects of Methyl Tertiary-Butyl Ether (MTBE)
 3. Human Health Effects of Other Ether Compounds
 4. Human Health Effects of Ethanol
 5. Conclusions and Recommendations

1. INTRODUCTION TO GASOLINE ADDITIVES

Since 1979, oxygenates as gasoline additives have been used in limited areas of the United States as octane enhancers to replace lead at levels around 2 to 8% by volume. During the 1980s, oxygenates came into wider use as some states implemented oxygenated gasoline programs for the control of carbon monoxide air pollution in cold winter. Oxygenates were added to conventional gasoline nationally at higher percentages with the passage of the 1990 Clean Air Act (CAA), which required that oxygenates be added either seasonally (15% MTBE by volume) or year-round (11% MTBE by volume) to gasoline in specific parts of the country where carbon monoxide in the winter or concentrations of ozone in the summer exceed their respective National Ambient Air Quality Standards (NAAQS). Title II of the 1990 CAA addresses mobile sources and mandates the following two types of oxygenated fuel programs:

- Section 211 (m) of the CAAA of 1990 directed the EPA to mandate that states establish a winter-time oxygenated fuel program in 42 geographic areas designated as moderate or serious CO non-

attainment areas (i.e., areas with CO concentrations exceeding the CO NAAQS) by November 1, 1992. The oxygenated fuel program was mandated during the winter months of November through February based on the fact that gasoline combustion is less efficient at colder temperatures, resulting in increased CO emissions, and that pollutants stay closer to ground level in winter. During these months, the oxygen content requirement for gasoline sold in the designated areas is a minimum of 2.7 percent.

- Title II of the 1990 CAA under Section 211(k) specified that only reformulated gasoline (RFG) with a 2% oxygen weight requirement can be sold in metropolitan areas with populations more than 250,000, classified as extreme or severe for ozone pollution. These included Los Angeles, New York, Baltimore, Chicago, Milwaukee, Houston, San Diego, Sacramento, Philadelphia, and Hartford. Less-polluted nonattainment areas for ozone opted to implement the RFG program since its initiation in January 1995. During Phase I of the RFG program, refiners were required to reduce vehicle emissions of ozone-forming volatile organic compounds (VOCs) and air toxics by 15 to 17% annually, compared to 1990 baseline averages. The benzene and the aromatic content of RFG could not exceed 1.0% and 25% by volume, respectively. Nitrogen oxides (NO_x) emissions were not allowed to increase relative to conventional "baseline" gasoline. During Phase II, a 25 to 29% reduction in VOC was required, as well as a 20 to 22% reduction in air toxics and a 5 to 7% reduction in NO_x emissions from the 1990 baseline.

Among the available oxygenates in the marketplace, methyl tertiary butyl ether (MTBE) is the most widely used because of its low vapor pressure, compatibility with the blending and distribution system for gasoline, high octane rating, availability

and ease of production, and low cost. MTBE had the second highest volume of all chemicals produced in the United States in 1997.

On the other hand, ethanol is the predominant oxygenate used in gasoline in some areas, particularly in the Midwest. Other fuel oxygenates that are in use or may potentially be used to meet the 1990 CAA oxygenated fuel requirements include ethyl tertiary butyl ether (ETBE), tertiary-amyl methyl ether (TAME), diisopropyl ether (DIPE), tertiary-butyl alcohol (TBA), and methanol. This article assesses the human health effects of MTBE, ETBE, TAME, DIPE, and ethanol based on potential for exposure to these compounds as a result of using oxygenates in the U.S. gasoline supply. The priority is given to human studies in analyzing health effects information related to each of these compounds in gasoline.

In the 1992–1993 CO season, MTBE-blended gasoline accounted for approximately two-thirds of the demand, with ethanol providing the remainder. According to the EPA, ethanol is the predominant oxygenate used in the oxyfuel program (about 7.5% and 15% by volume for ethanol and MTBE, respectively) to meet the 2.7% by weight oxygen requirement. On the other hand, MTBE at 11% by volume is the primary oxygenate used in RFG to meet the 2% by weight oxygen requirement. The EPA's Office of Mobile Sources estimated in 1994 that approximately 30 million people live in areas using wintertime oxygenated gasolines and about 45 million live in high ozone RFG areas. Later estimates by SRA Technologies suggest that 70 and 57 million people live in areas using oxyfuel and RFG, respectively. It is also reported that about 109 million Americans live in counties where MTBE is used as oxygenate in gasoline. There is some overlap between the areas implementing the winter oxyfuel program and the areas implementing the year-round RFG program. Estimated oxygenated volume consumed in RFG areas in the United States is 237 (MTBE), 0.3 (ETBE), 14.6 (TAME), and 24.2 (ethanol) thousand barrels per calendar day. These estimates suggest that a large fraction of the U.S. population may get exposed to oxygenates on a daily basis.

2. HUMAN HEALTH EFFECTS OF METHYL TERTIARY-BUTYL ETHER (MTBE)

Acute health symptoms attributed to methyl tertiary butyl ether (MTBE) were reported in various parts of

the United States soon after the addition of oxygenated compounds to gasoline during the winter of 1992–1993. Such complaints were not anticipated but have subsequently focused attention on possible health risks due to oxygenated gasoline. The health complaints were accompanied by complaints about reduced fuel economy and engine performance and by controversial cancer findings in long-term animal studies.

2.1 Epidemiological Evidence for Air Exposures

Following the introduction of MTBE-oxygenated gasoline in Fairbanks, Alaska, in November 1992, about 200 residents reported headaches, dizziness, irritated eyes, burning of the nose and throat, coughing, disorientation, nausea, and other acute health symptoms. The Alaska Department of Health and Social Services and the Center for Disease Control (CDC) conducted a number of field epidemiology studies to investigate the potential relationship between MTBE oxyfuel exposure and acute health symptom responses. A two-phase study was performed in Fairbanks, with Phase I occurring while MTBE-oxyfuel was in use and Phase II occurring 6 weeks after MTBE-oxyfuel use had been terminated. In this study, the relationship between MTBE oxyfuel exposure and blood levels of MTBE and a major metabolite of MTBE, tertiary butyl alcohol (TBA), was also investigated. The occupational and non-occupational exposure groups answered a questionnaire inquiring about the presence of certain symptoms, including the seven key complaints of headache, eye irritation, burning of nose/throat, cough, nausea, dizziness, and spaciness. Occupationally exposed individuals reported a greater prevalence of health complaints during Phase I compared to Phase II. Overall, the CDC's analysis of Fairbanks data showed that there was a statistically significant difference between the number of cases among the individuals who were exposed to higher concentration of MTBE (e.g., taxi drivers) compared to people exposed to lower exposure levels (e.g., students). The CDC cautioned that the results did not provide definitive evidence that the use of MTBE oxyfuels caused the reported symptoms due to potential confounding factors (e.g., negative publicity about MTBE-oxyfuel effects, small sample size, strong odor associated with MTBE, and lack of controls) and called for further research to determine whether there actually is an increase in acute health symptoms and, if so, whether they are causally associated

with MTBE-oxyfuel. However, the governor of Alaska suspended the program in Fairbanks within 2 months of the introduction of oxyfuel and a switch from MTBE-oxyfuel to ethanol-oxyfuel occurred in following years.

A desire to conduct a study similar to the Fairbanks study in the absence of confounding factors observed in Alaska (especially negative publicity) led to a field study in another part of the country. The CDC compared the prevalence of acute health symptom reports in Stamford, Connecticut, in April 1993 (where MTBE-oxyfuel was used) and Albany, New York, in May 1993 (with no oxyfuels). There were no known consumer health complaints in either location. For the Stamford study, CDC in cooperation with the State of Connecticut Health Department selected 221 individuals from four groups of workers and administered a questionnaire similar to that used in Fairbanks. The CDC reported that the 11 individuals with the highest blood MTBE levels were statistically significantly more likely to report one or more key symptoms than the other 33 people with lower blood levels. However, there was no apparent relationship between symptom prevalence by occupational category and the median blood MTBE or the TBA concentrations associated with each category. CDC again cautioned that the small numbers of subjects studied, the wide range of blood MTBE and TBA levels within each category, and the wide variability of response prevented a definitive conclusion. For the Albany study, 264 individuals were nonrandomly selected and categorized into three groups presumed to have tiered levels of exposure. Overall, the results indicated that the Albany residents with no MTBE-oxyfuel exposure had symptom prevalence as high as those in Stamford residents with MTBE-oxyfuel exposure. It was suggested that this outcome could potentially be attributed to the high prevalence of symptoms due to allergies and viruses at the time of the study in Albany.

In New Jersey, the EOHSI evaluated 237 state-employed garage workers from two groups in 1993: (1) 115 northern New Jersey workers sampled during the wintertime oxyfuel program and (2) 122 southern New Jersey workers sampled 10 weeks after the phase-out of the program in that area. All participants were asked to complete questionnaires about symptoms and their severity over the previous 30 days and about preshift and postshift symptoms. Essentially no differences in symptom reports were found between the northern (high-exposure) and southern (low-exposure) groups. In a separate study,

11 workers who pumped MTBE-oxyfuel for more than 5 hours per day were matched with 11 workers from the south who had low or nondetectable exposures to MTBE. No differences in symptom prevalence were found between the groups.

Another EOHSI study attempted to identify a sensitive subpopulation for the effects of MTBE-oxyfuel exposure. This involved 13 subjects with multiple chemical sensitivities (MCS), five subjects with chronic fatigue syndrome (CFS), and six healthy controls. Through a symptom questionnaire, these subjects were asked to assess symptoms associated with situations in which they could be exposed to MTBE such as refueling and driving. On a total symptom score range of 0 (no discomfort) to 28 (severe discomfort), the MCS subjects scored 5.3, compared to 4.8 for the CFS subjects and 1.2 for the healthy subjects. It was not possible to draw any definitive conclusions from this study due to (1) lack of evidence to support that an unusually high rate of symptoms or an increase of symptoms was occurring uniquely where MTBE was most prevalent, (2) small sample sizes, and (3) inability to distinguish between those subjects using and not using MTBE oxyfuel.

In response to variety of health concerns of Milwaukee citizens after the beginning of the RFG program in January 1995, the state of Wisconsin conducted a random-digit-dial telephone survey of 500 residents of the Milwaukee metropolitan area and equal numbers of residents in Chicago and in non-RFG areas of Wisconsin. In Milwaukee, oxygenates in the RFG samples were 49% MTBE, 38% ethanol, and 14% ETBE. Symptom prevalence was significantly higher in Milwaukee than in either Chicago or elsewhere in Wisconsin (23% compared to 6%). In Milwaukee, persons were more likely to report symptoms if they had experienced a cold or flu, smoked cigarettes, or were aware that they had purchased RFG. Because overall symptom prevalence was essentially equivalent in Chicago and non-RFG areas of Wisconsin, the authors concluded that factors other than RFG use such as an increase in price of gasoline, odors of MTBE, and negative media coverage significantly contributed to the differences in symptom prevalence between Milwaukee and the other two areas. However, due to study design limitations, the study could not "rule out subtle effects of RFG exposure, or the possibility that a relatively small number of individuals may have a greater sensitivity to RFG mixtures."

Hakkola and his coworkers investigated neurological effects reported by 101 Finnish male tank drivers with exposure to gasoline containing 10%

MTBE. Drivers with longer exposure scored higher for fatigue and hostility changes as compared to drivers with shorter exposure. In addition, 20% of drivers reported the following symptoms in the order of decreasing frequency: headache, nausea, increased salivation, dizziness, and dyspnea.

2.2 Controlled Human Inhalation Exposure Chamber Studies

Assessment of humans exposed directly in a chamber to MTBE was hoped to answer whether there was a direct causal relation between MTBE exposure and acute health symptoms, particularly in view of the inconclusive evidence gathered from the field epidemiology studies. In an EPA investigation, 37 healthy nonsmoking individuals were exposed for 1 h to both clean air and 1.4 ppm (5 mg/m³) "pure" MTBE in air at 75° F and 40% humidity. The endpoints selected, based on observations from Alaska, were the types of symptoms associated with organic solvent exposure: (1) nonspecific symptoms including headache, nasal irritation, throat irritation, cough, eye irritation, odor quality, and dizziness; (2) measures of behavioral response (neurobehavioral evaluation system test battery); (3) indicators of upper airways inflammation (nasal lavage); and (4) indicators of eye irritation. The study found that MTBE exposure had no statistically significant effects on any of the endpoints. Yale University researchers replicated the EPA study, with few minor modifications—that is, 43 subjects, slightly higher MTBE concentration of 1.7 ppm (6 mg/m³), and an additional experiment involving a 1-h exposure to a complex mixture of 16 VOCs commonly found in gasoline at a concentration of 7.1 ppm without MTBE. Similar to the EPA study, it was found that MTBE exposure had no statistically significant effects on any of the endpoints studied as compared to clean air exposure. The VOC exposure caused an increase in inflammatory cells in the nasal lavage on the day following exposure when the VOC exposure compared to clean air exposure. Taken together, the EPA and Yale investigations showed that controlled exposure of a small number of healthy humans to MTBE in air under the conditions studied did not cause increased symptoms or measurable responses. Exposure concentrations utilized in these controlled exposure studies (1.4 and 1.7 ppm) did not represent the worst-case exposure conditions based on measurements from the customer breathing zone samples taken in air during refueling (median exposure range: 0.3 to 6 ppm, occasionally higher). Nor were subjects

chosen who were seemingly susceptible to MTBE based on their complaints.

Another controlled exposure chamber study exposed normal subjects to much higher concentrations of MTBE (5, 15, and 25 ppm compared to 1.4 and 1.7 ppm) and a longer exposure duration (2 hours) than the EPA or the Yale study (1 hour). This study did not include neurobehavioral assessments. No effects associated with MTBE exposure were identified. In a separate study, 13 healthy male subjects were exposed to 25 and 75 ppm MTBE in air in an exposure chamber. Subjects with 75 ppm exposure developed irritation of mucous membranes and nonspecific symptoms indicative of effects on mood.

A controlled exposure study was conducted at the EOHSI to investigate the symptoms, psychophysiologic and neurobehavioral effects of exposure to the MTBE. The exposures were conducted at concentrations similar to those found while refueling a vehicle using both control subjects and those individuals self-described as sensitive to MTBE. In the double-blinded repeated-measures controlled exposure study, subjects were exposed for 15 min each to clean air, gasoline alone, gasoline with 11% MTBE, and gasoline with 15% MTBE. Relative to the control group ($n = 19$), sensitive individuals ($n = 12$) reported significantly more total symptoms only when exposed to gasoline with 15% MTBE and not to gasoline with 11% MTBE or clean air. However, the sensitive group did not exhibit significant differences in psychophysiologic responses from the control group at any exposure level. This study did not identify a dose-response relationship for MTBE exposure and symptoms were specific to MTBE. The EOHSI research group also studies the driving habits and vehicle maintenance patterns of these two study groups. Results showed that the individuals who had self-reported heightened sensitivity to oxygenated gasoline vapors drove their vehicles more often and fueled their vehicles more frequently than the control group. In addition, sensitive individuals more frequently drove their vehicles with compromised body integrity that could lead to higher exposures.

2.3 Health Effects Associated with Oral Exposure to MTBE

In the mid-1990s, the focus on MTBE began to shift from questions about air quality-related health risks and benefits to questions about water quality risks. The change was triggered by a number of

factors including the USGS studies demonstrating contamination of the nation's water supply with MTBE, and discovery of underground storage tank (UST) leaks resulting in high concentrations of MTBE in water causing the closure of drinking water wells in Santa Monica, South Tahoe, and San Jose, California. The emergence of widespread reports of MTBE-impacted drinking water resources in other states (e.g., New Jersey, Maine, Vermont, New York, Virginia) heightened interest in protection of water resources from MTBE. The presence of MTBE in ground water and in reservoirs used for drinking water in other parts of the country is now relatively widespread.

Faced with an acute potential public health problem and degradation of water quality (e.g., strong odor and taste associated with MTBE) resulting in loss of water supplies for consumer consumption, the California legislature enacted a number of bills calling to address issues pertaining to oxygenates including risks and benefits to human health and the environment of MTBE and its combustion byproducts in air, water, and soil and alternate oxygenates like ethanol and ETBE. In response to these legislative bills, the Office of Environmental Health Hazard Assessment (OEHH) of the California Environmental Protection Agency proposed a Public Health Goal (PHG) of 14 ppb for MTBE in drinking water in June 1998 to protect consumers against health risks over a lifetime of exposure. The proposed PHG for MTBE was based on carcinogenic potency of 1.7×10^{-3} (mg/kg-day)⁻¹, which is the geometric mean of the cancer slope factor estimates obtained from three cancer sites: combined male rat kidney adenomas and carcinomas in the 1992 inhalation study of Chun *et al.*, the male rat Leydig cell tumors and the leukemia, and lymphomas in female rats in the 1995 oral study by Belpoggi *et al.* In July 1998, the Department of Health Services (DHS) adopted a secondary maximum contaminant level (SMCL) for MTBE of 5 ppb based on available data on odor and taste thresholds. Another product was a University of California (UC) report on "Health and Environmental Effects of MTBE," submitted to the governor and legislature of the state of California on November 12, 1998. The UC report states "For the general population, the risk of exposure to MTBE through ingestion of MTBE-contaminated water is currently low." However, because MTBE is highly soluble in water, poorly adsorbs to soil, readily transfers to groundwater from LUSTs, and pipelines, travels much further than plumes commonly associated

with gasoline spills, and is highly resistant to biodegradation, the report concluded that "There are significant risks and costs associated with water contamination due to the use of MTBE." Most importantly, the UC scientists recommended phasing out MTBE over an interval of several years instead of an immediate ban. Based on significant risk to California's environment associated with the continued use of MTBE in gasoline, Governor Davis, on March 25, 1999, directed appropriate state regulatory agencies to devise and carry out a plan to begin immediate phase out of MTBE from California gasoline, with 100% removal achieved no later than December 31, 2002.

Faced with another MTBE problem, in 1996, the EPA began working on revising the 1992 drinking water lifetime health advisory for MTBE of 20–200 ppb to meet the needs of states that were struggling to endorse safe levels that ensure both consumer acceptability of MTBE-laden drinking water and public health protection against potential adverse health effects of ingested MTBE. Meanwhile, in the absence of enforceable federal drinking water standards, individual states began to adopt action levels or cleanup levels for MTBE in water. In December 1997, the EPA's Office of Water published "Drinking Water Advisory: Consumer Acceptability Advise and Health Effects Analysis on Methyl Tertiary-Butyl Ether (MTBE)." The advisory recommended that keeping MTBE levels in the range of 20 to 40 ppb or below to protect consumer acceptance of the water resource would also provide a large margin of safety from potential adverse health effects. The EPA could not develop a RfD for MTBE due to the persisting limitations of health effects data (e.g., lack of a pharmacokinetic model for converting inhalation exposure data for MTBE to oral exposure, limited reported pathology and historical tumor incidence, no toxicity studies with oral exposure to water). A blue-ribbon panel convened by the United States Environmental Protection Agency (EPA) recommended that Congress should act to provide clear federal and state authority to regulate or eliminate the use of MTBE and other gasoline additives that threaten drinking water supplies.

2.4 Carcinogenic Effects of MTBE

There are no human data on which carcinogenicity of MTBE can be evaluated. However, the evidence from chronic exposure by inhalation and oral exposure routes demonstrate that MTBE is

carcinogenic in rats and mice. Two chronic animal cancer bioassays with inhalation exposure to MTBE resulted in rare kidney tumors and testicular tumors in the mid- and high-dose groups of male rats, and liver tumors in mice at the high-dose group. Sprague-Dawley rats exposed to MTBE in olive oil as the vehicle for oral gavage for 2 years showed dose-related increases in lymphomas and leukemias in females and in interstitial cell tumors of the testes in males. The biological mechanism responsible for MTBE cancer production in animals is not known. The lack of mechanistic information limits our ability to determine whether the MTBE bioassay results are relevant to humans. In 1998, IARC found insufficient evidence to establish that MTBE poses cancer risks to humans. The EPA has also concluded that potential for MTBE's carcinogenicity for humans is of concern and requires further study.

3. HUMAN HEALTH EFFECTS OF OTHER ETHER COMPOUNDS

The other aliphatic ethers, ethyl tertiary butyl ether (ETBE), tertiary amyl methyl ether (TAME), and diisopropyl ether (DIPE) are proposed as potential alternatives to MTBE as an oxygenate in gasoline. In particular, TAME was considered to be an attractive alternative due to its lower vapor pressure as compared to MTBE despite its lower octane content. These compounds are being used in varying percentages in gasoline in the United States (e.g., 19.90% TAME in Sussex County, Delaware, 1.6% TAME in San Diego, California, in 1999).

Although a number of toxicity studies in animal models for other ether oxygenates of ethyl tertiary butyl ether (ETBE), tertiary amyl methyl ether (TAME), and diisopropyl ether have been conducted and published, human relevance of these findings is not known. There appears to be qualitative similarities to the MTBE health effects, with kidney and liver being the most sensitive target organs. Bio-transformation and pharmacokinetics of these compounds in animals and human volunteers have been studied except DIPE. A summary of animal studies and pharmacokinetic studies in animals and humans can be found in Ahmed's 2001 research.

The EPA has not been able to establish any quantitative cancer potency factor or noncancer toxicity values (i.e., reference dose [RfD] or reference concentration [RfC]) to date due to inability to establish dose-response relationships from the animal

data in a scientifically defensible manner. To fill this data gap, the EPA mandated the manufacturers to perform toxicity testing of evaporative emissions of these compounds under the CAA Section 211(b) by 2004. The EPA also called for "comprehensive evaluation of multimedia, life-cycle environmental impacts" of different fuels to assess viability of TAME, ETBE, and DIPE to be a safe replacement for ethanol or MTBE. A number of scientific bodies convened to evaluate health effects of oxygenated gasolines more than two decades also repeatedly called for more research into understanding the health effects of these compounds alone and in combination with gasoline. The California DHS proposed to add ETBE and TAME to the list of unregulated chemicals for which monitoring is required by public water systems.

4. HUMAN HEALTH EFFECTS OF ETHANOL

Ethanol used as oxygenate in gasoline is produced from grains or other renewable agricultural and forestry feedstocks. The U.S. Congress has approved a 5.4 cents/gallon federal subsidy for this biomass-based source of fuel in nations' gasoline supply. Ethanol production in 2000 was 1.5 billion gallons. Approximately 30% of this amount is used in RFG during the summer in ozone nonattainment areas. About 20% is used in the winter oxygenate program to control CO air pollution. The remainder 50% is used in conventional gasoline in the United States to enhance octane and extend fuel supplies. Ethanol is highly volatile when spilled; it will evaporate five times faster than MTBE. However, it is commonly used in many modern day products including pharmaceuticals, mouthwash products, alcoholic beverages, and cleaning products. Because humans frequently ingest fermented beverages containing 12% ethanol by volume, ingestion-related health effects of ethanol are well known. The Health Effects Institute (HEI) in its assessment of health effects of oxygenates stated that our knowledge on health effects of ethanol emanates from ingestion exposures. Ethanol was shown to be associated with increased rates of tumors of digestive track with oral exposure to moderate levels of alcohol when other carcinogens were present. Ethanol is classified as a Class I carcinogen by IARC based on human data for alcoholic beverage consumption. There is no data of such association for inhalation exposures. However,

ethanol, when blended in gasoline increases volatility of other compounds, results in an increase in evaporative emissions of VOCs, including benzene. Many of these VOCs are known or suspected carcinogens and also have adverse health effects associated with noncancer health endpoints. Ethanol exposure is also associated with cancers in the liver, esophagus, oral cavity, pharynx, and larynx. There have been reports of adverse reproductive and developmental effects of associated with exposure to ethanol in animal and human studies. Irvine analyzed the relevance of the developmental toxicity of ethanol in occupational settings and concluded that while developmental toxicity may result from drinking alcoholic beverages for which threshold levels have yet to be determined there was no evidence for developmental toxicity in occupational environment.

With the executive order to phase out of MTBE in gasoline distributed in California by December 31, 2002, Governor Davis also directed CARB and the State Water Resources Board to conduct an environmental fate and transport analysis of ethanol in air, surface water, and groundwater and the OEHHA to prepare an analysis of health risks associated with exposure to ethanol in gasoline. In terms of air quality effects, the substitution of ethanol for MTBE in CA's fuel supply was found to have not significantly impact air quality as long as RFG regulation in CA addressed the potential for ethanol to increase evaporative emissions and to cause more rail and truck traffic to bring ethanol to the state. The CA analysis found no substantial differences in the air-related public health impacts of different non-MTBE fuel formulations including ethanol. Although the replacement of MTBE by ethanol is expected to have some benefits in terms of water contamination due to ethanol's relative toxicity and rapid biodegradation, these could not be quantified. The overall estimated health risks for exposure scenarios with ethanol, MTBE, and nonoxygenated fuel did not differ between fuel formulations, but showed a significant improvement for 2003 in comparison to 1997.

The Northeast States for Coordinated Air Use Management (NESCAUM) also prepared an analysis in response to potential phase-out of MTBE in Northeast States due to adverse water quality impacts of MTBE and the need to protect environmental resources. The NESCAUM estimated that the use of ethanol instead of MTBE would increase emission rates of acetaldehyde, a combustion by-product, into the atmosphere in the Northeast by 50 to 70%. Acetaldehyde is also a probable

human carcinogen. Ambient levels of acetaldehyde in the Northeast exceed the health-protective levels of acetaldehyde at majority of the air monitoring locations throughout the northeastern states. NESCAUM estimates that ethanol at 10% by volume in conventional gasoline would increase the VOC emissions significantly. Therefore, NESCAUM stressed the importance of studying health effects of ethanol alone, in combination with gasoline, but importantly, its lifecycle, its degradation products, and concurrent VOC emission increases. Schifter *et al.* also estimated an increase of about 80% in acetaldehyde emissions with 6% by volume ethanol as a substitution for 5% by volume MTBE accompanied by 16% decrease in CO and a 28% decrease in formaldehyde motor vehicle emissions in Mexico City. However, according to the CARB analysis on impact of use of ethanol-containing gasoline on air quality in Denver, Colorado, Albuquerque, New Mexico, Brazil, and other areas, acetaldehyde levels were substantial only in Brazil, where the fuels contained either pure ethanol or 22% ethanol, much greater than the maximum 10% ethanol allowed in CA gasoline.

There are no EPA toxicity values established for cancer (cancer slope factor) or noncancer (RfD or RfC) health endpoints for ethanol. There is also no safe drinking water standard for ethanol. The lack of toxicity values for ethanol for various exposure routes and lack of exposure information in various exposure media do not enable us to quantitatively estimate health risks associated with exposure to ethanol under different exposure conditions for public exposed to ethanol due to its circulation in the environment related to its widespread use in gasoline. Such analysis should not only include ethanol, but also all environmental degradation products of ethanol, and health risks and benefits analysis focusing on multimedia exposures is essential for not repeating the history of MTBE episode.

5. CONCLUSIONS AND RECOMMENDATIONS

There have been significant air quality benefits since the implementation of the oxygenated fuels program. Although some of the improvement in air quality indices may be simply due to fleet turnover, increase in fuel efficiency over time, and improvements in vehicle testing and maintenance programs, the oxygenates have also been credited for air quality

benefits. For example, 10 to 14% reduction in ambient CO concentrations has been attributed to use of oxygenated in winter months in CO non-attainment areas. Although the National Research Council concluded that little air quality impact in terms of ozone reduction can be gained from the use of oxygenates in RFG, the EPA's blue ribbon panel acknowledged major reductions in VOCs, especially, some of the carcinogenic air toxics (e.g., benzene, 1,3-butadiene) in many urban environments where RFG is used during the summer months. However, health complaints associated with exposure to MTBE containing gasoline and discovery of degradation in water quality in national scale with contamination of groundwater, surface water, and drinking water resources with MTBE overshadowed the potential air quality benefits and associated air-quality-related health benefits in the decade since the enactment of the oxygenated fuel programs by the Congress. The course of the national debate on oxygenates has shifted to potential alternatives to MTBE. We know more about the potential human health effects associated with exposure to MTBE alone or in combination with gasoline than about its potential substitutes. Quantitative assessment of the public health inhalation risks of MTBE-oxyfuel in comparison to conventional gasoline has been difficult, in part due to uncertainties and data deficiencies in toxicity combined with the paucity of the population-specific human exposure data. Although both short-term field epidemiological studies and a few controlled human laboratory studies of healthy adults provide a better understanding of human exposures to MTBE and MTBE-oxyfuel, the large differences in terms of confounding parameters, study design, exposure conditions, and exposure parameters greatly complicate the interpretation of the results of these few studies. Although more animal toxicity data for MTBE have become available and have increased our scientific understanding of the toxicity of MTBE (especially the pharmacokinetics of MTBE), there are insufficient data to answer questions about the relevance of these animal data to humans. Additionally, in the absence of oral toxicity studies, health effects of ingested MTBE cannot be determined. Although neither the epidemiological studies nor the controlled human exposure studies of normal and sensitive subjects have been, in majority, inconclusive, they cannot rule out the possibility that for a small percentage of the population, those claiming that they were affected by MTBE, there is a causal relationship between their exposure and their symptoms. In conclusion, the

human health effects of MTBE are still unclear. For some other oxygenates, animal testing has not been thoroughly conducted and published yet. Any alternatives to MTBE, including TAME, ETBE, DIPE, and ethanol, should be studied with well-designed experiments to assess potential environmental and public health implications of using these compounds in large scale. This requires a substantial commitment and well-coordinated research effort by government and industry. The EPA scientists have evaluated retrospectively the governmental decision making with MTBE and concluded that

The most valuable lesson to be gained from the MTBE experience may be the importance of obtaining comprehensive understanding of the many and varied risks and benefits potentially offered by any given fuel option. This should entail a consideration of all facets of the life cycle of a fuel or additive, including the environmental fate of any emissions or releases, because the by-products of a chemical may be of greater toxicological or environmental significance than the parent compound itself.

Given the fact that ethanol has increasingly gained oxygenate market share in the absence of such comprehensive evaluation should remind the EPA and the scientific community about the importance of critical findings gleaned through the MTBE experience.

SEE ALSO THE FOLLOWING ARTICLES

Air Pollution from Energy Production and Use • Air Pollution, Health Effects of • Clean Air Markets • Climate Change and Public Health: Emerging Infectious Diseases • Consumption, Energy, and the Environment • Ethanol Fuel • Hazardous Waste from Fossil Fuels • Motor Vehicle Use, Social Costs of • National Energy Policy: United States • Suburbanization and Energy

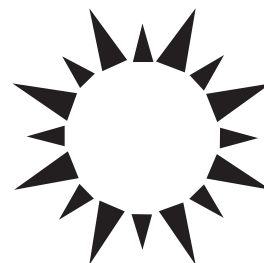
Further Reading

- Adler, K., Schaffer, S., and Peebles, J. E. (1993). Oxygenated fuel season starts smoothly. *Fuel Reformulation* 62, November/December.
- Ahmed, F. E. (2001). Toxicology and human health effects following exposure to oxygenated or reformulated gasoline. *Toxicol. Lett.* 123, 89–113.
- Anderson, H. A., Hanrahan, L., Goldring, J., and Delaney, B. (1995). An investigation of health concerns attributed to reformulated gasoline use in Southeastern Wisconsin. Wisconsin Department of Health and Social Services, Division of Health, Bureau of Public Health, Section of Environmental Epidemiology and Prevention. Madison, WI.

