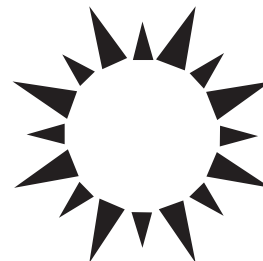


Mechanical Energy

JOSEPH PRIEST

Miami University
Oxford, Ohio, United States



-
1. Mechanical Energy
 2. Work
 3. Kinetic Energy
 4. Work–Energy Principle
 5. Conservative and Nonconservative Forces
 6. Potential Energy
 7. Conservation of Mechanical Energy
 8. Energy Conversion
 9. Applications of Mechanical Energy

Glossary

conservative force When the net work done by a force is zero for every path that ends up at the starting point.

energy The capacity or ability of an object to do work, with the joule (*J*) as the measuring unit.

Hooke's Law When an elastic object exerts a force proportional to the displacement of the object and in the opposite direction.

joule (J) The unit of work and energy; a force of 1 newton (N) acting over a distance of 1 m does 1 J of work.

kinetic energy The ability of an object to do work as a result of having mass and speed; in terms of mass (*m*) and speed (*v*), kinetic energy is $\frac{1}{2}mv^2$.

mechanical energy The sum of the kinetic energy and potential energy of an object.

nonconservative force When the net work done by a force is not zero in a path that ends up at the starting point.

potential energy The ability of an object to do work as a result of an advantageous position.

power The rate of doing work or converting energy; the watt (W) is the metric measuring unit, where 1 W is equal to a rate of 1 J per second.

restoring force A force on an object that tends to restore the object to its condition prior to application of the force.

watt (W) A rate of doing work or converting energy; the watt is the metric measuring unit, where a rate of doing work of 1 J per second is a watt ($P = \text{work}/t$).

work The result of a force acting on an object as the object moves from one position to another; in one-dimensional motion, work is the product of the force (*F*) and displacement (*d*).

The physical concepts of work, kinetic energy as energy due to motion, and potential energy as energy due to some favorable position are discussed in this article. Added together, kinetic energy and potential energy are called mechanical energy. In the absence of nonconservative forces (e.g., friction), the mechanical energy of an object does not change with time and is said to be conserved. The motion of a child on a swing is discussed in light of the conservation of energy. A wide variety of important types of kinetic energy are derived from the conversion of other forms of energy. Such is the case in a hydroelectric system that is discussed in some detail in the article.

1. MECHANICAL ENERGY

Energy has some meaning to everyone. A person often does not have energy following a bout with a cold. Concerns about depletion of our energy resources, solar energy, wind energy, and nuclear energy are common topics in newspapers and on television. Although energy has several meanings, in a physical sense it is considered “a capacity for doing work.” An object can have a capacity for doing work due to its motion and by virtue of an advantageous position. The combination of these two types of energy is called mechanical energy. The implications of mechanical energy are tied to the meaning of work.

2. WORK

A person does physical work when cleaning a room. A student does mental work when preparing for an exam. Whether physical or mental, work involves an effort directed toward producing some outcome. In the physical sense, effort is associated with force (push or pull), and work is done when a force acts on an object as it moves through some distance. The

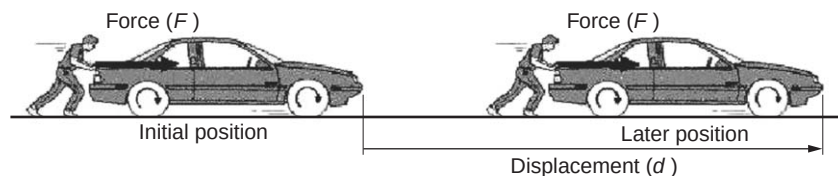


FIGURE 1 Illustration showing a car moving to the right that has been displaced by an amount d . During the displacement, the person exerts a force F to the right. The work (W) done by the person is the product of force and displacement ($W = Fd$).

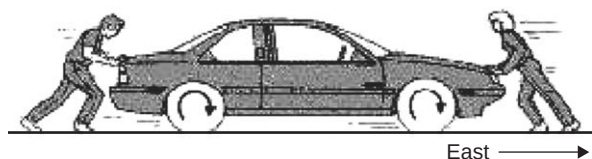


FIGURE 2 Illustration showing a car traveling east that is aided by a person pushing east and hindered by a person pushing west. The work done by the person pushing east is positive. The work done by the person pushing west is negative.

person cleaning a room does work on the broom by displacing it through some distance. In the simplest situation, the direction of the force is either the same as the direction of the displacement or opposite to the direction of the displacement. This being the case, the numerical value of the work (W) is the product of the force (F) and the displacement (d) (Fig. 1):

$$W = Fd \quad (1)$$

Measuring force in newtons (N) and displacement in meters (m) gives work the units of newton-meters. One newton-meter is called a joule (J).

The physical definition of work retains much of the popular notion with one important difference. No matter how much force (effort) is exerted, no work is done in the physics sense if the object does not move. A person may fret and sweat while pushing on a heavy box, but no work is done if the box does not move.

A force in the same direction as the direction of the movement produces an effect that is very different from that if the directions were opposite. A person pushing eastward on a car moving eastward tends to help the car along (Fig. 2). However, a person pushing westward on the same car tends to slow the car. We distinguish these two situations by labeling the work as “positive” when force and movement have the same direction and as “negative” when force and movement have opposite directions. The car is moving eastward under the action of one person pushing eastward and another person pushing westward. The person pushing eastward does positive work. The person pushing westward does negative work.

The net work on an object is the algebraic sum of the works done by each force acting on the object. Suppose that a sled moves 2 m eastward as a result of a girl pulling due east with a force of 100 N and a boy pulling due west with a force of 50 N. The girl does

$$+100 \text{ N} \times 2 \text{ m} = 200 \text{ J of work,}$$

the boy does

$$-50 \text{ N} \times 2 \text{ m} = -100 \text{ J of work,}$$

and the net amount of work is

$$+200 \text{ J} - 100 \text{ J} = +100 \text{ J.}$$

3. KINETIC ENERGY

A day does not pass without a person being involved in physical work. The forces that move one in walking and running perform work. Lifting food to one’s mouth and chewing food involves forces doing work. On some days, a person works more easily and more efficiently than on other days. One’s activities may be cast in terms of his or her ability to do work. Because physical work involves forces and movement, one might ask, “Under what conditions does something have a capacity for moving an object through some distance?” A car in motion has this capacity because if it slams into the rear of a car stopped at a red light, the struck car will surely move some distance. The energy associated with masses in motion is called kinetic energy. A big car colliding with a stopped car will “do more work” on the stopped car than will a smaller vehicle colliding with a stopped car. Similarly, a fast-moving car colliding with a stopped car will “do more work” on the stopped car than will a slow-moving car of the same type colliding with a stopped car. A numerical evaluation of kinetic energy should reflect these observations. Formally, the kinetic energy (K) of an object having mass m moving with speed v is defined as one-half the product of mass and square of speed:

$$K = \frac{1}{2}mv^2. \quad (2)$$

Both energy and work are measured in joules. Importantly, Eq. (2) shows that kinetic energy increases if the mass and/or the speed increase. Doubling the mass of an object while keeping its speed the same will double the kinetic energy. Doubling the speed of an object while keeping its mass the same will increase the kinetic energy by four times. The dependence on speed is especially significant.

4. WORK-ENERGY PRINCIPLE

At some time, an object such as a car may have speed v and kinetic energy $K = \frac{1}{2}mv^2$. Later, its speed may change to V so that its kinetic energy is $\frac{1}{2}mV^2$. The change in kinetic energy is the later value minus the earlier value, that is, $\frac{1}{2}mV^2 - \frac{1}{2}mv^2$. The work-energy principle states that the net amount of work on the object between the initial and later times is equal to the change in kinetic energy:

$$W_{\text{net}} = \frac{1}{2}mV^2 - \frac{1}{2}mv^2. \quad (3)$$

When a hockey player hits a stationary 0.16-kg puck with a hockey stick and imparts to it a speed of 45 m/s (100 miles per hour [mph]), its kinetic energy changes from zero to $\frac{1}{2}(0.16 \text{ kg})(45 \text{ m/s})^2 = +162 \text{ J}$. The change in kinetic energy is $+162 \text{ J} - 0 \text{ J} = +162 \text{ J}$, and the net work is $+162 \text{ J}$. The glove of a baseball catcher receiving a 0.15-kg baseball traveling 40 m/s (90 mph) reduces the speed of the baseball to zero. The change in kinetic energy of the ball is $0 \text{ J} - \frac{1}{2}(0.15 \text{ kg})(40 \text{ m/s})^2 = -120 \text{ J}$, and the net work is -120 J .

You see this principle in operation in many processes. The kinetic energy of water flowing over a dam increases as its speed increases. The increase in kinetic energy is the result of positive work done on the water by the gravitational force. When a car starting from rest is set into motion, its kinetic energy increases. This is due to work done by a force in the direction of motion of the car. Likewise, when the car slows down, its kinetic energy decreases. This is due to (negative) work by a force on the car acting in a direction opposite to the direction of motion.

5. CONSERVATIVE AND NONCONSERVATIVE FORCES

A box being lifted from the floor is acted on by an upward force provided by the lifter and a downward force due to gravity. Positive work is done by the

lifter because the force and movement are in the same direction. The gravitational force does negative work because the force and movement are in opposite directions. If the person lowers the box back to the floor, the gravitational force does positive work because the force and movement now have the same direction. The negative work done by gravity in the upward movement is equal in magnitude, but of opposite algebraic sign, to the work done by the gravitational force in the downward movement. Accordingly, the net amount of work done by the gravitational force in the round trip of going up and back down is zero. When the net amount of work done by a force in any trip that ends up at the starting point is zero, the force is said to be conservative. If the net amount of work done by a force in a trip that ends up at the starting point is not zero, the force is said to be nonconservative. A friction force is a nonconservative force. A friction force acting on an object always opposes its movement so that the work due to friction is always negative. Consequently, the net work due to friction in a trip that ends up at the starting point is never zero.

6. POTENTIAL ENERGY

When an object such as a box is in an elevated position, it has “a capacity for doing work” because if it is dropped and hits something, it can exert a force on that something and push it through some distance. The box at an elevated height is an example of potential energy. Potential energy is associated with conservative forces. By definition, the change in potential energy when moving from one position to another is the negative of the work done by the conservative force acting during the change in position. Labeling U the potential energy and ΔU the change in potential energy, the definition may be expressed as

$$\Delta U = -W_{\text{conservative}}. \quad (4)$$

6.1 Gravitational Potential Energy

The gravitational force on a box of mass m is mg , where g is the acceleration due to the gravitational force. If the box is raised a height h , the work done by gravity is $W = -mgh$. The change in potential energy of the box is then $\Delta U = -W_{\text{conservative}} = +mgh$. The positive sign means that the potential energy has increased. When the box falls from the height h , the work done by gravity is positive and the change in

potential energy becomes $\Delta U = -mgh$; the potential energy has decreased. The potential energy acquired by the box by placing it in an advantageous position can be recovered by letting it drop to its initial position on the floor. Water atop a dam in a hydroelectric plant has potential energy. When it falls toward the bottom of the dam, it loses potential energy but gains kinetic energy. At some point, the water does work by pushing on the blades of a water turbine, and the kinetic energy of the water is converted to rotational energy of the turbine.

6.2 Elastic Potential Energy

A material is said to be elastic if it returns to its original condition after being stretched or compressed. A spring, a rubber band, the bow of a bow and arrow, and a shock absorber on a car are examples of elastic objects. Somewhat like the earth pulling downward on a box that is being lifted upward, an elastic object pulls in an opposite direction to the force that is stretching or compressing it. The object has potential energy in the stretched or compressed condition due to its capacity to do work if it is released. For many elastic objects, the force exerted by the object, called the restoring force, is proportional to the extension or compression and in a direction opposite to the extension or compression. This being the case, the object is said to obey Hooke's Law. A force obeying Hooke's Law may be expressed as

$$F = -kx, \quad (5)$$

where F is the force exerted by the elastic object, x is the extension or compression, and k represents the strength of the force. For a linear spring, k is called the spring constant having units of newtons per meter. The stronger the spring, the larger the spring constant. The potential energy of a spring obeying Hooke's Law is given by

$$U = \frac{1}{2}kx^2. \quad (6)$$

Elastic potential energy is used in many ways. The kinetic energy acquired by an arrow has its origin in the elastic potential energy in the flexed bow. Pole vaulters acquire much of their vault from a bent pole. Toy guns expel projectiles by releasing the elastic potential energy of a compressed spring. Atoms in a molecule are held together by spring-like forces that lead to a form of potential energy. Release of that potential energy often leads to the emission of light.

7. CONSERVATION OF MECHANICAL ENERGY

Friction is always present in a mechanical system, but if it can be ignored, the mechanical energy (i.e., the sum of the kinetic energy and potential energy) does not change with time. To illustrate, a girl waiting to move down a playground slide has potential energy but no kinetic energy because she is at rest. Moving down the slide, she loses potential energy but gains kinetic energy. If friction between the girl and the slide can be ignored, the mechanical energy at any moment is unchanged. At the bottom of the slide, all of her initial potential energy would have been converted to kinetic energy. The constancy of mechanical energy is called the conservation of mechanical energy.

To the extent that friction can be ignored, the mechanical energy of a child on a swing is constant at any moment (Fig. 3). Held in an elevated position waiting for the swing to begin, the child has only potential energy. When released, the child gradually loses potential energy but gains kinetic energy. The sum of the two energies is unchanged at any moment. At the lowest portion of the swing, the potential energy is zero, making the kinetic energy a maximum. As the swing moves upward from the lowest position, the child gradually gains potential energy but loses kinetic energy. The child is momentarily at rest at the uppermost position of the swing, making the kinetic energy zero and the potential energy a maximum. If friction were absent, the back-and-forth motion would continue unabated and mechanical energy would be conserved at any moment. But as anyone who has taken a child to a playground knows, friction is always present and the motion gradually dies out unless the person pushes the swing to replace the energy lost to friction.

8. ENERGY CONVERSION

An object in an elevated position clearly has potential energy because if it is dropped and contacts something during its downward flight, it may do work on that something. Similarly, a compressed spring has potential energy because if it is released, it may strike something and do work on it. The general idea of potential energy as a capacity for doing work and the rearrangement of things when the potential energy is converted goes beyond these two mechanical cases. For example, gasoline has potential energy because if

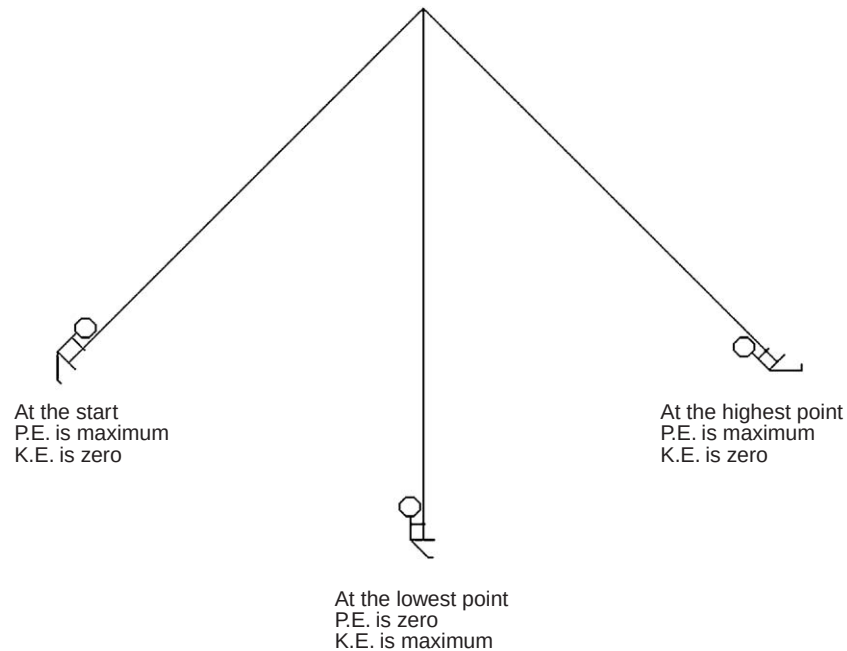


FIGURE 3 Illustration showing that, to the extent that friction can be ignored, the mechanical energy of a child in a swing is unchanged at any moment. P.E., potential energy; K.E., kinetic energy.

a gasoline vapor/air mixture is ignited in the cylinder of an internal combustion engine, the expanding gas pushes against a piston and does work on it. The root of the potential energy is found in the molecules from which the gasoline is formed. Energy is extracted from these molecules when their atoms are rearranged into different molecules during the combustion process. Similarly, the uranium fuel in a nuclear reactor has potential energy that is extracted from rearrangement of neutrons and protons through nuclear fission reactions. The energy produced from the nuclear reactions is then used to produce steam, which pushes against the blades of a turbine, producing rotational kinetic energy.

An industrial society finds myriad uses for kinetic energy. Pistons moving up and down in an internal combustion engine have kinetic energy. Wheels rotating on an automobile or a truck have kinetic energy. Water falling from atop a dam has kinetic energy. Around a home or in a factory, there are numerous motors providing rotational kinetic energy for a multitude of purposes. In all of these examples, the kinetic energy evolves from a conversion from some other form of energy. The force behind the movement of a piston comes from an expanding gas produced by the ignition of a gasoline/air mixture. The kinetic energy acquired by water rushing to the bottom of a dam comes from a conversion of

gravitational potential energy. Motors convert electric energy to rotational kinetic energy. Whether large or small, and whether simple or complex, converters producing kinetic energy all subscribe to the principle of conservation of energy. Each one converts energy into some form of kinetic energy regarded as useful, and each one diverts energy that is not immediately useful and might never be useful. Because energy is diverted, the efficiency defined as

$$\text{efficiency} = \frac{\text{useful energy}}{\text{total energy converted}}$$

can never be 100%.

9. APPLICATIONS OF MECHANICAL ENERGY

9.1 Pile Driver

A pile is a large metal or wooden post driven into the ground. A pile driver (Fig. 4) lifts a rather massive object (hammer) above the pile and drops it. Each drop of the hammer drives the pile farther into the ground until the required depth is reached. Lifting the hammer requires mechanical work. In so doing, the hammer acquires potential energy. When released, the hammer gradually loses potential energy

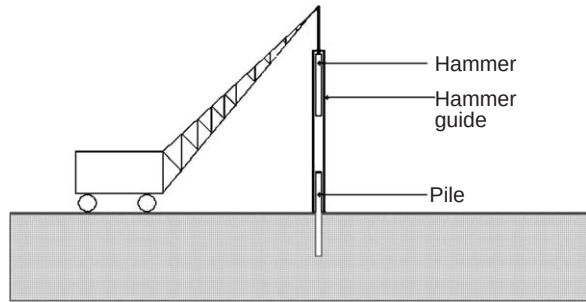


FIGURE 4 Basic components of a pile driver.

and gains kinetic energy. Work is done on the pile during contact, causing the kinetic energy to decline, eventually to zero.

9.2 Hydroelectric System

Whenever an object has gravitational potential energy, some agent had to do work on the object. In the case of a pile driver, there is a hoist of some sort. Water atop a dam has potential energy. Nature provides the means to fill the water reservoir through evaporation followed by condensation and rain. To see the energy transformations involved when the water falls from the top of the dam onto the blades of a turbine, it helps to visualize the motion of an object having boundaries. Accordingly, imagine the motion of a cube of water moving with the stream (Fig. 5). At the top of the dam, the cube of water of mass (m) and speed (V) has kinetic energy ($\frac{1}{2}mV^2$), and potential energy (mgh) due to its position at the top of the dam. As the gravitational force pulls the cube from the top of the dam to the bottom, it loses potential energy but gains kinetic energy. At the bottom of the dam, all of the potential energy (mgh) has been converted to kinetic energy ($\frac{1}{2}mv^2$). The total kinetic energy now includes its kinetic energy before falling plus the kinetic energy gained by falling:

$$\frac{1}{2}mv^2 = \frac{1}{2}mV^2 + mgh. \quad (7)$$

The cube of water arrives at the paddle wheel with energy ($\frac{1}{2}mv^2$). The force of the cube on the paddle wheel causes work to be done on the wheel. Accordingly, the paddle wheel rotates, acquiring energy (E) while the cube of water loses kinetic energy. The kinetic energy of the cube after it passes by the paddle wheel (K) equals the kinetic energy it had before striking the paddle wheel minus the energy acquired by the paddle wheel:

$$K = \frac{1}{2}mv^2 - E \quad (8)$$

In a hydroelectric power plant, the turbine would drive an electric generator and the energy (E) would be used for generating electricity.

9.3 Pumped Storage Hydroelectric System

The demand for electric energy by a community varies with the time of day and with the time of year. An electric power utility must be prepared to meet these demands. This poses an engineering problem because there is no practical method of storing electric energy on a scale that will meet the demands of a large community. To meet short-term increases in demand, electric utilities employ generators that can be turned on and off on short notice. For example, they may use a gas turbine similar to a jet plane engine to drive a generator. Another scheme is to use a pumped storage hydroelectric system. Such a system does not rely on nature to replenish the water in a reservoir but rather uses electrically run pumps. Importantly, the system can generate electricity on short notice. A schematic diagram of the system is shown in Fig. 6. Water is forced to an elevated reservoir by a motor-driven turbine. The water in the reservoir has gravitational potential energy by virtue of the work done on it. When electricity is needed, the water is allowed to flow downward into the turbine that drives the motor, which now functions as an electric generator. The energy required to elevate the water is never completely recovered in the process. Nevertheless, the system is economical because the reservoir can be filled when electric energy demands and costs are low. It is also possible to have a system in which water flows from ground level to underground turboelectric generators. In this case, work has to be done to restore the water to ground level.