- Beller, M., Schloss, M., and Middaugh, J. (1992). "Potential illness due to exposure to oxygenated fuel in Fairbanks, AK." State of Alaska, Department of Health and Social Services, Anchorage, AK.
- Belpoggi, E., Soffritti, M., and Maltoni, C. (1995). Methyl-Tertiary-Butyl Ether (MTBE)—A gasoline additive—causes testicular and lympho-hematopoietic cancers in rats. *Toxicol. Industrial Health* 11(2), 119–149.
- Cain, W. S., Leaderer, B. P., Ginsberg, G. L., Andrews, L. S., Cometto-Muñiz, J. E., *et al.* (1996). Acute exposure to low-level methyl tertiary-butyl ether (MTBE): Human reactions and pharmacokinetic responses. *Inhalation Toxicol.* 8, 21–48.
- Center for Disease Control and Prevention (CDC) (1993). An investigation of exposure to methyl tertiary butyl ether in oxygenated fuel in Fairbanks, Alaska. National Center for Environmental Health, Atlanta, GA.
- Center for Disease Control (CDC) (1993). An investigation of exposure to methyl tertiary butyl ether among motorists and exposed workers in Stamford, Connecticut. U.S. Department of Health and Human Services, National Center for Environmental Health, Atlanta, GA.
- Center for Disease Control (CDC) (1993). An investigation of exposure to MTBE and gasoline among motorists and exposed workers in Albany, New York. U.S. Department of Health and Human Services, National Center for Environmental Health, Atlanta, GA.
- Chandler, B., and Middaugh, J. (1992). "Potential illness due to exposure to oxygenated fuel in Anchorage, Alaska." State of Alaska, Department of Health and Social Services, Bull. No. 1, Anchorage, AK.
- Chun, J. S., Burleigh-Flayer, H. D., Kintigh, W. J. (1992). Methyl Tertiary Butyl Ether: Vapor inhalation oncogenicity study in Fischer 344 rats. BRRC Report 91N0013B. Union Carbide. Bushy Run Research Center. Export, PA.
- Davis, J. M. (2002). Fuel oxygenates: How do their health effects compare? Presented at the ASTSWMO State Symposium on Fuel Oxygenates: Analysis, Assessment, and Remediation. October 21. Arlington, VA.
- Davis, J. M., and Farland, W. H. (2001). The paradoxes of MTBE. *Toxicolog. Sci.* 61, 211–217.
- Environmental Protection Agency (EPA) (1992). Methyl-t-Butyl Ether. Drinking Water Health Advisory. Health and Ecological Criteria Division. Office of Science and Technology. Office of Water, Washington, DC.
- Environmental Protection Agency (EPA) (1993). Assessment of potential health risks of gasoline oxygenated with methyl tertiary butyl ether EPA/600/R-93/206. Office of Research and Development, Washington, DC.
- Environmental Protection Agency (EPA) 1994. Health risk perspectives on fuel oxygenates. Report No. EPA 600/R-94/217. Ann Arbor, Michigan, Office of Research and Development.
- Environmental Protection Agency (EPA) 1994. National air quality and emissions trends report, 1993. EPA 545/R-94-026. Washington, DC.
- Environmental Protection Agency (EPA) (1995). Proceedings of the Conference on MTBE and Other Oxygenates: A Research Update, July 26–28, 1993, Fairview Park Marriott, Falls Church, Virginia. EPA/600/r-95/134. U.S. Environmental Protection Agency. National Center for Environmental Assessment, Research Triangle Park, NC.
- Environmental Protection Agency (EPA) (1996). Oxyfuel research needs. Report No. EPA 600/R-96/069. Office of Research and Development, Washington, DC.
- Environmental Protection Agency (EPA) (1997). Drinking water advisory: consumer acceptability advice and health effects analysis on methyl tertiary butyl ether (MTBE) EPA-822-F-97-009. Health and Ecological Criteria Division, Office of Science and Technology, Office of Water, Washington, DC.
- Environmental Protection Agency (EPA) 1999. Achieving Clean Air and Clean Water: The Report of the Blue Ribbon Panel on Oxygenates in Gasoline. EPA420-R-99-021. September 15.
- Environmental Protection Agency (EPA) 1999. SB 521 Research Project Review. Executive Summary, Review of Volumes I–IV. "Health and Environmental Assessment of MTBE." Office of Mobile Sources, Office of Air and Radiation, Washington, DC.
- Environmental Protection Agency (EPA) 2000. Control of MTBE in gasoline. Office of Transportation and Air Quality. March
- Erdal, S., and Goldstein, B. D. (2000). Methyl tert-butyl ether as a gasoline oxygenate: Lessons for environmental public policy. *Annu. Rev. Energy Environ.* 25, 765–802.
- Federal Register (1998). National Toxicology Program; Call for Public Comments on 11 Agents, Substances, Mixtures and Exposure Circumstances Proposed for Listing in or Delisting from the Report on Carcinogens, 9th ed. 63(239), 68783–68785.
- Fielder, N., Kelly-McNeil, K., Mohr, S., Lehrer, P., Opiekun, R. E., *et al.* (2000). Controlled human exposure to methyl tertiary butyl ether in gasoline: Symptoms, psychophysiologic and neurobehavioral responses of self-reported sensitive populations. *Environ. Health Persp.* 108, 753–763.
- Fielder, N., Mohr, S. N., Kelly-McNeil, K., and Kipen, H. M. (1994). Response of sensitive groups to MTBE. *Inhalation Toxicol.* 6, 539–552.
- Hakkola, M., Honkasalo, M. L., and Pulkkinen, P. (1997). Changes in neuropsychological symptoms and moods among tanker drivers exposed to gasoline during a work week. *Occupational Med.* 47(6), 344–348.
- Irvine, L. F. H. (2003). Relevance of the developmental toxicity of ethanol in the occupational setting: A review. *J. Applied Toxicol.* 23, 289–299.
- Johanson, G., Hihlen, A., and Lof, A. (1995). Toxicokinetics and acute effects of MTBE and ETBE in male volunteers. *Toxicol. Lett.* 82/83, 713–718.
- Lidderdale, T. (1999). Oxygenate demand in reformulated and oxygenated gasoline control areas. Report. Department of Energy. Energy Information Administration. August 6.
- Lioy, P. J., Weisel, C. P., Jo, W., Pellizzari, E., and Raymer, J. H. (1994). Microenvironmental and personal measurements of methyl-tertiary butyl ether (MTBE) associated with automobile use activities. *J. Exposure Anal. Environ. Epidemiol.* 4, 427–441.
- Mohr, S. N., Fielder, N., Weisel, C., and Kelly-McNeil, K. (1994). Health effects of MTBE among new jersey garage workers. *Inhalation Toxicol.* 6, 553–562.
- Moolenaar, R. L., Hefflin, B. J., Ashley, D. L., Middaugh, J. P., and Etzel, R. A. (1994). Methyl tertiary butyl ether in human blood after exposure to oxygenated fuel in Fairbanks, Alaska. *Arch. Environmental Health* 49(5), 402–409.
- National Research Council (NRC) (1996). Toxicological and Performance Aspects of Oxygenated Motor Vehicle Fuels. National Academy Press, Washington, DC.
- National Research Council (NRC) (1999). Ozone-Forming Potential of Reformulated Gasoline. Committee on Ozone-Forming Potential of Reformulated Gasoline. Commission on Geosciences, Environment, and Resources, National Academy Press, Washington, DC.
- National Science and Technology Council (1997). Interagency Assessment of Oxygenated Fuels. National Science and

- Technology Council, Committee on Environment and Natural Resources (CENR) and Interagency Oxygenated Fuels Assessment Steering Committee. White House Office of Science and Technology Policy (OSTP) through the CENR of the Executive Office of the President, Washington, DC.
- Newman, A. (1995). MTBE detected in survey of urban groundwater. *Environ. Sci. Tech.* 29(7), 305A.
- Office of Science and Technology Policy (OSTP) (1996). Air Quality Benefits of the Winter Oxyfuel Program. Prepared for Interagency Oxygenated Fuel Assessment. Coordinated by Office of Science and Technology Policy, Executive Office of the President, Washington, DC.
- Opiekun, R. E., Freeman, N. C. G., Kelly-McNeil, K., Fiedler, N. L., and Lioy, P. J. (2001). Effect of vehicle use and maintenance patterns of a self-described group of sensitive individuals and nonsensitive individuals to methyl tertiary-butyl ether. *J. Exposure Anal. Environ. Epidemiol.* 11, 79–85.
- Peebles, J. E. (1994). EPA finalizes reformulated gasoline regulations—Finally! *Fuel Reformulation*, March/April.
- Prah, J. D., Goldstein, G. M., Devlin, R., Otto, D., Ashley, D., *et al.* (1994). Sensory, symptomatic, inflammatory and ocular responses to and the metabolism of methyl tertiary butyl ether in a controlled human exposure experiment. *Inhalation Toxicol.* 6, 521–538.
- Rihimäki, V., Matikainen, E., Akila, R., Teravainen, E., Mutanen, P., *et al.* (1998). “Central Nervous System Effects of the Gasoline Additive Methyl-Tert-Butyl Ether (MTBE).” Finnish Institute of Occupational Health, Finland.
- Schifter, I., Vera, M., Diaz, L., Guzman, E., Ramos, F., and Lopez-Salinas, E. (2001). Environmental implications on the oxygenation of gasoline with ethanol in the metropolitan area of Mexico City. *Environ. Sci. Technol.* 35, 1893–1901.
- Squillance, P. J., Pope, D. A., and Price, C. V. (1995). “Occurrence of the Gasoline Additive MTBE in Shallow Ground Water in Urban and Agricultural Areas. U.S. Geological Survey. Fact Sheet FS-114–95.” U.S. Geol. Surv, Rapid City, SD.
- UCRL (1999). Health and Environmental Assessment of the Use of Ethanol as a Fuel Oxygenate. Prepared by the California Air Resources Board, the State Water Resources Control Board, and the California Environmental Protection Agency’s Office of Environmental Health Hazard Assessment December 30.
- White, M. C., Johnson, C. A., Ashley, D. L., Buchta, T. M., and Pelletier, D. J. (1995). Exposure to methyl tertiary butyl ether from oxygenated gasoline in Stamford, Connecticut. *Arch. Environ. Health.* 50, 183–189.
- Zogorski, J. S., Delzer, G. C., Bender, D. A., Baehr, A. L., Stackelberg, P. E., *et al.* 1998. MTBE: Summary of Findings and Research by the U.S. Geological Survey. Proceedings of the 1998 Annual Conference of the American Water Works Association.

Geographic Thought, History of Energy in



BARRY D. SOLOMON

Michigan Technological University
Houghton, Michigan, United States

MARTIN J. PASQUALETTI

Arizona State University
Tempe, Arizona, United States

-
1. Early Research on the Coal Industry, 1934–1955
 2. Fossil Fuels and Electric Power Generation, 1955–1971
 3. The Birth of Commercial Nuclear Energy, 1951–1973
 4. Modeling the Spatial Development and Transportation of Fossil Fuels and Power Plant Siting
 5. Hazards, Risks, and Technological Diffusion: Behavioral Research by Geographers Since the TMI Accident
 6. Development of Sustainable Energy Resources

Glossary

econometric analysis A test of economic theory with linear or nonlinear regression and other multivariate statistical methods using socioeconomic and technical data.

geography of energy The study of energy development, transportation, markets, or use patterns and their determinants from a spatial, regional, or resource management perspective.

location theory Spatial theory developed to solve optimization problems for selecting a preferred zone or location for a facility, which involves trading off transportation costs with production and processing costs, production balances, and market-delivery levels.

mathematical programming model An operations research technique designed to determine an optimal solution to a well-defined problem by maximizing or minimizing an objective function subject to technical, economic, and often environmental constraints.

regional input–output analysis An analysis of the economy at the urban, state (provincial), or multistate (provincial) levels based on a fundamental identity that equates supply and demand using fixed proportions of intermediate inputs.

Three Mile Island (TMI) The name of the nuclear power plant near Harrisburg, Pennsylvania, that was the site of the most serious nuclear accident in the United States on March 28, 1979.

Since the field of geography encompasses broad study of interactions between humans and the environment, the concept of energy has been operationalized and studied in several ways. These approaches have included energy balance studies in geographical climatology, energy resource availability and conservation in applied climatic research, cultural geography, and economic geography of energy resources. In the past few decades, new geographic perspectives on energy have been introduced from spatial analysis and modeling to applied, political, transportation, hazards, geographic information systems, global environmental change, and ecological economics. Some of these areas are even considered separate fields.

Understanding and modeling solar radiation and the energy budget of the climate at a variety of spatial scales have long been central concerns of several types of climatological studies: modeling, synoptic and dynamic climatology, climate change, and applied climatology. Analytical approaches have become increasingly mathematical over time, and this work has been somewhat separate from other branches of geography. It is consequently more convenient to trace the evolution of the energy concept in geographic thought by focusing on the contributions of human geography, which initially considered energy from economic and cultural perspectives. The article begins with a discussion of pre-World War II studies of the coal industry.

Subsequent sections consider the rise of petroleum as a global political force, commercial nuclear power, and additional work on the location and spatial distribution of fossil fuels. A new period in geographic thought followed the Three Mile Island nuclear power plant accident in 1979, when geographers gave attention to technological hazards and risks of energy systems, the back end of the nuclear fuel cycle, the diffusion of energy conservation technologies, and the increasing scarcity of fossil fuels. The article concludes with a discussion of the growing need to develop sustainable energy resources in the context of concerns about global climatic change and rising energy demands in developing nations.

1. EARLY RESEARCH ON THE COAL INDUSTRY, 1934–1955

Coal, as the fossil fuel of greatest abundance and widest distribution, has long attracted attention from geographers. In the early years, this attention focused on the location of deposits and their transport and much of it mirrored the major coal fields of Britain, Germany, Poland, and the United States. The accounts in geography journals were largely descriptive, economic, and cartographic. Little or no interest was evident regarding the environmental costs. Later, when the public mood shifted and impacts were widespread, more attention was paid to environmental impacts, especially air pollution and its dispersal, impacts of acid rain, landscapes altered by strip mining and waste heaps, and the forced relocation of residents. This emphasis came with the increasing knowledge not only of the damage that was created by our demand and use of coal but also that we should and could do something to reduce or reverse the impact it was having on air, land, and water. From this point forward, research articles were much more widely dispersed, within a range of environmental journals and books, and the type of writing was more analytical.

2. FOSSIL FUELS AND ELECTRIC POWER GENERATION, 1955–1971

Parallel to the growth and development of the oil and natural gas industries in the United States and other advanced economies, geographers began to give

more attention to the other fossil fuels, their development, and use patterns in the mid-1950s. Another exciting new energy source gained attention with the commercialization of atomic power plants in Britain, Russia, and the United States during this period. Nuclear power development attracted considerable attention among geographers in North America and Europe. This topic is considered separately since the economics, spatial diffusion, risks, social acceptance, waste problems, and other dimensions of nuclear power were highly uncertain in the 1950s and thus it was treated differently.

Although a quantitative revolution in geography first appeared in the mid-1950s there was little evidence of a change in methods in energy research until the early 1960s. Indeed, most geographical publications on energy issues in the 1950s were studies on the regional geography of energy (i.e., “areal differentiation”). Although important, this work was overly descriptive and often handicapped by scant understanding of the complex interplay between resources and technical, economic, and social factors in energy markets. Several detailed analyses were completed on the oil, gas, and electric power industries of the United States, Britain, Europe, and the former Soviet Union during this time. A few researchers went beyond the single-region scale and assessed the development of power interconnections in Europe, North America, and the Soviet Union. A 1960 paper addressed the conflict between hydroelectric power generation and salmon fisheries in the Pacific Northwest, which is still relevant today. This study suggested the use of (but did not apply) optimization criteria to help resolve the conflict. The same year marked the inaugural publication of *Soviet Geography: Review and Translation*, which was to become a major scholarly outlet for energy studies of the former Soviet Union and its various regions.

Neoclassical studies of the regional geography of energy continued throughout the 1960s at a variety of spatial scales. Several books and articles on the economic geography of the oil and gas industries analyzed national and international patterns of location, economic growth, development, and trade but did not suggest general market patterns. The first truly international energy study, a book on the economic geography of the world oil industry by the British geographer Peter Odell, was published in 1963, just a few years after the formation of the Organization of Petroleum Exporting Countries (OPEC). Articles on coal geography also assessed economic growth and development but considered

declining or less prosperous coal fields and their socioeconomic effects as well. Political factors (i.e., national policies, government regulation, political stability, etc.) began to be mentioned more frequently in this literature but were still secondary to economic considerations.

Of special significance was the work by Gerald Manners in the 1960s to incorporate location theory into energy geography. General locational principles were identified and applied to the oil, gas, and electric power sectors. Transportation costs were shown to be of fundamental importance because they determine the energy facility location between a resource market(s) and final demand center(s). Also important are factor costs of production, resource availability, weight and quality, scale economies, technical efficiency, modal and interfuel competition, and the load factor(s). Similar principles were used to explain the changing pattern of pit location within a coal field. The first such analytical, general approach for studying the industrial geography of energy was published in book form by Manners in 1964. (A 1950 book on the geography of energy by Pierre George received little notice outside of France.) Owing to data availability and the scale of the economies, his research focused on Western Europe and North America. National studies and atlases for the United States and United Kingdom were to come later.

As the 1960s closed, a few prominent authors completed seminal works on the international patterns of energy geography. A landmark book by Odell addressed oil and world power; this book was first published in 1970 and updated seven times. The monograph considered this leading industry's production, economic growth, and trade in the context of its prominent role in national and international political affairs. By analyzing the global structure of the oil industry along with regional perspectives and company–nation state relationships, Odell provided the most complete story of this vital resource. Three more general energy papers by geographers in the September 1971 issue of *Scientific American* showed the close connection between energy resource type and technology at the various life-cycle stages and the pattern of energy flows in primitive and advanced industrial economies.

Two final (although brief) energy books from the perspective of world economic geography closed out this period, both of which were written by Nathaniel Guyol. One of these monographs provided, for the first time, a system of energy accounting that made it possible to compare energy statistics in multiple

countries on a consistent basis, a subject he had worked on since the late 1940s. An accounting worksheet, suggestive of a first cut of a simple input–output table, was developed for The Netherlands as an example. The other volume was narrower but more detailed: a systematic survey of the electric power industry in 162 nations. Important observations were made by Guyol on the regional variations in electricity supply and use.

3. THE BIRTH OF COMMERCIAL NUCLEAR ENERGY, 1951–1973

The use of commercial nuclear power grew up as the stepchild of the military program. After spending an enormous amount of money to develop and construct the three bombs that were detonated in 1945, there was a concerted effort to turn this research to peaceful use. There followed almost a decade of research and development that led to the construction of electrical generating stations at Calder Hall, England, Shippingport, Pennsylvania, Eureka, California, and eventually more than 400 other commercial units.

Geographers showed little interest in the dispersed siting of the nuclear research program of the war years, but afterwards they addressed the possibility of using nuclear bombs to reshape the earth's surface, the location of nuclear fuels, and the siting of nuclear power plants. It was electricity generation that was to achieve the greatest momentum as commercial enterprises began appearing in several countries: Britain, the United States, the Soviet Union, and later France, Germany, and Japan. Researchers during this period published a steady, if thin, series of geographical studies, only rarely turning their attention to the spatial pattern of the fuel cycle or dispersal of the technology.

For more than 15 years, there seemed little uncertainty or public misgivings about the wisdom of a nuclear power program, and the geographic literature did not contradict this trend. Industry, government, and the public seemed to all be moving in the same direction, one that looked to nuclear power to generate a growing, perhaps dominant, proportion of the great amounts of new generating capacity that many countries were demanding. With U.S. growth rates for electricity exceeding 7% per year in the 1960s, nuclear power not only seemed like the most likely supplement to coal-fired power plants but also seemed to be the eventual replacement

technology as concerns grew over the pollution that such fossil power plants produced. Moreover, it was recognized that by themselves coal stations would not be able to keep ahead of the tidal wave of demand for electricity. The public, bolstered by the reassurances of government and contractors, regarded the use of nuclear power as the only way to meet the steep rise in need.

During the early civilian period, power plant siting was influenced mostly by seismic stability and the availability of cooling water. Little consideration was given to isolating power plants from urban areas; indeed, many were located close to load centers, the better and more conveniently to service their needs. Thus, the pattern that was emerging favored the densely settled northeastern United States and Chicago. Gradually, additional power plants appeared in the southeast. In Britain, all but one (in Wales) were sited along coasts. France used both coastal and riverine sites, starting mostly with the latter locations, and in contrast to the United States and the United Kingdom, the distribution of nuclear power plants had no obvious locational concentration. However, the benign public reaction to nuclear power was to be short-lived.

4. MODELING THE SPATIAL DEVELOPMENT AND TRANSPORTATION OF FOSSIL FUELS AND POWER PLANT SITING

By the early 1970s, the quantitative revolution was firmly entrenched in energy geography. Researchers began to apply several mathematical and statistical techniques to problems of spatial energy resource development and transportation, power plant siting, and socioeconomic and environmental impact analysis. These important research developments, spanning the 1970s through the mid-1990s, are discussed in turn.

4.1 Resource Development and Transportation

Beginning in the 1970s, geographers developed and applied a variety of simulation, mathematical programming, and econometric models to oil and gas issues. Among the early work was a rather extensive analysis of North Sea oil development through 2080. A less ambitious study assessed the optimal oil

pipeline network in the British sector of the North Sea. North American oil industry studies also used mathematical programming models, with the site selection and storage levels at the Strategic Petroleum Reserve in Louisiana as an example. Since the mid-1980s, a series of related papers have used econometric methods to update the oil industry models of geologist M. King Hubbert. This work often differentiated resource quality and used more than 50 years of historical data to assess the bleak future prospects of the U.S. oil industry. Somewhat less attention has been given to oil development in other regions. A few econometric studies have analyzed foreign investment, locational analysis, and resource-based industrialization in the Middle East, other developing countries, and the former Soviet Union. Additional studies have highlighted the main result of these developments for the United States: growing dependence on foreign sources of oil.

Mathematical programming methods also began to dominate research on the development and marketing of natural gas in North America and the Soviet Union. Since the production of this fossil fuel generally lagged that of petroleum in the United States and was decontrolled starting in the late 1970s, new research opportunities were pursued for modeling greater development, especially in Alaska, and trade with Canada and Mexico. Similarly, natural gas emerged as the fuel of choice in the Soviet Union in the 1980s. In order to advance this research area, it was necessary to improve upon previous models of the pipeline network and capacities. Following the decline in oil and gas prices and plentiful gas supplies in the United States since 1986 this research area has been largely abandoned. Only one major analysis of U.S. natural gas development has been completed since then, an econometric analysis of the declining yield per effort and resource scarcity.

Not surprisingly, the quantitative research methods that were first used in the 1970s for oil and gas studies also were applied to coal problems, along with the use of cluster analysis and spatial interaction models. Given the wide distribution of coal reserves in the United States and their often long distances from markets, transportation bottlenecks have occasionally arisen. Such problems are ideal cases for the use of operations research methods, such as linear, nonlinear, mixed integer, and multiobjective programming. An early 1970s study of the Great Lakes region was a landmark in the modeling of the whole coal resource delivery system. This was followed by a series of coal studies: regional industry analyses, such

as for Britain, western U.S. states, and Pennsylvania; transshipment problems involving railroads and ports; coal exports; competition for water resources; other environmental constraints; and structural and technological change.

Geographers, along with scholars in other fields, have also studied the development potential of synthetic fuels from fossil fuels. This work has analyzed the spatial distribution patterns and development potential of coal gasification and liquefaction plants, oil shale, tar sands, and alcohol fuels. The huge water and other resource requirements and environmental effects were given prominent attention. These studies tracked the government policy interest and short-lived massive subsidies to potential synthetic fuels industries in North America in the late 1970s and early 1980s, which were briefly considered a viable alternative to the growing dependence on foreign oil supplies. Among these resources, only alcohol fuels (from ethanol and methanol) and the Canadian tar sands industry have established commercial viability.

4.2 Power Plant Siting

The siting patterns of electric power generating stations have been a major research area among energy geographers. Although early papers on this subject were published in 1960, the vast majority of these studies were done in the mid-1970s through the mid-1980s. Perhaps surprisingly, a retrospective discriminant analysis of U.S. power plant siting patterns from 1912 to 1978 showed that fuel choice was more important than regional differences in the siting process. Development of site suitability models for fossil-fueled and nuclear power plant location analysis was pioneered at the Oak Ridge National Laboratory in the late 1970s. These and later siting models were applicable to efforts to avoid or mitigate inevitable locational conflicts surrounding proposals for many new energy facilities. Geographers developed land use screening procedures to generate local site suitability scores by estimating a land use compatibility index and using importance weights offered by decision makers. A large range of engineering, ecological, and socioeconomic criteria were analyzed. These models were applied to the Mid-Atlantic states. Despite the value of these methods for initial site screening, it has been noted that the approach largely failed to be implemented because of a lack of need for new power plant sites at that time. On the other hand, the screening models amount to a stepwise heuristic

and can select inferior sites. Improvements to the method were made through the development and application of optimal transshipment and multi-objective programming models, which were also applied to the Mid-Atlantic region. Another study considered the challenges in siting coal-fired power plants in western states, where there are stricter air quality standards.

The siting and safety of nuclear power plants received the most attention by geographers of energy during the early to mid-1980s. Ironically, following the Three Mile Island (TMI) accident in 1979, North American energy geographers shifted their research from power plant siting to reactor decommissioning and waste disposal, whereas British colleagues focused on nuclear plant siting. Computer programs were written in England to analyze alternative power plant siting criteria while emphasizing remote areas and the potential human dose from radioactive releases during accidents. Government policy was shown to downplay the siting criterion of proximity to human population centers, and thus arguably safety, despite the existence of ample remote sites in Britain. Cluster analysis was used to show that local population density received greater attention with regard to siting in the United States than in Britain. Broadening this discussion from a Canadian perspective, it was shown that existing nuclear power expansion plans in the mid-1980s were not as robust as construction of small-scale hydroelectric stations. Expansion of cheap hydroelectric power in eastern Canada, whether small or large, fits in well with plans for increased electricity trading with the United States. Indeed, further study has highlighted the parallel socioeconomic and environmental challenges accompanying the recent "outsiting" of nonnuclear power plants in Mexico and Canada that have been dedicated to U.S. markets.

Since U.S. electric utilities had a greatly diminished need for power plant siting in the late 1980s as independent power producers and demand-side management picked up the slack, energy geographers abandoned regional siting analysis and developed broader, more detailed, and sophisticated industry resource planning models. A few major, large linear programming (LP) models were built with a full range of technical options for meeting future electricity demand under environmental constraints. One of these used multicriteria decision making and applied it to demand-side management planning in British Columbia, Canada. Most of the LP models

were applied to sulfur dioxide (SO₂) emissions control. There was some limited regional detail to them, which overlapped with the work of regional scientists and regional economists.

4.3 Socioeconomic and Environmental Effects

Research by geographers on the mixed effects of coal mining dates back to at least the early 1950s. Initially, this work focused on the eastern United States, but by the late 1970s and 1980s it had shifted to the arid West (along with an increasing share of the coal production). This research analyzed the potential socioeconomic impacts of “boom-town” development of coal or synthetic fuels. The declining economic fortunes of some of the eastern coal mining areas such as anthracite towns of Pennsylvania continued to receive attention in the 1980s. Regional input–output (IO) analysis was demonstrated to be an especially useful tool for analyzing economic effects of spatial shifts in the coal industry. Similarly, most coal research in Britain during this period focused on the uneven economic development or shift between declining and developing regions and a short-lived effort to revive the industry. When U.S. western energy resource development plans were scaled back, research emphasis shifted to the economic and ecological effects and control of acid rain emissions of SO₂ and NO_x. This subject continued to receive the attention of a few geographers through the 1990s as the Acid Rain Program of the Clean Air Act Amendments was implemented, including (inter-state) SO₂ allowance trading.

Many energy geographers have examined the socioeconomic and environmental effects and planning issues surrounding petroleum development in disparate regions such as the Middle East, Scotland, Norway, and the U.S. Gulf Coast. The U.S. studies have also considered impact mitigation in the context of the National Environmental Policy Act and the macroeconomic effects of a future large supply disruption by OPEC. Econometric analyses by geographers for the United States have shown since 1989 that policies to expand domestic oil production would have adverse economic and ecological effects. Although less sweeping in its findings, a regional economic analysis in Canada of potential oil resource development in the nearby Beaufort Sea found that little or none of the net economic benefits would accrue to Canada’s Northwest Territories. This research used a multiregional IO model.

5. HAZARDS, RISKS, AND TECHNOLOGICAL DIFFUSION: BEHAVIORAL RESEARCH BY GEOGRAPHERS SINCE THE TMI ACCIDENT

Public disinterest in nuclear power was shaken by the successful campaign to raise awareness about its purported safety risks. It was a movement that slowly gained momentum, but by 1976 an anti-nuclear referendum in California received ample public notice and approximately 40% of the votes. Although the referendum failed, it was a watershed for the nuclear industry because it raised many more questions than the industry could successfully answer about the safety and the financial feasibility of the technology. In addition, it prompted more restrictive and demanding siting criteria, and more than anything else it put nuclear power on the map of regular public debate. Like colleagues from other disciplines, geographers became more involved, and it was a natural topic of interest because most of the issues were spatial in nature.

The increased attention by geographers first manifested itself in questions about the adequacy of the siting criteria that had been unquestioned during the previous decade. Since much of the impetus for the nuclear debate had come from California, it was not surprising that the strongest opposition to power plant siting appeared there as well. Geographers played a quiet role at first, mostly related to siting, but their participation grew as the public debate expanded to other spatial issues, such as radioactive waste transport and disposal, dispersal of radioactivity by air currents, and identification of evacuation routes that might be needed in the event of an accident. These developments were all part of a simmering debate about nuclear power in the United States and elsewhere, but they would come to a boil with the TMI Unit 2 accident near Harrisburg, Pennsylvania, on March 28, 1979. Ironically, Pennsylvania had enthusiastically commissioned the earliest U.S. nuclear power plant 22 years earlier.

It was the TMI accident that firmly placed geographers within the nuclear power debate. For several years, many geographers turned their attention to the myriad questions that the accident raised, including siting criteria, the dispersal pattern of possible contamination, evacuation behavior, and the adequacy of emergency planning and preparedness. The effects that the accident had on nuclear power were fundamentally geographical. Most

noticeably, no new operating licenses were granted afterward, largely because of the elevated public concerns about the appropriateness of virtually every suggested site. Emergency requirements that were promulgated became so demanding that power plants already completed (e.g., the Shoreham plant on Long Island) were unable to develop an approvable evacuation plan.

Geographers participated in these studies and were most heavily involved in evaluating and measuring public risk perceptions. They asked questions about relative risk, the level of risk required to trigger evacuation, and the use of distance as a buffer to risk. These issues came into play because of the obvious inadequacy of emergency preparations at TMI, but the interests of geographers were not limited to them. Issues that had long been beneath the surface came into full view, and they included all forms of "nuclear geography," including a striking increase in the attention paid to radioactive waste transport and disposal.

Although attention to these matters was initially concentrated in the United States, it did not remain within its borders. Geographers in other countries, particularly Britain, heard the call as well, and geographers there had been making similar intellectual preparation to participate. There was a different attitude in the United Kingdom, however, in large part because their system of government had not been as receptive to an open debate on nuclear power as in the United States. As it soon became clear, its citizens had more practical reason for concern. Great Britain is a country smaller than Arizona, without the vast distances and isolation possible in the United States. Moreover, when the inquiry into the 1957 Windscale accident in northwest England confirmed public suspicions about nuclear hazards and ill-considered official secrecy, a phalanx of geographers was to lead the charge for a much wider nuclear debate.

All the nuclear plants that currently exist in the United States had already been sited by 1979, but this was not the case everywhere. In the United Kingdom, for instance, great attention was paid to siting and safety issues of the planned addition to the Sizewell nuclear station on the Suffolk coast. There was already a Sizewell A, but Sizewell B was to be a pressurized-water reaction of American design, ostensibly safer than the Magnox and Advanced Gas-Cooled reactors that comprised all the early British reactors. In response to the work of many groups, and especially geographers in the United States and Britain, the government conducted the

longest and most detailed public inquiry in the history of the country.

Like the TMI accident, the Sizewell inquiry galvanized geographical attention. Soon, geographers in England were writing about waste transport through London, waste disposal under the Irish Sea, and elevated leukemia frequencies in Cumbria. Perhaps due to the smaller size of the country, much of the work of geographers in Britain attracted more attention than similar work did in the United States. Geographers regularly appeared on TV and in various newspapers and magazines, which gave their comments more visibility. Eventually, geographers from Britain and the United States began collaborating on articles and books. Even France, which operates nuclear power plants that are closer to London than those in Wales, was not immune from criticism, and although the debate was more subdued there, geographers tracked and reported on various acts of public opposition.

Attention of geographers to nuclear issues peaked in the early 1980s and then began to subside, only to be shaken into a higher degree of agitation 7 years later by the accident at the Chernobyl nuclear power plant in Ukraine. This was a far more serious accident, and it had global implications. Climatologists tracked the movement of radioactive materials, but access to data was difficult owing to long-standing policies of the Soviet Union. Nevertheless, several geographers focused their attention on the disaster, with some examining issues of the power plant's location adjacent to the vulnerable Pripyat Marshes and 90 miles upwind from the heavy concentration of humans around Kiev, others observing how land had to be reshaped and forests leveled for safety reasons, and still others focused on the forced evacuation of more than 135,000 people who lived near the plant. Others were intent on observing the various political ramifications, including how the accident helped hasten the collapse of communism.

The Chernobyl accident also pounded several more nails in the nuclear coffin. In 1988, Swedish voters reaffirmed that nation's decision to phase out nuclear power, and Germans voted to take much the same direction. The RBMK design used at Chernobyl was largely abandoned, and many of those that had been built in the former Soviet Union and satellite countries have been closed. Moreover, attention to the various elements of nuclear power plant decommissioning gained increasing momentum.

Although it was obvious that nuclear plants would eventually be "turned off," the questions were when, by what process, and to what effect. There had been

little public attention to these questions prior to 1985, but geographers helped bring attention to them during the next decade, emphasizing postoperational land use, transport and storage of the resulting low-level radioactive waste, the spatial and temporal sequence of targeted plants, and the socioeconomic consequences of the decommissioning process.

At approximately the time that interest in nuclear power was on the decline among geographers, attention was increasing in the area of energy conservation and efficiency. It was not causative, but when problems began appearing and proliferating with nuclear power, several electric utility demand-side management strategies were waiting in the wings. The logic was that if the electricity that nuclear power had been expected to supply was not actually going to be available, shortfalls would occur unless demand could be cut. It was cut, across the board and across the country, in every end-use sector, with the most noticeable changes occurring in the commercial and industrial parts of the economy and gradual progress in the residential sectors as well. Geographers worked with other professions to identify where these changes could be made and evaluated program effectiveness. There were several obvious uses to emphasize, including increased efficiency in lighting, insulation, appliances, and transportation.

6. DEVELOPMENT OF SUSTAINABLE ENERGY RESOURCES

Although physical geographers have long included solar energy in its various manifestations in their work, only since the late 1970s have geographers actively researched renewable forms of energy. Even so, this research interest tapered off in the late 1980s, seemingly coincident with the crash in world oil prices in 1986. This hiatus was short-lived because by the early 1990s, there was renewed interest in fossil fuel scarcity, namely that of oil and gas. Geographers also began to explore the energy policy implications of the growing concern regarding global climate change, with an emphasis on energy efficiency and renewable energy development.

6.1 Oil and Gas Scarcity

Although acknowledgment of the ultimate scarcity of petroleum resources dates back to the 1956 work of M. King Hubbert, who predicted with startling

accuracy the ultimate size and production peak of oil resources for the United States and the world, there is much greater concern with its "net energy" from production rather than ultimate depletion. Cutler Cleveland and Robert Kaufmann, among others, in a series of papers since 1991 have convincingly forecast using econometric analyses that the amount of oil added to proven reserves per effort from additional well drilling (yield) in the United States will cease to be a net energy source early in this century. This line of research builds on the work of Hubbert, who was the first to recognize that the historical production pattern would follow a classic bell-shaped curve, reflecting the interplay of resource depletion, real oil prices, technical innovation, and political decisions on long-term production costs. Consequently, domestic oil discovery, production, and proven oil reserves are on an inexorable downward path. These findings call into question energy policy decisions based on assumptions that these geophysical realities of the U.S. resource base can be altered.

With the conventional oil resource base increasingly scarce and dependence on it unsustainable, the cleaner option of natural gas has received more attention, especially since estimates of its proved reserves were greatly increased in the early 1990s. Here too, analysis by Cleveland and Kaufmann is instructive. Using a double-log econometric model of yield per effort of gas drilling in the United States, they found an exponential decline for gas discoveries not associated with oil fields in the lower 48 states from 1943 to 1991. Their model accounts for cumulative drilling, real gas price, changes in drilling effort, and shifts between onshore and offshore gas. Thus, natural gas development lags oil but is subject to similar factors controlling its scarcity.

6.2 Greenhouse Gas Reduction and Renewable Energy Development and Use

Coal, the third fossil fuel, is well-known to have the highest emission rate of carbon dioxide (CO₂), the principal greenhouse gas. By the early 1990s, a growing cadre of geographers began to actively work on energy policy responses to the threat of planetary warming and climate change. Indeed, they recognized that many preferred policy options also reduce conventional air pollutants and have a variety of other ancillary benefits. These energy options include switching to gas, energy efficiency and energy conservation, and rapid development of renewable and possible nuclear energy sources (although no

major energy geographers have promoted the latter option through their research).