9.4 Warning

Steel is a very hard metal that behaves like a very stiff spring when compressed or stretched. Bolting two steel plates together compresses each plate to some extent. Although the compression may be small, the elastic potential energy ($\frac{1}{2}kx^2$) in the plates can be large because the spring constant (k) is large. If the nut on the bolt holding the plates together is released gradually, the elastic potential energy declines gradually. But if for some reason the nut cannot be turned and the bolt is freed by chiseling the bolt in two, the elastic potential may be released suddenly, causing the nut to spring away. So violent is the separation that a person may be seriously injured if

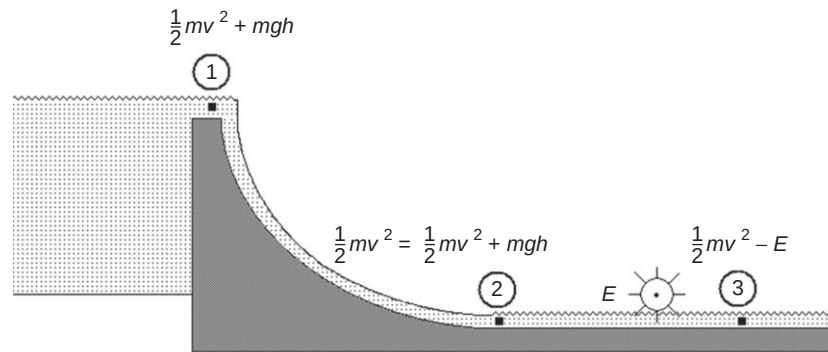


FIGURE 5 Energy transformations involved in water flowing over a dam. (Position 1) At the top of the dam, the cube of water has kinetic energy ($\frac{1}{2}mv^2$) due to its motion and potential energy (mgh) due to its position above the bottom of the dam. (Position 2) At the bottom of the dam, all of the potential energy (mgh) has been converted to kinetic energy. The energy of the cube of water is kinetic ($\frac{1}{2}mv^2$) and includes the kinetic energy it had at the top of the dam plus the kinetic energy acquired by falling over the dam. (Position 3) Passing by the paddle wheel, the cube of water transfers energy (E) to the wheel. Its energy after passing the wheel is still kinetic and is equal to its energy before impinging on the wheel ($\frac{1}{2}mv^2$) minus the energy (E) imparted to the wheel.

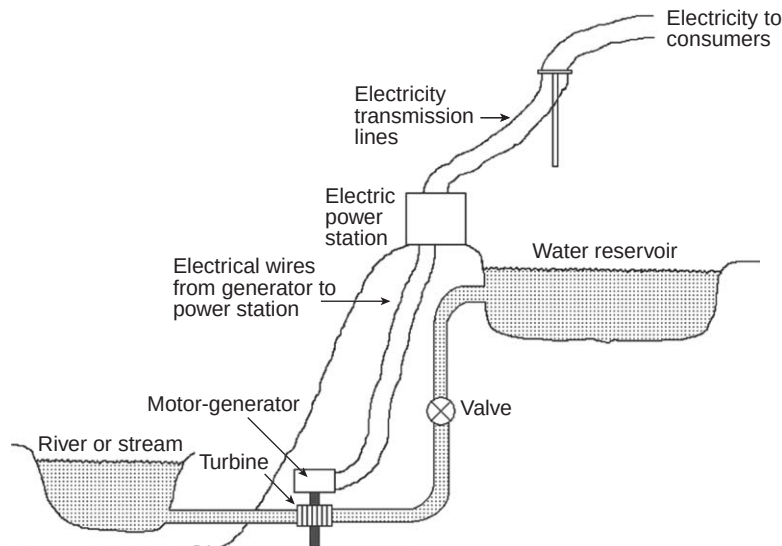


FIGURE 6 Principle of a pumped storage hydroelectric system. Water in an elevated reservoir has potential energy as a result of being pumped from a river or stream. Electricity is generated when the water flows through a turbine-generator combination on its way back to the river or stream.

he or she has the misfortune of encountering the nut. Such incidences have actually happened and have prompted warnings to workmen who may have to free a frozen nut.

SEE ALSO THE FOLLOWING ARTICLES

Conservation of Energy Concept, History of • Electrical Energy and Power • Energy in the History and Philosophy of Science • Forms and Measurement of Energy • Heat Transfer • Hydropower Technology

• Storage of Energy, Overview • Thermodynamic Sciences, History of • Thermodynamics, Laws of • Work, Power, and Energy

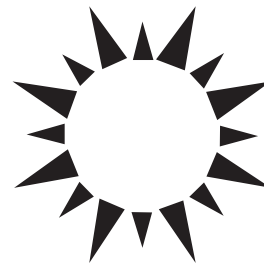
Further Reading

- Hobson, A. (2002). "Physics: Concepts and Connections," 3rd ed. Prentice Hall, Upper Saddle River, 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.

Media Portrayals of Energy

JAMES SHANAHAN

Cornell University
Ithaca, New York, United States



-
1. Introduction
 2. Media Coverage of Energy: History
 3. Public Opinion about Energy
 4. Media Effects on Public Opinion and Policy

Glossary

environmentalism A social movement of the 20th century focusing on the threats to human health posed by a variety of pollutants. Includes a broad swath of concerns, such as air and water pollution, climate change, wilderness protection, endangered species protection, and sustainability.

mass media The system of communication by which large, commercial organizations produce entertainment and news content for large, heterogeneous audiences. The term usually includes television, radio, newspapers/magazines, popular music, publishing, and films.

muckraker A name for a type of investigative newspaper or magazine journalist of the early 20th century. Muckrakers examined social problems, working conditions, environmental pollution, and other threats associated with the activity of large industry.

public opinion A collective indication or measurement of how the public feels about given issues. Usually measured using scientific random sampling techniques and opinion questionnaires or surveys. The term also refers to a more general, impressionistic perception of public sentiment on a given issue.

The mass media, especially newspapers, magazines, television, and movies, play a role in the portrayal of energy issues. Since the energy crisis of 1973, the attention of scholars has been turned to how media shape, frame, and influence audience perceptions of energy and the environment. During the period following the energy crisis, especially throughout the 1980s and early 1990s, issues of energy conservation were frequent topics of scholars' attention. Oil crises and the perceived dangers of nuclear

power sensitized society to the potential importance of conserving energy. These issues were also frequently mentioned in the media. In these earlier days of the environmental movement, it was believed that mass media could be prominent tools to encourage conservation of energy, through adoption of energy-saving measures in the home, driving smaller and more fuel-efficient vehicles, and promoting alternative energy source use. The attitudes of this period highlighted the fact that media can play an important role in people's individual choices. Influences such as advertising, journalism, television programs, and other media need to be considered when determining how people make choices in energy use behavior.

1. INTRODUCTION

In general, the influence of media on energy use has been considered within the wider field of study of media effects on environmental concerns. This field of research has shown that media do have both direct and indirect impacts on how people develop conceptions of the environment. However, these effects are not always consistent; they do not always move in the same direction. Thus, it has been shown repeatedly that information derived from the media is an important factor in how people develop awareness about the environment. Because people directly experience only a small portion of the environment, knowledge about global environmental issues and problems must come from mediated sources. People who are more attentive to newspapers and other journalistic sources are more likely to be informed and knowledgeable about environmental issues and problems. Those who are particularly attentive to environment-specific media sources (such as environmental magazines or television programs) are better informed about environmental issues and are more concerned about them. On the

other hand, general media attention (especially to entertainment media) is associated with less knowledge and concern about environmental issues. Frequent viewers of television programs, for instance, are less concerned about environmental issues in general. Further, they are more supportive of materialist ideologies that tend to promote higher energy consumption.

These countervailing examples show that it can be dangerous to lump all “media” together under one umbrella. Sometimes media outlets promote conservation of energy and concern for the environment; specific campaigns aimed at reducing energy consumption are often studied with this as the goal. Meanwhile, a more general attitude of promotion of consumption seems to guide the overall media atmosphere, which is built on advertising-supported ideologies of material growth. Given these tendencies, it is necessary to try to disentangle the various media effects, in order to get some idea about the key questions of interest. The main question that is the focus in this article concerns how the media influence public conceptions of energy within an environmental framework.

2. MEDIA COVERAGE OF ENERGY: HISTORY

Before discussing current issues in media coverage of energy, some of the main historical aspects of how media have treated the energy issue are summarized. This discussion will be contextualized within the wider theme of how the media treat the environment.

It is a common conception that environmentalism in the public mind and in the media frame is a phenomenon of the latter 20th century. However, there are clear examples of how the mass media provided a stage for a discussion of environment and energy issues well back into the 19th century. Preservationists such as John Muir, for example, were using the national media in their attempts to set aside lands for national parks. Muir’s articles in magazines such as *Century* attracted the attention of Eastern elite audiences to issues that were mainly situated in the West. These strategies helped develop national attention for environmental issues in the context of a nationally emerging mass media structure. Though a national mass media system had emerged as early as the 1830s, especially with the development of mass newspapers in major cities served by wire services, this structure was most

firmly cemented after the Civil War. How did this affect coverage of energy issues?

2.1 Coal and Oil

Attitudes about sources of energy such as coal and oil are recoverable from publications of the 19th century, which followed the development of new energy sources with close attention. However, as with media coverage of any major issue, it is not possible to determine a single prevailing attitude toward an energy source such as coal or oil. Certainly, many of the major media sources adhered to a generally nationalistic ideology, oriented toward goals such as economic progress, westward expansion, and even imperial aspirations. Within this context, fossil fuel sources obviously played an important role in the development of America as a world power. Thus, one major current of the media mainstream was oriented toward a view of fossil fuels as expanding the possibilities for America, without worrying too much about potential negative consequences.

But worrisome themes also appeared in the press. The economic booms and busts of oil discoveries were linked to boom-and-bust media cycles: high-hoped optimism followed by cautionary tales was a not uncommon cycle in the press. Stories told of those who made and lost fortunes were as characteristic of news coverage in the oil years as they were in the days following the Internet bubble. Oil was often presented as something ridiculously easy to obtain and sell, violating all previous rules of the marketplace. After a boom had gone bust, the national dailies turned to tales of woe and ghost towns in their coverage.

As oil became an industry, it attracted a different kind of attention, that of the “muckrakers.” Muckrakers were the forerunners of today’s “investigative journalist.” Their reformist spirit led them to attack industry on issues ranging from child labor to food safety to environmental issues. Ida Tarbell (see Fig. 1), writing in the pages of *McClure’s*, a national magazine, attacked the Standard Oil Company for its monopolistic and predatory practices. Her family life in the Pennsylvania oil patch had been disrupted by Rockefeller’s South Improvement scheme. Published serially from 1902 to 1904, Tarbell’s “History of Standard Oil” is considered a founding piece of investigative journalism. But not all media controversies resulted in immediate change. Concerns about ethyl leaded gasoline were covered fairly extensively, especially in the New York City area, from 1924 to 1926. Worker deaths from lead



FIGURE 1 Ida Tarbell (1857–1944).

poisoning led crusading scientists to point out that lead additives could be harmful to the public at large, due to risks from lead deposition in exhaust. Although a committee was formed by the Surgeon General to look at the issue, no action was taken, even with significant attention from newspapers such as the *New York World* and *The New York Times*. It was not until the 1970s that the Environmental Protection Agency (EPA) would take action to remove lead from gasoline. Still other themes were developed earlier than one might expect. Recognizing American dependence on oil led to journalistic calls for conservation and development of renewable energy sources long before the onset of the environmental era in the 1970s. The concern was not environmental in nature, but stemmed from a worry about maintaining America's preeminence despite dependence on oil.

Typically, media attention to environmental issues is driven by both journalistic cycles and events. Events are spectacular disasters or phenomena with widespread impact. A "killer smog" in Donora, Pennsylvania in 1948 drew enough media attention to catalyze the development of smoke-abatement programs in cities around the United States. This was an early precursor to the more far-reaching programs of air pollution control that were eventually enacted in the 1970s. Indeed, air pollution themes had received sporadic attention throughout the 20th

century, but the atmospheric impacts of fossil fuel consumption began receiving a closer look in the media in the 1960s. This peaked with the passage of national environmental legislation (the National Environmental Protection Act) in 1970.