Three major types of greenhouse-related energy studies have dominated the work of energy geographers in the past few years: detailed technology assessments at the U.S. national level, sectoral-specific studies, and foreign case studies. One leading study provided an integrated assessment of the potential benefits of advanced energy technologies for greenhouse gas reductions in the United States. The authors found that large CO₂ reductions are possible at incremental costs below the value of energy saved, especially efficiency measures. Many sectoral-specific studies by geographers have highlighted the technological improvements and potential roles of energy efficiency, wind, solar, and solar-hydrogen technologies in buildings, transportation, manufacturing, and electric power. Much of this research has been conducted at the Oak Ridge National Laboratory.

Since global warming and climatic changes do not respect political boundaries, case studies of other nations can be valuable. For instance, several geographers have studied the broad range of policies that would be required to significantly reduce CO₂ emissions in Germany, Canada, and China. These studies not only highlight the great difficulty of significant emission reductions but also emphasize the necessity to implement policy reforms outside of the energy sectors (e.g., addressing urban design and transportation policies). To accelerate the process of emissions reduction in developing countries and economies in transition, the 1992 Framework Convention on Climate Change approved a pilot program in joint implementation of the treaty, whereby two or more nations cooperatively reduce or sequester emissions in their countries. Implementation of this pilot program was slowed, however, by concerns among developing countries that national sovereignty would be violated and that Western nations or Japan would gain a competitive economic advantage. Geographers have generally been sympathetic to these concerns, questioning the policy's fairness and cost-effectiveness, while identifying constructive solutions that involve compromises by the applicable countries.

Renewable energy technologies have different spatial characteristics than those of fossil or nuclear fuels (e.g., with respect to availability, transportability, reliability, and storage), which have yet to be fully explored. Among conventional energy resources, the two renewable sources that are well established are hydroelectricity and biomass. Hydro-

electric installations, however, are not always sustainable because gradual siltation of their reservoirs may limit facility lifetime to 50–200 years. Since the 1990s, hydroelectric power has attracted increasingly negative attention in the United States and elsewhere, where the large part of its potential has already been tapped. Identified shortcomings include forced migration of people, displacement and destruction of salmon stocks, and downstream ecological changes.

Geographers have written a modicum of papers and monographs on hydropower facility construction in the past decade, which have focused on trends in North America and a few mammoth dam projects in developing countries such as Brazil, China, and India. Research emphasis in the United States has shifted from dam construction to dam removal and downstream restoration. Several dams have been suggested for demolition, especially smaller structures. Of the large dams, the most attention has been given to the proposed removal or "neutralization" of Glen Canyon Dam, located upstream from the Grand Canyon.

Biomass has long been the major (potentially) renewable energy source in developing nations, and it has received the most attention among the renewables from geographers. It has also received research attention in Scandinavia, where it is commonly used but somewhat controversial in Sweden. Along with its relative abundance, biomass also has challenging limitations owing to its wide variety of forms, ranging from wood to crops and garbage.

Wind energy has been the fastest growing energy source in the world in the past decade, and in the late 1990s Germany overtook the United States as the global leader. An important book printed in 2002 includes 10 chapters by geographers that compare and critique the varying prospects and policy approaches to renewable energy development, especially wind, biomass, and solar energy, in the United States and northern Europe. Wind energy use and public perception in California, the location of its most extensive U.S. development, have been examined by geographers in articles and a 2002 book that examines the issue of aesthetic intrusion on the landscape. Extensive study of wind energy has also been conducted by climatologists. For example, a 1997 paper analyzed the influence of climatic variability on wind resources in the U.S. Great Plains, the area of greatest wind power potential.

Like wind and hydropower, immobility presents geographic challenges for the development of geothermal energy, particularly the pressure it puts

on natural resources. Geographers have given the least research attention among the renewable energy sources to geothermal (because this research area is dominated by geologists and engineers), after significant contributions were made by Pasqualetti in the late 1970s and 1980s.

6.3 Energy in the Developing World

Faster than other professions, geography is predisposed to grasp differences between the developed and the developing world—differences that are particularly apparent in considerations of energy. The developed world relies on fossil fuels, whereas the developing world relies on renewables. The developed world searches for energy to maintain a high standard of living, whereas the developing world must sell its resources to stave off poverty. The developed world worries about the environmental impacts of energy supply, whereas the developing world worries about having any energy supply at all. The developed world has achieved a high standard of living by consuming prodigious amounts of energy, whereas the developing world resents any attempts to curtail their upward mobility by doing the same, even though the environmental impacts are now so much more obvious, troubling, and dangerous. The developed world is in the minority, whereas the developing world is in the majority. Such differences form the basis for much of the work geographers have pursued in developing countries.

Although many developing countries share a need to trade environmental quality for energy supply, they are in other ways profoundly different. For example, China and Mexico are industrial countries, whereas Haiti and Ghana have primarily agrarian economies. In the former, government concentrates on how to increase gross domestic product; in the latter, government tends to worry more about matters of day-to-day survival. The work of geographers has focused on issues of environmental quality, economic growth, equity, and sustainable development and in a more speculative mode the consequences of bringing per capita energy consumption in these countries closer to that of the United States or the European Union.

The work of energy geographers in developing countries is often closely tied to environmental considerations. China, for example, is intent on using its great reserves of coal to help fuel its climb toward greater economic prosperity. As a consequence of pursuing this goal, air pollution has

increased in many locations within the country, as has concern about the possible impact of such additional coal burning on global warming. Geographers, among many scientists, have monitored such trends and have expressed opinions ranging from how one might minimize such impacts to suspecting that the impacts on climate change are insignificant.

Other topics of interest to geographers include energy trade and transportation; the use and scarcity of biomass; hydropower development; and the development of geothermal, wind, and solar power. Of these topics, energy transport and trade have had perhaps the most natural and longest interest. The focus of such attention by geographers evolved from matters of shipping routes and port facilities to the location of lines for the movement of energy by rail, pipe, and wire. Shipping routes tend to change infrequently, and then usually as the result of military conflicts and new discoveries. Modern examples include the closing of the Suez Canal and the development of oil fields in the Gulf of Campeche, both in the 1970s.

Geographers continue to track the discovery and development of new sources of energy, although such activities trend toward the expansion of existing areas of extraction rather than new developments. Exceptions to this rule include oil developments in African countries such as Chad and South American countries such as Ecuador, Peru, and Bolivia. In the latter country, interest has been drawn especially to natural gas fields on the eastern side of the Andes, with the prospect of piping through Chile to the Pacific Coast for liquefaction and transport to markets. One market is the United States, and Bolivia is considering shipping to a planned regasification facility in Baja, California, and then piping the gas to California. In so doing, they will be avoiding the environmental encumbrances of trying to locate regasification plants within the United States while still having a U.S. market. In all cases, these and similar projects hold the hope for improving the economic prosperity of citizens of the producing countries, although the usual questions exist as to whether a significant proportion of the extracted wealth will remain local. Geographers have noted this ethical issue and in a few cases have proffered recommendations.

Although these examples rest on matters of resource sales and environmental impacts, geographers have sustained an interest in the transportation routes, including the economic costs and benefits of the developing countries that are physically touched along the way. Chad needs to move its oil through

Nigeria, for example, and the vast resources of the landlocked Caspian Sea must cross through several countries to gain access to the sea, which is a diplomatic challenge. Locally, Mexico and the United States have been working to establish closer ties through increased land-based trade of electricity, coal, oil, and natural gas, in both directions, across their common border. This has caught the attention of geographers in both countries, who have pointed to the shared advantages. They have also identified some of the shared environmental costs when the impacts of energy development cross international borders, such as when air pollution wafts northward from power plants in the Salinas district south of Eagle Pass, Texas.

Hydropower projects always hold a particular appeal for developing countries. Not only do they provide electricity, flood protection, and, often, irrigation water but also they create jobs and national pride. In developing countries, however, they also produce some problems that have not been as publicized elsewhere. Often, these problems have been associated with the displacement of large numbers of people as reservoirs fill. In the past, the most notable examples of this pattern have been in Africa and Brazil, but currently world attention has focused on the Yangtze River in China, the Naranda River in India, and the Euphrates River in Turkey. Even in places where few people live, hydroelectric projects are often controversial, either for reasons of environmental disruption, such as on the border of Uruguay and Brazil, or for reasons of native land claims (as with the huge projects on James Bay in Canada). Although all these hydro projects have inherently geographic applicability, only a few geographers are involved in studying them.

The oldest and most traditional energy resource is biomass, usually in the form of wood from shrubs and trees, crops, and animal dung. Biomass can be used for heating, cooking, and transportation. Wood fuel use and sustainable development have been extensively researched in case studies of India, Pakistan, Tanzania, Kenya, Zimbabwe, Ghana, and Haiti. Brazil, in contrast, has supplemented wood and charcoal use with sugarcane and bagasse for ethanol fuel and cogeneration. The heat content of biomass is low by weight compared to coal and other fossil fuels, so for areas of quickly growing population many geographers have reported how such dependency quickly transforms landscapes, which often do not recover their original appearance or productivity. The relentless and growing dependency on such resources has secondary impacts on soil

erosion, microclimates, and other aspects of local and regional ecological balance, which geographers have long noted and studied, especially as it pertains to the creation and implementation of government policies.

Developing countries are among the most logical candidates for the development of renewable energy resources, especially direct solar and wind power. From a geographical perspective, direct solar power, either for hot water or for electricity, holds the greatest advantage because it is not site specific; photovoltaic panels can be installed and used with little need for local technical competence or development of a delivery infrastructure. Geographers have written about the relative advantages of such sources of power for isolated people in India and several other developing countries in Latin American and Africa, and the experiences in these locations can be transferred to just about any place in the tropics. Regarding wind power, some developing countries, such as Mongolia, have great potential, but wind (like geothermal) is spatially concentrated and immobile. The number of studies should increase in the coming decades as the quest for sustainable development continues.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • Economic Geography of Energy • Economic Thought, History of Energy in • Energy in the History and Philosophy of Science • Geopolitics of Energy • Nuclear Power: Risk Analysis • Public Reaction to Energy, Overview • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

- Cleveland, C. J., and Kaufmann, R. K. (1991). Forecasting ultimate oil recovery and its rate of production: Incorporating economic forces into the model of M. King Hubbert. *Energy J.* 12(2), 17–46.
- Cook, B. J., Emel, J. L., and Kasperson, R. E. (1990). Organizing and managing radioactive-waste disposal as an experiment. *J. Policy Analysis Management* 9(3), 339–366.
- Cook, E. F. (1976). “Man, Energy, Society.” Freeman, San Francisco.
- Cuff, D. J., and Young, W. J. (1986). “The United States Energy Atlas.” 2nd ed. Macmillan, New York.
- Elmes, G. A., and Harris, T. M. (1996). Industrial restructuring and the changing geography of the United States coal-energy system, 1972–1990. *Ann. Assoc. Am. Geogr.* 86(3), 507–529.

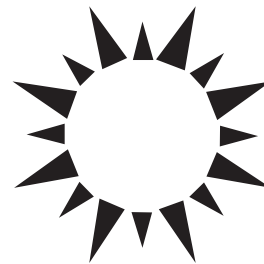
- Guyol, N. B. (1971). "Energy in the Perspective of Geography." Prentice-Hall, Englewood Cliffs, NJ.
- Hosier, R.H. (Ed.). (1993). Urban energy and environment in Africa [Special issue]. *Energy Policy* 21(5), 434–558.
- Jackson, E. L., Kuhn, R. G., Macey, S. M., and Brown, M. A. (1988). Behavioral energy research in geography. *Can. Geogr.* 32(2), 162–172.
- Majumdar, S. K., Miller, E. W., and Panah, A. I. (Eds.). (2002). "Renewable Energy: Trends and Prospects." Pennsylvania Academy of Science, Easton, PA.
- Manners, G. (1971). "The Geography of Energy." 2nd ed. Hutchinson, London.
- Odell, P. R. (1986). "Oil and World Power." 8th ed. Penguin, Harmondsworth, UK.
- Pasqualetti, M. J. (ed.). (1990). "Nuclear Decommissioning and Society: Public Links to a New Technology." Routledge, London.
- Pasqualetti, M. J. (2000). Morality, space, and the power of wind-energy landscapes. *Geogr. Rev.* 90, 381–394.
- Sagers, M. J., and Green, M. B. (1986). "The Transportation of Soviet Energy Resources." Rowman & Littlefield, Totowa, NJ.
- Smil, V. (1988). "Energy in China's Modernization." Sharpe, Armonk, NY.
- Smil, V. (1998). "Energies: An Illustrated Guide to the Biosphere and Civilization." MIT Press, Cambridge, MA.
- Solomon, B. D., and Ahuja, D. R. (1991). International reductions of greenhouse-gas emissions: An equitable and efficient approach. *Global Environ. Change* 1, 343–350.
- Wilbanks, T. J. (1994). Sustainable development in geographic perspective. *Ann. Assoc. Am. Geogr.* 84(4), 541–556.
- Zeigler, D. J., and Johnson, J. H. Jr. (1984). Evacuation behavior in response to nuclear power plant accidents. *Professional Geogr.* 36(2), 207–215.

Geopolitics of Energy

AMY MYERS JAFFE

Rice University

Houston, Texas, United States



-
1. Introduction
 2. The Stubborn Reality of the Middle East
 3. World Demand Trends
 4. Enter Russia
 5. The Importance of U.S. Growth to Oil Producers
 6. OPEC Capacity Expansion
 7. Consumer Country Power in a Deregulated Setting
 8. Consumer Country Strategies

Glossary

Anthropologic Belonging to the scientific study of the origin, behavior, and physical, social, and cultural development of humans.

biomass Plant material, vegetation, or agricultural waste used as a fuel or an energy source.

carbon intensive Having a high content of carbon, a primary element that when burned produces carbon dioxide, which in great amounts can produce changes to the global climate.

externality An intended or unintended cost resulting from a government policy or market operation, coming or obtained from an independent source.

fossil fuel Fuel formed from the geological past and dug out of the earth.

monopoly Exclusive possession in the trade of some commodity.

monopsony Near exclusive possession in the trade of some commodity held by a small group.

Organization of Petroleum Exporting Countries (OPEC) A cartel of oil-producing countries that work together as a group to regulate oil market prices.

renewable Able to be renewed; not depletable.

sustainable productive capacity The daily amount of oil that an individual oilfield or a group of oilfields can produce at a rate that can be sustained for more than 90 days.

World energy use is expected to increase dramatically over the next three decades, reaching some 15.3

billion tons of oil-equivalent (toe) by 2030, up from 9.2 billion toe in 2000, according to the International Energy Agency's *World Energy Outlook 2002*. This increase is the equivalent of two-thirds of current demand and will be driven by strong economic growth in the developing world, especially Asia, and by population growth, which is expected to expand by one-third during the same time period. The developing countries' share of world demand will rise from roughly one-third today to 40% by 2030. Meeting such huge requirements will be a major challenge, requiring massive investment at every stage of the energy supply chain, much of it in developing countries that currently lack adequate resources to meet their current needs. Mobilization of this investment will be a major test for policy-makers into the future.

1. INTRODUCTION

More than a quarter of the world's population has no access to electricity, and close to half of the world's people rely on traditional biomass—wood, crop residues, and animal waste—as their main cooking and heating fuels. This situation is unlikely to change, barring the development of a cheap breakthrough energy technology. Current projections are that fossil fuels will remain the primary sources for energy, at approximately 90% of the increase of demand, for at least the next three decades, potentially with dire environmental consequences. Energy emissions of carbon dioxide (CO₂) are expected to grow roughly in line with energy demand unless there is faster deployment of more efficient and cleaner technologies or a larger than expected, policy-driven shift toward less carbon-intensive fuels.

According to projections made in 1993 by the Australian Bureau of Agriculture and Research Economics (ABARE), the growth in CO₂ emissions

for the developing world is projected to increase by five times that for the developed world in the next decade or so. Currently, the developed countries emit approximately 60 to 70% of total anthropogenic CO₂ emissions and are also responsible for 80% of current concentrations. During the coming decades, the contribution to emissions from the developing world will gain pace, even in the event that Western industrial countries begin to curb emissions in line with the Kyoto accords. In particular, the per capita emissions of four key developing countries—China, India, Indonesia, and Brazil—are expected to grow rapidly during the coming decade and beyond. In 1995, the per capita emissions of these four countries represented only 5 to 10% of the per capita emissions of the United States and were significantly lower than the world average. Over the long term, rising energy use in these countries can be expected to have a dramatic effect on global greenhouse gas (GHG) emissions (Table I).

The desire to wean the global economy—industrial and developing worlds alike—off of fossil fuels to safeguard the environment and provide a fuel source affordable across the globe poses many barriers, including technical cost, available infrastructure, and long lead times for new investments. Many forecasts predict a lowering of the growth in coal demand during the coming three decades as some societies, such as China and India, favor a switch to other sources (e.g., natural gas) for environmental reasons. But coal remains in ample supply in much of the developing world, and production costs are low. World coal reserves are enormous and, compared with oil and natural gas, are widely dispersed. Faced with the insecure nature of imported oil supply and potentially rising prices,

even countries seeking environmental improvement may find themselves unable to move away from heavy reliance on coal in power and industrial sectors. As natural gas and crude oil prices have risen during the past few years, a reemergence of rising coal use has been registered. So, although it is expected that coal use will wane during the coming decades for environmental reasons, such forecasts may be deemed too optimistic (Fig. 1).

Environmental advocates would like to see global energy politics move in the direction of renewable fuels such as solar, wind, and hydropower. According to the International Energy Agency's (IEA) *World Energy Outlook 2002* (WEO 2002), whereas natural gas is expected to see its world share rise from 23% today to close to 30% by 2030, renewable energy is unlikely to capture a large part of the world energy market during the coming two decades or so (Fig. 2).

The reality is that even with massive investment, renewable resources are unlikely to overtake fossil fuels to any significant extent during the coming decades. Within the "green" community, advocates would like to see the relationship between oil and renewable energy flip during the next 50 years or so such that oil becomes a marginal fuel and renewable energy becomes the major supply source (Fig. 2). But such a green scenario would require massive government intervention and incredible investments of capital to level the current playing field, where fossil fuels maintain a pivotal price advantage in most markets. Thus, without major changes from technological advances or policy intervention, investment in renewables, although highly beneficial to environmental concerns, is unlikely to emerge at large enough rates to make a significant dent in rising demand for oil and natural gas.

TABLE I

Projected Emission Levels in 2010 (Millions of Metric Tons of Carbon)

Country	Year	Residential/ Commercial	Transport	Industrial and other	Final consumption	Primary requirement	Per capita metric tons of CO ₂
China	1995	131.77	36.97	487.90	656.64	798.17	2.44
	2010	222.42	92.55	826.24	1141.21	1387.48	3.77
India	1995	18.14	32.77	85.95	136.86	228.28	0.90
	2010	39.14	60.23	153.78	253.15	422.49	1.37
Indonesia	1995	9.91	12.55	14.48	36.94	59.39	1.13
	2010	17.68	21.64	24.47	63.79	102.59	1.59
Brazil	1995	4.27	25.70	26.97	56.94	64.68	1.49
	2010	8.17	43.26	42.03	93.46	110.24	2.13

Source. Hartley, Medlock, Warby: Rice University, Baker Institute.

Note. U.S. per capita emissions in metric tons of CO₂ were 19.13 in 1992 and 19.69 in 1996.

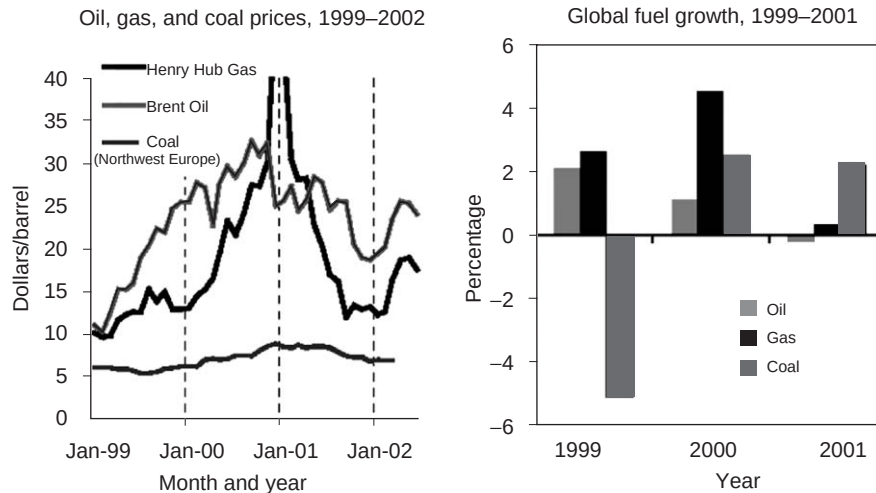


FIGURE 1 Graphs illustrating the reemergence of coal through the process of rising demand. From BP.com.

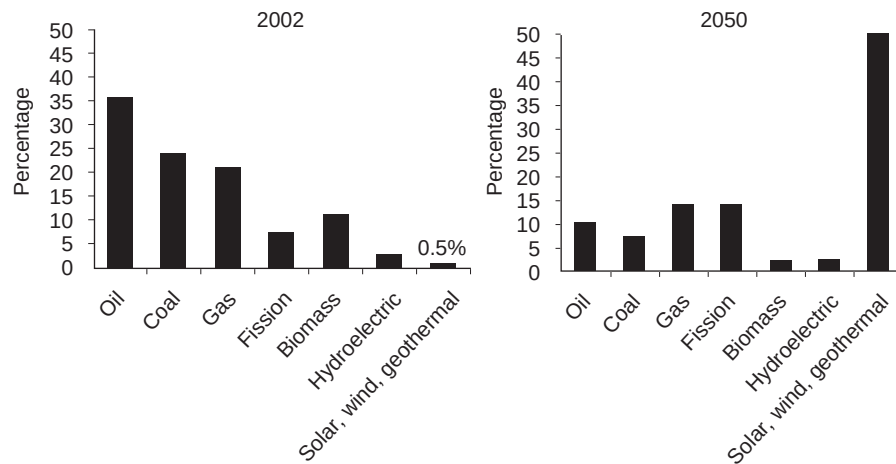


FIGURE 2 Bar graphs depicting the environmental vision for 2050. From Rice University.

For example, the share of nonhydro renewable resources in Organization for Economic Cooperation and Development (OECD) electricity generation, despite relatively proactive green policies by world political standards, is expected to rise from 2% in 1997 to only 4% in 2020 under current programs and is likely to achieve only a 9% share with significant government intervention. According to the IEA's *World Energy Outlook 2001* (WEO 2001), to accomplish this 9% share would require new investments of \$228 billion and would result in a reduction of CO₂ emissions of only 6% over the period. Wind power capacity has grown by more than 70% since 1997 but still represents less than 1% of the world's total installed electricity capacity, and 69% of world wind facilities are located in Europe.

Nuclear power could be a significant contributor to the long-term global sustainable energy mix, but

to play this role successfully, nuclear power will have to overcome serious economic and political challenges, including concerns over nuclear waste, safety, and nonproliferation. In many countries where nuclear power plays a major role, governments are exiting the generation business, with some exceptions (e.g., Finland, South Korea). As electricity markets become increasingly deregulated, nuclear facilities will have to compete on commercial grounds with generators using other fuel sources, and this may be increasingly difficult given the long lead times for siting, design, and other regulatory approvals and grassroots plant costs.

Although third-generation nuclear reactors will continue to play a major role for the next decade or so, the future rests with new fourth-generation designs that might be able to provide safer, more reliable nuclear power through smaller modular units that

will better conform to commercial needs and competitive pressures. In the meantime, until nuclear power is able to offer a competitive challenge to traditional fuels, it is assumed that the bulk of the increase in energy demand will be met with a rise in oil and natural gas production. According to *WEO 2001*, global oil demand is expected to rise by approximately 1.6% per year, from 77 million barrels/day in 2002 to 120 million barrels/day by 2030. Near three-quarters of this increase in demand will come from the transport sector, where renewable energy and nuclear energy are not expected to play a significant role, barring a major technological breakthrough.

2. THE STUBBORN REALITY OF THE MIDDLE EAST

Much ink has been spilled chronicling the world's rising dependence on Middle East oil. According to *WEO 2001*, more than 60% of the world's remaining conventional oil reserves are concentrated in the Middle East. A quarter of these reserves sit in Saudi Arabia alone. The Middle East is currently supplying more than one-third of world oil demand (Fig. 3).

This percentage could rise significantly in the future, depending on policies in consumer countries and on the pace of development of new resources and technologies. The IEA forecasts that the Middle East share of world oil supply will rise to more than 40% by 2020 under its *World Energy Outlook 2000* (WEO 2000) reference case. The U.S. Department of Energy, in one forecast, even predicts that the need for OPEC oil could rise from 28 million barrels/day in 1998 to 60 million barrels/day in 2020, with the majority of supply coming from the Middle East, especially Saudi Arabia.

Iran, Iraq, Syria, Sudan, and Libya currently produce approximately 8 million barrels per day or roughly 10% of the world oil supply. Saudi Arabia

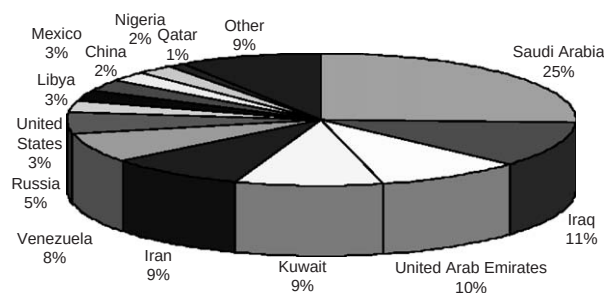


FIGURE 3 Pie chart depicting countries containing proven world resources in the year 2000.

alone is responsible for nearly 10% of the world oil supply and holds a unique position in oil markets. It maintains the largest share of spare idle production capacity of any other nation in the world. The kingdom is the only oil producer in the world that can single-handedly replace, within a short period of time, the total loss of exports for any other oil producer on the globe. No other nation currently has enough spare capacity to claim this role. Saudi Arabia is also the world's largest exporter, in past years selling nearly 100% more than its next largest export competitor, Russia (Fig. 4).

Saudi Arabia's cushion of spare capacity, which represents more than two-thirds of all global spare capacity, has provided security and stability to world oil markets for two decades. The kingdom derives its international clout from this custodial role and is unlikely to relinquish it, especially in the face of weakening oil market trends. The kingdom must worry about market share considerations in the face of rising investment in Russia, Africa, Iran, and Canada as well as long-term economic considerations. Oil revenues are critical to the health and well-being of the Saudi economy. Therefore, Saudi Arabia can ill afford to jeopardize its coveted position as a stable and reliable supplier from which its geopolitical influence and long-term economic interests derive.

3. WORLD DEMAND TRENDS

World oil demand is expected to rise at a rate of roughly 1.5 to 2.0% per annum over the next two decades, from about 76.4 million barrels/day in 2001

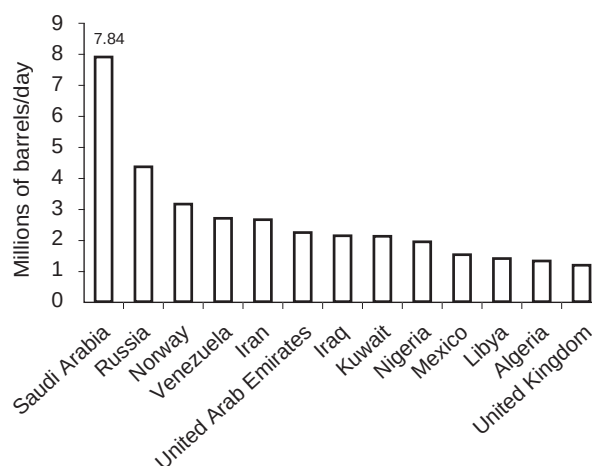


FIGURE 4 Bar graph depicting top world oil net exporting countries for the year 2000.

to 89 million barrels/day in 2010 and 104 million barrels/day by 2020, according to *WEO 2002*. During this same period, Asian demand is anticipated to increase from 28 to 35% of total world demand. Asia's oil consumption, at more than 19 million barrels/day, already exceeds that of the United States. Currently, only approximately 40% of Asia's oil supplies are produced inside the region.

By 2010, total Asian oil consumption could reach 25 million to 30 million barrels/day, of which 18 million to 24 million barrels/day will have to be imported from outside the region. Asian oil demand averaged 21.5 million barrels/day during the first half of 2003, whereas local oil production totaled only some 6.2 million barrels/day, leaving a gap of more than 15.3 million barrels/day that was met by imports, mainly from the Middle East and to a lesser extent from Africa, according to the *Petroleum Intelligence Weekly* database. This is up from a gap of more than 11 million barrels/day in 1998. As Asia develops, oil demand is expected to grow two to three times more rapidly than in the industrialized West, reaching approximately 29 million barrels/day by 2010, according to the "business as usual" forecast scenario of the IEA. Regional oil supplies are not expected to keep pace in any corresponding fashion. The IEA forecasts that Asian oil production will average only 6.4 million barrels/day by 2010, only a modest gain from current levels. In a study by Jaffe and Barnes, analysts expected somewhat larger gains of 1 million barrels/day or so but far less than the 6 million to 11 million barrels/day projected gain in demand.

More than 60% of Middle East exports already go to Asia and nearly 70% of all Asian imports come from Middle East oil producers. Some 84% of all crude oil refined in Singapore comes from the Middle East, whereas 78% processed in Japan originates from that region. The IEA projects that the Asia Pacific region will be importing 20 million to 24 million barrels/day from the Middle East by 2020. These oil linkages could present new political and strategic challenges in light of the emerging geopolitics of the post-September 11, 2001, period. The defense of oil supplies from the Persian Gulf is now achieved nearly single-handedly by the United States. Although it is hard to predict how the U.S. "War on Terror" will affect this role during the coming decades, expanding oil ties between major Asian powers and Middle East exporters will certainly influence East Asian strategic thinking, especially in relation to regional sea lanes issues. The result is likely to be some alteration in big power relations in

East Asia during the coming years that might test traditional relationships.

Saudi Arabia and other Middle East oil producers have already grasped the growing importance of the Asian markets and have been building bridges to the region. In July 2001, Saudi Arabia overtook the United Arab Emirates as Japan's primary supplier of oil. Saudi Arabia's exports to Japan grew by 20.7% to 27.40 million barrels that month, whereas the United Arab Emirates' exports to Japan dropped by 19.7% to 26.25 million barrels, according to a *Kyodo News Service* report. Pricing flexibility will be key to the kingdom's strategy to win over any market to which it lays claim.

China's current refinery configurations are currently not suitable to refine large amounts of the lower quality supplies that are produced in Persian Gulf countries such as Iraq, Iran, Saudi Arabia, and Kuwait. This explains the relatively small amount of crude imports from these countries (Fig. 5). Without refinery upgrades, China is projected to have only a little more than 1.35 million barrels/day of capacity to process this lower quality Persian Gulf oil by 2005 (Table II).

However, low-sulfur supplies from Yemen and Oman are expected to fall during the coming years as production from older reservoirs begins to decline. Thus, China is seeking light crude oil supplies from other sources, including Africa, and is seeking to make new upstream investments to lock up such supplies. The state oil company, CNOOC, secured Indonesian supplies recently with its \$585 million purchase of the Widuri and Cinta fields in Indonesia from Repsol-YPF, thereby adding 100,000 barrels/day of light crude oil that can be processed quickly through China's existing refinery capacities.

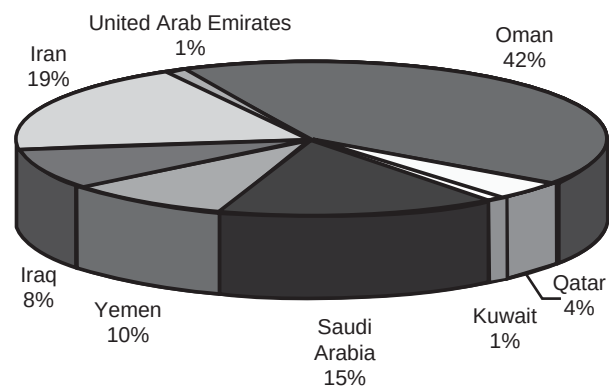


FIGURE 5 Pie chart depicting Chinese oil imports from Middle East countries for the year 2000.

TABLE II

China Refinery Capacity by Crude Tolerance (Thousands of Barrels/Day)

Crude Type	1998	2002	2005
Sweet: <0.9% S	4350	4050	3700
Medium sweet: 0.90–1.15% S	160	550	750
Sour: 2% S or higher	240	600	750–1350
Total	4750	5200	5200–5800

Source. Asia Pacific Energy Consulting. "China's Oil Strategy towards the Middle East." Working paper, APEC.

SINOPEC is expected to follow suit and has been negotiating for an exploration deal in Nigeria. Chinese state firms are also seeking exploration and production deals in other African states as well as Libya.

To protect its piece of the growing Chinese market, Saudi Arabia has shown a willingness to help China upgrade its refining capacity to use more Saudi crude oil. In late 2001, Saudi Aramco, ExxonMobil, and Fujian Petrochemical took another step toward the ultimate completion of a 240,000-barrels/day upgrade of a major Fujian refinery by signing an agreement for a joint feasibility study.

In addition, Saudi Aramco is now the largest shareholder in the Thalín refinery and has been negotiating with SINOPEC to expand the refinery at Maoming. Also, China has made a deal for 10 million tons of Saudi oil annually for a 50-year period, with SINOCHEN processing part of the oil at the new joint venture Qingdao refinery and the rest at its Dalian facility, signaling a close and growing import/export relationship between the two countries.

4. ENTER RUSSIA

Mirroring an important international trend, Saudi and other Middle East producers will find competition for growing Asian markets from Russia, which in the aftermath of the September 11 attacks has made clear its intentions to increase oil and gas production with an eye to enhancing exports to the West and Northeast Asia.

Russian firms have been able to marshal capital on their own to revive the Russian oil industry, and Russian oil production rose by more than 640,000 barrels/day from September 2000 to September 2001. The following year, Russian production rose nearly 600,000 barrels/day, from 6.95 million

barrels/day in December 2001 to 7.55 million barrels/day in August 2002. Russian production continued to rise to 8.385 million barrels/day by September 2003. Moscow's political leaders, as well as leaders from Russia's corporate oil and gas sector, are portraying the country's oil firms as stable sources of supply, willing to add output to the market to sustain reasonable prices. In the view of Russian leaders, this new geopolitics of energy "can help Moscow gain both economically and politically," according to Morse and Richard.