Fossil fuels and American industry in general came under further media scrutiny in the 1960s. Rachel Carson's *Silent Spring* (1962), though not focused directly on fuels or energy sources, drew attention to the chemical industry and its promotion of dangerous chemicals such as dichlorodiphenyl trichloethane (DDT). *Silent Spring* sensitized Americans to risks both known and unknown in and about the general environment. Carson arguably inaugurated the age of environmentalism, in which the media played an important role in turning people's attention to a variety of environmental issues, with fossil fuel consumption at or near the top of the list of problems. Carson's book, serialized in the *New Yorker* magazine, received wide media attention, building on the earlier model for journalistic crusading that had been successfully tested by both John Muir and Ida Tarbell.

This touched off a frenzy of popular publishing activity focused on energy consumption and environmental issues. Books such as *The Waste Makers* (Vance Packard, 1960), *The Quiet Crisis* (Stewart Udall, 1963), *The Population Bomb* (Paul Ehrlich, 1968), *The Closing Circle* (Barry Commoner, 1971), *The Limits to Growth* (Donella Meadows *et al.*, 1972), *The Poverty of Power* (Barry Commoner, 1975), *The Eco-Spasm Report* (Alvin Toffler, 1975), and *The Fate of the Earth* (Jonathan Schell, 1985) are examples of popular publications that made some impact on public sensibilities and opinions about energy consumption problems. These efforts were strengthened by press attention to the oil crises of the 1970s, which caused a great deal of concern about energy dependence. With energy issues bouncing around the press throughout the 1970s, the issue was often at the top of the public agenda, particularly when gasoline prices were rising. With a cast of villainous characters [the Organization of Petroleum Exporting Countries (OPEC) oil sheiks] and an issue striking at the heart of the American economy, foreign oil was the biggest news item in 1973, and to a lesser extent in 1979. However, specific events also masked cycles of press concern about energy issues. These cycles are generated not by specific external motivating factors, but by tendencies to adhere to narrative structure embedded within journalistic practice. For instance, in the late 1980s and early 1990s, journalists in both newspapers and magazines

and on television turned their attention to a broad range of environmental issues. The attention was initially brought about by an extremely hot and dry summer in 1988. Scientists' claims about global warming suddenly found a foothold in the media, when James Hansen of the National Aeronautics and Space Administration (NASA) claimed that global warming could be definitively tied to human agency. Bill McKibben published a widely read book, *The End of Nature* (1989), which argued that climate change meant that human influence was suddenly spread throughout the entire environment, and there was no more pristine "nature" to be found. The book was also serialized in the *New Yorker*.

Suddenly, the press was covering a wide range of issues, from climate change to acid rain to radon gas to ozone depletion. The years 1988–1991 represented an unprecedented period of public concern and attention to environmental issues. The Exxon Valdez accident (1991) was yet another event that encouraged even more environmental debate in the media. It was during these years that the public and policy debate began to give more credence to the idea that the media could be a positive force for social change, especially in the big, important areas such as climate change. Environmental concern was so strongly present in the mediated discourse of the United States that companies such as McDonald's and Burger King adopted new forms of packaging to counter charges that styrofoam hamburger containers consumed too much energy in production and were persistent in the waste stream. Even entertainment television began to manifest signs of a green consciousness. Hollywood celebrities increasingly focused on environmental issues such as rain forest conservation. Public relations agencies with the purpose of placing environmental references in television programs and feature movies came into being. Movies with explicitly and implicitly environmental themes were more common. *Dances With Wolves* (1990) was a prototypical environmentally themed movie of the period, focusing on the connection of the Sioux Indians to the land. Movies with more directly energy-related themes include *The China Syndrome* (1979), about a nuclear disaster uncovered by an inquiring reporter; the *Mad Max* series of movies (1989–2000) about a postapocalyptic Australian outback gripped in wars for the dwindling supply of oil; *Waterworld* (1999), a story about a post-climate-change world that has been inundated with water; and *Ferngully* (1992), a children's story about a destroyed forest that threatens a race of environmentally conscious sprites.

On television, children's shows such as *Captain Planet* (1990–1996) often focused on themes of energy conservation.

After 1991, however, general media attention began to subside (see Fig. 2). Vice President Al Gore published *Earth in the Balance* in 1992, but he missed the peak of media and public excitement. The eruption of Mount Pinatubo in the Philippines led to cooler temperatures globally, which also seemed to cool press interest. Indeed, research has shown that news attention to climate change issues varies with the actual temperature, such that journalists are less likely to cover climate change during cooler periods. In any case, coverage for all environmental issues was declining throughout the mid-1990s. The pattern was following that predicted by social scientist Anthony Downs, who argued that press attention to environmental issues would always be cyclical due to the inherent nature of the issues. Because environmental issues are difficult to solve, public attention will fade when the costs of achieving environmental gains are calculated. Also, the issues are seen as neither interesting nor fascinating from a journalistic standpoint. Other theorists also pointed out that the news agenda has a carrying capacity. When one issue rises in salience, others decline. The first Gulf War (1991), even though it had an important energy motivation, turned attention away from the environmental aspects of energy issues.

By 1995, environmental issues were neither as much in the media nor in the scope of public attention. The celebration of "Earth Day," which in 1990 had received a lot of media attention, went barely noticed in 1995. A strain of antienvironmental

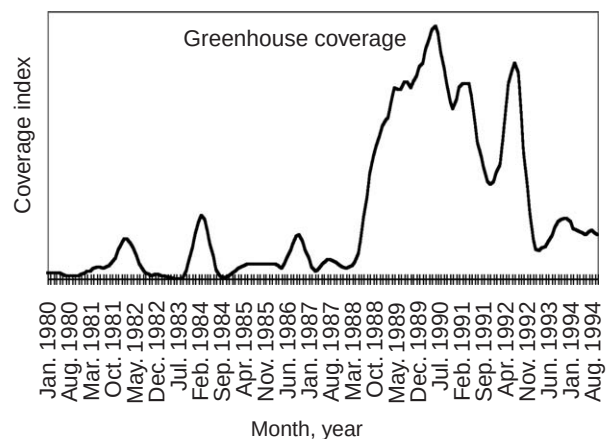


FIGURE 2 Coverage of climate change reported in *The New York Times*. From Shanahan and McComas (1999).

thinking was also becoming evident in the works of authors such as Julian Simon, who preached environmental optimism in works such as *The Ultimate Resource*. Readers of popular literature could see these ideas in books such as Greg Easterbrook's *A Moment on the Earth* (1995). Also, the energy industry was not at all inactive. Industry public relations groups such as the Western Fuels Association had engaged in active public communication campaigns to discredit the science associated with predictions about climate change. To some extent, these campaigns were successful, and public opinion eventually turned away from concern about climate change. Although the scientific community was not dissuaded in its opinions about climate change, lack of public resolve and media attention has made it difficult to develop stronger U.S. policies on climate change.

Since the mid-1990s, energy issues have struggled to receive major attention. Cheap oil has fueled the popularity of sport utility vehicles and trucks, which have been advertised heavily by car producers. Indeed, the reliance of the media on the automobile and energy-producing industries, through advertising revenues, is an important factor to consider when examining coverage of energy issues. Although the major media certainly did not shy away from criticism during disasters such as the Exxon *Valdez* incident, some critics have pointed out that mainstream news organizations are too reliant on funds from "Big Oil," or from Detroit.

Energy consumption, thus, is not currently viewed as a major issue on the American agenda, with the minor exception of concerns about dependence on oil and its relation to issues of susceptibility to terrorism. September 11, 2001 and the Iraq War (2003) activated certain sectors of society to begin thinking again about U.S. reliance on foreign oil. Slogans such as "No blood for oil" encapsulate the thinking of the social protest sectors that have sought to mobilize wider opposition to U.S. foreign policy in the Middle East. But the first Gulf War and the Iraq War were both very popular. The U.S. media sector assisted the war efforts in both cases by focusing on issues of patriotism and nationalism; their efforts were rewarded with strong opinion support. Energy issues were widely disregarded in both cases, except insofar as U.S. forces achieved successes in "protecting" energy resources such as the Iraqi oil fields. The media, having passed through a period of concern about environmental issues, revealed little of that concern in recent times.

2.2 Nuclear Energy

Nuclear energy, compared with traditional energy sources such as coal or oil, has received much different treatment in the media. Nuclear energy has always been portrayed in a more bipolar fashion, as offering either fantastic possibilities or horrifying risks. Coverage has tended to swing between these poles without resting much in the middle. As a consequence, public opinion has been more mistrustful.

After the detonation of atomic bombs at the end of World War II, there were attempts to promote peaceful uses of atomic energy. It would be fair to say that pro-atomic energy campaigns played a large role in shaping the media atmosphere. During the 1950s, the government encouraged a positive strain of thinking about nuclear energy that permeated the media. Dwight Eisenhower's "Atoms for Peace" speech of 1953 is one example of an attempt to create a rubric for public communication about nuclear energy that could escape from the dread perceptions engendered by its military history. But the fear brought about by the Cold War was made manifest in media coverage of various types. It was difficult for nuclear power to escape negative associations. Although many nuclear reactors were built and used, the attitude of the entertainment media in the 1960s toward nuclear power was focused either on science fiction or on doomsday scenarios. Movies such as *Failsafe* (1964: "It will have you sitting on the brink of eternity!") both informed and characterized the public fascination with and fear of nuclear weapons; also notable was Stanley Kubrick's *Dr. Strangelove* (1964), a satire of nuclear militarism based on *Failsafe*. Transfer of public fear from nuclear weapons to nuclear energy was probably inevitable. At a more quotidian level, nuclear energy had been portrayed as the source of horrific mutations and science-fiction accidents in more movies than can be recounted here [though they include *Attack of the Crab Monsters* (1957), *The Beast From 20,000 Fathoms* (1953), *The Creation of the Humanoids* (1962), *Godzilla, King of the Monsters!* (1956) and *Them!* (1954)]. And many movies (as well as television programs such as *The Twilight Zone* and *The Outer Limits*) used post-nuclear-holocaust scenarios as a standard setting or as a formulaic plot device.

Mediated public debate about nuclear energy as a power source emerged more strongly in the media in the 1970s. One major factor was the emergence of antinuclear activism. Media coverage of environmental activism can be a two-edged sword. Oftentimes,

specific stories about environmental issues present the environmental perspective as out of the mainstream. Activists and their causes are not at the core of public opinion concern. However, the salience derived from media coverage has been very beneficial to environmental causes. The nuclear activism issue is an excellent case in point.

Risk studies have shown that the risk from nuclear energy is perceived as an “imposed” risk. People are more likely to express concern about risks that they do not normally assume in the course of their daily lives. When such risks are covered heavily in the media, it may make little difference that actual risk ratios are relatively low. Such was certainly the case for nuclear power, which experienced negative attention from all sides, and never recovered from negative associations with the atomic bomb. The key media event that affected perceptions of nuclear power was Three Mile Island. Walter Cronkite (“the most trusted man in America”) and other mainstream media were arguably quite alarmist in their coverage, making connections to the recently released movie *The China Syndrome*. Supporters of nuclear power have been extremely critical of the media model that was established with this incident; they blame it for the fact that citizens estimate risks from nuclear power as much higher than risks known in fact to be more dangerous, such as cigarette smoking.

With Chernobyl, a similar dynamic was introduced, although the media took a different tack given that the accident occurred within the Soviet Union. Studies showed that news agendas were dominated by the accident in the U.S. and especially in Europe. Though some coverage emphasized the relative safety of U.S. reactors compared to Soviet ones, this incident added yet another negative to the list of achievements of nuclear power. In the time between the two accidents at Three Mile Island and Chernobyl, the broadcast by the American Broadcasting Corporation (ABC) of *The Day After* (1983) was arguably the biggest entertainment media event of the 1980s that dealt with nuclear energy. It told the story of a U.S. city after a nuclear attack. It was watched by half of the U.S. population, and was seen as a feature film in many foreign countries. The Reagan White House was concerned about its antinuclear bias. The broadcast’s effects on public opinion and knowledge about nuclear power were also studied, with inconclusive results, but undoubtedly few positive effects accrued for the nuclear energy industry, which was increasingly under fire.

3. PUBLIC OPINION ABOUT ENERGY

Public opinion polls have frequently asked citizens to comment on their thoughts, feelings, fears, and attitudes about energy sources. In terms of attitudes about energy in general, opinion and concern have varied. Through the 1970s and into the 1980s, in a variety of polls, about 80% of U.S. respondents said that the U.S. energy situation was “serious.” This reflected general concern about energy prices and reliance on foreign oil. When assessing blame for the energy situation, respondents have been most likely to finger oil companies, although Arab states were often blamed as well. Oil companies, as is true of most energy institutions, receive very low levels of public support (nuclear power institutions receive the lowest ratings of all) in public opinion polls.

In the late 1980s and early 1990s, the U.S. moved beyond a debate on the energy supply, which was largely not an issue anymore. In the late 1980s, more consumers turned toward larger vehicles, and questions about greenhouse emissions and the role of fossil fuels in climate change moved to the front of the public agenda. These issues were foregrounded by the media, resulting in unprecedented levels of public concern about fossil fuels into the early 1990s. However, these concerns faded with the onset of the Gulf War and the economic downturn that followed. Coverage of climate change and other issues followed a classic cyclical pattern, as was noted above. How did this affect public opinion? One series of data showed that public concern about governmental spending on environmental issues was strongest during the time of peak media coverage. Concern was stronger among those who were heavier consumers of news media. In the late 1990s, concern about fossil fuels tended to drop, along with general concern about environmental issues. Although major international conferences on climate change, such as those held in Kyoto in 1997, did provide opportunity for new spates of energy-related coverage, these often focused on U.S. refusal to go along with world climate policy.

In general, as already noted, compared with fossil fuels, nuclear energy has sparked more negative feelings. In September of 1945, 93% of respondents to a National Opinion Research Center poll felt that there was a real danger of most people living on Earth being killed by nuclear bombs if a new world war broke out. But still, citizens generally strongly felt that it was acceptable to have used the bomb in 1945; they were also generally mixed on whether

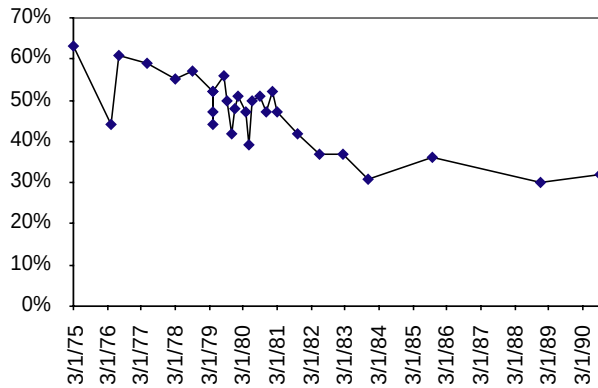


FIGURE 3 Percentage of poll respondents favoring building more nuclear plants in the United States (Harris poll data).

splitting the atom would turn out to be a net positive or negative for society.

In the 1970s, after the first oil shock, U.S. citizens felt that it was extremely important to develop alternative sources of energy. A 1976 Harris poll found that 80% of respondents favored developing nuclear energy sources, along with other sources such as solar energy and more oil exploration. But then, among other influences, news about nuclear power accidents in Three Mile Island and Chernobyl had effects on public opinion, though not initially as drastic as one might expect. Some researchers argued that nuclear power could “rebound” from the negative effects of such disasters. But over time, this proved not to be the case, as the public moved consistently toward negative views of nuclear power in the late 1980s and 1990s (Fig. 3). In addition to the perceived dangers evident in the disasters, other factors affecting support were antinuclear activism (often media based) and the perceived high expense associated with constructing nuclear plants. Also, the fictional portrayals mentioned above (*The China Syndrome* and *The Day After*) probably played a role.

4. MEDIA EFFECTS ON PUBLIC OPINION AND POLICY

To what degree do the media affect public attitudes about energy? Although many studies have examined relationships between media use and environmental attitudes, no clear finding has emerged. On a case-by-case basis, it is often difficult to show how specific media messages change public opinion, apart from extremely salient mass spectacles such as *The Day After*. For instance, studies of the effects of messages such as President Carter’s “moral equivalent of war”

speech on energy conservation showed minimal effects in terms of citizens actually conserving. Studies that attempted to tie proconservation media coverage or campaigns to specific energy consumption habits (such as lowering the thermostat) have also not shown much effect. At the micro level, it is clear that there are a myriad of forces that determine individual energy decisions, among which the media may play only a small role.

At a macro level, however, it is difficult to imagine a more powerful institution than the media for setting the social agenda and defining terms. Repeated exposure to messages favoring increased material consumption does play a role in actual consumption, as studies have shown. Viewers who spend a lot of time watching television (“heavy” viewers) are more likely to believe that the world is more affluent than it really is, they desire more material possessions, and they are less likely to be concerned about environmental issues. Such television viewers are even more likely to answer opinion questions about material wealth more quickly than “light” viewers; the implication drawn by psychologists is that heavy viewers work from a store of media-cultivated images that allow them to heuristically view the world as wealthier than it actually is. Media images also play a major role in determining people’s risk perceptions and fears, particularly about nuclear energy. Thus, we can look at media effects at two levels. Environmental activists, energy producers, and other active participants in social policy focus most often on the day-to-day slate of messages that a citizen receives. These messages convey a welter of often conflicting impressions about the state of energy, what the polls refer to as the “energy situation.” “Actors” with significant resources can play a major role in attempting to tip the balance of these conflicting messages. Issue-advertising models along the lines of those developed in the 1970s by Herbert Schmerz of Mobil Oil Company are one way that energy producers attempt to influence the policy discourse. As well, since the 1990s, most major energy producers have added environmental divisions to their corporate structure, in recognition of the increasing public relations importance of environmental issues. On the other hand, the incredible development and growth of environmental activist organizations have balanced out the claims and arguments of energy producers. Groups such as Greenpeace have been persistent and successful in developing media strategies. Groups with specific media missions include the Environmental Media Association, which strives to

make Hollywood productions more environmentally relevant.

Still, the fundamental tone of the media is oriented toward growth, toward what some sociologists called the "Dominant Social Paradigm." This paradigm values economic growth and encourages confidence in the ability of new technologies to fuel such growth. Scholars who have examined media from a broad cultural perspective have been most impressed with contributions of advertising to the creation and maintenance of this paradigm. Even though sociologists have detected the outlines of an emerging "New Environmental Paradigm," there is little doubt that the media system is still fundamentally structured to encourage growth in consumption.

Bill McKibben summed up these views in his *The Age of Missing Information* (1992), which dealt with a comparison of world views obtainable in the media versus those obtained in the real world. To McKibben, the power of television was not its specific messages, but rather the dizzying amount of content, most of which was focused on material consumption in one form or another. He complemented the views of scholars such as George Gerbner, who argued that the true effect of the media is the extent to which it cultivates perceptions of reality. The efforts of media researchers to document that the media do have an influence on perceptions of reality have not been in vain. In the energy sphere, consumption of media is tied to materialism, support for the dominant social paradigm, and less concern about environmental issues such as energy conservation.

Thus, the media have not changed much since the days of early coverage of oil booms. In the media, energy is the fuel of our technologically materialist culture. At the same time, media will sporadically play a watchdog role in suggesting environmental safeguards as far as our use of energy is concerned. Thus, even despite the various disasters, risks, problems, and accidents associated with fossil fuel consumption, the media do not fundamentally question our reliance on these fuels. Perhaps this is because everyone realizes we have no real alternative. On the other hand, with nuclear energy, we see

the veto power of the media in action. In that nuclear power has not been an essential energy source, it has also been fair game for media criticism.

SEE ALSO THE FOLLOWING ARTICLES

Climate Change and Energy, Overview • Conservation Measures for Energy, History of • Consumption, Energy, and the Environment • Environmental Change and Energy • Geopolitics of Energy • Global Energy Use: Status and Trends • Lifestyles and Energy • Oil Crises, Historical Perspective • Public Reaction to Energy, Overview • Public Reaction to Nuclear Power Siting and Disposal • Public Reaction to Renewable Energy Sources and Systems

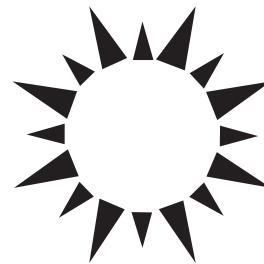
Further Reading

- Allen, C., and Weber, J. (1983). How Presidential media use affects individuals' beliefs about conservation. *Journalism Q.* 68(1), 98–110.
- de Boer, C. (1977). The polls: Nuclear energy. *Public Opin. Q.* 41(3), 402–411.
- Downs, A. (1972). Up and down with ecology—The "issue attention cycle." *Public Interest* 28, 38–50.
- Erskine, H. (1963). The polls: Atomic weapons and nuclear energy. *Public Opin. Q.* 27(2), 155–190.
- Farhar, B. (1994). Trends: Public opinion about energy (in the polls). *Public Opin. Q.* 58(4), 603–632.
- McKibben, B. (1989). "The End of Nature." Random House, New York.
- McKibben, B. (1992). "The Age of Missing Information." Random House, New York.
- Neuzil, M., and Kovarik, W. (1996). Conflict management and scientific controversy. In "Mass Media and Environmental Conflict: America's Green Crusades," Chap. 6. Sage, Thousand Oaks, California. The importance of dramatic events. *Ibid.*, Chap. 7.
- Rosa, E., and Dunlap, R. (1994). Poll trends: Nuclear power: Three decades of public opinion. *Public Opin. Q.* 58(2), 295–324.
- Shanahan, J., and McComas, K. (1997). Television's portrayal of the environment: 1991–1995. *Journalism Mass Commun. Q.* 74(1), 147–159.
- Shanahan, J., and McComas, K. (1999). "Nature Stories." Hampton Press, Cresskill, New Jersey.
- Shanahan, J., Morgan, M., and Stenbjørre, M. (1997). Green or brown? Television's cultivation of environmental concern. *J. Broadcast. Electron. Media* 41, 250–268.
- Tarbell, I. (1902). The history of the Standard Oil Company. *McClure's Mag.* 20(1), 3–17.