If concrete plans of the Russian oil firms and their Central Asian counterparts come to fruition during the coming years, rising oil exports from the former Soviet Union could rise by another 2 million barrels/day during the next 5 years or so and move closer to rivaling those of Saudi Arabia. This threat has grabbed the imagination of Middle East policy-makers inside OPEC, who called on Russia late in 2001 to cease and desist from stealing OPEC's market share. Moscow initially responded with conciliatory talk toward the end of 2001 and again in March 2002, but Russian oil production rose 600,000 barrels/day during the first three quarters of 2002 versus 2001 levels. And Russian officials have publicly announced Moscow's intentions to increase its oil production during the coming years to help the West diversify its oil supply. Russian government officials and private company chairmen also posited the possibility of Russia establishing an international strategic stockpile of Russian oil that could be used to serve countries that are not currently able to amass strategic stocks. The Russian stockpile would be used to supply markets during times of market emergency and to dampen price run-ups.

Four Russian oil giants have announced plans to build a proposed \$2 billion pipeline and deepwater year-round port at Murmansk, in the north of Russia, that could be used to export oil in very large crude carriers (VLCCs) to distant markets such as the United States. A feasibility study of the port project, which is slated to handle up to 1 million barrels/day of exports, is being financed by Lukoil, Yukos, Sibneft, and TNK. The project has been considered for U.S. Exim Bank support. Russian companies see the U.S. sales as a means to diversify its markets so as to avoid oversupplying Europe with growing sales from the former Soviet Union.

Traditionally, Russian oil exports have been bound for Europe and have not confronted Middle East oil producers' efforts to build market share in the United States and Asia. But more recently, there has been a political push in both Moscow and Beijing

to facilitate growing cooperation between Russian and Chinese firms in oil and gas investments and trade. A similar warming of relations on energy trade issues has taken place between Russia and Japan.

The view from Beijing now is that Russia is a key future hydrocarbon supplier. China hopes to receive supplies from new investments by Western firms at Sakhalin and from new field developments within Russia, especially by Yukos. A natural gas pipeline plan involves some 9.5 billion cubic meters (bcm)/year of hydrocarbon exports to Japan and potentially Korea and China from the Sakhalin Islands. Within the next 5 years or so, Russian supply to East Asia from the Sakhalin oilfields under current development are also likely to provide upward of 500,000 barrels/day of incremental supply.

Beyond Sakhalin, there are other prospects from East Siberia, where Yukos, the Russian firm holding perhaps the largest potential reserves in the area, is already in discussion with CNPC/Sinopec on joint projects.

5. THE IMPORTANCE OF U.S. GROWTH TO OIL PRODUCERS

So much attention has been paid to the increase expected from Asian markets that analysts often forget to note the continued importance of the U.S. market to OPEC. U.S. net oil imports rose from 6.79 million barrels/day in 1991 to 10.2 million barrels/day in 2000, while global oil trade (i.e., oil that was exported across borders from one country to another) rose from 32.34 million to 42.67 million barrels/day. The U.S. share of the increase in global oil trade over the period was a substantial 33%. In OPEC terms, the U.S. import market was even more significant—more than 50% of OPEC's output gains between 1991 and 2000 (Table III).

Saudi Arabia, in particular, has been keenly aware of the importance of the U.S. market and careful to keep nagging political issues between the United States and the kingdom on a separate track from oil policy. The kingdom has been diligently maintaining its U.S. market share during recent years and threw down the gauntlet in 1997 to punish Venezuela for challenging its grip on the important U.S. market. Following the oil price collapse of 1997–1998, newly elected Venezuelan President Hugo Chavez rescinded state oil firm PDVSA's plans to expand capacity to 7 million barrels/day by 2007. Instead, Venezuela announced revised plans to grow its production

TABLE III
Importance of U.S. Growth to Oil Producers

	1991	2000
U.S. and Asia Pacific share of world oil demand		
United States	24%	25%
Asia Pacific	22%	28%
U.S. net oil imports as a share of global oil trade		
Total	32.34 million barrels/day	42.67 million barrels/day
Net U.S. imports	6.79 million barrels/day	10.20 million barrels/day
U.S. share of increase in global oil trade, 1991–2000	33.02%	
U.S. share of increase in OPEC's output, 1991–2000	55.58%	

Source. Morse, E. (2002, February 21). "The Price of Crude Oil: Markets, Politics, and Geopolitics." Lecture, Columbia University.

more gradually, to 5.5 million barrels/day by 2006. However, prolonged political strife inside the country and a major oil workers strike in 2003 not only discouraged further investment by foreign companies (which were already discouraged by rising tax rates) but also left much of Venezuela's oil industry infrastructure damaged, lowering its potential to produce and refine crude oil. Venezuela's oil production capacity was already dropping during recent years given constraints on investment capital and a tough geological decline curve at aging fields, but damage sustained during the oil workers strike and near civil war in the country led to a more rapid reduction in sustainable capacity to 2.2 million barrels/day at the end of 2003, down from a peak of 3.7 million barrels/day in 1998.

Saudi Arabia may face a new challenge to its U.S. market share from a regional exporter, but this time north of the border. Analysts for the kingdom's oil industry say that they are watching with concern the rising output profile for Canada. The start-up of new tar oil sands projects and other offshore projects has put Canada on track for the largest jump in oil output in several years, despite lower capital spending by industry that is often cited as evidence that production will decline. Canada's production has increased from 2.155 million barrels/day in 2000 to 2.3 million barrels/day in mid-2003.

Companies have spent more than \$7 billion in oil sands projects since 1996 and are expected to

allocate an additional \$25 billion by 2007. Production, which has already reached 800,000 barrels/day, could rise to 1.5 million barrels/day or more by 2010 if currently proposed projects meet their targets (Table IV).

The potential for Canadian heavy crude and oil sands development is huge. Of the 2.5 trillion barrels of crude bitumen (oil sands) resources in place in Canada, approximately 12% (or 300 billion barrels) is thought to be recoverable, a figure comparable to the proven reserves of Saudi Arabia. This is in addition to extra heavy oil resources. In Venezuela, where some U.S. oil companies are working on heavy crude upgrading projects, more than 1.2 billion barrels of bitumen are thought to exist.

During past decades, the high costs of mining and processing tar sands and extra heavy crude oil were somewhat prohibitive at approximately \$21 per barrel. But engineering and technological advances are lowering costs rapidly. Analysts estimate that oil sands projects "all in" costs (i.e., investment and operating) have fallen to as little as \$8 to \$9 per barrel and could fall to \$6 to \$7 per barrel during the coming years as industry continues to make new advances. However, cost variables in Canada have been somewhat unstable given a tight labor market, and cost overruns have been seen at several of Canada's top existing oil sands projects, including Syncrude, Athabasca, and Millennium.

African supply may also become a key focus of efforts by the West to diversify away from the Middle East. Of the 14.5 million barrels/day of exports projected to be needed from Africa or the Middle East by 2010 in the Atlantic Basin market to fill its supply gap, at least half will be able to be sourced from Africa and North Africa. Angola, a non-OPEC producer, is projected to see increased production, from 750,000 to more than 2 million barrels/day by

2010, as new fields are developed in its deep offshore zone at full-cycle costs of between \$3 and \$5 per barrel. Chinese national oil companies are also showing interest in African upstream investment, with CNPC already firmly planted in Sudan and Chinese companies showing expanded interest in Nigeria, Chad, Niger, Equatorial Guinea, Algeria, Tunisia, and Libya.

6. OPEC CAPACITY EXPANSION

Over the past few years, there has been a dramatic shift in the internal politics of OPEC, reflecting changes of political leadership at the highest levels and of broader policies of key members. The result has been greater cohesion inside the producer group and a clearer articulation and implementation of goals and aspirations.

The mid-1990s were characterized by OPEC disunity and overproduction, so much so that the cartel members themselves were questioning the future viability of the organization. Venezuela's ambitious campaign to increase oil productive capacity from 2.8 million barrels/day in 1991 to 7 million barrels/day by 2010, combined with the gradual rise in Iraqi oil export rates through the auspices of the UN Oil for Food program, made it nearly impossible for the producer group to agree on a workable production sharing agreement.

During this period of the mid-1990s, market share considerations were paramount for most OPEC countries with spare capacity. Saudi Arabia, in particular, had to concern itself with both the short- and long-range implications of Venezuela's market expansion plans, which were expected to have a significant bearing on the kingdom's ability to maintain its sales to the United States. Venezuela's government was committed to supplementing state revenue by increasing oil exports and in 1992 began a new policy to allow international oil companies (IOCs) to invest in the country's once nationalized oilfields. This campaign contributed to a rise in Venezuelan output, from roughly 2.5 million barrels/day through the 1980s to between 3.5 million and 3.7 million barrels/day by 1997. Venezuela refused to comply with any OPEC production sharing agreements, greatly damaging the producer group's ability to manage oil markets and providing a strong disincentive for Saudi Arabia to contribute a large cutback in output to defend oil prices.

This political stumbling hit a dramatic snag in during 1997–1998 when an unexpected economic

TABLE IV
Approved Canadian Oil Sands Projects

Project	2001	2002	2005	2010
Syncrude	225	250	350	465
Suncor	130	220	260	400
Shell/Chevron	—	155	155	155
Albian/Koch/T.N.			95	190
Canadian Natural Resources				300
Gulf Resources				100 ^a

Source. *Petroleum Argus* newsletter.

^aProposed for 2013.

meltdown hit much of Asia and precipitated an unexpected drop in Asian oil use. Asian oil demand fell to 18.18 million barrels/day in 1998, down from 18.53 million barrels/day, a drop of 1.9%. This compared with a 4.5% rise in Asian oil demand between 1996 and 1997. The slowdown in Asia contributed to a major change in the global oil supply–demand balance, limiting the growth in worldwide oil demand to 74.24 million barrels/day in 1998, up only 0.5% from 1997 compared with more typical annual growth rates of 2 to 3% during the mid-1990s (Table V).

By 1997–1998, these circumstances led oil prices to collapse below \$10, requiring extraordinary efforts by both OPEC and non-OPEC countries such as Mexico and Norway. The intense financial suffering of nearly all OPEC countries and key changes in government leadership in Venezuela paved the way for a major agreement among oil producers to trim output and propel oil prices back above \$22 per barrel. OPEC set a price band of \$22 to \$28 per barrel as its target that remained in place as of this writing.

OPEC's new, more unified dynamic is rooted in several factors:

- A rise in democratization, freedom of the press and political debate, and a growing tide of anti-Americanism are bringing a greater concern for popular opinion inside OPEC countries, especially in the Middle East Gulf region, than was the case in the past. This new concern for popular sentiment is restricting the options of regional leaders to accommodate Western interests. Populations, as well as some leaders, remain bitter about the suffering that took place when oil prices collapsed during the late 1990s.
- Rising populations and economic stagnation in many OPEC countries indicate that revenue pressures have been taking precedence over other considerations.
- Lack of investment in infrastructure and oilfields over the years due to tight state treasuries and rising social pressures has greatly curtailed OPEC's spare productive capacity, rendering it much easier for cartel members to agree to restrain output, at least in the short term.

Given this new dynamic and colored in part by the rise of Chavez as president of Venezuela, OPEC rhetoric during recent years has taken a more strident turn, with oil ministers defending their choices to attain a “fair” price for OPEC oil despite ramifications for global economic growth rates. The debate

TABLE V

Oil Demand per Region, 1996–1998 (Million Barrels/Day)

	1996	1997	Percentage change	1998	Percentage change
World	72.03	73.9	2.6	74.24	0.5
United States	18.25	18.62	2.0	18.7	0.4
European Union	13.03	13.10	0.5	13.39	2.2
Asia	17.74	18.53	4.5	18.18	−1.9
China	3.55	11.5	11.5	3.95	−0.3

has become one of economic struggle for “rents” between oil producers who demand high revenues and major consumers whose economies can grow faster with low oil prices.

OPEC's rhetoric has generally been directed at OECD consumer governments that capture rents from oil sales through high national energy taxes. OPEC's anti-tax, anti-Western rhetoric comes against the backdrop of popular domestic sentiment inside OPEC countries that their governments are not doing enough to deliver economic benefits to a substantial portion of their populations. Leaders in OPEC countries cannot be seen as delivering benefits to Western consumers at the expense of their own citizens because such perceptions would leave regimes more vulnerable to public attack and more susceptible to the efforts of opposition parties and groups.

Although political factors have contributed to OPEC's unity over the past 2 years, the more limited amount of spare capacity that has to be held back by most member states is also a supportive factor. Indeed, oil prices tend to track changes in OPEC spare capacity quite closely (Fig. 6).

In 1985, when oil prices collapsed, OPEC was estimated to have some 15 million barrels/day of shut-in production capacity, equal to perhaps 50% of its theoretical capacity (Iran and Iraq were at war with one another at the time) and 25% of global demand. By 1990, when Iraq invaded Kuwait, spare capacity globally was still 5 million to 5.5 million barrels/day, the amount of oil taken off the market by the UN embargo. That was approximately 20% of OPEC's capacity, and 8% of global demand, at the time. During the winter of 2003, before OPEC's seasonal cuts, spare capacity was a negligible 2% of global demand, a level that left oil markets highly susceptible to disruption and supported a run-up in prices to more than \$30 per barrel.

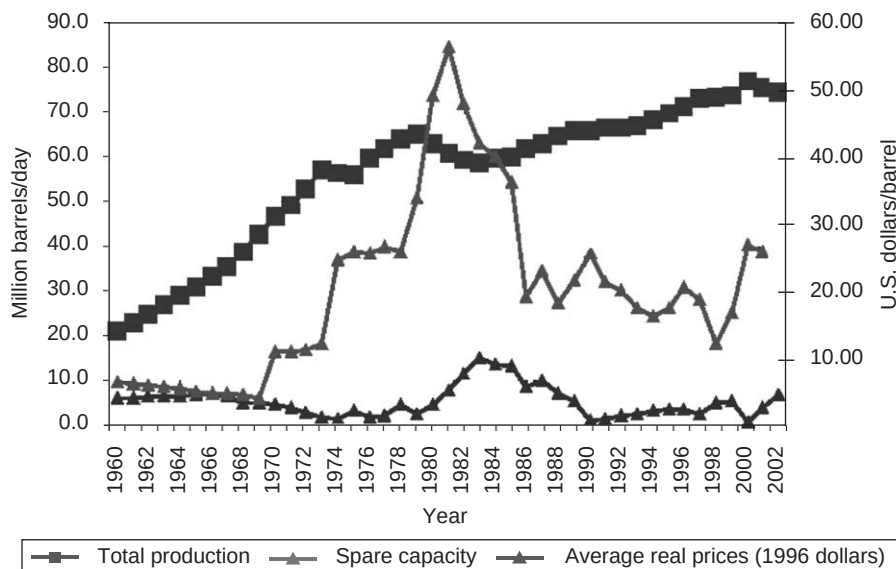


FIGURE 6 Line graph depicting the relationship between oil prices and the OPEC spare capacity. From Hetco Trading.

OPEC has been helped during recent years by irregularity in Iraqi production rates given the regime in Baghdad's aggressive posture to use the withdrawal of its UN-monitored oil exports as a political tool. In 2003, the U.S. military campaign in Iraq greatly curtailed the country's oil exports when facilities were looted and sabotaged in the immediate aftermath of the war. Iraq's export rates fell from just over 2 million barrels/day prior to the U.S. campaign to approximately 500,000 barrels/day in late 2003. In addition, other OPEC members have also benefited from a drop in capacity in Venezuela and Kuwait, both of which have been struggling against technical problems and natural declines in key fields. Venezuelan oil production capacity has fallen from 3.7 million barrels/day in 1998 to approximately 2.2 million barrels/day currently. Kuwait has lost capacity in certain western and northern fields such as Raudhatain, limiting its output capacity by several hundreds of thousands of barrels per day. Sustainable Kuwait capacity has recently been pegged at 2 million barrels/day.

However, the challenge to OPEC will increase during the coming years as OPEC member countries begin to anticipate an expected restoration of Iraqi capacity over time. Several individual OPEC countries have a backlog of new fields that can be brought on line soon. Several countries, such as Nigeria, Algeria, and Iran, have been quietly expanding capacity and anticipate possible production rises over the next few years. Countries such as Algeria

and Nigeria, whose new fields involve foreign oil company investment, will increasingly be under additional pressure from these IOC investors to obviate OPEC agreements and allow new fields to move forward at optimum production rates. Meeting these pressures will be difficult for OPEC if market demand for its oil does not increase significantly as expected due to a slowdown in economic growth or other inhibiting factors.

Iran is forging ahead with capacity expansions and could reach more than 5 million barrels/day by 2008 based on investments through buy-back agreements with IOCs and National Iranian Oil Company investment. However, this ambitious plan will be dependent on the success of major IOC field developments, including Azadegan, South Pars, Sirri, Gagh Saran, and the Ahwaz area. Nigeria has plans to expand production in its offshore area, and this could (if successful) raise production to more than 3.5 million barrels/day by 2006.

Prior to the U.S. campaign in early 2003, Iraq had commercial export capacity of approximately 2.2 million barrels/day via Turkey and the Gulf, with additional smuggling capacity through Syria, Iran, and Jordan. U.S. officials were hopeful that Iraq would be able to resume exports at prewar levels by late 2003. Iraq has expressed a desire to expand its productive capacity to 6 million barrels/day, and analysts believe that Iraq could quickly raise its oilfield productive capacity back to 3.6 million barrels/day if security can be restored across the

country and Iraq's upstream oilfield sector were opened to foreign oil company investment. However, for Iraq to be able to export these higher volumes, extensive repairs will be needed to both its oilfields and its export infrastructure.

There will be pressure inside OPEC to bring Iraq back into the production quota system as its production capacity is restored. Key members may insist that Iraq be initially limited to its historical allocation of a quota at parity with that of Iran and then hold its export increases to levels in line with market demand. However, Iraq will have pressing economic reconstruction needs to take into consideration as well, raising questions, depending on the future politics inside the country, as to whether it will be in a position to opt for policies that emphasize solidarity with OPEC.

OPEC's current capacity is estimated at 29.5 million barrels/day but could rise to as much as 36.4 million barrels/day by 2005 if new fields under development come on-line as expected. If OPEC were to expand access to foreign IOC participation, particularly in key countries now off limits to Western investment, capacity could expand to as much as 44 million barrels/day by 2010, leaving very little room for growth from non-OPEC supplies to take place without pressuring oil prices over the next 8 years or so.

7. CONSUMER COUNTRY POWER IN A DEREGULATED SETTING

As discussed previously, the United States is a major buyer of imported crude oil, and its imports represent a large share of the market for internationally "traded" oil. Given the large scale of U.S. purchases, incremental U.S. acquisitions of oil affect the overall international market price of oil. Stated another way, the cost of each marginal barrel is higher than the price paid for that barrel given that this additional purchase affects the costs of all oil consumed. From the perspective of the United States, this constitutes an externality in the sense that consumers will base their decisions regarding the quantity of oil (and oil products) purchased on the average, rather than the marginal, cost of oil. On the other hand, the fact that the United States faces a rising supply curve for oil gives it monopsony power. To the extent that the United States—or a group of consuming countries that include the United States or is of a comparable scale—take concrete actions to reduce the size of its purchases, it can lower the market price of oil.

It has been well established that OPEC frequently changes its price targets in response to changes in market demand. Discussion of the size of a monopoly (power) wedge that is the "difference between the current price of oil and the marginal cost of adding a barrel of demand" was the subject of an Energy Modeling Forum working group and other exercises.

OPEC's response to efforts by the OECD to reduce oil demand has varied over time. Initially, during the 1980s as the OECD demand receded, OPEC maintained its output path, letting oil prices decline. This result produced the highest buying power wedge size and reinforced the viability of consumer country policies that sought to exercise this monopsony power. However, OPEC can also maintain its price preference path and allow its market share to decline. This appears to be its strategy during recent years. But the continuation of the production sharing discipline needed to maintain this approach may become harder to sustain in the future.

Consuming countries have implemented two key approaches to dealing with OPEC given the increasing trend toward oil price deregulation inside the OECD. One is that governments have organized to hold strategic stocks of oil. The second is the imposition of consumer taxes on oil, thereby reducing demand for oil. As discussed previously, OPEC has engaged in a public relations battle against Western consumer taxes on oil.

Prior to deregulation, during the 1970s, crude oil was sold mainly through exclusive, long-term, fixed-priced contract arrangements with a handful of major suppliers. Crude prices throughout the world were regulated by national governments at fixed levels. Under this system, oil sellers could pose a clear threat to any individual buyer. A clear thread linked a disruption in supply to the shortfall of actual deliveries. A particular barrel lost in exports was neatly matched to a particular barrel lost to a particular buyer that had a contract for that barrel. Little recourse was available. Barrels were moved around the world like pieces on a chessboard, and there was no other *deus ex machina* to mitigate or alter the impact of any supply cutoff if one's particular barrels were lost. Lucky buyers whose supply chains remained intact could continue to receive their oil at prices locked by long-term fixed-price contracts at precrisis levels. In a word, some buyers were safe and others were not. This system created a psychology that enhanced OPEC's monopoly power and left individual consuming nations vulnerable to the possibility of a supply cutoff. The

oil crisis of 1973 confirmed fears that oil producers might use this leverage to adverse political ends.

Given the high costs of a cutoff under this fixed price system (i.e., regulated price system), large consumers, oil distributors, and middlemen had direct incentives to carry ample inventories of spare supplies. The possibility that oil from any particular supplier could be disrupted without recourse was reason enough to pay the costs to carry extra supplies on hand. Prices that could be passed on to consumers were in many cases inflexible and regulated, and competitors that were not subject to a cutoff would have a clear cost advantage.

Today's world is radically different. Crude oil prices have been "deregulated," that is, set by market forces through market auction practices that allocate supplies to the highest bidder. Price arbitrage will ensure that a price increase for one buyer or location will be a price increase for all buyers and locations.

In the new free-wheeling marketplace for oil, buyers cannot insulate themselves from price jumps by holding long-term contracts given that prices in these contracts are pegged to the spot market. But in today's market, one can always get the oil one desires, even in a major supply disruption, by simply paying a higher price.

At the end of the day, as the market clears, such as it did during the Iraqi invasion of Kuwait in 1990, buyers with imports from countries that were not subject to disruption are in no better position than buyers that had all of their imports subject to interruption. In an unfettered spot market, shortages will manifest themselves not in a physical inability to find supplies but rather in the rising price of oil. The market will "clear" or allocate available physical supplies to the highest bidder.

In this new commoditized market, commercial players such as distributors and middlemen have less incentive to carry inventory, especially if supply disruptions are infrequent. Higher costs during times of supply cutbacks can simply be passed on to end users, reducing the business risks that motivated the accumulation of commercial inventories previously. Although the existence of a large spot market trading at unregulated prices certainly appears to work against OPEC's ability to "blackmail" any particular buyer, structural reductions in commercial inventories can augment the producer group's power to the extent that alternative supplies from countries outside of OPEC are reduced.

The "public goods" aspect of oil market stability elevates the accumulation of oil inventories to a domain beyond the commercial industry. Individual

players in the oil market will not take into account the externalities of low commercial stocks such as increased energy price volatility and its negative distributive effects on poorer consumers or its aggregate negative macroeconomic effects. Thus, government plays a role in ensuring that adequate inventories are on hand to maintain orderly markets and to counter the temptation of suppliers with monopoly power from taking advantage of short-term tightness in oil markets. To achieve these ends, government-held stock levels must be credible to convince oil producers that efforts to exploit temporary market tightness by further cuts in production to achieve even higher oil prices will not be successful. Such attempts to extract additional rents from consuming countries would be countered by the release of sufficient inventories to offset any cuts in production contemplated by producers. The larger the government-held stocks and the more consuming governments that participate in such a stock holding program, the more effective it is likely to be in serving as a deterrent to OPEC's monopoly power in deregulated markets.

Market deregulation has left consumers more susceptible to the ill effects of energy price volatility. However, consuming country governments have been able to reduce the negative effects of price variability by increasing energy efficiency and reducing dependence on oil through the use of hefty consumer taxes on oil. The net effect of such taxes is to discourage wasteful use of energy by consumers while collecting some of the rents that would otherwise accrue to oil producers. Furthermore, large oil consumption taxes can force OPEC to accept lower prices, as occurred throughout most of the late 1980s and 1990s.

When OPEC's monopoly power strengthens due to short-term market tightening, the incentive to exploit that power is tempered by the fact that increases in monopoly rents will not accrue entirely to producers but rather must be shared with consuming countries that have high energy taxes.

8. CONSUMER COUNTRY STRATEGIES

OPEC's ability to raise prices in the short run to levels that may be damaging to the economies of major consuming countries poses a major policy challenge to the governments of consuming countries. Rising oil prices threaten consumer country national interest in several ways. Oil price volatility

can inhibit investment, inhibit economic growth, and spur inflation in major economies that purchase consumer country exports.

The burden of rising energy import costs also threatens social stability in key regional consuming countries such as India and Pakistan and the South-east Asia region. Moreover, supply constraints also make it easier for governments or subnational groups to threaten vital interests of the United States, Japan, and their allies.

Thus, consuming countries have a clear interest in undertaking policies that will undermine OPEC's short- and long-term ability to act as a cartel to inflate oil prices. Policies undertaken in conjunction with other consuming nations are likely to be more effective than policies undertaken individually by increasing the strength of the monopsony wedge.

Policy options include the following:

- Expand and restructure the IEA to better reflect the change in consumer country demand trends.
- Encourage investment in diverse non-OPEC oil resources.
- Develop and deploy alternative energy technologies.
- Develop federal policies to continue enhancing energy efficiency.
- Maintain or expand taxes on oil and gasoline.

The IEA was created a quarter of a century ago as a mutual protection society of OECD countries. Designed as a political grouping to prevent any oil-producing countries from using oil exports as a political instrument to influence the foreign policies of IEA members, the IEA was formed at a time when the OECD countries dominated global energy consumption. Today, the IEA excludes the most rapidly growing energy-consuming countries in the world, including China, India, and Brazil. And, as a result, these new consumers become vulnerable economically during times of disruptions and become potentially vulnerable to political pressures of producers.

During recent years, there has been discussion about increasing the number of member countries inside the IEA, and South Korea has joined the organization. Other countries, such as China and India, are investigating creating national strategic oil stockpiles. The IEA invited both China and India to participate as observers in recent meetings and is pursuing options for finding mechanisms for major non-IEA oil-consuming countries to participate in joint stockpiling emergency programs.

However, the extent of effectiveness of the IEA system will depend on oil market developments,

including Asian demand trends. The member countries of the IEA now represent a smaller portion of the oil market than they did at the time of the IEA's formation in 1977. As oil demand growth in Asia expands during the coming decade, new strains could come to the international system if new policies are not put into place. The omission of key consumer countries from Asia in the global emergency stockpiling system will increasingly put pressure on the effectiveness of limited existing stocks in the OECD countries. Moreover, tensions created by Asian "free riding" or possible "hoarding" actions during a crisis could hinder the IEA's ability to stabilize international oil markets in the future.

The OECD countries comprising the IEA accounted for 42.3 million barrels/day out of a total world oil use of 60.6 million barrels/day in 1977 or approximately 70% of world oil demand. The United States alone consumed 30% of the world's oil used in 1977. Asia Pacific demand at that time, at 10.1 million barrels/day or roughly 16% of world oil demand, was a less critical component in the world oil use situation.

By 2001, the OECD share of world oil use had declined to 62% of total world demand, while Asia Pacific use had grown to 28%, overtaking the U.S. share of 25%. Asian economic powers Japan, South Korea, Australia, and New Zealand are OECD members and, as such, are part of the IEA system now. A few Asian countries, such as Pakistan and China, have announced plans to create strategic stocks, but other key Asian oil consumers, such as India, Taiwan, Thailand, and The Philippines, are not. As their share of world oil demand grows, this disconnect between Asia's size and importance as a consumer region and its lack of energy policy coordination with other large oil-consuming countries (and/or the IEA) will create new problems and challenges for international oil markets and the international economic system. In particular, it is important that large Asian consuming countries not purchase and "hoard" oil during an IEA stockpile release because this activity would reduce the effectiveness of a stock release to calm markets and prevent oil supply shortages. The best possible win-win scenario would be for new links to develop between the IEA and other large consumer countries or consumer country groups.

Despite the obvious benefits from an Asian stockpiling group or system, Asian countries have been reluctant to commit to the expense involved in creating such a system. Although an Association of Southeast Asian Nations (ASEAN) delegation visited

Japan in 2000 to study the industrialized country's emergency stockpiling system, the cost per capita was considered "high" by many ASEAN delegates. Press reports indicate that ASEAN energy ministers have focused on security of sea lanes and international gas pipelines, rather than on joint stockpiling arrangements, at recent meetings.

Meetings between IEA officials and Chinese state planning officials have similarly failed to produce an agreement of coordination. The Chinese government has announced its intention to build a strategic petroleum reserve, but so far the government has failed to define the reserve's size or possible location publicly. It is considering requiring state oil companies to hold stocks comparable to 10% of their import levels. There has also been no official verification that China would like to work cooperatively with its Asian neighbors or OECD countries to develop measures to safeguard against disruptions in oil supplies from the Middle East. In March 2001, the State Council and the State Development Planning Commission included a "strategic petroleum reserve" as a key project in the 10th 5-year plan, but continued debate about its size, location, and funding among party leaders, State Council agencies, National People's Congress deputies, and the three central-owned state oil and gas companies has called into question the quick construction of such a reserve.

Although informal programs to encourage stockpiling by developing world countries would have a positive impact, such efforts cannot replace the more effective tool of centralized coordination with the IEA. Centralized efforts are needed so that international norms and standards can be met during a crisis. The United States and Japan are reviewing ways in which the IEA can work with key countries that are not members of the IEA to encourage them to define their strategic oil stockpile requirements and to build strategic stocks (or create minimum inventory requirements for industry). Russia has also made pronouncements recently, offering to establish a Russian strategic oil stockpile to be used to supply countries lacking in their own strategic stocks. This concept was aired at the U.S.–Russia Energy Summit in Houston, Texas, in October 2002.

In the post-September 11 climate, consumer governments are increasingly discussing enhancing development of backstop technologies or promoting alternative energy sources that can serve to reduce the need for fossil fuel. In this practice, backstop technologies create an incentive for oil producers to avoid oil price shocks and supply disruptions for fear

that the new technologies would be released and used, permanently eliminating sales markets. Alternative energy supplies provide ready substitutes if an increase in the price of oil is too extreme, and they can shield the economy from the negative impact from disruption of any one fuel source.

The deployment of improved car technology could have a dramatic effect on future oil demand trends as well as play a major role in lowering CO₂ emissions by advancing fuel efficiency. The benefits to the role that enhanced energy efficiency has played in protecting the domestic Japanese economy from oil price variability are well known. Japan did not experience a severe recession after the 1979–1980 price shock, whereas the United States, United Kingdom, and Germany, all of which were less energy efficient at the time, did experience a painful loss in economic output.

It has been shown that the lower a country's energy consumption to gross domestic product (GDP) ratio or the shorter the period that oil prices will remain higher, the lower the cost of the trade-off between inflation and GDP loss. New technologies on the horizon could allow more gains in energy efficiency. Such technologies may include microturbines for distributed power markets, improved car technologies, and household solar technologies. OECD governments should encourage the deployment of these technologies in the marketplace through tax incentives or other vehicles in an effort to reduce their individual exposure to OPEC's monopoly power.

SEE ALSO THE FOLLOWING ARTICLES

Economic Geography of Energy • Environmental Injustices of Energy Facilities • European Union Energy Policy • External Costs of Energy • Geographic Thought, History of Energy in • Global Energy Use: Status and Trends • Nationalism and Oil • OPEC, History of • World Environment Summits: The Role of Energy

Further Reading

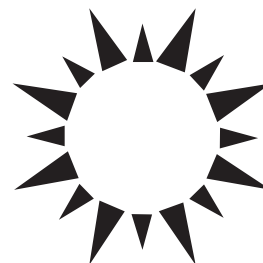
- Bohi, D. R., and Toman, M. A. (1996). "The Economics of Energy Security." Kluwer Academic, Boston.
- Chinese energy majors in long march upstream. (2002). *Petroleum Intelligence Weekly*, January 28, p. 3.
- Fujian Petrochemical, ExxonMobil, Aramco sign agreement for submission of Fujian project joint feasibility study. (2001). *Business Wire*, November 15.

- Gately, D. (1982). OPEC and the buying power wedge. In "Energy Vulnerability" (J. L. Plummer, Ed.), pp. 37–57. Ballinger, Cambridge, MA.
- Hartley, P., Medlock, K., and Warby, M. (n.d.) "First Things First: Development and Global Warming." Working paper. Baker Institute, Rice University.
- Huntington, H. G. (1998). Crude oil prices and U.S. economic performance: Where does the asymmetry reside? *Energy J.* 19(4), 107–132.
- Jaffe, A.M., and Barnes, J. (n.d.). "Energy Trade in Asia." Policy forum paper, Johns Hopkins University, School of Advanced International Studies.
- Jaffe, A. M., and Lewis, S. W. (2002). China's oil diplomacy. *IISS Survival* 44(1), 115–133.
- Jaffe, A. M., and Soligo, R. (2002). The role of inventories in oil market stability. *Q. Rev. Econ. Finance* 42, 401–415.
- Japan's crude oil imports fall for 4th month in July. (2001, August 24). *Kyodo News Service*.
- Manning, R. (2000). "The Asian Energy Factor." Palgrave, New York.
- Maugeri, L. (2003). Not in oil's name. *Foreign Affairs* 82(4), 165–174.
- Morse, E. L., and Richard, J. (2002). The battle for energy dominance. *Foreign Affairs* 81(2), 16–31.
- Ng, E. (2001, November 7). ASEAN not in favour of "expensive" oil stockpile. *Agence France Presse*.
- Russia's TNK pushes for new markets, oil reserve sites for urals. (2002). *Oil Daily*, October 4.
- Saudi Arabia's petrochemical industry diversifies to face challenges. (1999). *Oil and Gas Journal*, August 16, p. 65.
- Yergin, D. (1991). "The Prize." Simon & Schuster, New York.
- Yergin, D., and Hillenbrand, M. (Eds.). (1982). "Global Insecurity." Houghton Mifflin, Boston.