Microtechnology, Energy Applications of

RICHARD B. PETERSON

Oregon State University
Corvallis, Oregon, United States



-
1. Introduction and Overview
 2. Unit Operations
 3. Systems
 4. Materials, Fabrication, and Costs

Glossary

coefficient of performance (COP) A figure of merit for cooling systems; it is defined here as the amount of thermal energy removed from a cooled space divided by the amount of work or heat supplied to the cooler to accomplish the heat removal.

fuel cell An electrochemical device for directly generating electricity by combining a fuel, such as hydrogen, with oxygen to form a reaction product, such as water; because theoretical performance is not tied to the thermodynamic Carnot efficiency, higher chemical-to-electrical energy conversion can occur.

fuel processor A chemical reactor specifically for converting a fuel from a complex mixture of hydrocarbons to a less complex, and often pure, form of usable fuel (e.g., hydrogen).

heat exchanger A device for transferring thermal energy from a hotter fluid to a colder one; heat exchangers come in a variety of configurations, including (but not limited to) parallel flow, counter-flow, cross-flow, compact, shell-and-tube, plate-and-frame, regenerative, recuperative, unmixed, and mixed streams.

logistics fuel Any number of liquid fuels widely used by the military for transportation and power generation; examples include JP-4, JP-6, diesel, gasoline, and kerosene.

microchannel array An array of channels with characteristic dimensions of less than 1 mm designed for conveying a heat and/or mass transfer fluid; when used as the basis for heat exchangers, boilers, and condensers, high rates of heat transfer result.

Micro Electro Mechanical Systems (MEMS) A class of devices typically made from silicon or employing it in the fabrication process; devices integrate electronics and

mechanical devices onto a single substrate with typical feature sizes in the 1- to 50- μm range.

microreactor A miniaturized chemical reaction system characterized by small size (typically $<1\text{ cm}^3$) and fabricated, at least partially, using methods of chemical or electrochemical etching, laser machining, electro discharge machining (EDM), or other microfabrication processes.

Microtechnology-Based Energy and Chemical Systems (MECS) A class of devices integrating heat and mass transfer components along with chemical reactor technology in a single integrated system; feature sizes are typically 50 μm to 1 cm.

nano-, micro-, and mesoscale Dimensional regimes for describing various levels of feature and/or device size; definitions vary, but a typical one would be nanoscale (10 nm to 1 μm), microscale (1–100 μm), and mesoscale (0.1–10.0 mm).

platelet architecture A fabrication scheme using a variety of micromachining techniques to place structural features in thin plates; several plates are then stacked together in a registered manner and bonded, producing a single part having an intricate array of embedded features.

process intensification The enhancement of heat and mass transfer rates as the characteristic length defining the process is decreased.

unit operation A process characterized by a single function; several unit operations can be combined together to produce, or generate, an end result.

The trend toward miniaturization has branched out into many fields of engineering. The area of Micro Electro Mechanical Systems (MEMS) has been established for well over a decade and is focused primarily on sensors and actuators, although other components such as gears, linkages, valves, and fluid mixers are common research and development topics. The energy area, including chemical processing, power generation, refrigeration, and heat

pumping, is in the nascent stages of process and systems miniaturization. The developing area of Microtechnology-Based Energy and Chemical Systems (MECS) holds great potential for spawning new commercial sectors to the economy. The primary aim of the MECS area is to miniaturize and integrate the components necessary for advanced energy and chemical systems. Miniaturization will be especially important in areas where portability, compactness, weight, reliability, and point application are the driving considerations.

1. INTRODUCTION AND OVERVIEW

At the beginning of the 21st century, the trend toward miniaturization has branched out into many fields of engineering. The area of MEMS has been established for well over a decade and is focused primarily on sensors and actuators, although other components, such as gears, linkages, valves, and fluid mixers, are common research and development (R&D) topics. MEMS have feature sizes between 1 and 50 μm , with development in the area being heavily dependent on fabrication processes common in the electronics industry. The energy area, including chemical processing, power generation, refrigeration, and heat pumping, is in the nascent stages of process and systems miniaturization. This developing area is called by many terms—Micro Systems or Micro Systems Technology (MST) in Europe and Micro Chemical and Thermal Systems (MicroCATS) or MECS in the United States—and holds great potential for spawning new commercial sectors to the economy. The primary aim of the MECS area is to miniaturize and integrate the components necessary for advanced energy and chemical systems. Researchers in the field view this idea as a way of improving traditional energy systems while addressing the challenges of new applications. Miniaturization will be especially important in areas where portability, compactness, weight, reliability, and point application are the driving considerations.

Systems based on this technology rely on the extraordinary rates of heat and mass transfer associated with microcomponent architecture. That is, as a system's characteristic heat transfer or mass transfer path is reduced, rates of transfer increase significantly. Depending on the type of system considered and the performance metric

chosen, the increase can scale as $1/L$ or $1/L^2$, where L is the characteristic length defining the transfer path. An example is the fabrication of heat exchangers from microchannel arrays where the thermal diffusion path is very small and the total surface area per unit volume is high. Early work on microchannel heat exchangers demonstrated heat transfer rates of 20 kW/cm^3 , albeit under extreme flow conditions. This is the rate of heating needed by an average home on a cold day in the northern latitudes.

Other important effects are present with a reduction in scale. As with heat transfer, microscale devices have high mass diffusion rates. This leads to very fast and complete mixing in small volumes. Another feature present in microsystems is precise control over chemical reactions and biological processes by rapidly controlling system temperature. With flow-through microscale devices, temperature gradients can be very large (on the order of 100,000 K over a distance of a few microns or a time period of a few microseconds). Finally, by virtue of scale, the stress in miniaturized structural devices is lower for a given operating pressure. Therefore, it is practical to operate at higher pressures than in conventional processing situations. This can increase efficiency substantially or skew a chemical reaction to higher yields of a desirable product.

Enhancement of the primary transport processes by scale reduction is responsible for what researchers in the area call process intensification. This term captures the idea of using microscale structures in mesoscopic devices to enhance performance in small or miniaturized systems. However, size reduction cannot go on indefinitely. This is due to the increasing difficulty of transporting fluids and maintaining temperature differences within a miniaturized system. Excessive pressure drops, parasitic internal heat conduction, and mechanical friction are just a few of the existing problems when miniature energy systems are developed.

Integration is also a central issue in MECS research. To take advantage of mass production techniques leading to economical devices, it is desirable to have most if not all of the critical components of a system integrated together and fabricated simultaneously. Realistic implementation of this concept will probably rely on a number of additional steps, such as “pick and place” (common in the electronics industry), to complete the device. Aside from lowering production costs for single systems, additional capability and functionality can result from processing multiple devices at once. For

example, integrated arrays of components open up new approaches to solving challenging problems in the energy and chemical processing fields. In high-capacity applications such as chemical plants, production schemes based on massively parallel arrays could be used. The principal advantage to this approach is the higher production efficiency associated with microscale processing and the inherent reliability of parallel architecture where the failure of one or even several individual devices would have a small effect on the overall production rate. If a few individual components required replacement, they can be isolated from the array, removed, and then replaced without shutting down the overall system.

A clear definition of an energy system must be given to elucidate the application areas for this new technology. In the context used here, the term "energy systems" generally involves energy generation, use, distribution, or conversion from one form to another. Furthermore, microtechnology-based energy systems involve the concept of microfabricated internal features providing process intensification over standard practices in common use today.

Component and system size spans the mesoscale/microscale regime, as shown in Fig. 1. Nanoscale features may be important to this field during the coming decades but are beyond the scope of this discussion.

Intermediate size applications are the most likely to be commercially attractive in the short term, especially where MECS technology makes possible previously impractical activity. For instance, a MECS reformer to strip hydrogen atoms from hydrocarbon-based fuels will allow use of fuel cells in automobiles (with higher efficiencies and lower pollution). For this idea to work, high-temperature steam reformers and membrane separation are needed in small package configurations. Thermal management will also be critical to the development of practical systems. Heat loss from palm-sized reformers must be minimized through use of high-performance insulation, and heat exchangers are necessary to conserve thermal energy in the various internal gas streams. With the process intensification afforded by MECS technology, palm-sized fuel reformers for fuel cell-powered automobiles become feasible.

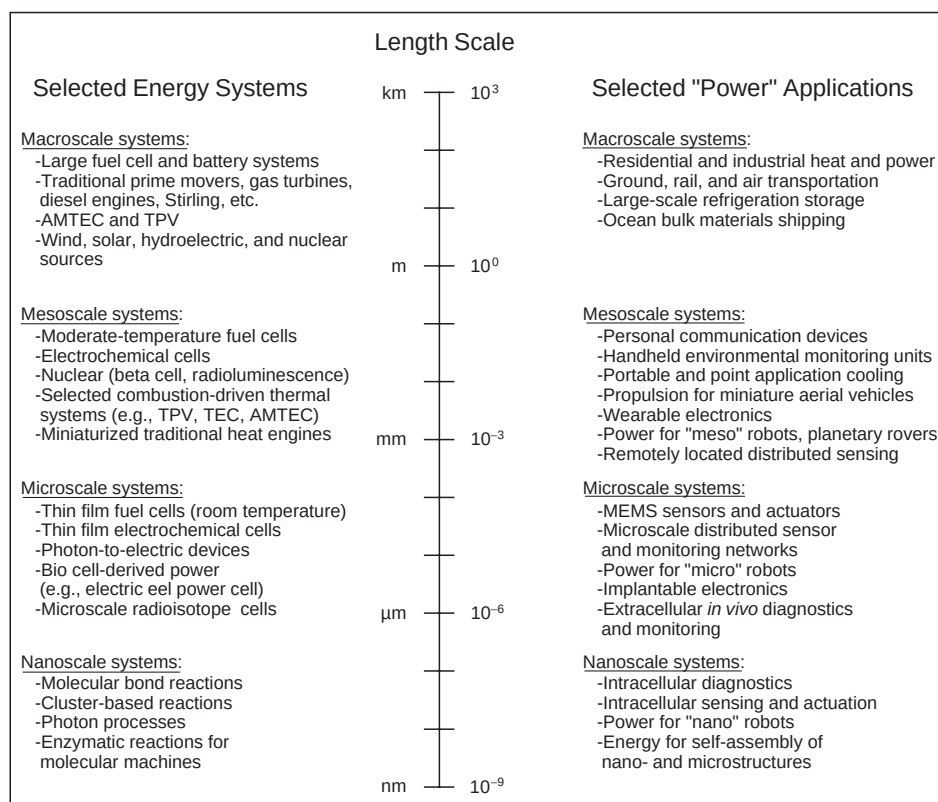


FIGURE 1 Length scales important to the energy area.

Two other areas where this technology can make an impact are small-scale power production and resource processing. As an example of the first application area, consider the chemical energy stored in a liquid hydrocarbon fuel. When burned with oxygen from the surrounding air, the heat generated can be two orders of magnitude higher than the energy contained in an electrochemical battery. If the heat can be converted to electricity with 20% or higher efficiency, stored energy densities for portable power generation can be 5 to 20 times those of current battery technology. This would require miniaturized and integrated components to achieve practical implementation in small-scale packages, but no fundamental limitations exist in developing a battery replacement operating from hydrocarbon fuels. Another application area would involve resource processing at remote locations such as on the surface of Mars. If human exploration of Mars is to take place, it makes sense to process native materials for fuel, oxygen, and water. This type of resource processing would be accomplished most reliably by large numbers of small devices working to generate the necessary power and performing the required chemical processing of the indigenous materials.

Other application areas are possible and include decentralization of heating, ventilation, and air-conditioning (HVAC) systems (super zoning), cooling of protective suits for hazardous environments, and micropropulsion of miniaturized air, land, and sea vehicles. Although the true value of MECS will become known only during the next few decades, the vision of bringing about both evolutionary and revolutionary changes to the energy area is one of the driving forces in this developing field.

2. UNIT OPERATIONS

A unit operation is a single process critical to the functioning of a larger integrated system. Several unit operations are typically joined together, either sequentially or in parallel, to accomplish the overall system function. For MECS, this overall function could include (but is not limited to) heat generation through combustion, heat-activated heat pumping, or power generation. The unit operations terminology derives from chemical engineering where filtration, evaporation, distillation, and batch reaction processing are relevant. In this work, a broadening of the term is implied to include other elementary operations important for the generation and conversion of energy.

2.1 Heat Exchangers, Evaporators, and Condensers

The most fundamental process from an energy systems standpoint is heat transfer to (or from) a working fluid. This unit operation typically occurs in small channels, the characteristic diameter of which is less than 1 mm. Other heat transfer configurations, such as fluid flowing through a network of posts or through a mesh, are also possible. Specific examples of this fundamental unit operation include phase change occurring in boilers, evaporators, and condensers. Single-phase systems where the fluid undergoes a temperature change only also falls into this category.

A slightly more complicated arrangement for transferring thermal energy from one fluid to another occurs in a heat exchanger. Many different configurations exist for this type of device. Classifications include parallel flow, counter-flow, and cross-flow for the primary arrangement of the flow as well as shell-and-tube, plate-and-frame, single pass, multiple pass, and other terminology specifying the physical arrangement. Regardless of the type and configuration, heat exchangers are designed to recover thermal energy from one fluid and deliver it to another to enhance the overall performance of a system (by way of conserving thermal energy).

Use of engineered microstructures in heat and mass transfer has the potential of enhancing transfer rates, thus leading to process intensification in thermal/mass diffusion limited situations. The typical microstructure used is the microchannel array. Characteristic dimensions of these structures are small enough so that laminar conditions exist throughout. Consequently, diffusional processes are responsible for thermal and species mixing and result in a time for process completion proportional to d^2/α , where d is the channel dimension (the diameter if circular) and α is the thermal diffusivity (for thermal mixing) of the fluid. Thus, smaller channels lead to higher rates of thermal and mass transfer, resulting in process intensification. Surface area per unit volume can also be increased in these devices. If special design accommodations are not made in microchannel systems, the penalty for enhanced transfer rates is often higher pressure drops across the device. One of these accommodations would be to array a larger number of parallel microchannels so that a shorter length of fluid channel results in a lower pressure drop, whereas the increased number of parallel paths maintains the required throughput. It is not uncommon for heat fluxes exceeding 100 W/cm^2 to be achieved with water flowing through microchannel arrays.

2.2 Valves

Microvalves have been developed for a variety of applications suitable for MECS. Valve classification can be divided into passive and active categories. The passive type is the most fundamental and is configured as a basic check valve where a thin flapper plate covers an orifice. Microvalves of this type have been fabricated using a variety of techniques, including lithography and etching, laser micromachining, and electro-discharge machining. Materials have included metal on metal, metal on polymers, and silicon with photoresists such as PMMA and polyimides. Soft lithography has also been used to create check valves. Work in this area has resulted in miniature peristaltic pumps driven by compressed air. Sizes of these microvalves vary from the 50- μm range (for orifice diameters) up to 1 mm. Larger sizes are typically classified as conventional macroscale devices.

Active valves have designs that rely on electric actuation in some specific form. Direct actuation by a piezoelectric element has been achieved where the sealing element in the valve is a simple piezoelectric bending element, or the so-called bimorph, in place of a passive flapper in the check valves discussed previously. Electromagnetic solenoid actuation has also been developed, but this activity borders on conventional solenoid valves found commercially. A new development in the electromagnetic actuation area is arrays of active valves with characteristic sizes of a few millimeters and integrated together on substrates containing manifolding and sensor placement. This level of sophistication is only beginning to emerge from laboratory work in the commercial sector.

Secondary electrical effects using heating of bimetallic strips or shape memory alloy elements are being used for actuating valves. Electrical current passing through the active element (a metal strip) itself, or through a nearby resistor for I^2R heating, is the most common way of actuating valves of this type. For the bimetallic valves, heating causes a differential thermal expansion of the two metals, leading to a bending or distorting in a suitably designed element. Shape memory elements can also be designed as a composite actuator to provide an active opening and closing of an orifice on electrical current flow. Thermopneumatic forces have also been harnessed for valve actuation. This method relies on the expansion of a material, often through phase change, on heating. A common problem with all valve designs is leakage in the adverse direction, that is, the direction opposite to the desired flow when closed. Due to scaling of the sealing area, which goes

as the characteristic length of the valve squared, valves become progressively more leaky as the overall scale is reduced. This can be mitigated somewhat by design; however, effective low-leakage valves at the microscale remain problematic.

2.3 Pumps and Expanders

Many small-scale energy systems require effective miniaturized pumps and expanders. These devices are responsible for conveying a working fluid through various components to execute a cycle that may result in power production, heat pumping, or refrigeration. As an example, a pump is required for pressurizing the working fluid in Rankine power cycles. Miniaturized pumps could also cycle a heat transfer fluid through a cooling loop for heat rejection purposes. On the other hand, expanders are necessary for any cycle employing a working fluid to generate power. Specific examples include micro-turbines, miniaturized piston-based machines, and roots-type expanders.

Most candidate techniques for pumping and expanding are ineffective for MECS applications. However, there are a number of ongoing projects concerned with development of pumps, compressors, and expanders. Micro turbo machinery has been the focus of one such project. Its targeted application is to develop a gas turbine engine the size of a shirt button. Many spinoff applications will result if the project is successful, including compressors, pumps, and expanders—all based on high-speed rotating machinery. Note that scaling turbo machinery down to the several millimeter size regime requires careful attention to machining tolerances, heat transfer effects, and operating speed. It is anticipated that rotational rates of up to 2 million revolutions per minute (rpm) will be needed in such devices. Various piston-based expander and compressor configurations can also be miniaturized at least down to the several millimeter range, but sub-millimeter devices will be a challenge. Seals, friction, and heat transfer effects all have adverse scaling characteristics; hence, novel design and fabrication methods are needed.

Pumps in the true microscale regime can be fabricated using photolithographic techniques where flexible diaphragms are used and electrostatic or electrothermal mechanisms are used for driving the diaphragms. Simple check valves of the flapper kind can be used to control the flow. Current state-of-the-art pumps of this type have low pumping speeds and limited pressure increases. Other microscale pumps working with weak electro-fluid interaction, such as

electrohydrodynamic and even magnetohydrodynamic effects, have also been evaluated. Low pumping speeds and minimal pressure rises have been observed. Until effective designs are developed, mesoscale pumps, compressors, and expanders may have to be coupled with many parallel flow microscale components to be effective. Hybrid mesoscale/microscale devices would result from this. Recent developments in the actuator area could help to bring about practical mesoscale/microscale pumps and compressors. New polymer-based electro-active materials and magnetic shape memory actuators are now becoming available to MECS researchers.

2.4 Mixers

Mixing is a process where two or more liquids or gas phase components are brought together and combined to form a uniform composition. Both turbulent and diffusive processes are typically employed to achieve the mixing action. As the scale of the process is reduced, diffusion becomes dominant. Microtechnology mixers are designed to bring together individual components for mixing with the smallest diffusional scales possible. Mixing is an important unit operation in the chemical and energy industry and can be found in combustors, microreactors, and adsorbing processes.

Micromixing technology has been developed in several different directions, but the common theme with most concepts is to rapidly bring together the constituents where a very small diffusive length scale exists for the final mixing step. Mixing in microchannel geometries has been studied in “Tee” sections, opposed microjets, channel flow through arrays of posts, and various interdigitated designs. Mixing in converging channels and jet-in-cross-flow geometries has also been studied. Practical micromixers have been developed and proven out using deep-etch photolithography (in silicon) where an interdigitated arrangement yielded a length scale of $100\text{ }\mu\text{m}$ (distance between two streams). Note that the time for diffusional mixing is proportional to L^2/D , where L is the distance between two mixing streams and D is the diffusion coefficient. Hence, it is important to reduce the device size to achieve rapid and complete mixing.

2.5 Separation Units

Separation is a unit operation used in absorption heat pumps and chemical processing applications such as solvent extraction and product separations.

A typical separation process in a heat pump application involves the desorption of ammonia from a water–ammonia solution. Although a number of configurations have been studied for this process at the macroscale, most are based on gravity and have relatively low rates of desorption. For developing small heat-activated absorption heat pumps, microtechnology can be applied through the use of thin “sheet” channels where desorption of ammonia is accompanied by heat addition. As the sheet film in the channel is made thinner ($<100\text{ }\mu\text{m}$), diffusional processes become exceedingly fast and high rates of desorption can be achieved.

Practical implementation of the thin sheet geometry for desorption has taken the form of mechanical membrane gas–liquid contactors. This approach has been dictated by the consequences of capillary forces present at small-length scales. Surface tension and flow in the “lubrication” regime of fluid dynamics conspire to prevent unconstrained films from flattening out to the dimensions necessary for rapid desorption. With the use of a mechanically constrained liquid film, progress on desorber units has progressed to the point of making miniature heat-activated heat pumps feasible. To realize the potential of the concept, an integrated approach to the design of a thin film desorber must be used. Thus, combining a microchannel heat exchanger with a mechanically constrained thin film desorber has resulted in high-performance units ideal for small-scale systems.

2.6 Microreactors and Combustors

MECS will most likely employ combustion for driving processes such as vapor generation, endothermic chemical reactions, and (most notably) fuel reforming. Both fuel reformers and combustors will be of a miniature design relying on embedded catalysts for promoting chemical reactions at moderate temperatures ($350\text{--}750^\circ\text{C}$). Many potential configurations exist depending on the application and constraints on the design. Microchannel arrays are a potential configuration; mesh and post architecture is another to achieve the desired surface area and small diffusional lengths necessary.

Small-scale fuel reforming is an important area of research in the microreactor area. Hydrogen production for fuel cells is the main driver for this activity. Fuels such as methanol, ammonia, and gaseous hydrocarbons have been tested in laboratory settings with reaction volumes on the order of 1 mm^3 . These devices are constructed using several different

techniques, but etched silicon and platelet technology (described later) are two methods being explored. Note that thermal management of microreformers is exceedingly important given that heat loss can represent a significant fraction of the energy transfer rates present in the device. Consequently, small reformers will require both high-performance insulation and gas stream configurations that recover thermal energy in the exiting flows. A systems approach to design optimization incorporating thermal management, flow pressure drops, and component interfacing will be necessary for practical systems to emerge from the laboratory.