Geothermal Direct Use

JOHN W. LUND

Geo-Heat Center, Oregon Institute of Technology
Klamath Falls, Oregon, United States



-
1. Utilization
 2. Equipment
 3. Economic Considerations
 4. Future Developments

Glossary

agribusiness application Agriculture and aquaculture developments; in this article, it includes the heating of the various applications, such as greenhouses and fish ponds.

aquaculture pond/raceway heating Heat energy provided to ponds and raceway in order to optimize the growing of fish, shrimp, alligators, or other aquatic species.

balneology The use of mineral waters for therapeutic purposes.

direct use The use of the heat energy rather than its conversion to other forms such as electricity; for example, a furnace used for space heating.

district heating Energy provided to a group of buildings or even to an entire community, generally in the form of hot water or steam supplied from a central plant to customers through a pipeline distribution system.

energy cascading The multiple use of an energy source by various applications, with each successive one using an increasingly lower temperature in order to maximize the efficiency of the system.

geothermal energy Energy that can be extracted from the earth's internal heat. The heat is produced by the radioactive decay of thorium, potassium, and uranium that exist in the earth's molten core.

greenhouse heating Heat energy provide to a glass or plastic enclosure in order to optimize the growing of tree seedlings, flowers, and vegetables.

heat pumps A device used to transfer heat from a low-temperature resource to a high-temperature reservoir, thus providing the higher temperature for space heating. The process can be reversed to provide space cooling.

industrial application Manufacturing, machining, assembling, and producing products for consumption; in this article, it includes the heating of the various applications.

space heating Energy provided to a room and building for heating purposes.

Direct or nonelectric utilization of geothermal energy refers to the immediate use of the heat energy rather than to its conversion to some other form such as electrical energy. The primary forms of direct heat use are for swimming, bathing, and balneology (therapeutic use); space heating and cooling, including district heating; agriculture (mainly greenhouse heating and some animal husbandry), aquaculture (mainly fish pond and raceway heating), and industrial processes; and heat pumps (for both heating and cooling). In general, the geothermal fluid temperatures required for direct heat use are lower than those for economic electric power generation.

Most direct use applications use geothermal fluids in the low to moderate temperature range between 50 and 150°C, and in general, the reservoir can be exploited by conventional water well drilling equipment. Low-temperature systems are also more widespread than high-temperature systems (above 150°C), so they are more likely to be located near potential users. In the United States, for example, of the 1350 known or identified geothermal systems, only 5% are higher than 150°C; whereas 85% are lower than 90°C. In fact, almost every country in the world has some low-temperature systems, whereas only a few have accessible high-temperature systems.

1. UTILIZATION

Traditionally, direct use of geothermal energy has been on a small scale by individuals. Recent developments involve large-scale projects, such as district heating (Iceland and France), greenhouse complexes (Hungary and Russia), or major industrial use (New Zealand and the United States). Heat exchangers are also becoming more efficient and better adapted to geothermal projects, allowing use of lower temperature water and highly saline fluids. Heat pumps utilizing very low-temperature fluids have extended geothermal developments into

TABLE I
Summary of Regional Geothermal Use in 2000

Region	Direct use			
	MWt	%	GWh/year	%
Africa	125	0.8	504	1.0
Americas	4,355	28.8	7,270	13.7
Central	5		38	
North ^a	4,308		7,013	
South	42		219	
Asia	4,608	30.4	24,232	45.7
Pacific Rim	3,489		18,242	
Central/Mid-East ^b	1,119		599	
Europe	5,715	37.7	18,905	35.7
Central/east	1,284		4,405	
West/north ^c	3,872		11,307	
CIS ^d	559		3,463	
Oceania	342	2.3	2,065	3.9
Total	15,145		52,976	

^aIncludes Mexico.

^bIncludes Turkey.

^cIncludes Azores and Guadeloupe.

^dIncludes Russia, Georgia, and Armenia.

traditionally nongeothermal countries, such as France, Switzerland, and Sweden, as well as areas of the midwestern and eastern United States. Most equipment used in these projects is of standard, off-the-shelf design and needs only slight modifications to handle geothermal fluids.

Worldwide in 2002, the installed capacity of direct geothermal utilization was 15,145 MWt and the energy use was approximately 190,699 TJ/year (52,976 GWh/year) utilizing at least 52,746 kg/s of fluid distributed among 58 countries. A summary by region is presented in Table I. This amounts to saving an equivalent of 13.5 million tonnes of fuel oil per year (TOE). The distribution of the energy use among the various types of use is shown in Fig. 1 for the world, and for comparison that of the United States is shown in Fig. 2. The installed capacity in the United States (2000) is 4000 MWt and the annual energy use is 20,300 TJ (5640 GWh), saving 3.94 million TOE. Internationally, the largest uses are for space heating (37%) (three-fourths of which is due to district heating) and for swimming, bathing, and balneology (22%), whereas in the United States the largest use is for geothermal heat pumps (59%). In comparison, Iceland's largest geothermal energy use (77%) is for space heating [15,600 TJ/year (4,334 GWh/year)], primarily with district heating systems.

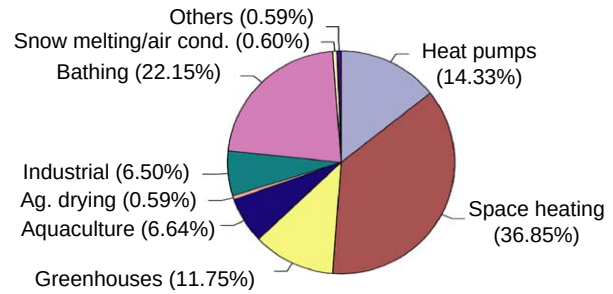


FIGURE 1 Distribution of geothermal energy use throughout the world.

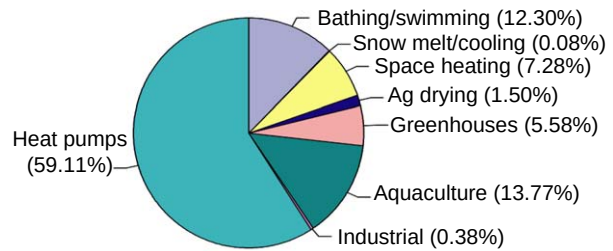


FIGURE 2 Distribution of geothermal energy use in the United States.

The Lindal diagram, named after Baldur Lindal, the Icelandic engineer who first proposed it, indicates the temperature range suitable for various direct use activities (Fig. 3). Typically, agricultural and aquacultural uses require the lowest temperatures, from 25 to 90°C. The amounts and types of chemicals in the geothermal water, such as arsenic and dissolved gases such as boron, are a major problem with regard to plants and animals; thus, heat exchangers are often necessary. Space heating requires temperatures in the range of 50–100°C, with 40°C useful in some marginal cases and ground-source heat pumps extending the range to 4°C. Cooling and industrial processing normally requires temperatures higher than 100°C. In terms of market penetration, the leading user of geothermal energy is Iceland, where more than 86% of the population enjoys geothermal heat in their homes from 26 municipal district heating services and 50% of the country's total energy use is supplied by direct heat and electrical energy derived from geothermal resources.

1.1 Historical Development

Although the direct use of geothermal energy has a much longer history than that of electric power generation, the data on utilization has been under

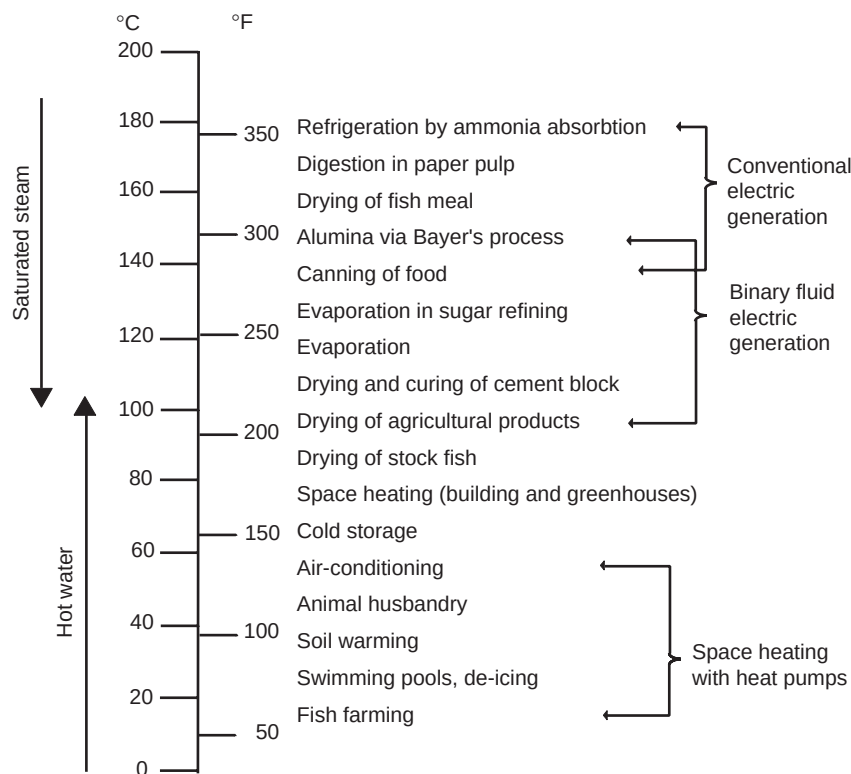


FIGURE 3 Lindal diagram.

reported. In fact, it is difficult to compare installed capacity and annual use due to the inclusion or exclusion of bathing, swimming, and balneology data. This has not been consistent; in the past, this use was not included, but in current reports it is included but not in a consistent manner. Also, values prior to 1970 were not summarized and up to 1980 could only be estimated from country descriptions in rapporteur reports. The early reports did not include China, a large user of geothermal energy for direct use, due to the political situation at the time, and they also did not include the United States even though a geothermal district heating system had been installed in Boise, Idaho, in 1892 and individual wells had been utilized in Klamath Falls, Oregon, since the 1930s for home heating. Finally, since many direct uses are small and not concentrated in one place, they are often overlooked by researchers reporting on their own country.

As a result, the 1961 United Nations (UN) conference in Rome reported only developments in Iceland, New Zealand, Italy, Japan, and Kenya. This report described district heating of 45,000 houses in Reykjavik; use of 1000 wells in Rotorua, New Zealand, for space heating; heating of 95,000 m² of

greenhouses in Iceland; production of 21,000 tons/year of salt in Japan; the pulp and paper plant at Kawerau, New Zealand; the chemical industry at Larderello, Italy; pig raising in New Zealand; and chicken hatching in Kenya. The 1970 report of the UN meeting in Pisa included descriptions from Hungary, Iceland, Italy, Japan, New Zealand, and the USSR. As mentioned previously, China and the United States were not included. The data in Table II are based on information from the 1970 UN conference in Pisa, a report by Lawrence Livermore Laboratory in 1975, the second UN conference in San Francisco, papers by Lund in 1979 and 1982, and reports from the Geothermal Resources Council (GRC) annual meetings in 1985 and 1990, the 1995 World Geothermal Congress in Italy, and the 2000 World Geothermal Congress in Japan. Starting in 1995, geothermal heat pumps (ground-source heat pumps) were included in the reports and now comprise a significant percentage of the totals.

The large increase in installed capacity between 1980 and 1985 is due to the inclusion of pool heating at spas in Japan (onsen) and the first available data from China. The annual growth rate (based on MWt) from 1970 to 1980 was 9.3%, from 1980 to

TABLE II

Worldwide Development of Geothermal Direct Heat Utilization

Year	Installed energy		No. of countries	Participants reporting
	MWt	GWh/year		
1960			5	Iceland, Italy, New Zealand, Japan, and Kenya
1970	800 est.	2,200 est.	6	+ Hungary, and USSR
1975	1300 est.		10	+ France, the Phillipines, Turkey, and the United States
1980	1950		14	+ Austria, Czechoslovakia, Germany, and Taiwan
1985	7072	23,960	24	+ Australia, Canada, China, Columbia, Denmark, Mexico, Poland, Romania, Switzerland, and Yugoslavia
1990	8064		30	+ Algeria, Belgium, Bulgaria, Ethiopia, Greece, Guatemala, Thailand, and Tunisia; some countries not reporting
1995	8664	31,236	30	Argentina, Georgia, Israel, Macedonia, Serbia, Slovakia, and Sweden; some countries not reporting
2000	15,145	52,976	58	See Lund and Freeston (2001)

1990 it was 15.2% (which was strongly influenced by data from Japan and China), and from 1990 to 2000 it was 6.5%. The overall growth rate during the past 30 years has averaged 10.3% annually. The large increases from 1970 to 1990 (average annual increase of 12.2%) and the reduction from 1990 to present were influenced by the availability of cheap fossil fuels and the economic slowdown in Southeast Asia.

1.2 Swimming, Bathing, and Balneology

Romans, Chinese, Ottomans, Japanese, and central Europeans have bathed in geothermal waters for centuries. Today, more than 2200 hot springs resorts in Japan draw 100 million guests every year, and the “return-to-nature” movement in the United States has revitalized many hot spring resorts.

The geothermal water at Xiaotangshan Sanitarium, northwest of Beijing, has been used for medical purposes for more than 500 years. Today, the 50°C water is used to treat high blood pressure, rheumatism, skin disease, diseases of the nervous system, ulcers, and generally for recuperation after surgery. In Rotorua, New Zealand, at the center of the Taupo Volcanic Zone of North Island, the Queen Elizabeth Hospital was built during World War II for U.S. servicemen and later became the national hospital for the treatment of rheumatic disease. The hospital has 200 beds, an outpatient service, and a cerebral palsy unit. Both acidic and basic heated mud baths treat rheumatic diseases.

In Beppu on the southern island of Kyushu, Japan, hot water and steam meet many needs: heating,

bathing, cooking, industrial operations, agriculture research, physical therapy, recreational bathing, and even a small zoo. The waters are promoted for “digestive system troubles, nervous troubles, and skin troubles.” Many sick and crippled people come to Beppu for rehabilitation and physical therapy. There are also eight Jigokus (“burning hells”) in town that show various geothermal phenomena that are used as tourist attractions.

In the area of the former Czechoslovakia, the use of thermal waters has been traced back before the occupation of the Romans and has had a recorded use of approximately 1000 years. Today, there are 60 spa resorts located mainly in Slovakia that are visited by 460,000 patients annually, usually for an average stay of 3 weeks each. These spas have old and well-established therapeutic traditions. Each sanitarium is designated for the treatment of specific diseases depending on the chemical composition of the mineral waters and spring gas, availability of peat and sulfurous mud, and climatic conditions. The therapeutic successes of these spas are based on centuries of healing tradition (balneology), systematically supplemented by the latest discoveries of modern medical science.

Bathing and therapeutic sites in the United States include Saratoga Springs, New York; Warm Springs, Georgia; Hot Springs, Virginia; White Sulfur Springs, West Virginia; Hot Spring, Arkansas; Thermopolis, Wyoming; and Calistoga, California. These sites were originally used by Indians for bathing and recuperating from battle. There are more than 115 major geothermal spas in the United States with an annual energy use of 1500 TJ.

1.3 Space Conditioning

Space conditioning includes both heating and cooling. Space heating with geothermal energy has widespread application, especially on an individual basis. Buildings heated from individual wells are popular in Klamath Falls, Oregon; Reno, Nevada; and Taupo and Rotorua, New Zealand. Absorption space cooling with geothermal energy has not been popular because of the high temperature requirements and low efficiency. Geothermal heat pumps (groundwater and ground coupled) have become popular in the United States, Sweden, and Switzerland for both heating and cooling.

An example of the use of space heating and cooling with low- to moderate-temperature geothermal energy is the Oregon Institute of Technology in Klamath Falls, Oregon (Fig. 4). Here, 11 buildings (approximately 62,000 m² of floor space) are heated with water from three wells at 89°C. Up to 62 liters/s of fluid can be provided to the campus, with the average heat utilization rate more than 0.53 MWt and the peak at 5.6 MWt. In addition, a 541-kW (154 tons) chiller requiring up to 38 liters/s of geothermal fluid produces 23 liters/s of chilled fluid at 7°C to meet the campus cooling base load (recently decommissioned).

1.4 District Heating

District heating originates from a central location and supplies hot water or steam through a network of pipes to individual dwellings or blocks of buildings. The heat is used for space heating and cooling,

domestic water heating, and industrial process heat. A geothermal well field is the primary source of heat; however, depending on the temperature, the district may be a hybrid system, which would include fossil fuel and/or heat pump peaking.

Geothermal district heating systems are in operation in at least 12 countries, including Iceland, France, Poland, Hungary, Turkey, Japan, and the United States. The Warm Springs Avenue project in Boise, Idaho, dating back to 1892 and originally heating more than 400 homes, is the earliest formal project in the United States. The Reykjavik, Iceland, district heating system (Fig. 5) is probably the most famous. This system supplies heat for approximately 160,000 people. The installed capacity of 830 MWt is designed to meet the heating load to approximately -10°C; during colder periods, the increased load is met by large storage tanks and an oil-fired booster station.

In France, production wells in sedimentary basins provide direct heat to more than 500,000 people from 61 projects (200,000 housing units). These wells provide 40–100°C water from depths of 1500–2000 m. In the Paris basin, a doublet system (one production and one injection well) provides 70°C water, with the peak load met by heat pumps and conventional fossil fuel burners (Fig. 6).

1.5 Agribusiness Applications

Agribusiness applications (agriculture and aquaculture) are particularly attractive because they require

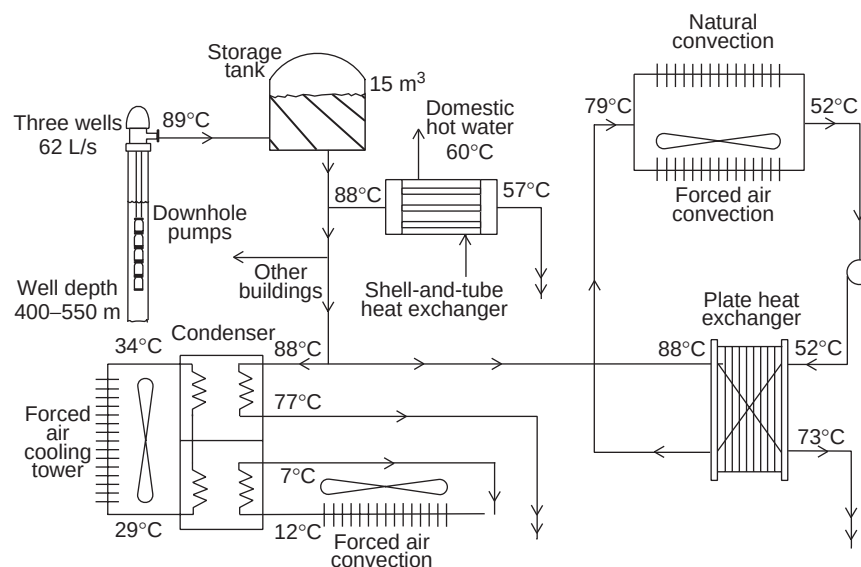


FIGURE 4 Oregon Institute of Technology heating and cooling system.

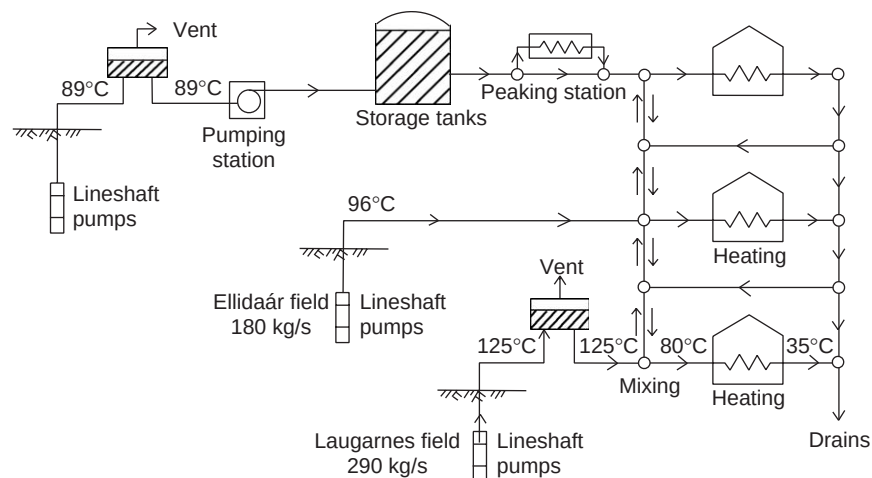


FIGURE 5 Reykjavik district heating.

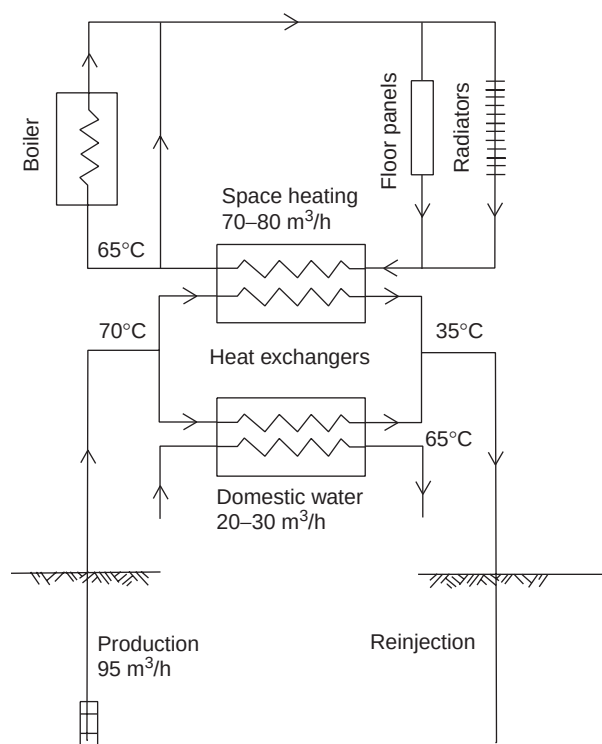


FIGURE 6 Melun l'Almont (Paris) doublet heating system.

heating at the lower end of the temperature range, for which there is an abundance of geothermal resources. Use of waste heat or the cascading of geothermal energy also has excellent possibilities. A number of agribusiness applications can be considered: greenhouse heating, aquaculture and animal husbandry, soil warming and irrigation, mushroom culture, and biogas generation.

Numerous commercially marketable crops have been raised in geothermally heated greenhouses in Tunisia, Hungary, Russia, New Zealand, Japan, Iceland, China, and the United States. These include vegetables such as cucumbers and tomatoes, flowers (both potted and bedded), house plants, tree seedlings, and cacti. Using geothermal energy for heating reduces operating costs (which can account for 35% of the product cost) and allows operation in colder climates where commercial greenhouses would not normally be economical.

The use of geothermal energy for raising catfish, shrimp, tilapia, eels, and tropical fish has produced crops faster than by conventional solar heating. Using geothermal heat allows better control of pond temperatures, thus optimizing growth (Fig. 7). Fish

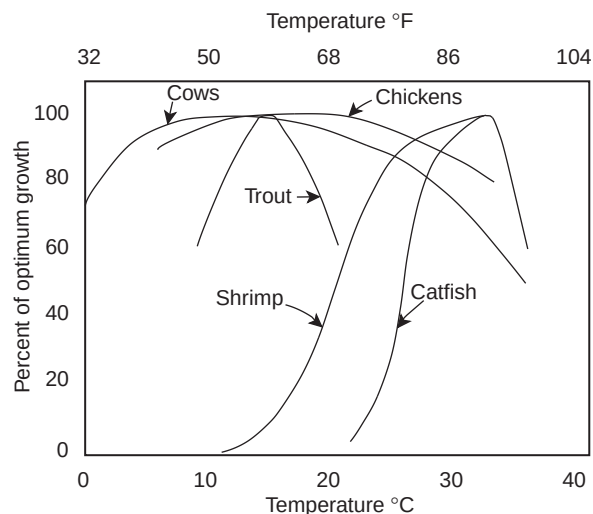


FIGURE 7 Effect of temperature on animal and fish growth.

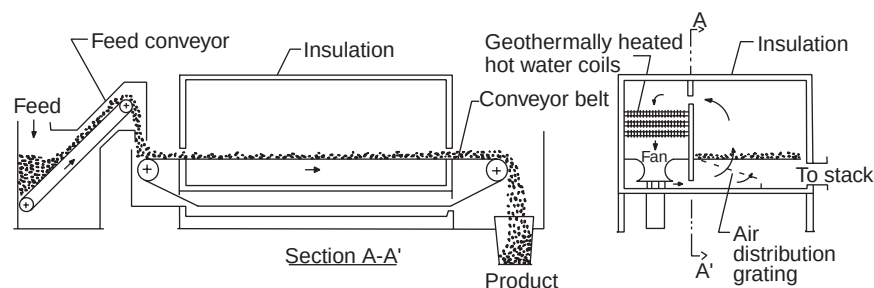


FIGURE 8 Continuous belt dehydration plant.

breeding has been successful in Japan, China, and the United States. A very successful prawn-raising operation, producing 400 tons of giant Malaysian freshwater prawns per year at \$17–27/kg, has been developed near the Wairakei geothermal field in New Zealand. The most important factors to consider are the quality of the water and disease. If geothermal water is used directly, concentrations of dissolved heavy metals, fluorides, chlorides, arsenic, and boron must be considered.

Livestock-raising facilities can encourage the growth of domestic animals by a controlled heating and cooling environment. Use of an indoor facility can lower the mortality rate of newborns, enhance growth rates, control diseases, increase litter size, make waste management and collection easier, and, in most cases, improve the quality of the product. Geothermal fluids can also be used for cleaning, sanitizing, and drying of animal shelters and waste as well as assisting in the production of biogas from the waste.

1.6 Industrial Applications

Although the Lindal diagram shows many potential industrial and process applications of geothermal energy, the world's uses are relatively few. The oldest industrial use is in Larderello, Italy, where boric acid and other borate compounds have been extracted from geothermal brines since 1790. Today, the two largest industrial uses are the diatomaceous earth-drying plant in northern Iceland and a pulp, paper, and wood processing plant in Kawerau, New Zealand. Notable U.S. examples are two onion dehydration plants in northern Nevada and a sewage digestion facility in San Bernardino, California. Alcohol fuel production has been attempted in the United States; however, the economics were marginal and thus the industry has not been successful.

Drying and dehydration are important moderate-temperature uses of geothermal energy. Various

vegetable and fruit products are feasible with the use of continuous belt conveyors (Fig. 8) or batch (truck) dryers with air temperatures from 40 to 100°C. Geothermally drying alfalfa, onions, pears, apples, and seaweed is an example of this type of direct use. A new development in the use of geothermal fluids is the enhanced heap leaching of precious metals in Nevada by applying heat to the cyanide process. Using geothermal energy increases the efficiency of the process and extends the production into the winter months.

2. EQUIPMENT

Standard equipment is used in most direct-use projects, provided allowances are made for the nature of geothermal water and steam. Temperature is an important consideration, as is water quality. Corrosion and scaling caused by the sometimes unique chemistry of geothermal fluids may lead to operating problems with equipment components exposed to flowing water and steam. In many instances, fluid problems can be designed out of the system. One such example concerns dissolved oxygen, which is absent in most geothermal waters except perhaps the lowest temperature waters. Care should be taken to prevent atmospheric oxygen from entering district heating waters, for example, by proper design of storage tanks. The isolation of geothermal water by installing a heat exchanger may also solve this and similar water quality-derived problems. In this case, a clean secondary fluid is then circulated through the user side of the system as shown in Fig. 9.

The primary components of most low-temperature, direct-use systems are downhole and circulation pumps, transmission and distribution pipelines, peaking or backup plants, and various forms of heat-extraction equipment (Fig. 9). Fluid disposal is either surface or subsurface (injection). A peaking

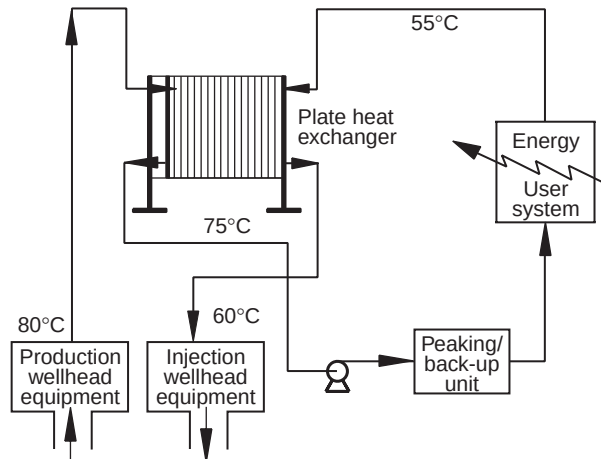


FIGURE 9 Geothermal direct-utilization system using a heat exchanger.

system may be necessary to meet maximum load. This can be done by increasing the water temperature or by providing tank storage (such as is done in most of the Icelandic district heating systems). Both options mean that fewer wells need to be drilled. When the geothermal water temperature is warm ($<40^{\circ}\text{C}$), heat pumps are often used.

The equipment used in direct-use projects represents several units of operations. In the following sections, the major units are described in the order in which they are used to transform geothermal waters for district heating.

2.1 Downhole Pumps

Unless the well is artesian, downhole pumps are needed, especially in large-scale, direct utilization systems. Downhole pumps may be installed not only to lift fluid to the surface but also to prevent the release of gas and the resultant scale formation. The two most common types are lineshaft pump systems and submersible pump systems.

The lineshaft pump system (Fig. 10) consists of a multistage downhole centrifugal pump, a surface-mounted motor, and a long driveshaft assembly extending from the motor to the pump. Most are enclosed, with the shaft rotating within a lubrication column, which is centered in the production tubing. This assembly allows the bearings to be lubricated by oil since hot water may not provide adequate lubrication. A variable-speed drive set just below the motor on the surface can be used to regulate flow instead of simply turning the pump on and off.

The electric submersible pump system (Fig. 11) consists of a multistage downhole centrifugal pump,

a downhole motor, a seal section (also called a protector) between the pump and motor, and electric cable extending from the motor to the surface electricity supply.

Both types of downhole pumps have been used for many years for cold water pumping and, recently, in geothermal wells (lineshafts have been used on the Oregon Institute of Technology campus in 89°C water for 45 years). If a lineshaft pump is used, special allowances must be made for the thermal expansion of various components and for oil lubrication of the bearings. Lineshaft pumps are preferred over the submersible pumps in conventional geothermal applications because lineshaft pumps cost less and they have a proven track record. However, for setting depths exceeding approximately 250 m, a submersible pump is required.

2.2 Piping

The fluid state in transmission lines of direct-use projects can be liquid water, steam vapor, or a two-phase mixture. These pipelines carry fluids from the wellhead to either a site of application or a steam-water separator. Thermal expansion of pipelines heated rapidly from ambient to geothermal fluid temperatures (which may vary from 50 to 200°C) causes stress that must be accommodated by careful engineering design.

The cost of transmission lines and the distribution networks in direct-use projects is significant. This is especially true when the geothermal resource is located a great distance from the main load center; however, transmission distances of up to 60 km have proven economical for hot water (e.g., the Akranes project in Iceland, where asbestos cement covered with earth has been successful).

Carbon steel is the most widely used material for geothermal transmission lines and distribution networks, especially if the fluid temperature is higher than 100°C . Other common types of piping material are fiberglass-reinforced plastic (FRP) and asbestos cement. The latter material, used widely in the past, cannot be used in many systems today due to environmental concerns; thus, it is no longer available in many locations. Polyvinyl chloride (PVC) piping is often used for the distribution network and for uninsulated waste disposal lines in which temperatures are well below 100°C . Conventional steel piping requires expansion provisions, either bellows arrangements or by loops. A typical piping installation has fixed points and expansion points approximately every 100 m. In addition, the

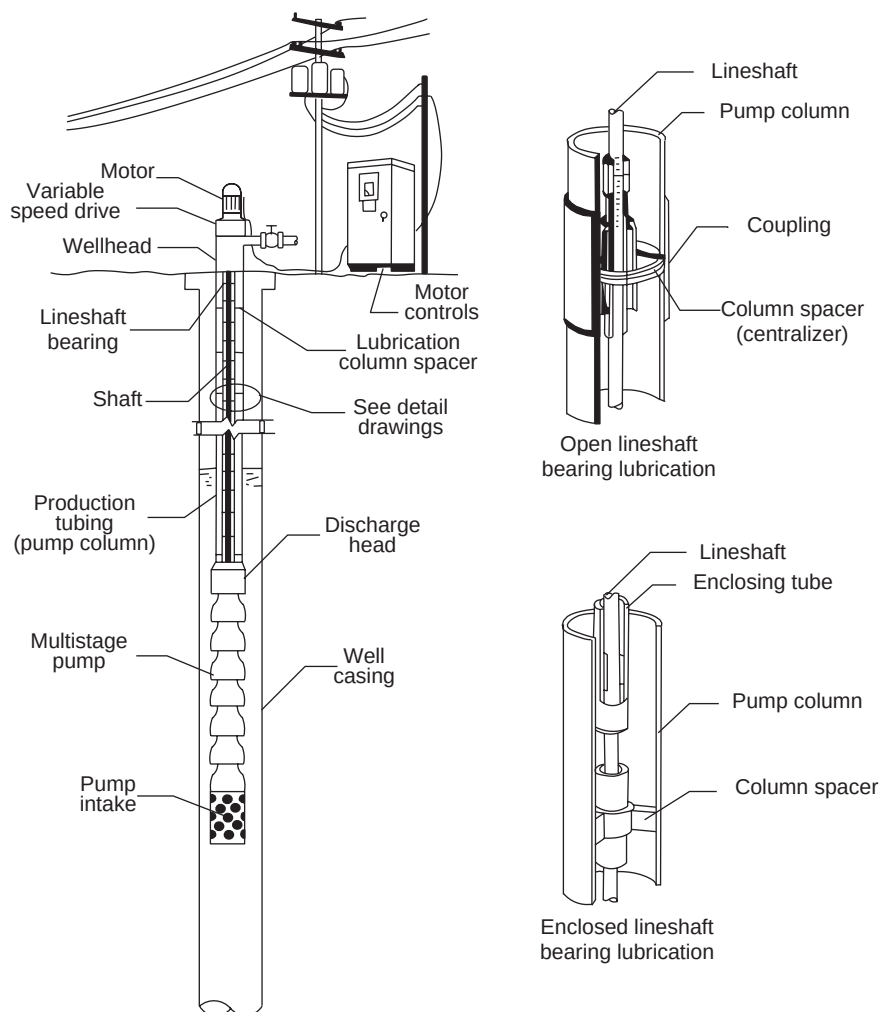


FIGURE 10 Lineshaft pump.

piping has to be placed on rollers or slip plates between points. When hot water pipelines are buried, they can be subjected to external corrosion from groundwater and electrolysis. They must be protected by coatings and wrappings. Concrete tunnels or trenches have been used to protect steel pipes in many geothermal district heating systems. Although expensive (generally more than \$300 per meter of length), tunnels and trenches have the advantage of easing future expansion, providing access for maintenance and a corridor for other utilities, such as domestic water, waste water, electrical cables, and phone lines.

Supply and distribution systems can be either single-pipe or two-pipe systems. The single-pipe system is a once-through system in which the fluid is disposed of after use. This distribution system is generally preferred when the geothermal energy is

abundant and the water is pure enough to be circulated through the distribution system. In a two-pipe system, the fluid is recirculated so that the fluid and residual heat are conserved. A two-pipe system must be used when mixing of spent fluids is necessary and when the spent cold fluids need to be injected into the reservoir. Two-pipe distribution systems typically cost 20–30% more than single-piped systems.

The quantity of thermal insulation of transmission lines and distribution networks depends on many factors. In addition to minimize the heat loss of the fluid, the insulation must be waterproof and watertight. Moisture can destroy any thermal insulation and cause rapid external corrosion. Aboveground and overhead pipeline installations can be considered in special cases. Considerable insulation is achieved by burying hot water pipelines. For example, burying

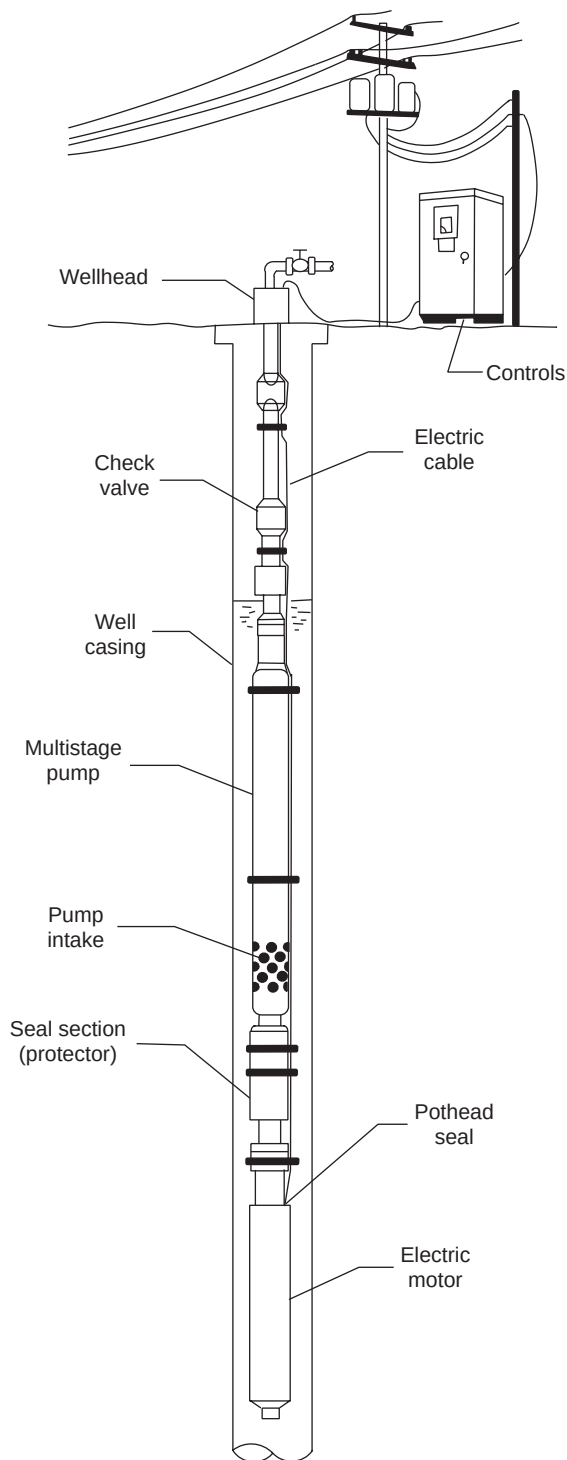


FIGURE 11 Submersible pump.

bare steel pipe results in a reduction in heat loss of approximately one-third compared to aboveground steel pipes in still air. If the soil around the buried pipe can be kept dry, then the insulation value can be

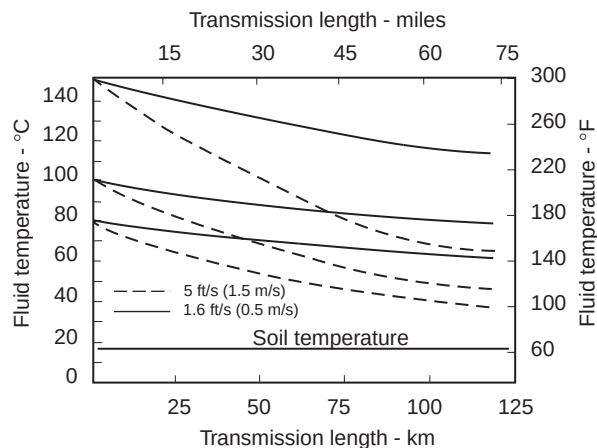


FIGURE 12 Temperature drop in hot water transmission line.

retained. Carbon steel piping can be insulated with polyurethane foam, rock wool, or fiberglass. Belowground, such pipes should be protected with PVC jacket; aboveground, aluminum can be used. Generally, 2.5–10 cm of insulation is adequate. In two-pipe systems, the supply and return lines are usually insulated, whereas in single-pipe systems only the supply line is insulated.

In flowing conditions, the temperature loss in insulated pipelines is in the range of 0.1–1.0°C/km, and in uninsulated lines the loss is 2–5°C/km (in the approximate range of 5–15 liters/s flow for 15-cm diameter pipe). It is less for larger diameter pipes; for example, a <2°C loss is experienced in the aboveground, 29-km long and 80- to 90-cm-diameter line (with 10 cm of rock wool insulation) from Nesjavellir to Reykjavik, Iceland. The flow rate is approximately 560 liters/s and 7 h are required to cover the distance. Uninsulated pipe costs approximately half that of insulated pipe and thus is used where temperature loss is not critical. Pipe material does not have a significant effect on heat loss; however, the flow rate does. At low flow rates (off-peak), the heat loss is higher than that at greater flows. Figure 12 shows fluid temperatures, as a function of distance, in a 45-cm-diameter pipeline insulated with 50 mm of urethane.

Several examples of aboveground and buried pipeline installations are shown in Fig. 13. Steel piping is used in most cases, but FRP or PVC can be used in low-temperature applications. Aboveground pipelines have been used extensively in Iceland, where excavation in lava rock is expensive and difficult. In the United States, belowground installations are more common to protect the line from vandalism and to eliminate traffic barriers.

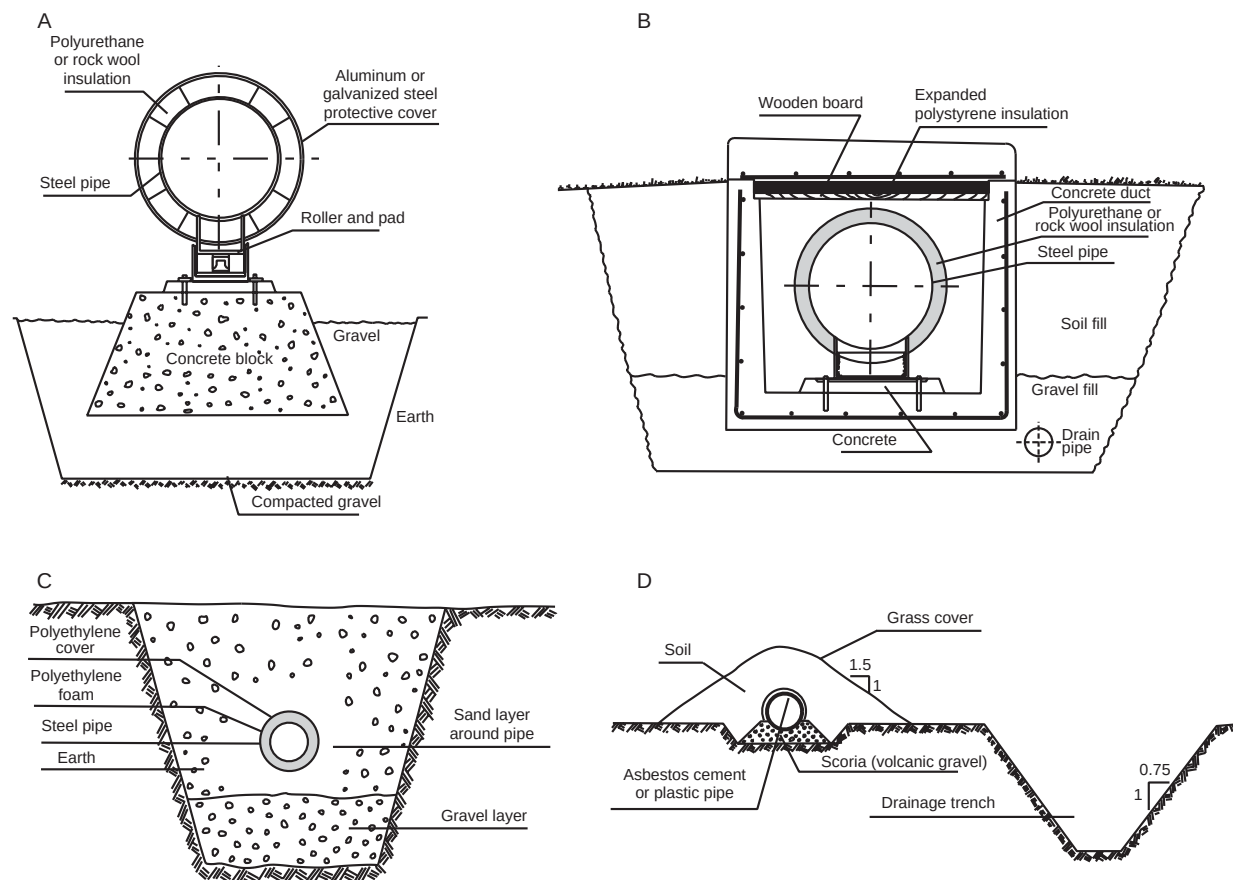


FIGURE 13 Examples of above- and belowground pipelines: (A) aboveground pipeline with sheet metal cover; (B) steel pipe in concrete tunnels; (C) steel pipe with polyurethane insulation and polyethylene cover; and (D) asbestos cement pipe with earth and grass cover.

2.3 Heat Exchangers

The principal heat exchangers used in geothermal systems are the plate, shell-and-tube, and downhole types. The plate heat exchanger consists of a series of plates with gaskets held in a frame by clamping rods (Fig. 14). The countercurrent flow and high turbulence achieved in plate heat exchangers provide for efficient thermal exchange in a small volume. In addition, compared to shell-and-tube exchangers, they have the advantages of occupying less space, are easily expandable when addition load is added, and cost 40% less. The plates are usually made of stainless steel, although titanium is used when the fluids are especially corrosive. Plate heat exchangers are commonly used in geothermal heating situations worldwide.

Shell-and-tube heat exchangers may be used for geothermal applications but are less popular due to problems with fouling, greater approach temperature

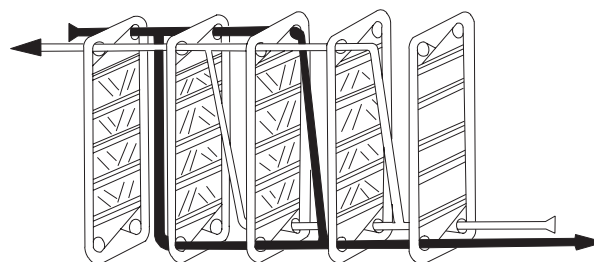


FIGURE 14 Plate heat exchanger.

(the difference between incoming and outgoing fluid temperature), and larger size.

Downhole heat exchangers eliminate the problem of disposing of geothermal fluid since only heat is taken from the well. However, their use is limited to small heating loads, such as heating of individual homes, a small apartment house, or business. The exchanger consists of a system of pipes or tubes suspended in the well through which secondary

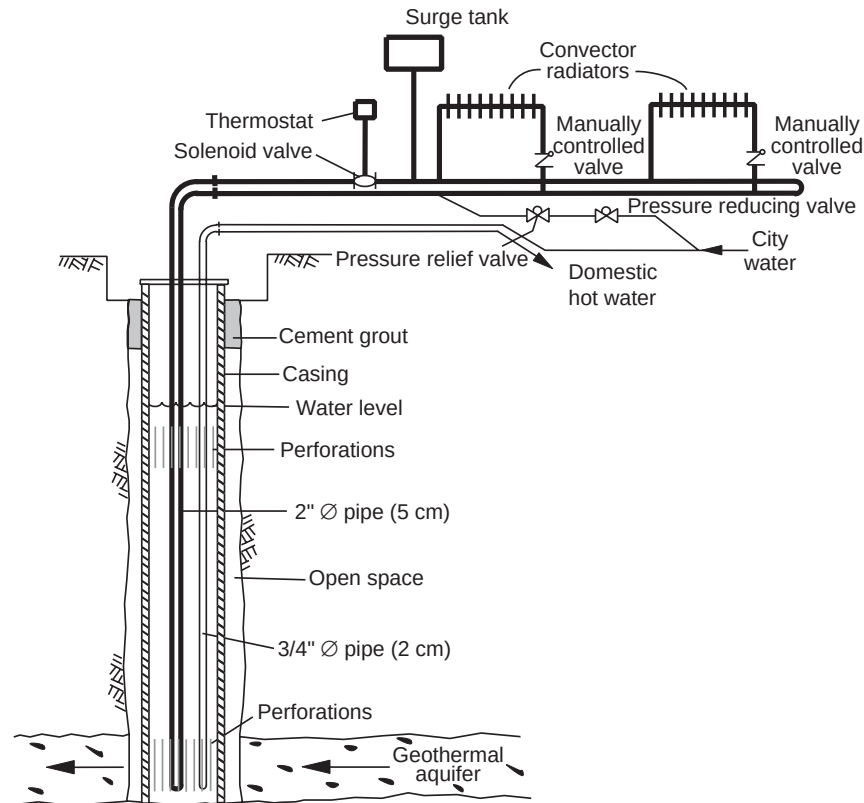


FIGURE 15 Downhole heat exchanger (typical of Klamath Falls, OR).

water is pumped or allowed to circulate by natural convection (Fig. 15). In order to obtain maximum output, the well must have an open annulus between the wellbore and casing, and perforations above and below the heat exchanger surface. Natural convection circulates the water down inside the casing, through the lower perforations, up into the annulus, and back inside the casing through the upper perforations. The use of a separate pipe or promoter to increase the vertical circulation has proven successful in older wells in New Zealand.

2.4 Heat Pumps

Ground-coupled and groundwater (often called ground-source or geothermal) heat pump systems are being installed in great numbers in the United States, Sweden, Switzerland, and Germany. Groundwater aquifers and soil temperatures in the range of 5–30°C are being used in these systems. Ground-source heat pumps utilize groundwater in wells or by direct ground coupling with vertical heat exchangers (Fig. 16). Nearly every state in the United States, especially in the midwestern and eastern states,

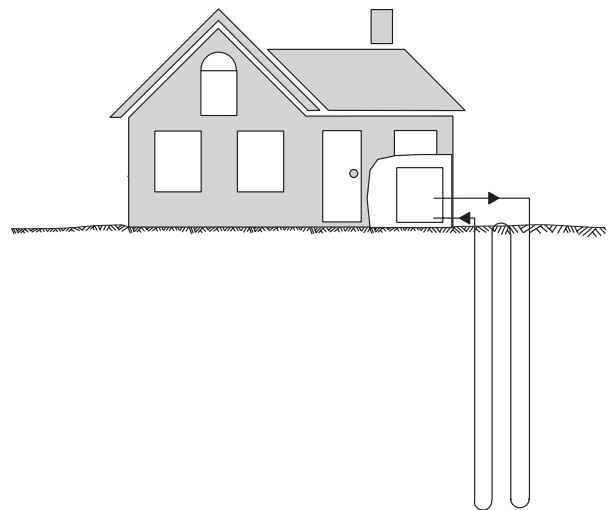


FIGURE 16 Typical ground-source heat pump installation.

utilizes these systems, which are in part subsidized by public and private utilities. It is estimated that approximately 70,000 groundwater systems, and more than 210,000 closed-loop vertical and 170,000 horizontal systems, are in use.

Like refrigerators, heat pumps operate on the basic principle that fluid absorbs heat when it evaporates into a gas and likewise gives off heat when it condenses back into a liquid. A geothermal heat pump system can be used for both heating and cooling. Water-to-air and water-to-water heat pumps are adaptable to geothermal energy. Heat pumps with heating capacities ranging from less than 3 kW to more than 1500 kW are available.

2.5 Convectors

Heating of individual rooms and buildings is achieved by passing geothermal water (or a heated secondary fluid) through heat convectors (or emitters) located in each room. The method is similar to that used in conventional space heating systems. Three major types of heat convectors are used for space heating: forced air, natural air flow using hot water or finned tube radiators, and radiant panels (Fig. 17). All these can be adapted directly to geothermal energy or be converted by retrofitting existing systems.

2.6 Refrigeration

Cooling can be accomplished from geothermal energy using lithium bromide or ammonia absorption refrigeration systems. The lithium bromide system is the most common because it uses water as the refrigerant. However, it is limited to cooling above the freezing point of water. The major application of lithium bromide units is the supply of chilled water for space and process cooling. They may be either one- or two-stage units. The two-stage units require higher temperatures (approximately 160°C), but they also have high efficiency. The single-stage units can be driven with hot water at temperatures as low as 77°C (such as at Oregon Institute of Technology; Fig. 4). The lower the temperature of the geothermal water, the higher the flow rate required and the lower the efficiency. Generally, a condensing (cooling) tower is required, which will add to the cost and space requirements.

For geothermally driven refrigeration below the freezing point of water, the ammonia absorption system must be considered. However, these systems are normally applied in very large capacities and have seen limited use. For the lower temperature refrigeration, the driving temperature must be at or higher than approximately 120°C for a reasonable performance. Figure 18 illustrates how the geothermal absorption process works.

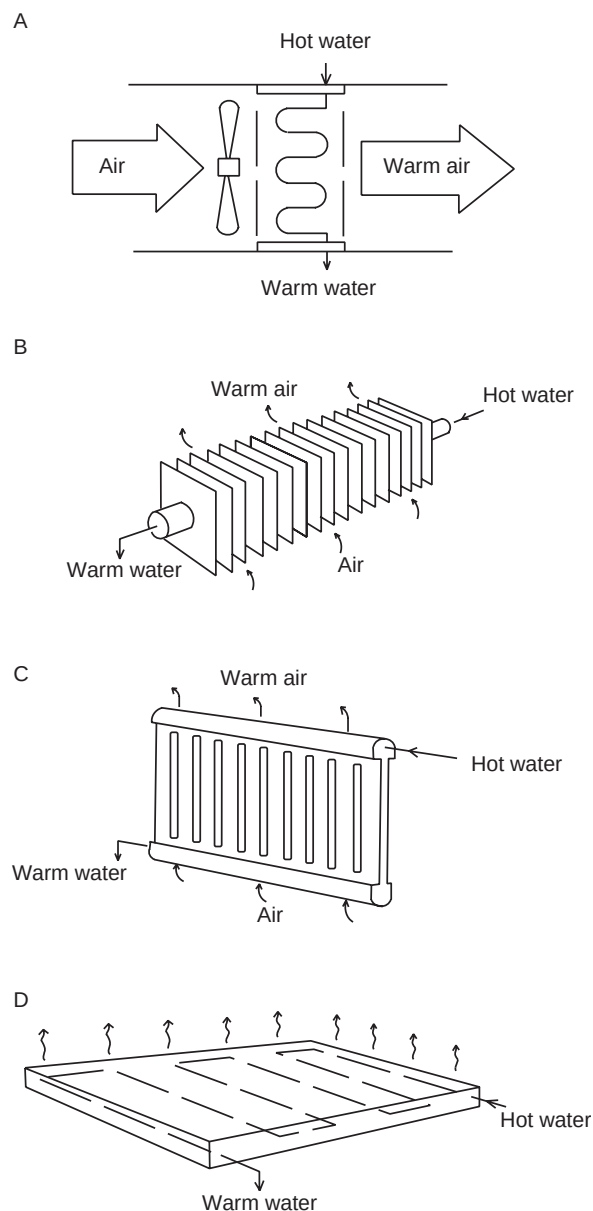


FIGURE 17 Convectors: (A) forced air; (B) material convection (finned tube); (C) natural convection (radiator); and (D) floor panel.

3. ECONOMIC CONSIDERATIONS

Geothermal projects require a relatively large initial capital investment, with small annual operating costs thereafter. Thus, a district heating project, including production wells, pipelines, heat exchangers, and injection wells, may cost several million dollars. In contrast, the initial investment in a fossil fuel system includes only the cost of a central boiler and distribution lines. The annual operation and

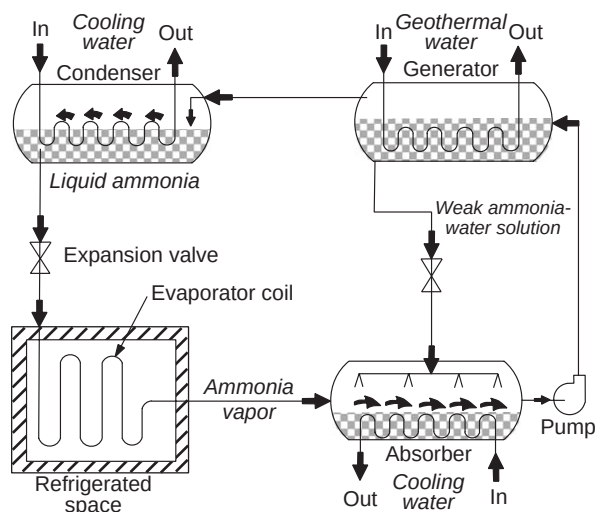


FIGURE 18 Geothermal absorption refrigeration cycle.

maintenance costs for the two systems are similar, except that the fossil fuel system may continue to pay for fuel at an ever-increasing rate, whereas the cost of the geothermal fuel is stable. The two systems, one with a high initial capital cost and the other with high annual costs, must be compared.

Geothermal resources fill many needs, including power generation, space heating, greenhouse heating, industrial processing, and bathing. Considered individually, however, some of the uses may not promise an attractive return on investment because of the high initial capital cost. Thus, one may have to consider using a geothermal fluid several times to maximize benefits. This multistage utilization, where increasingly lower water temperatures are used in successive steps, is called cascading or waste heat utilization. A simple form of cascading employs waste heat from a power plant for direct-use projects (Fig. 19).

Geothermal cascading has been proposed and successfully attempted on a limited scale throughout the world. In Rotorua, New Zealand, after geothermal water and steam heat a home the owner often uses the waste heat for a backyard swimming pool and steam cooker. At the Otake geothermal power plant in Japan, approximately 165 tonnes per hour of hot water flows to downstream communities for space heating, greenhouses, baths, and cooking. In Sapporo, Hokkaido, Japan, the waste water from the pavement snow-melting system is retained at 65°C and reused for bathing.

Current district heating costs are 0.23–0.42 ¢/1000 kcal (0.27–0.49 ¢/kWh) in Turkey compared to 3.4 ¢/kWh for natural gas and 11.2 ¢/kWh for electricity-based heating. The Klamath Falls district

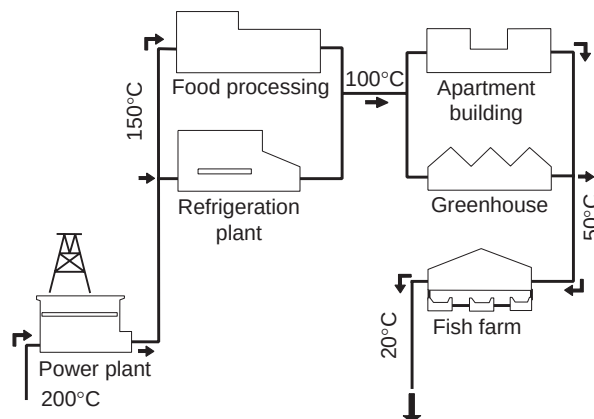


FIGURE 19 An example of cascading.

heating system charges 1.6–2.0 ¢/kWh. This is 50–80% of the natural gas cost, depending on the efficiency of the gas conversion, and the comparable cost for electricity in the city is 5.5 ¢/kWh. Construction costs for heating in Turkey are \$850–1250/kW and the cost per residence is approximately \$2000, an investment that is amortized over 5–10 years. Stefansson reported that an average consumer heating cost in 1995 for four European countries was 2.4 ¢/kWh.

Other estimates (1990 data) of the capital cost for various direct-use projects in the United States are as follows:

- Space heating (individual): \$463/kW of installed capacity
- District heating: \$386/kW of installed capacity
- Greenhouses: \$120/kW of installed capacity
- Aquaculture: \$26/kW of installed capacity

International data indicate a rate of \$270/kW of installed capacity for all projects reported, with a range of \$40–1880/kW. In the United States, the annual operation and maintenance cost is estimated to be 5% of the installed cost.

4. FUTURE DEVELOPMENTS

There appears to be a major potential for the development of low to moderate enthalpy geothermal direct use throughout the world, which is not being exploited due to financial constraints and the low price of competing energy sources. Given the right environment, and as gas and oil supplies dwindle, the use of geothermal energy will provide a competitive, viable, and economic alternative source of renewable energy.

Future development will most likely occur under the following conditions:

1. Co-located resource and uses (within 10 km).
2. Sites with high heat and cooling load density ($> 36 \text{ MWt/km}^2$).
3. Food and grain dehydration (especially in tropical countries, where spoilage is common).
4. Greenhouses in colder climates.
5. Aquaculture to optimize growth, even in warm climates.
6. Ground-coupled and groundwater heat pump installation (for both heating and cooling).

SEE ALSO THE FOLLOWING ARTICLES

Cogeneration • District Heating and Cooling • Geothermal Power Generation • Ground-Source Heat Pumps • Ocean Thermal Energy • Solar Thermal Power Generation • Thermal Energy Storage

Further Reading

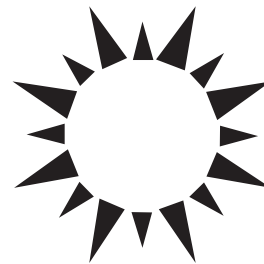
- Boyd, T. L. (1999). The Oregon Institute of Technology Geothermal Heating System—Then and now. *Geo-Heat Center Q. Bull.* 20(1), 10–13.
- Geo-Heat Center (1997.). Geothermal direct-use equipment. *Geo-Heat Center Q. Bull.* 19(1).
- Geo-Heat Center (1999.). Downhole heat exchangers. *Geo-Heat Center Q. Bull.* 20(3).
- Geo-Heat Center Web site: <http://geoheat.oit.edu>.
- Gudmundsson, J. S., and Lund, J. W. (1985). Direct uses of Earth heat. *Energy Res.* 9, 345–375.
- Kavanaugh, S., and Rafferty, K. (1997). “Ground-Source Design of Geothermal Systems for Commercial and Institutional Buildings.” ASHRAE, Atlanta, GA.
- Lund, J. W. (1996a). Balneological use of thermal and mineral waters in the USA. *Geothermics* 25(1), 103–148.
- Lund, J. W. (1996b). Lectures on direct utilization of geothermal energy, United Nations University Geothermal Training Programme, Report 1. Orkustofnun, Reykjavik, Iceland.
- Lund, J. W., and Boyd, T. L. (2000). Geothermal direct-use in the United States, update: 1995–1999. In “Proceedings of the World Geothermal Congress 2000, Japan.” International Geothermal Association, Pisa, Italy.
- Lund, J. W., and Freeston, D. H. (2001). Worldwide direct uses of geothermal energy 2000. *Geothermics* 30(1), 29–68.
- Lund, J. W., Lienau, P. J., and Lunis, B. C. (eds.). (1998). “Geothermal Direct-Use Engineering and Design Guidebook.” Geo-Heat Center, Klamath Falls, OR.
- Muffler, L. P. J. (ed.). (1979). “Assessment of geothermal resources of the United States—1978, USGS Circular No. 790.” U.S. Geological Survey, Arlington, VA.
- Rafferty, K. (2001). “An Information Survival Kit for the Prospective Geothermal Heat Pump Owner.” Geo-Heat Center, Klamath Falls, OR.
- Ragnarsson, A. (2000). Iceland country update. In “Proceedings of the World Geothermal Congress 2000, Japan,” pp. 363–376. International Geothermal Association, Pisa, Italy.

Geothermal Power Generation

KENG CHOON LEE

Geopower NZ Limited

Mount Roskill, Auckland, New Zealand



1. Geothermal Resources
2. Geothermal Power Cycles
3. Geothermal Drilling
4. Geothermal Steam Production Fields
5. Geothermal Power Stations
6. Environmental impacts and other technical problems
7. Commercial development of geothermal power projects

Glossary

Annubar A device (trade name) that measures steam flow rate in a pipe.

bleed A small flow through a side valve below the master valve to keep a well hot.

casing A steel pipe that forms a well.

casing head flange The first flange that is attached to a well casing above ground.

directional (deviated) drilling Drilling a well away from the vertical line.

drain pot A short pipe at the bottom of a steam pipe to collect condensate flowing in it.

drawdown Downhole pressure loss causing a well unable to sustain constant discharge.

exergy The maximum thermodynamic work that can be converted from heat.

feed zone The location where geothermal fluids enter a well.

Kuster gauge A device (trade name) that measures pressure and/or temperature in a well.

master valve The first shut-off valve above the casing head flange of a well.

nozzle A short pipe attached to a pressure vessel or pipe.

reinjection The process of putting spent or unused geothermal fluids back into the ground.

sinker bar A long metal bar that provides counterweight when an instrument is lowered into a flowing well.

spinner A device that measures fluid flow direction and/or velocity in a well.

stimulation The process of removing the cold column of water sitting in a well so that the well can self-discharge.

superficial steam velocity Steam velocity assuming that the steam phase of a two-phase fluid occupies the entire cross section of the pipe.

tracer A chemical injected into a well to detect the water flow path.

turbine nozzle A wheel of fixed blades to direct steam flow to a wheel of rotating turbine blades.

wellhead The area around the casing head flange.

wellhead pressure (WHP) Fluid pressure in a well just below the master valve.

Energy can be converted from one form to another. Geothermal power generation is the production of electrical energy from the heat (thermal) energy of the earth (geo). Electricity is a much more convenient, clean, and useful form of energy than is heat. Geothermal energy has been used directly for its heat and therapeutic effect for thousands of years by the Romans, Chinese, and Turks. Its indirect use was realized in 1904 when Prince Piero Ginori Conti of Italy successfully produced electricity in an experimental “indirect cycle” using pure steam produced from a heat exchanger. This led to the first 250-kWe geothermal power station in 1913 at Larderello in Tuscany, Italy. Nowadays, geothermal energy is used directly for many purposes, but it is the generation of electricity that provides humankind with the most benefits and challenges. Whereas Larderello is the first dry steamfield that was developed for power, Wairakei in New Zealand is the first geothermal power station that was developed from a wet field. Energy conversion is subjected to the laws of thermodynamics. Geothermal power generation is similar to the conventional fossil fuel-fired power generation using steam turbines. The main difference is the lack of a manmade boiler in a geothermal power station; hence, steam supply requires more careful scientific assessment. However, geothermal power generation has other advantages. It has environmental benefits due to its low carbon dioxide emissions; it is “renewable” if managed in a sustained manner; it has cultural, social, and economic benefits in terms of tourism and recreational values; and because it is indigenous, it lessens the dependence on

imported oil. Developing countries such as The Philippines, El Salvador, and Kenya have much of their electricity generated by geothermal energy, saving these countries a great deal in terms of oil imports.

1. GEOTHERMAL RESOURCES

The earth is an enormous sphere with more than a 6000-km radius. A solid brittle outer shell approximately 100 km thick, consisting of the continental crust (5–50 km thick) and the outermost part of the mantle, floats on a dense liquid metal core. Between the solid shell and the liquid core is the plastic layer of the mantle. The innermost core is believed to be solid metal due to the enormous pressure there. The continental crust is not one piece; rather, it is made up of many pieces known as tectonic plates. The earth has a heat source that is mainly from within the core due to radioactive materials. Heat is also generated by friction in the continental crust and the outermost mantle as the various tectonic plates move relative to each other. The consequences of this are the often catastrophic earthquakes and volcanic eruptions.

The temperature on the earth's surface ranges from about -50°C at the poles to about 50°C in the deserts. The earth's innermost core is believed to be approximately 5000°C ($\pm 1000^{\circ}\text{C}$). However, the earth's temperature gradient is not linear from the surface to the center of the earth but rather increases very steeply with depth from the earth's surface to approximately 1000°C at 100 km. To put things in perspective, we are looking at roughly 10 km of the continental crust for geothermal resources.

Geothermal resources can be defined as all of the heat energy stored in the earth's crust that is accessible locally. The word "locally" is important because a 20°C resource is usable at a location with -20°C ambient temperature but not 20°C (unless a heat pump is used). This is because heat transfer requires a temperature difference between a heat source and a heat sink. Geothermal resources from which energy could be extracted economically in the near future (<100 years) are parts of the useful accessible resource base. Geothermal resources that can be extracted economically today are defined as geothermal reserves. These identified reserves are further divided into possible, probable, and proven reserves depending on how much scientific work (geological, geophysical, geochemical, and drilling) has been carried out on them.

Whether a geothermal resource is technically suitable for geothermal power generation depends very much on the type and quality of the geothermal system and on the available technology. Currently, the geothermal system that is suitable for power generation is the hydrothermal type. This consists of a reservoir that contains hot water under high pressure. When this water is accessed by a drillhole, it ascends and some flashes to steam as the pressure releases.

The thermal water is mainly meteoric, so geothermal energy can be considered "sustainable" if the quantity of water withdrawn does not exceed recharge and if the heat added to the replenished water is at least equal to that extracted. However, geothermal energy is technically not strictly "renewable" because it is heat in the rock that heats the water, and this is not renewable. Most geothermal systems have lifetimes of less than 1 million years. Also, in general, heat that is extracted with the water from wells is much higher than that replenished from the reservoir rock because the heat capacity of water is much greater than that of the rock.

Hydrothermal systems are commonly divided into vapor- or steam-dominated and liquid- or water-dominated systems. In a steam-dominated system, the fluid in the reservoir is such that steam is the continuous phase that controls the pressure. When this fluid reaches the surface, the steam produced is normally superheated. In a water-dominated system, water is the continuous pressure-controlling phase. When this fluid reaches the surface, less than 50% by weight is steam. Although much of the world geothermal power generation comes from vapor-dominated systems, they are rare. The most famous are Larderello in Italy and The Geysers in California. Most geothermal systems are water dominated and are one-tenth the size of The Geysers in electrical power capacity. Examples of famous water-dominated systems include Wairakei in New Zealand and Cerro Prieto in Mexico.

Geothermal resources are commonly classified according to the average fluid temperatures in their reservoirs. Unfortunately, the demarcation lines are not agreed on (Table I). Exergy is also used to classify geothermal resources according to their specific exergy index,

$$SE \times I = (h - 273s)/1192,$$

where h is the specific enthalpy in J/g and s is the specific entropy in J/gK . Geothermal resources with $SE \times I > 0.5$ are high exergy or quality, those with $SE \times I < 0.05$ are low quality, and those in between are medium resources (Fig. 1).

TABLE 1
Classification of Geothermal Resources by Temperature (°C)

	(a)	(b)	(c)	(d)
Low enthalpy	<90	<125	<100	<150
Intermediate	90–150	125–225	100–200	
High enthalpy	>150	>225	>200	>150

Source: Dickson and Fanelli (1990).

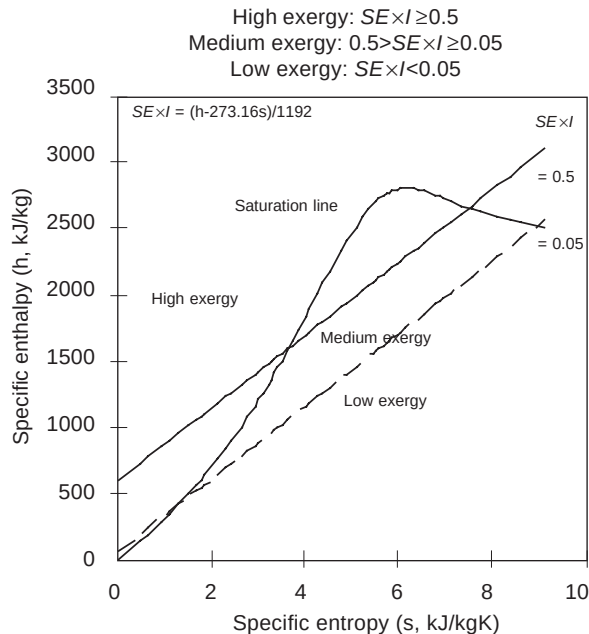


FIGURE 1 Classification of geothermal resources by exergy. Reproduced from Lee (2001).

The average earth's temperature gradient for the top 10 km from the surface is approximately 30°C/km. Places with gradient greater than 40°C/km are designated as thermal areas. Thermal fields are thermal areas with subsurface permeability that allows a fluid to convey deep-seated heat to the surface. Some thermal areas have 90°C/km gradient. The gradient can be higher in active volcanic areas.

Indicators of geothermal resources are surface thermal manifestations. However, a lack of surface manifestations does not imply a lack of geothermal resources (e.g., hot dry rock [HDR] system, geopressed system). Surface manifestations can be divided into active and inactive manifestations. Active manifestations are associated with heat and fluid discharges, whereas inactive ones are cold-altered ground and sinter deposits. Examples of active manifestations include hot springs, geysers, fumar-

oles, solfataras, mud pools, hot pools, warm ground, steaming ground, hydrothermal eruption craters, and volcanoes. Altered ground has rocks whose minerals have changed chemically due to interactions with thermal fluids. Sinter deposits are chemicals precipitated out of the thermal fluids and persist when the thermal fluids stop discharging.

An ideal geothermal system consists of a heat source and a reservoir of geothermal fluid in a confined space accessible by drilling. Heat is transferred from the heat source to the reservoir, which in turn transfers heat to the earth's surface. The heat source is normally a magmatic intrusion at shallow depths (5–10 km). The reservoir is a volume of hot rock that is permeable so that water can flow in and out of it. Because of the high pressure and temperature in the reservoir, the water tends to absorb gases such as carbon dioxide and hydrogen sulfide and to dissolve solid chemicals such as silica, calcite, and salts. When the geothermal fluid is withdrawn from the reservoir, pressure in the reservoir is likely to decrease, and this will cause more cold water to flow into the reservoir to be heated up again. This is the ideal renewable or sustainable geothermal system.

An HDR geothermal system contains a reservoir of hot rock that is not permeable. Research has been carried out to create permeability in the rock by hydrofracturing with water at very high pressure after drilling into the reservoir. Temperatures of 200°C have been measured in HDR at a depth of 3 km, and HDR systems of up to 10 km deep are being researched.

A geopressed geothermal system consists of a reservoir that is totally confined and contains water at very high pressure. There is no recharge of the reservoir water, which is usually saline. The presence of a geopressed system is most likely in a deep sedimentary basin. Pressures of 400 to 1000 bar abs and temperatures of 100 to 200°C at a depth of approximately 3 km are said to be typical.

A geothermal resource is normally "discovered" by geoscientific exploration. The exploration program normally follows the sequence of reconnaissance, prefeasibility, and feasibility studies. Geologists identify any geothermal surface manifestations and make geological and hydrological studies. These studies are important for all subsequent work, including siting of exploration and production wells. Geochemists study the chemistry of geothermal fluids on the surface to determine the type of geothermal system and estimate the reservoir fluid temperature. Geophysicists measure physical parameters such as heat flow, temperature, electrical conductivity, gravity, and

magnetism of the rocks. The information can be used to deduce the shape, size, depth, and capacity of the deep geothermal reservoir. Exploration wells are then drilled to confirm the power potential of the geothermal resource.

In a geothermal power project, the most important factor is the capacity and life of the resource. Unlike an oil reservoir that is exhausted when the oil is completely extracted, a geothermal reservoir can have its reservoir fluids replaced. This recharge of the geothermal reservoir fluids is an important process that controls the behavior of the reservoir. The study of recharge behavior is by means of geothermal reservoir engineering. Geological and well test data are normally analyzed using reservoir modeling computer software to estimate the capacity and life of the resource.

2. GEOTHERMAL POWER CYCLES

When a geothermal resource has been deemed suitable for power generation, a decision is required to select a suitable power cycle. A power cycle consists of a series of thermodynamic processes that converts heat energy to electricity. The most common geothermal power cycle is the Rankine cycle, consisting of two isothermal and two adiabatic processes. More than one power cycle can be used for a geothermal resource. Indeed, other direct uses can be combined with power generation.

Geothermal power cycles can be classified into five major types: steam Rankine cycle, organic Rankine cycle (ORC), total flow cycle, Kalina cycle system (KCS), and trilateral flash cycle (TFC). ORC, KCS, and TFC are binary cycles that use geothermal fluids as the primary heat source to vaporize a secondary fluid for power generation. Currently, nearly all geothermal power plants use steam or ORC or a combination of the two. There are several total flow machines, but they are not widely used. The first pilot Kalina geothermal power plant was commissioned in Iceland in 2000. TFC is still at the experimental stage.

The following subsections examine the various power cycles based on the types of geothermal resources.

2.1 Dry Steam (Vapor-Dominated) Systems

Dry steam systems are rare, but they have high power capacity and are currently producing appro-

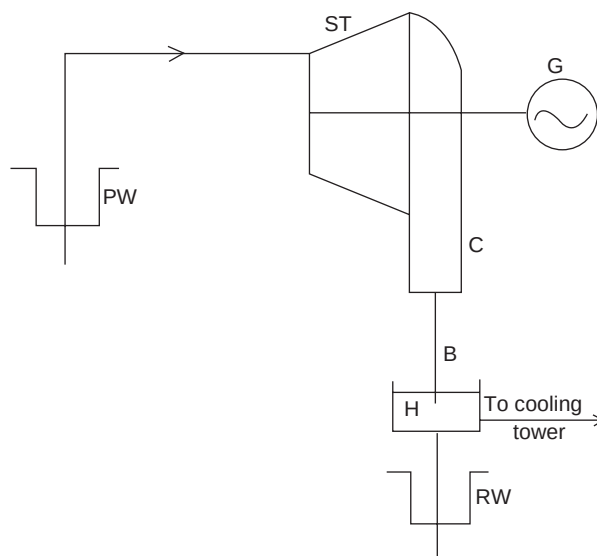


FIGURE 2 Power cycle for dry steam systems with condenser. B, barometric leg; C, condenser; G, generator; H, hot well; PW, production well; RW, reinjection well; ST, steam turbine.

ximately 50% of total world geothermal electrical power.

Conceptually, the power cycle for dry steam geothermal systems is the simplest: saturated or slightly superheated dry geothermal steam from wells is piped directly to condensing steam turbines (Fig. 2). The condensate is reinjected for environmental reasons. Other problems associated with solids and gases carried in the steam have to be addressed. Solids are removed using a trapping device before they reach the turbines, and gases are removed from the condensers by a gas extraction system. The gases are mainly carbon dioxide (>90%) and some hydrogen sulfide. Depending on the environmental protection law, the hydrogen sulfide might need to be treated by a chemical process.

2.2 Two-Phase (Liquid-Dominated) Systems

Two-phase systems are liquid-dominated systems. In two-phase geothermal systems, the geothermal fluids contain both water and steam at roughly an equal weight ratio (i.e., dryness, $x \sim 0.5$) at the wellhead. The water phase needs to be removed by a separator, and the steam is piped to the steam turbines (Fig. 3). The separated water can be reinjected directly or, depending on the fluid pressure, can be flashed to produce a lower pressure steam for power generation before being reinjected (Fig. 4).

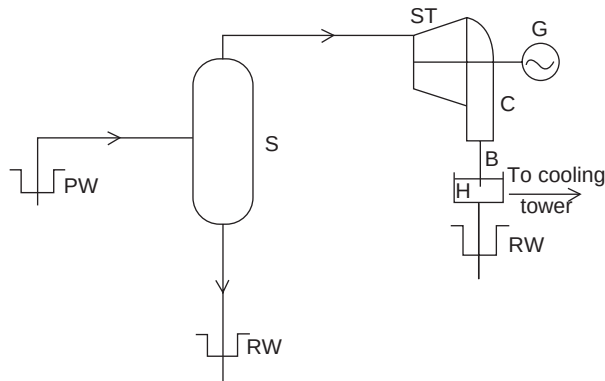


FIGURE 3 Power cycle for single-pressure hot water or two-phase geothermal systems. B, barometric leg; C, condenser; G, generator; H, hot well; PW, production well; RW, reinjection well; S, separator; ST, steam turbine.

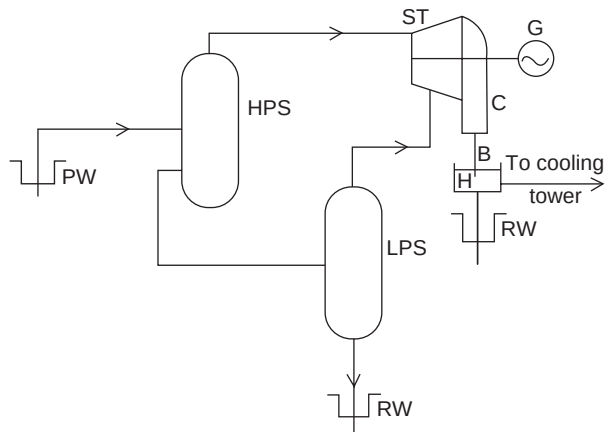


FIGURE 4 Power cycle for dual-pressure hot water or two-phase geothermal systems. B, barometric leg; C, condenser; G, generator; H, hot well; HPS, high-pressure separator; LPS, low-pressure separator; PW, production well; RW, reinjection well; ST, steam turbine.

2.3 Hot Water (Liquid-Dominated) Systems

Hot water systems are the most common geothermal systems suitable for power generation. In a hot water system, both the liquid and steam phases appear at the wellhead because the water flashes as it ascends the wells. The dryness fraction at the wellheads is normally less than 0.3. Depending on the wellhead pressures, the power cycle selected can be a single-pressure steam system (Fig. 3) or a dual-pressure one (Fig. 4). It is common to flash the water phase to produce more steam in a low-pressure flasher or separator. In practice, there is no real distinction between a flasher and a separator. Both are pressure

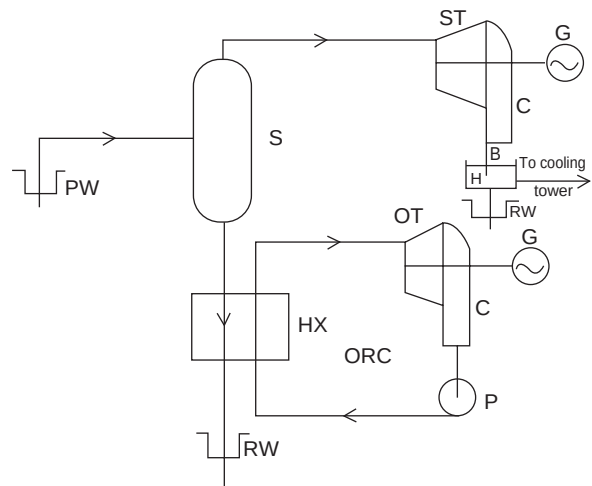


FIGURE 5 Power cycle for hot water or two-phase geothermal systems with binary ORC. B, barometric leg; C, condenser; G, generator; H, hot well; HX, heat exchanger; OT, organic vapor turbine; P, pump; PW, production well; RW, reinjection well; S, separator; ST, steam turbine.

vessels, and both have the function of separating the liquid from the steam phases. The flashing is normally done by a throttling orifice or valve installed upstream of the flash vessel. However, to maintain the large volume of low-pressure flashed steam, a flasher is normally larger than a high-pressure separator. Wairakei originally had a three-pressure steam system, but it was down-rated to a two-pressure system in 1983 because the steamfield drawdown could not sustain the high-pressure steam.

Nowadays, it is common to use the separated water for an ORC (Fig. 5). Exhaust steam from a back pressure turbine can also be used for a separate ORC (Fig. 6). For a hot water system that has a low dryness fraction, high gas content, or corrosive fluids, the steam Rankine cycle is unsuitable. A total flow or binary cycle (ORC, KCS, or TFC) may be used.

In a total flow cycle, the entire geothermal fluids from the wells are fed to a total flow machine that is robust enough to accept two-phase fluids (Fig. 7). The materials of construction can be selected to withstand the corrosive fluids. The helical screw expander is the first total flow machine to be developed and is now also used to generate electricity from waste heat in other industries. Other total flow machines are the biphasic turbine and the Robertson engine.

In an ORC, the geothermal fluids are used to vaporize a low-boiling point (30–40°C) fluid, such as isopentane or isobutane, to very high pressure (15–20 bars), and this fluid is then used to provide the motive force for a vapor turbine. The advantage

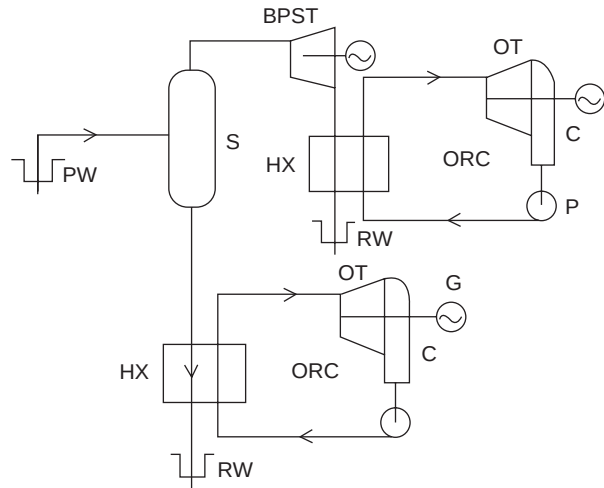


FIGURE 6 Power cycle for hot water systems with back pressure steam turbine (BPST) and binary ORC. See Fig. 5 caption for abbreviations.

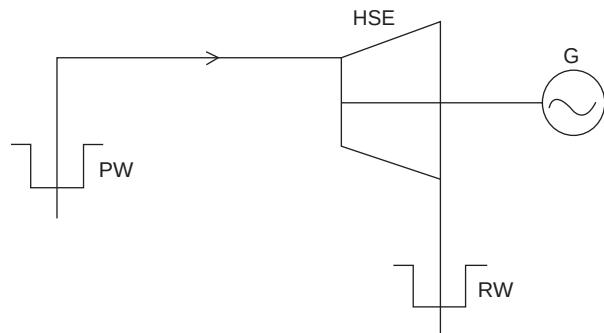


FIGURE 7 Total flow power cycle for hot water geothermal systems. G, generator; HSE, helical screw expander; PW, production well; RW, reinjection well.

of the binary ORC is that geothermal fluids with a temperature as low as 80°C can be used to generate electricity (Fig. 8).

A Kalina power cycle makes use of a mixture of two fluids in the secondary loop of the power cycle. It is considered a family of power cycle systems because not only can the ratio of the mixture of fluids vary with the temperature of the heat source, but different fluids also can be used for the mixture. An ammonia/water mixture appears to be the most suitable for geothermal applications. A pilot Kalina cycle geothermal power plant was commissioned in Iceland in 2000 and tested successfully using a geothermal fluid temperature of 121°C .

TFC, developed in 1993, is the latest binary power cycle. The concept of TFC is to improve the heat transfer process between the geothermal fluid and the secondary working fluid by keeping the working fluid

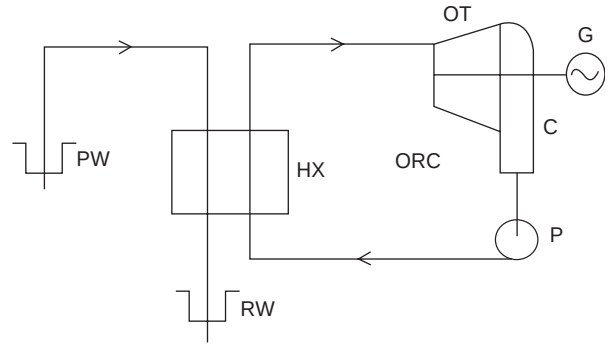


FIGURE 8 Binary ORC power cycle for hot water geothermal systems. See Fig. 5 caption for abbreviations.

as pressurized liquid, which is then used to power a helical screw expander. The best working fluid is still being researched, but low hydrocarbons appear to be most suitable.

2.4 Hot Dry Rock

HDR geothermal technology is still in the research stage. The concept is to drill two wells into hot rocks at a depth of 5 to 10 km, where the temperature gradient is high ($70\text{--}90^{\circ}\text{C}/\text{km}$). Cold water is pumped into one well and pressurized to cause fractures in the hot rocks so that pressurized hot water flows up the other well. This is equivalent to an artificial liquid-dominated geothermal system. Power cycles suitable for natural hot water systems would be suitable for the HDR system.

2.5 Geopressured Geothermal System

A geopressured system of 1000 bar abs and 100°C is equivalent to an average hot water system in its ability to produce electricity. Because this is still at an early research stage, no power cycles have been proposed for a geopressured system, although the total flow power cycle using a helical screw expander appears to be most suitable.

3. GEOTHERMAL DRILLING

Geothermal drilling technology is adapted from the oil industry. Most of the geothermal wells are drilled with a rotary-type drilling rig. Because of the hard, abrasive, and high-temperature volcanic rocks of geothermal systems, the drilling bits require hard abrasive resistant teeth such as diamond or tungsten carbide. Hence, geothermal drilling is an expensive

activity in geothermal development. Most geothermal wells are drilled 1 to 2 km deep to intersect faults and fractures so as to increase the success rate of production from permeable zones. This is enhanced by directional or deviated wells as opposed to vertical wells of earlier development. Directional drilling also allows drilling of more than one well on a single well pad, especially in difficult terrains. A geothermal well typically costs U.S. \$1 million and increases exponentially with depth. Geothermal drilling may be divided into three types: exploration, production, and reinjection.

Exploration drilling is carried out to assist earth scientists investigating a geothermal resource. Geologists will study the drill cuttings and drill cores of exploration wells that are normally drilled vertically and in small diameters (100–150 mm or 4–6 inches). It is not uncommon to make continuous drill cores from exploration wells so that the subsurface geology of the geothermal area can be confirmed. Exploration wells may be productive, and geothermal fluids discharged from the wells will confirm the chemistry of the fluids from earlier geochemistry studies.

Production wells are drilled after data from exploration wells have proven that a geothermal resource is economically viable in terms of its expected generating capacity and life. Geothermal fields of earlier development, such as Wairakei and Cerro Prieto, have vertical wells drilled to approximately 1 km deep. The fields have flat terrains, and wells are spaced as close as 20 m from each other. Nowadays, it is common to have several production wells drilled from a single well pad using directional drilling. Modern production wells are typically 2 km deep and may be deviated 1 km away from the well pad. Typical production wells have 250-mm ($9\frac{5}{8}$ -inch) diameter production casing. Some large geothermal fields in The Philippines have large (500-mm or 20-inch) diameter production casing. (Note: As drilling is adapted from the oil industry dominated by the United States, the unit used is imperial.)

For environmental reasons, modern practices reinject most, if not all, geothermal fluids produced from the production wells back into the ground after the energy has been used. This is achieved by piping the spent geothermal fluids to reinjection wells. Reinjection wells are drilled similarly to the production wells. Good reinjection wells require as much permeability as do production wells to swallow the reinjected fluids. Hence, a good production well makes a good reinjection well. The reinjection may require pumping to achieve the required injection rate. However, reinjection of cooler geothermal fluids

back into the ground may cool the hot fluids from production wells. Hence, reinjection wells are normally located either downstream of the flow direction of the geothermal fluids in the reservoir or at the edges of the production field so that reinjected fluids flow away from production zones.

A well is drilled by a rotating drilling bit (commonly a tricone bit with a tungsten carbide insert), with an additional load applied to it from the weight of a thick wall tube called the drill collar. The rotating motion comes from a rotating table with a square hole that transmits torque to a square tube known as the kelly. The kelly is connected to the drill collar by drill pipes to form the drill string. Drill pipes have a thinner wall than does the drill collar. The drill string is smaller in diameter than the hole being drilled, allowing a drilling fluid to be pumped down inside the drill string, clearing the cuttings around the drill bit, and carrying the cuttings up via the annulus between the hole and the drill string.

The drilling of a typical production well (Fig. 9) commences by drilling a large diameter hole (24–26 inches) up to approximately 50 m. A 20- to 22-inch outer diameter pipe (casing) is then cemented in the hole. The cementation is achieved by pumping a cement slurry down inside the casing and flowing up the annulus. The cement inside the casing is then displaced by water to just above the bottom of the casing (shoe). When the cement has set, a drilling wellhead with blowout preventer (BOP) is erected, controlling accidental discharge of geothermal fluid during drilling. Drilling continues with a smaller diameter hole (17.5 inches) for a $13\frac{3}{8}$ -inch casing from the top of the cement near the casing shoe to the next casing depth. This sequence is repeated until three or four casings are cemented to a depth assessed to be the top of the desired production zone in the reservoir. The final casing is called the production casing, usually $9\frac{5}{8}$ inches in a $12\frac{1}{4}$ -inch hole. In general, the shallow outer casings are required for drilling operations, whereas the inner casings are for production purposes. These casings not only control the geothermal fluids flowing inside the production casing but also prevent cold fluids outside from entering the well.

When the final section of the hole (8.5 inches) is completed in the production zone, a 7-inch casing with holes (slotted liner) is usually used to line the well. Unlike other casings, the liner is not cemented into place. It can be sitting at the bottom of the well or hanging from the production casing shoe. The liner prevents the production formation at the hole from collapsing and allows the geothermal fluids to

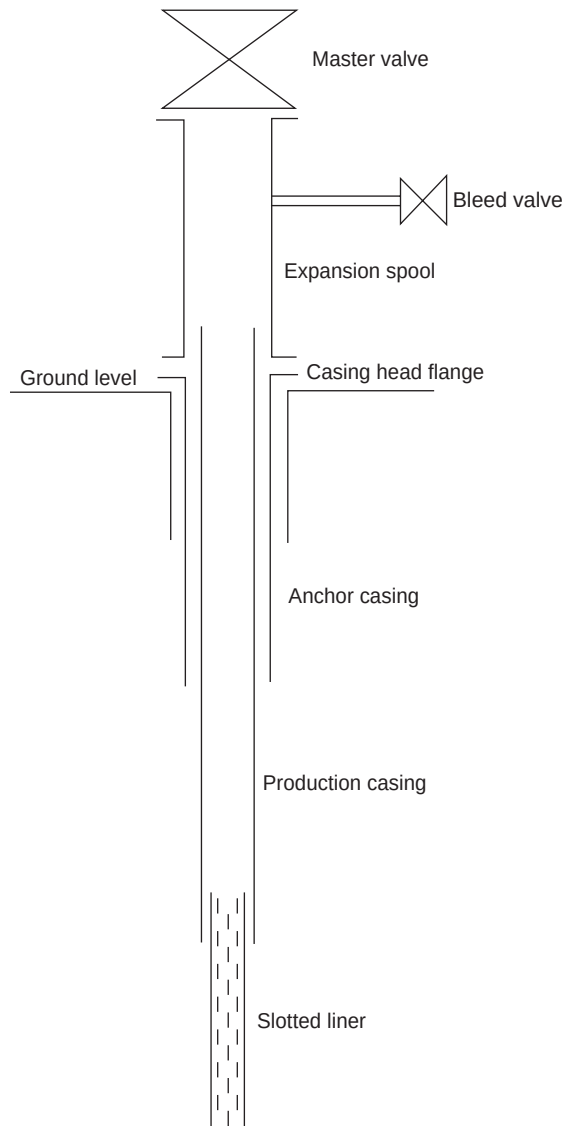


FIGURE 9 A typical well casing layout.

enter the well and flow up inside the production casing. A wellhead consisting of flanges, valves, and pipe spools is usually attached to the top of one of the intermediate cemented casings. The production casing is allowed to expand into a wellhead pipe spool during production of the hot fluids.

Because the safety of the drilling rig and the completed well can be at risk for an unstable well site, a complete geotechnical investigation might be necessary before site construction. Of particular concern is the possibility of a shallow blowout during drilling in areas with shallow permeable rock formation. To prevent this, consolidated grouting of the drilling site is carried out. This involves drilling and cementing slim holes to depths of 10 to 30 m

around the drilling site. This will direct any accidental blowout discharge of geothermal fluids away from the drilling rig.

A reliable supply of water is essential for safe drilling operation. Water is required for the drilling fluid—usually mud, cementing, quenching, or controlling a well from discharging. Water is also used as the drilling fluid when total loss of circulation is encountered, especially in the production zone, to prevent sealing of the production zone by drilling mud. A typical production geothermal well requires a reliable water supply of $2 \text{ m}^3/\text{min}$.

An important factor in drilling a geothermal well is to prevent the well from discharging during drilling. This is achieved by having the column of drilling fluid create a pressure greater than the fluid pressure in the reservoir formation. To increase the density of the drilling fluid, bentonite is commonly added to the drilling fluid to form drilling mud. There are other advantages in drilling with mud. It can keep drill cuttings in suspension in the event of stoppage during the drilling operations such as connection of additional drill pipes. However, drilling with mud also has its drawback. When a production zone is hit, the drilling mud, having higher pressure than the formation, will flow into the production zone fractures, effectively sealing the production zone. Hence, alternative drilling fluids that will not seal the production zone should be used when drilling in the production zone. These can be aerated water, foam, or air. In such situations, the practice of balance drilling is best such that the drilling fluid pressure balances the formation pressure.

4. GEOTHERMAL STEAM PRODUCTION FIELDS

Geothermal wells need to undergo a test program before they are used. This is so that the baseline conditions of both the wells and the geothermal aquifers that they tap can be determined. This baseline data are critical because all future information is compared against them.

After drilling has been completed and before the initial discharge, the well downhole conditions are measured. The temperature and pressure are measured by using a clockwork Kuster gauge or electronic logging tools. Standard practice is to initially do an injection (or completion) test, that is, sometime at a series of flows, where the temperature, pressure, and possibly flows using a spinner tool are

measured at intervals down the open hole section of the well. The well is then shut in and allowed to fully heat up. This can take less than 1 day (e.g., for a steam well) to more than 6 months (e.g., for a deep well with poor permeability). The downhole conditions during heat-up are also measured at intervals. The well is shut in to minimize circulation so that conditions in the cased-off section can be measured and so that the gas column can depress the liquid in the well to the upper feed zone. The downhole measurements will reveal the production zones in a well because hot geothermal fluids circulating at the production zones will cause these areas to heat up faster. The pressure profile of a well can reveal the water level in the well. The water level can also be determined by lowering a float into the well.

Once the new well is fully heated, it needs to be discharged to rid it of drilling mud, fluid, and debris before the well can be put into service. The initial discharge is normally a vertical one due to the erosive nature of the materials. The well is then tested to determine its output characteristics, that is, mass flow rate and enthalpy as a function of wellhead pressure (Fig. 10). Chemical sampling is also done during this output test. These data are very important because they comprise the baseline information.

The well output can be measured by several methods. For a small well discharge, a simple calorimeter can be used. The well output is discharged into a large vessel containing a mass of cold water. A heat balance will determine the output.

A method unique to the geothermal well test is the James lip pressure method. This method makes use of discharging two-phase fluid at choked flow conditions through an open pipe. By measuring the static "lip" pressure at the open end of the pipe, the flow rate can be determined from a correlating equation. The equation now is in several forms. The

original equation in imperial units is

$$G = 11400 P^{0.96} / h^{1.102},$$

where G = flow in pounds/(s feet²) of the lip pressure discharge pipe cross-section area, P = lip pressure in psia, and h = fluid enthalpy in Btu/pound.

In metric units, the equation is

$$G = 0.184 P^{0.96} / h^{1.102},$$

where G = tonnes/(s cm²), P = bar absolute, and h = kJ/kg.

The lip pressure method requires knowledge of the fluid enthalpy. However, the equation is relatively insensitive to enthalpy, so it can be estimated if not known. From experience, it is known that the discharge shape of the fluid at the lip pressure pipe will give some indication of the enthalpy. An estimate can also be made if the downhole conditions are known. If the fluid is known to be saturated water at a certain pressure during drilling, the enthalpy is also known. If the discharge is made into a silencer and the water flow rate is measured by a weir plate, the total mass flow can be determined. A quick estimate of the heat discharge or power potential of a well can be made by the lip pressure method during the first well vertical discharge. This number is useful for planning of the surface equipment for further tests.

A well output is most accurately measured by the separator method. The well is discharged into a separator. The steam phase is measured in a steam pipe using an orifice plate or Annubar, and the water is measured by a weir plate in a weir box at the silencer. The water flow rate needs to be corrected for that flashed in the silencer. However, this method is expensive. Production well surface equipment is normally designed to allow regular output tests with minimum disruption.

There are other well tests, such as downhole well pressure and temperature measurements, that are normally done with the well on bleed. Once a well has been initially discharged, it should not be completely shut off because a column of cold water may form in the well, preventing the well from discharging again without stimulation.

The discharge of a well may interfere with the flow of a nearby well. Interference tests can be done by either measuring the downhole pressures or measuring the discharge from a well while a nearby well is discharging at different flow rates.

After enough wells have been tested to confirm the geothermal field's capacity and suitability for the production of electricity, steamfield design needs to be carried out. It can be carried out in parallel with

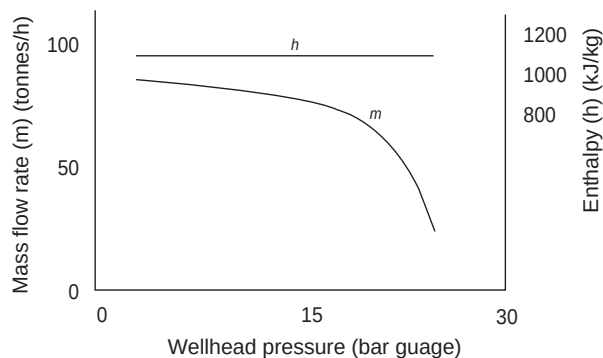


FIGURE 10 A typical geothermal well output characteristics.

further production well drilling. This is to determine the best method for collecting the geothermal fluids from the wells and processing them, especially when the geothermal fluids are two-phase ones. In general, a geothermal steamfield consists of the following major plants and equipment: the wells, the wellhead, the flasher/separator for a wet field, the silencer, fluid transmission pipelines, and a reinjection system.

4.1 The Wells

The use of a geothermal resource depends entirely on the output from the wells. A well's output can vary significantly from that of a typical well (Fig. 10). An example is the cycling wells that have fluctuating mass output with a cycle time as short as a few minutes. This type of well is not suitable for power production unless it can be balanced by stable output wells. Another type of problem well is one that continuously discharges solids with the well fluids. Sometimes this problem can be solved by throttling, but generally a dirty well is not used for production.

4.2 The Wellhead

A typical wellhead consists of several valves to provide service functions required (Fig. 11). The master valve is for completely shutting off the well. A small side valve is installed below the master valve to "bleed" the well when the main flow is shut off. It is not a good practice to completely shut in the well because the well may cool down and a column of cold water may sit in the well, preventing the well from discharging when the master valve is reopened. There may also be casing and cement damage from the thermal cycling. The well can flow again only when the column of water is removed by stimulation. Well stimulation is normally carried out by pumping air at the bottom of the well to airlift the flow. It can also be done by pumping steam or air into the well to depress the water column and then releasing the pressure quickly. Bleeding allows the well to stay hot.

It is difficult and costly to replace a master valve. A procedure must be in place to reduce damage during operation. A master valve is normally replaced by plugging the well below the master valve level, and this process normally requires a drilling rig. Some wells can be quenched by cold water injection through the bleed valves, allowing the valve to be changed. Because the master valve is so critical, it should be opened or shut only when there is no flow and should always be in either a completely shut or a fully open position. A sacrificial service valve is installed down-

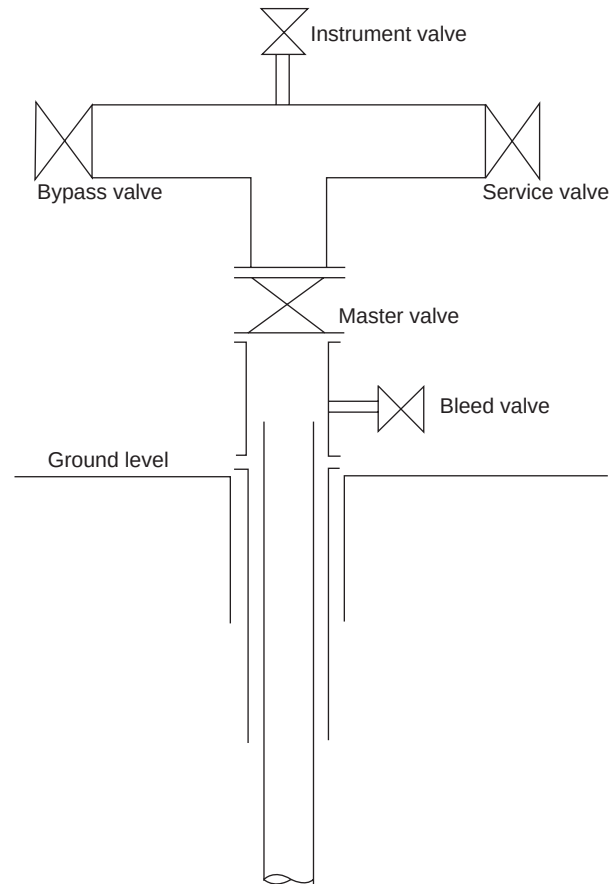


FIGURE 11 A typical wellhead and its valves' arrangement.

stream of the master valve. The service valve can be used for throttling, but long-term throttling of a well is normally done using a throttling orifice.

The service valve is commonly installed vertically above the master valve. It can be used to admit an instrument such as a Kuster gauge. The instrument is first lowered to sit above the shut master valve. The instrument is hung on a stainless steel wire that runs through a gland packing. The master valve is opened only after the gland packing is in place. The flanged spool with the gland system is called the recovery tube.

It is not common to make downhole measurements while a well is flowing. The instrument needs a long sinker bar to counter the buoyancy of the discharge. This means that a long recovery tube is needed. The flowing survey data are generally meaningful only if the well has a liquid feed zone of good permeability.

The bypass valve is required when the output of the well is not required for power generation for a short period, or for other reasons (e.g., tests, emergencies), and the well is not shut off due to possible interference. The flow is diverted through the bypass valve to a

silencer. For environmental reasons, discharge to waste via the silencer is not done unless necessary.

The arrangement of the wellhead valves normally results in the top service valve being several metres above the casing head flange. Earlier designs were to install the casing head flange in a deep cellar or at ground level, and a platform was built to access the service valve. Nowadays, a deep cellar is not preferred because hydrogen sulfide and carbon dioxide, being heavier than air, can accumulate in such a cellar, causing a lethal hazard. The preferred option is to install the casing head flange at ground level and have a tee piece just above the master valve and the service valve on the horizontal pipework. A small instrument valve or a blind flange is installed on the tee piece directly above the master valve.

4.3 The Flasher/Separator

Although sometimes a flasher is differentiated from a separator, the two are the same pressure vessel in practice. A flasher allows high-pressure geothermal water to flash at a lower pressure. In practice, the lower pressure is achieved by a throttling orifice or valve upstream of the flasher/separator. The vessel then separates the two phases by centrifugal and gravity effects.

The modern separators differ in many aspects from those first used at Wairakei during the 1950s. They differ in size, location, body arrangement, separated water vessel, two-phase inlet nozzle, water outlet nozzle, steam outlet nozzle, and efficiency. The original Wairakei separators were wellhead separators. Each wellhead has its own small separator handling the capacity of a single well. It can be as small as 750 mm diameter and 2 m high, handling about 20 tonnes/h of steam. A modern separator can be larger than 3 m in diameter and 15 m high, and it operates as a separation plant, accepting well output of several large wells. It can handle 600 tonnes/h of steam and is rated at more than 60 MWe.

The earliest separators at Wairakei were the Woods type, and some are still operating there today. A Woods separator has a U-bend-type separator operating in series with a top steam outlet cyclone separator that has a tangential two-phase inlet nozzle. The later design was the significantly cheaper Weber type with a bottom steam outlet (Fig. 12). A modern separator has a spiral inlet nozzle. This spiral inlet design increases the capacity of the separator by increasing the inlet superficial steam velocity from 20 to 40 m/s before the separation process breaks down. The water outlet nozzle is installed at the bottom of

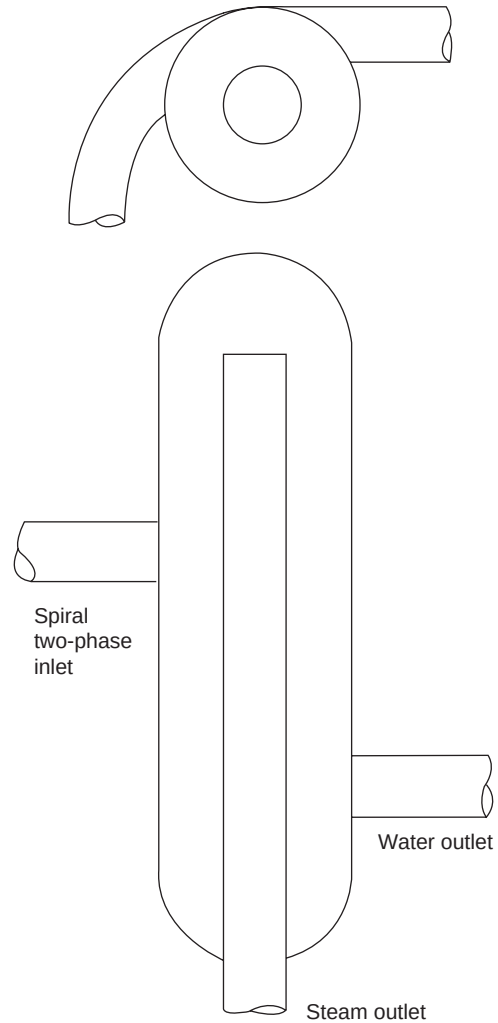


FIGURE 12 A common Weber-type geothermal separator.

the body exiting from the side. This is to improve the efficiency of water discharge to the water vessel, minimize erosion if the water is laden with solids, and minimize the risk of debris entering or blocking the water outlet pipe. The Weber separator has an efficiency that produces steam dryness exceeding 99.9% at design flow rates.

A modern separator design is an integral design that does not have a separate water vessel. Some modern designs incorporate an integral horizontal water vessel to reduce the height. To reduce height and earthquake effect further, large horizontal separators can be used.

4.4 The Silencer

A geothermal silencer is like a separator operating at atmospheric pressure. It is normally connected directly to a well and a separation plant for bypass

or emergency discharge. The silencer reduces the high-frequency noise discharging from an open pipe to a low-frequency rumbling noise that is more bearable to human ears. The earlier Wairakei silencer design consisted of twin barrels, and the body was made of timber. The modern silencer design consists of a single barrel made of corrosion-resistant steel, sometimes internally lined with concrete (Fig. 13). Fiberglass barrels worked well for silencers not in constant use. Some new silencer designs have a cone fitted at the top of the barrel to increase the steam velocity so that it is dispersed farther away from the silencer area. The noise can be further attenuated with an annulus full of scoria rock in the silencer barrel.

For a dry steamfield, a silencer is commonly just a pit full of rocks. A bypass pipeline discharges the dry steam at the bottom of the pit.

4.5 Fluid Transmission Pipelines

Many kilometers of pipelines are required in a geothermal steamfield to transport geothermal fluids from the wells and ultimately to the power plant. The geothermal fluids can be a two-phase steam/water mixture, separated steam, or separated water. All of the pipelines together with the separators, valves, and other plants and equipment in a geothermal steamfield have been collectively known as the energy gathering system (EGS), steam gather-

ing system (SGS), steam collection system (SCS), or fluid distribution and collection system (FDCS).

When wet steamfields, such as Wairakei and Cerro Prieto, were first developed, geothermal two-phase flow was not well understood and two-phase pipelines were kept to a minimum. So, separators were installed at the wellheads, and separated water and steam were transported as single-phase fluids from the separators to the desired destinations. So, each wellhead had a complicated pipe network.

Nowadays, two-phase pipelines are used extensively to transport geothermal fluids from several wells to feed a separation plant. This design greatly simplifies the pipe network at each wellhead and the whole steamfield because it allows the use of fewer large diameter pipelines in the steamfield. Another more important advantage is that large diameter pipes cause less pressure loss of the fluid being transported, thereby conserving energy.

The location of a separation plant is important because it affects well outputs and the total pressure loss and, ultimately, the economics of the power project. Two other factors affect the location of the separation plant: steam quality and steam pressure control.

The efficiency of a separator can be 99.99%. For a large separation plant, there may be 500 kg/h of water carryover in the separated steam. This water carryover needs to be removed as much as possible before it reaches the steam turbines because the water carryover in droplets traveling at high velocity erodes the turbine blades or solids precipitated from the water deposit in the turbine nozzles, ultimately choking the turbine output. There are two ways in which to remove the water carryover. One is to install a demister or scrubber at the power station end. The other is to install drain pots along the steam pipelines. The drain pots do not work if the steam velocity is high and the water carryover travels as droplets with the steam. There is also a minimum number of drain pots spaced at a minimum distance apart for the drain pots to function properly. This means that there must be a minimum length of steam pipeline downstream of the separation plant to allow for the scrubbing of the steam phase.

The second factor is that a minimum pipe length, and hence volume, is also required for the pressure control system to function properly.

4.6 Reinjection System

A major problem with a liquid-dominated steamfield is the large quantity of separated water that requires

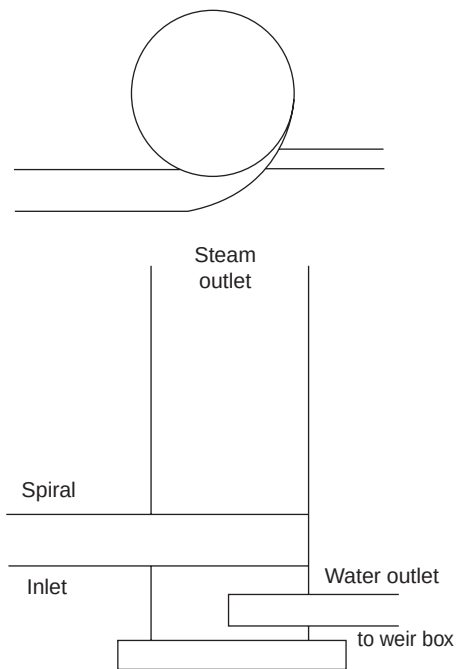


FIGURE 13 Modern single-barrel silencer.

disposal. When Wairakei was developed roughly 50 years ago, environmental concerns were not an issue and the separated water system was designed to simply drain into the nearby Waikato River. Nowadays, the only acceptable method is reinjection (unless the field is situated near the sea). Either the water flows down the well by gravity or, more often, a pump is required to increase the injectivity of the well. A control system is generally required to stop flooding or steam carryover.

The location of a reinjection well is important. A reinjection well within a steamfield can replenish the geothermal fluid that has been withdrawn, but cooler reinjected fluid may break through to the production wells and cool the production fluids. Nowadays, the strategy is to reinject at the perimeter of the steamfield to minimize the risk of the cooling effect. A problem with this strategy is that the permeability on the margins is generally poor, so more wells are needed. The aquifer flow paths can also change under exploitation so that reinjection wells can have rapid returns to production areas where initial tracer tests showed no flow. In the steam-dominated field of The Geysers, waste sewage water is reinjected to replenish the drawdown of the field.

5. GEOTHERMAL POWER STATIONS

The total installed geothermal power generating capacity in the world is approximately 9000 MWe from 21 countries, with the United States leading at nearly 3000 MWe and The Philippines with nearly 2000 MWe (Table II). Other major countries are Italy, Mexico, Indonesia, Japan, and New Zealand, with between 400 and 800 MWe each.

Geothermal power stations have very high availability, load, and capacity factors (>90%). They are most suitable as base load power stations. Liquid-dominated reservoirs are not suitable for peak load generation because two-phase flow and the separation process make their operations and control at changing flows difficult.

A geothermal power station that uses steam turbines to generate electricity is similar to a conventional fossil fuel-fired power station. The main differences are:

- Geothermal steam is usually saturated steam at less than 10 bar abs and 200°C (occasionally superheated up to ~50 bar abs and 300°C), but in

TABLE II

World Installed Geothermal Power Generating Capacity in 2000

Country	MWe
United States	2898.0
The Philippines	1909.0
Italy	785.0
Mexico	755.0
Indonesia	589.5
Japan	546.9
New Zealand	437.0
Iceland	170.0
El Salvador	161.0
Costa Rica	142.5
Nicaragua	70.0
Kenya	57.0
Guatemala	33.4
China	32.0
Russia	25.0
Turkey	20.4
Portugal (Azores)	16.0
Ethiopia	8.5
France (Guadeloupe)	4.2
Thailand	0.3
Australia	0.2
Total	8661.0

Source. Institute of Geosciences and Earth Resources, Pisa, Italy.

a fossil fuel-fired power station, the steam can be supercritical to as high as 300 bar abs and 600°C.

- Geothermal steam normally contains impurities of noncondensable gas (NCG) at up to a few percentage points by weight of steam (mainly carbon dioxide [$>90\%$] and a few percentage points of hydrogen sulfide) and dissolved solids (mainly silica or calcite) that precipitated from flashing of the liquid phase.
- A fossil fuel-fired power station has a boiler and the steam/water is in a closed loop, whereas a geothermal power station does not have a boiler and the steam/water is in an open loop.

Because of the low-pressure saturated steam, a geothermal steam turbine unit is limited to a maximum capacity of approximately 55-MWe condensing unit due to the maximum length of the last-stage turbine blades limited by the material metallurgy. Two of these turbines can be installed on a single shaft to a 110-MWe generator.

The NCG is allowed to pass through the steam turbines but must be removed from the condensers so

that the low subatmospheric pressure in the condensers can be maintained. A gas extraction system is used to remove the NCG from the condensers. A steam jet ejector is suitable for low gas contents of less than 1% weight of steam. A steam jet ejector has no moving parts, so it is easy to operate and has a low maintenance cost. However, the steam flow used to drive it is high ($\sim$ five times the NCG flow rate). Hence, steam jet ejectors are suitable only for low NCG levels. The NCG removed from the condensers is normally exhausted to the atmosphere.

For economic reasons, a gas exhauster is required when the gas contents exceed 2%. This is normally a multistage centrifugal compressor with an intercooler. Recirculation valves between the stages are necessary for control. It is typical to drive the compressor by electric motor, but the compressor can be driven by the steam turbine.

A liquid ring pump can be used to replace the final-stage steam jet ejector to improve efficiency and economy. Liquid ring pump systems are used in fossil fuel-fired power stations to maintain condenser vacuum. They have a small capacity, so they are suitable only for low gas contents.

The condensers mentioned previously are direct-contact condensers. Shell and tube condensers can also be used to better control NCG emissions. However, shell and tube condensers are expensive and less efficient.

When a geothermal resource cannot produce a sufficient quantity of clean enough steam to drive steam turbines, other modes of power generations need to be used. If the steam contains too much NCG ($> 10\%$ weight of steam), a condensing turbine unit becomes uneconomical. In such a case, a back pressure turbine that does not have an after-condenser can be used. The steam is exhausted directly to the atmosphere, but this mode of generation is unlikely to be acceptable nowadays for environmental reasons. Either total flow or binary mode of generation needs to be used.

Many geothermal ORC binary plants are in operation around the world. A heat exchanger (shell and tube type) is used to transfer heat from the geothermal fluids to a low-boiling point secondary fluid (isopentane or isobutane). The secondary fluid superheated vapor drives a vapor turbine that exhausts to an air-cooled shell and tube condenser. The secondary fluid is in a closed loop, and the spent geothermal fluids are reinjected.

Whereas an ORC binary plant usually uses an air-cooled condenser, a typical condensing steam turbine unit uses a direct-contact condenser. The resultant

condensate/cooling water mixture is normally removed from the condenser via a barometric leg or a pump. This condensate mixture flows to a holding pond known as the hot well, and from there it is pumped to a cooling tower to be cooled for recirculation to the condensers. As the geothermal steam from the steamfield eventually becomes a part of the recirculating cooling water, the total volume of the recirculating water tends to increase, even though some is evaporated at the cooling tower. Hence, some of the condensate mixture is reinjected from the hot well. Because this condensate mixture is warm ($\sim 40^\circ\text{C}$) and oxygenated, it is quite corrosive to carbon steel. Stainless steel pumps, pipes, and well casing are required for condensate reinjection.

Two types of cooling towers are commonly used for a geothermal power station: natural draft and mechanical draft cooling towers. For economic reasons, a natural draft cooling tower is built for an approximately 100-MWe capacity geothermal power station. Natural draft cooling towers also function better in a cooler climate. For warm climates, mechanical draft cooling towers are more suitable. Because the fan of a mechanical draft cooling tower is limited by the length of the its blades due to the strength of the construction material, a mechanical draft cooling tower has a small capacity. So, mechanical draft cooling towers are normally built as rows of cells.

6. ENVIRONMENTAL IMPACTS AND OTHER TECHNICAL PROBLEMS

Geothermal power stations have their own unique developmental, operational, technical, and environmental problems. Many geothermal resources occur in or near national parks or remote uninhabitable locations because they are normally associated with volcanoes and have beautiful surface manifestations and so become public reserves and tourism hot spots. Most of these geothermal resources do not get developed for power generation. Examples include Yellowstone Park in the United States and Orakei Korako Geyserland in New Zealand. However, geothermal resources that exhibit little surface manifestation in national parks do get developed. Examples include Hatchobaru and Otake in Japan.

Volcanoes, and hence geothermal resources, are also associated with mythical power and beliefs. The development of these geothermal resources requires much consultation and negotiation with

the natives inhabited in the areas. The final development of these resources normally includes some forms of partnership and reparation for the natives. Examples include Mount Apo in The Philippines and Ohaaki in New Zealand.

The development of a geothermal power project will have environmental impacts on the physical surroundings as well as on the socioeconomics and health of the surrounding communities. The physical environment includes the land, water, air, flora, and fauna. Most countries have strict environmental legislation such as the National Environmental Policy Act (NEPA) in the United States and the Resource Management Act (RMA) in New Zealand. The most important requirement of the environmental laws is the environmental impact report (EIR). This report is also called the environmental assessment (EA), environmental impact assessment (EIA), environmental impact statement (EIS), and assessment of environmental effects (AEE). The acts mandate processes that must be followed for the assessment and suggestions for any preventive and mitigating measures. Any disagreement may be redressed in an environment court. The EIR normally follows a checklist of environmental impacts. The positive result of an EIR is an environmental permit or consent, approved and issued based on the technical, environmental, and social acceptability of the proposed project.

Every geothermal resource is unique; hence, the actual physical environmental impacts of a geothermal power project development and its subsequent operation vary from one project to another. Some examples of the common physical environmental impacts are described next.

The land may be unstable, so any development may trigger landslide or soil erosion that could be made worse in areas of heavy rainfall. This is particularly true for many geothermal systems in difficult terrain associated with volcanic environments. Examples of geothermal systems with difficult terrain include The Geysers in the United States and Leyte in The Philippines. The actual land area required for a geothermal power plant is relatively small compared with that needed for other types of power plants when the areas of the resources are taken into account. A 100-MWe geothermal power plant requires approximately 1 km^2 of land; a coal-fired power plant requires approximately 4 km^2 , including the coal mining area; a hydropower station requires roughly 10 to 30 km^2 ; and a wind power station requires approximately 40 to 70 km^2 . The land uses that are unique to a geothermal power

plant are the pipeline corridors and geothermal well pads. Some geothermal areas are used for farming (e.g., Mokai and Rotokawa in New Zealand), but geothermal power plants and steamfields can coexist very well with other land uses such as farming and agriculture. Indeed, some development of geothermal systems provides aquaculture (e.g., prawn farm in Wairakei) and recreation (e.g., swimming pool at Svartsengi, Iceland).

Land subsidence is a common occurrence in geothermal development. Wairakei, after more than 40 years of operation, has a general subsidence of 10 to 30 mm/year over the whole steamfield. The maximum subsidence is approximately 250 mm/year, creating a small lake 18 m deep. Fortunately, the area of maximum subsidence is confined to a small area at the perimeter of the steamfield, causing minimal effects. Subsidence has also occurred at the Ohaaki after more than 20 years. It is difficult to predict land subsidence before the development of a geothermal steamfield. So, a monitoring program of the ground levels is required from the start of the development to ensure the safe operation of the steamfield as well as for environmental reasons. In general, the withdrawal of liquid from wet fields appears to cause greater subsidence than does withdrawal of steam only from dry fields.

The fluid production from a steamfield may also cause the surface manifestations to change, causing hydrothermal eruptions and creating new mud pools, fumaroles, and ground steam discharges. The changes to the "Craters of the Moon" near Wairakei are believed to be the result of fluid production at Wairakei. The withdrawal of geothermal fluids may also cause local micro earthquakes.

Gaseous discharge from the operation of a geothermal power plant is relatively small compared with that from the operation of a fossil fuel-fired power station. In general, the combustion of fossil fuels to generate electricity produces 3 to 20 times more carbon dioxide (kg/kWh) than that produced by a geothermal power plant. Although geothermal steam normally contains a small quantity of hydrogen sulfide that is not present in fossil fuels, the combustion of fossil fuels produces much more sulfur dioxide and nitrogen oxides that are not present in geothermal steam.

Water is required for the well drilling and construction phases. A circulating water system is required for the power plant during operation. Wet geothermal fields produce much separated geothermal water that requires disposal. The takings and discharges of water during the development,

construction, and operation stages all must be assessed for environmental impacts to the sources and sinks of water. All of these require water rights that are reviewed and monitored at regular periods.

Because the amount of land required for a geothermal power plant is relatively small, the effects on flora and fauna are also relatively small. However, being unique in nature, geothermal areas may also contain unique forms of flora and fauna. These flora (mostly fungi) and fauna (small organisms) may survive only in the warm areas of the geothermal activities. Their ecosystems might need to be protected and maintained.

Noise pollution of a geothermal power project will occur during the well drilling, construction, and operation phases. Because many geothermal areas are in remote areas where the background noise is very low (~ 40 dBA), any additional noise from geothermal activities will be heard by the local residents immediately. Impacts of noise on animals are reported to affect their behavior, physical size, and reproductive activities. Well drilling produces noise of 80 to 110 dBA, and a well discharging without a silencer produces 120 to 125 dBA of noise, which is at the threshold of pain for the human ear. The noise of a well discharging into a silencer is reduced to approximately 90 dBA.

Visual effects from a geothermal development include the tall drilling rig during drilling phases and the tall (~ 100 -m) natural draft cooling tower if adopted. Wet mechanical draft cooling towers are much lower, but their plumes are much more visible. The plumes can be minimized with a wet and dry cooling tower. Visual impact of geothermal pipelines has been minimized by painting them the color of the background.

Socioeconomic impacts are of most interest to the federal and local governments and the surrounding communities. The costs, benefits, and impacts on the local and wider communities will ultimately determine the success or failure of a geothermal project. Since the 1992 Earth Summit of the UN Conference on Environment and Development (UNCED) in Rio de Janeiro, Brazil, the emphasis has been on sustainable development, defined by the World Commission on Environment as “development that meets the need of present generation without compromising the need of future generations to meet their needs.”

A fundamental purpose of a socioeconomic impact study of a geothermal development is to identify the three key sectors that are affected by the project: investor, beneficiary, and victim. The final outcome of

the project must have the equity of benefit and compensation to all affected sectors. The main aim of a socioeconomic study is to determine the changes in conditions within, adjacent to, and in proximity to the project area as well as the direct and indirect impacts of these changes. This requires past and current baseline data, usually going back 10 years. These data are gathered from records, surveys, and consultations with the local communities and other interest groups. Social acceptability of projects are normally obtained if no extreme changes result and the affected communities can adjust their lifestyles to the new benefits that are gained. These benefits may include employment, lower electricity prices, and funding for local charities.

The most difficult problem that a geothermal power plant faces is the reliability of the steam supply. Well characteristics can and do change with time after a period of production. Many geothermal power stations (e.g., Wairakei, The Geysers) have never been loaded to their full capacities due to the shortfall of well output. The initial logical solution to this problem was to drill more wells, but this practice tends to worsen the problem with rapid drawdown. This problem is due to the overestimation of the reservoir power potential. The solution to this problem is staged development. The expected drawdown was incorporated during the design stage at Ohaaki, but the drawdown was faster than predicted.

A major operational problem that a geothermal power plant faces is the solid deposition of silica or calcite in wells and equipment exposed to the geothermal fluids. Although the chemistry of the deposition can be predicted with reasonable accuracy, the solution to the deposition problem is not always clear-cut. The common solution is to inject a chemical inhibitor into the problem wells. Suitable calcite inhibitors include polyacrylate, polymaleic acid, and phosphonate, depending on the application temperature. Many acids or alkali are suitable silica inhibitors.

The inhibitor is injected just below the fluid flashing point where the solid precipitates out of the liquid phase during flashing (Fig. 14). The inhibitor injection is very effective if applied correctly, although it does not actually stop deposition altogether. Physical removal of the solid deposits is still required once every few years instead of a few times each year without inhibitor. Not only is there a loss of production due to well shutdown to remove the deposits, but the cleaning may require a drilling rig or expensive acids.

Another hazard associated with a geothermal power plant is earthquake and/or volcanic eruption.

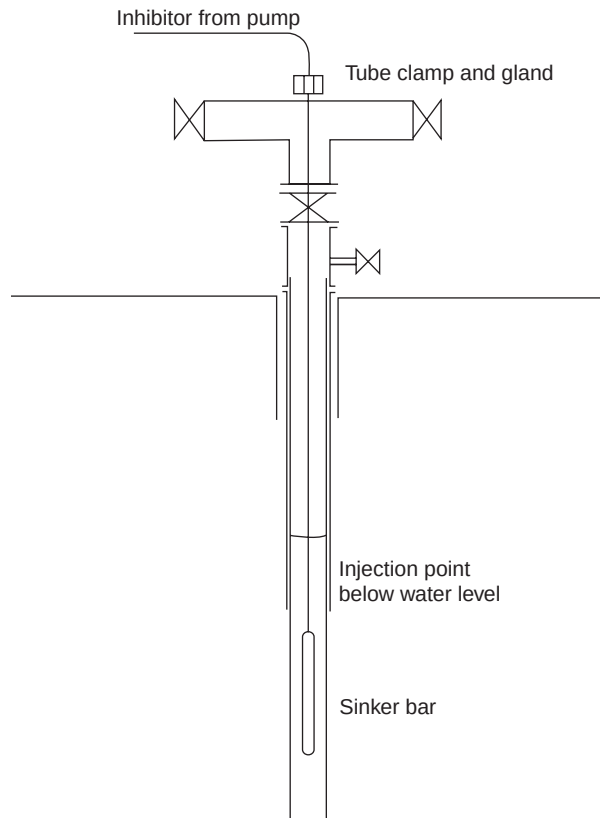


FIGURE 14 A typical inhibitor injection arrangement in a production well.

All plants and equipment can be easily designed to resist the expected seismic forces. Design for volcanic hazard is relatively easy if the discharge is slow-moving lava because it is simply a matter of locating the plant outside the lava flow path. It is much more difficult if the discharge is of the ash type.

7. COMMERCIAL DEVELOPMENT OF GEOTHERMAL POWER PROJECTS

In most countries, geothermal resources are classified as natural resources belonging to the government. Since the 1980s, there has been a worldwide trend to corporatize and/or privatize government commercial activities. Whereas the geothermal resources in the United States have been commercially developed by the private sector since the 1960s at The Geysers, other countries, notably The Philippines, Indonesia, and New Zealand, did not start private commercial geothermal projects until the 1980s. These private power developers are known as independent

power producers (IPPs). There is a perception that the private sector is more efficient than the public sector. The efficiency arises from the increased flexibility of the private sector and its accessibility to commercial funding.

There are several types of contracts in the commercial development of geothermal power projects. In a BOT (build, operate, and transfer) contract, a private developer builds the whole power station (including the steamfield), operates the power station for an agreed period of time to recoup its investment, and then transfers the power station back to the host country. In a BOO (build, own, and operate) contract, the developer builds and operates the power station as its own business. In a BOOT (build, own, operate, and transfer) contract, the developer transfers the power station back to the host country after an agreed period.

Developing a geothermal resource carries higher commercial risks than does building and operating a power plant. From a private developer's viewpoint, it is more favorable to just have an energy conversion agreement (ECA) contract in which the private developer builds and operates a power plant, converting the geothermal steam supplied by the host country and producing electricity for the host country (Fig. 15). The host country develops the geothermal resource. An ECA contract will involve a steam sales agreement (SSA) and a power purchase agreement (PPA). In a more traditional EPC (engineer, procure, and construct) contract, an engineering consulting company designs and builds the power plant and/or steamfield for a fee.

When private developers first entered the commercial development of geothermal power projects, they developed the resources, operated the steamfields, and sold steam to power companies operated by government agencies. The classic IPP operation is for the private sector to develop both the geothermal resource and the power plant and to sell electricity to the host utility that deals with the electricity market. The latest trend is the ECA operation.

An IPP, like any other investor, looks for a good return on its investment. To minimize the investment risk, the IPP wants certainty in costs and revenues over the investment period. A geothermal power project is like a hydropower project—capital intensive but with a low running cost. A geothermal project is commercially viable only if it creates a positive net present value (NPV) over its economic life for the investors. This requires a guaranteed stream of revenue in the future from the power sales, and the risks in the steam supply from the

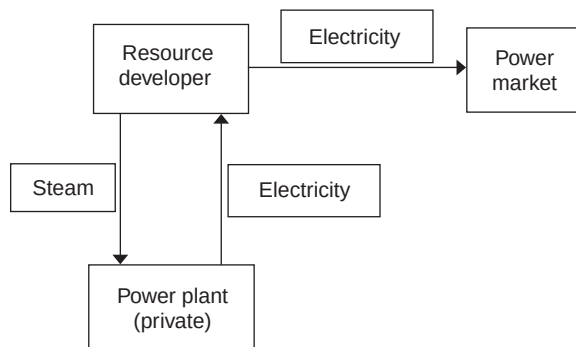


FIGURE 15 ECA project structure.

geothermal resource must be manageable. Therefore, the selection of the exploitation process, power cycle, and power plant is critical to minimize capital cost and maximize power output.

Risks are inherent in the resource and power plant development. Resource exploration risk is relatively high, with a corresponding higher return potentially at moderate investment. This type of risk is familiar to the mining and oil exploration companies. The steamfield and power plant development risk is low. Most IPPs and bankers are familiar with this risk. Risks are best managed by those who are best at controlling them. A developer is best responsible for the project implementation, and contractors are best at managing risks within their own areas of work. Some risks, such as resource and steamfield process risks, must remain with the owner. Risk allocation has driven the move toward EPC contracts.

There is an apparent conflict between the financial goal of maximizing return with a rapid development program and the risk mitigation strategy to minimize risk with a slow program. The solution to this apparent dilemma is to adopt a staged development.

The production drilling stage is a major expenditure phase, with each well costing U.S. \$750,000 to \$3 million. Well outputs are highly variable, from a dry well to one that may produce 40 MWe. Once production drilling starts, the aim is to produce electricity, and hence return on investment, as soon as possible. Cost controls are paramount during the production drilling stage.

A decision is required for the generating capacity of the power plant. For a staged development, the capacity of the first stage can be 25% of the estimated power potential. Because there are economies of scale in turbo-generator unit size, the tendency is to go for the largest single unit of approximately 55 MWe if possible. However, this requires optimization against operational flexibility and power demand. In general,

the power plant costs approximately 50 to 55% of the total cost of a geothermal power project, the production drilling costs 25%, and the steamfield costs 15%, with the remaining 5 to 10% of costs due mainly to exploration drilling. Currently, a geothermal power project costs between U.S. \$1000 and \$2000/kWe, depending on the quality of the steam. An electricity unit cost of U.S. 5 to 8 cents/kWh is achievable. This is very competitive compared with conventional thermal and hydropower stations.

Acknowledgments

I thank the following people for their support and assistance: D. H. Freeston, P. R. L. Browne, C. Morris, L. Bacon, K. L. Brown, and C. Z. Lee. I acknowledge the following permissions: M. Dickson and M. Fanelli for Table I, International Geothermal Association for environmental course materials convened by K. L. Brown (Geothermics) for Fig. 1, Geothermal Energy Association for Table II, H. Hole for drilling lecture notes, J. B. Randle for commercial development lecture notes, K. Koorey for Fig. 13, and K. C. Foong for steamfield lecture notes. I also thank the Geothermal Institute for office space.

SEE ALSO THE FOLLOWING ARTICLES

Geothermal Direct Use • Ground-Source Heat Pumps • Ocean Thermal Energy • Thermal Energy Storage

Further Reading

- Armstead, H. C. H. (1983). "Geothermal Energy: Its Past, Present, and Future Contributions to the Energy Needs of Man," 2nd ed. E & FN Spon, London.
- Clark, I. F., and Cook, B. J. (Eds.). (1992). "Geological Science: Perspectives of the Earth." Australian Academy of Science, Canberra.
- Dickson, M. H., and Fanelli, M. (Eds.). (1990). "Small Geothermal Resources: A Guide to Development and Utilization." UNITAR/UNDP Center on Small Energy Resources, Rome.
- Goff, F., and Janik, C. J. (2000). Geothermal systems. In "Encyclopedia of Volcanoes" (H. Sigurdsson, Ed.). Academic Press, San Diego.
- Grant, M. A., Donaldson, I. G., and Bixley, P. F. (1982). "Geothermal Reservoir Engineering." Academic Press, New York.
- Hochstein, M. P., and Browne, P. R. L. (2000). Surface manifestations of geothermal systems with volcanic heat sources. In "Encyclopedia of Volcanoes" (H. Sigurdsson, Ed.). Academic Press, San Diego.
- James, R. (1970). Factors controlling borehole performance. In "Proceedings of UN Symposium on Development and Utilization of Geothermal Resources" (Vol. 2, pt. 2, pp. 1501-1515). United Nations, Pisa, Italy.
- Kestin, J. (Ed.). (1980). "Sourcebook on the Production of Electricity from Geothermal Energy." Brown University Press, Providence, RI.

- Lee, K. C. (1997). Geothermal power technologies for different types of geothermal resources. *In* "New Energy Technology for Asia Pacific" (World Energy Council Asia Pacific Regional Forum, Vol. 2, pp. 189–203). WEC China National Committee, Beijing.
- Lee, K. C. (2001). Classification of geothermal resources by exergy. *Geothermics* 30, 431–442.
- Morris, C. (2003). "Operation and Maintenance Experiences of Wairakei Geothermal Power Station" (preprint). Contact Energy Ltd., New Zealand.
- Smith, I., Stosic, N., and Aldis, C. (1995). Trilateral flash cycle system: A high efficiency power plant for liquid resources. *In* "Proceedings of the World Geothermal Congress" (Vol. 3, pp. 2109–2114). WGC, Florence, Italy.