A heat source is required to drive endothermic reactions of fuel reforming and shift reactions. Evaporators and boilers also require a heat source. Of all the conventional sources of heat, combustion of liquid hydrocarbons has the highest stored energy densities. Compared with electrochemical cells, liquid fuels have energy densities 35 to 300 times greater than current battery technology. This assumes that the fuel has an energy density of 42 kJ/g (with air coming from the surroundings) compared with a zinc–air battery at 1.2 kJ/g or a lead–acid battery at 0.125 kJ/g. The development of moderate-temperature (450–1000°C depending on the type of fuel) combustors in a miniaturized form has led to a practical method of releasing this energy in a small overall component size. Microcombustors have been developed in a size range of a few cubic millimeters where a platinum catalyst has been used to promote reactions. Catalytic operation is a necessity because miniature and microscale devices have much larger surface/volume ratios, and hence a high degree of heterogeneous reactivity, compared with conventional macroscopic combustors. Also, true microscale operation typically takes place at length scales much smaller than the quench distance associated with the fuel being burned. However, high-temperature operation can mitigate the need for catalytic surfaces in some cases.

As the characteristic size of combustors is reduced from mesoscale to microscale, thermal management plays an increasingly important role in efficient combustor operation. Work currently taking place in miniaturized combustors includes development of small excess enthalpy burners fabricated using a three-dimensional printing process by stacking hundreds of individually patterned layers. A toroidal combustor has been developed in the form of a “Swiss roll” where the hot region of the burner is isolated from the surroundings by the inward spiral of reactants and the outward flow of products.

Excellent heat recovery can be achieved with this design. There has also been work accomplished on the development of microscale combustors based on a bayonet-style counterflow heat exchanger where the outside surface of the combustor is insulated by either vacuum conditions or high-performance insulation. Hydrogen and propane have been catalytically burned in volumes of less than 0.5 mm³ (Fig. 2) at temperatures in excess of 650°C. Other work has also taken place on small-scale diffusion flames and insulated burners. The current technology offers validation of the ability to construct small-scale heat sources for integration with other components to build miniaturized energy systems.

2.7 Balance of Plant

A critical issue with MECS development is the size and complexity of the “balance of plant.” This includes subcomponents such as air movers, fuel delivery components, sensors, valves, and other associative concerns (e.g., power for start-up). These issues are being explored by a few researchers concerned with the development of total systems, but all too often this aspect of energy systems developed is relegated to a minor consideration when in fact it can be the deciding factor for practical systems. The balance of plant problem must be explored in detail and resolved for each specific case being examined. Over the coming years, solutions may begin to emerge through ongoing efforts in related microtechnology development.

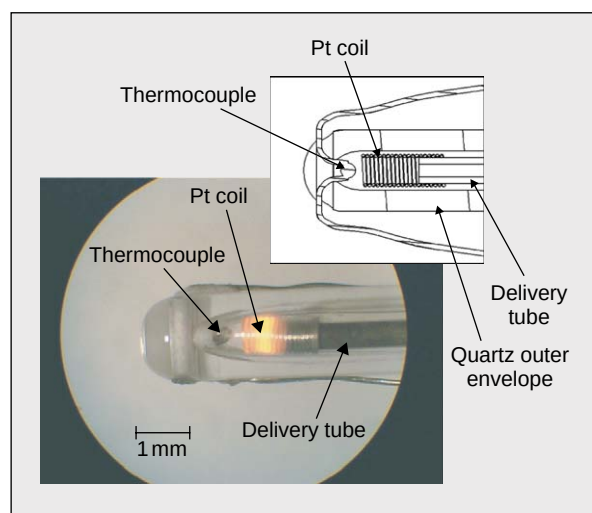


FIGURE 2 Example of catalytically promoted microcombustor. (Courtesy of Oregon State University.)

Specifically, microscale engineering and MEMS-type fabrication techniques may offer effective balance-of-plant solutions in an integrated packaging scheme.

3. SYSTEMS

Applications abound for microtechnology-based energy systems. Although not exhaustive, Fig. 1 conveys important developmental areas and provides a listing of technologies important in each size class. The current topic is concerned with the mesoscale/microscale size regime; overall systems of palm size or smaller are the main focus here. These systems are based on subcomponents that have microscale elements for process intensification but can fall into either the mesoscale or microscale regime. The focus here is concerned with power generation and cooling. These two applications will be the main areas for MECS during the coming years and could represent multi-billion-dollar industries by the end of the first decade of the new century.

3.1 Power Generation

Power generation can take the form of shaft work, electricity, or a propulsive effect (for applications involving flight or submersibles). Electrical power generation using fuel cells or microengines will become a direct competitor to batteries as the technology in MECS is developed into reliable systems. As mentioned earlier, stored liquid hydrocarbons have a large advantage over electrochemical storage in terms of energy density. However, batteries enjoy significant advantages in ease of use and reliability. Once MECS are developed to their full potential, they should provide portable applications capability with substantial benefits for both military and commercial uses.

3.1.1 Fuel Cells and Fuel Processing

Fuel cells are direct energy conversion devices that combine two reactants to produce electrical power. The reactants are typically a fuel such as hydrogen, or methanol, and oxygen from the air. Fuel cells require an electrolyte capable of passing an ionic charge carrier across an electronic conduction barrier where the ions are driven by a concentration gradient. Fuel cells also need a catalytic-based anode and cathode for reactant preparation. For mesoscale/microscale systems, fuel cells are best fabricated in thin film form. Depending on the desired power output of the system, the "footprint" may well be

relatively large to supply the required power. Systems arranged in this manner are referred to as mixed scale systems on the basis that one critical dimension is small (the thickness), whereas the extent of the device (its footprint) can be the requisite size to satisfy a particular power application.

Thin film fuel cells operate across a broad range of temperatures. Proton exchange membrane (PEM) cells based on Nafion or similar material can operate at room conditions but provide better performance at elevated temperatures. The upper practical temperature limit for Nafion is approximately 100°C, although pressurized systems can go higher. The reason for this limitation is the requirement for keeping the membrane saturated with water to promote ion passage. Fuels for PEM cells include hydrogen and methanol as well as other fuels if reforming takes place. Direct methanol fuel cells have received much attention recently as a possible power source for portable electronics. Developments in PEM cells have resulted in new membrane materials operating near 200°C, where power densities can be higher and where catalysts on the fuel side (the anode) have less susceptibility to carbon monoxide poisoning. This is critical for cells consuming a reformer gas because carbon monoxide in low concentrations is usually present even after filtering.

Higher temperature systems are also a possible choice for mesoscale power systems. Solid oxide fuel cells (SOFCs) have traditionally been made with electrolytes of yttria-stabilized zirconia having a thickness greater than 100 μm . This has dictated operating temperatures approaching 1000°C due to low ion mobility through the electrolytes. Research on thin film SOFCs over the past decade or so has shown the possibility of operating at temperatures as low as 500°C. This makes them attractive for small-scale systems. Attributes of the thin film SOFCs at these lower temperatures include tolerance to many types of fuels (including carbon monoxide), no water management issues, and the possibility of operating with either internal reforming or direct fuel oxidation. If the power density of the thin film devices can be maintained at the lower temperatures, practical small-scale systems may result.

As mentioned earlier, practical PEM fuel cells require hydrogen to operate (with the exception being the work on direct methanol fuel cells). However, to carry sufficient quantities of hydrogen gas for cell operation, high-pressure canisters are required. This can be a safety hazard, especially if pressures greater than 100 atm are needed. An alternative is to extract hydrogen from a fuel rich

in this constituent (e.g., methanol, ammonia, hydrocarbons). Fuel reforming is the process by which this extraction occurs and has recently been the focus of much research. Note that other hydrogen sources have been considered, such as the so-called hydrogen-on-demand systems using borohydrides, but are not examined here.

Reforming takes place at temperatures from approximately 350°C for methanol to over 750°C for hydrocarbons and ammonia. The reactions are endothermic and require a source of heat to drive them to near completion. Small-scale reformers are configured so that fuel with water vapor passes over a catalyst within a microchannel array. The fuel molecules are “reformed” so that hydrogen, carbon dioxide, carbon monoxide, and excess water emerge from the device. Trace amounts of unreacted hydrocarbons can also contaminate the exit stream. The consequence of not having a pure hydrogen stream from the reformer is severe; if small amounts of carbon monoxide are present in the hydrogen supplied to a PEM fuel cell, poisoning of the anode catalyst results. Hence, reformers require a rigorous cleanup and filtering of the fuel stream. Nevertheless, mesoscale/microscale reformers with metallic membrane filters having internal volumes of less than a few cubic millimeters are being developed. As with all mesoscale/microscale devices operating at elevated temperatures, thermal management is important for conserving the energy of the process.

3.1.2 *Miniature and Microscale Heat Engines*

Many characteristics of traditional engines make them attractive for use in power generation and propulsion. They tend to be self-aspirating and rely on combustion, which at the macroscale is a very robust form of heat generation. Fuel is plentiful and inexpensive, with storage easily realized. The energy density of the fuel (or fuel plus container), when compared with electrochemical sources, is high. Along with these advantages come a number of drawbacks, especially where miniaturization is concerned. For example, an engine is thermodynamically restricted (by the Carnot efficiency) in its conversion of chemical energy to work due to the intermediate heat-generating step. This is in contrast to the direct energy conversion of fuel cells. However, conversion efficiency is respectable in macroscopic engines and can approach the 30 to 40% range at design speed and power output. Note that the overall conversion efficiency of fuel cells rarely exceeds 50% because of the cell's internal electrochemical irreversibilities and losses due to power conversion electronics, fuel use,

and reforming. Thus, thermal engines provide a practical and reliable way of extracting chemical energy bound up in hydrocarbon fuels.

Most common thermal engines rely on compressing a cooler, condensed working fluid and expanding a hot, volume-expanded fluid. This is often accomplished through mechanisms that have rubbing or clearance seals, hot regions separated from cold structures, and bearings to allow differential motion between parts. All three of these defining characteristics suffer from adverse scaling effects. Although scaling laws are not discussed here, consider the case of maintaining the required temperature difference for the Brayton cycle (gas turbine). By its very nature, this engine exploits a difference in temperature to generate work from heat. This temperature difference is the driving force for engine operation, and as the temperature difference increases, better thermal efficiency results. However, as the size of the engine is reduced, its internal structure acts as a thermal pathway shunting a portion of the usable heat through two unproductive pathways. First, some of the heat is transferred to the surroundings without producing work. Second, heat is conducted through the connection between the turbine and the compressor, heating the inlet air stream. At smaller sizes, it becomes progressively harder to insulate the hot section of the engine to prevent heat flow through the two leakage paths. Simple calculations for palm-sized gas turbines having a 300 K inlet temperature and a turbine operating at 1800 K show the parasitic heat loss to be comparable to the overall power output. The parasitic effects become more pronounced as the engine is further miniaturized until no practical conventional design is possible.

Microthermal engines, as characterized by miniature Wankel engines, Stirling engines, and micro gas turbines, all are practical macroscopic engines but suffer from significant sealing problems, reduced subcomponent efficiency, friction, and thermal management issues when scaled to the mesoscopic size regime. To date, significant work has been invested in micro gas turbines and Wankel engines. The former is contemplated to have a rotor diameter of approximately 5 to 10 mm, a combustor temperature in excess of 1200°C, and rotation rates of up to 2 million rpm. Wankel engines and compressors have received considerable interest recently. The particular design features making this approach attractive are a simple overall design and no valves. Furthermore, a near two-dimensional layout of the engine (Fig. 3) would permit MEMS-type fabrication (in silicon or other suitable material) if workable designs emerge



FIGURE 3 Lithographically deined 1-mm Wankel rotor in its housing. (Courtesy of University of California, Berkeley.)

from the laboratory. As with most miniature internal combustion engines, the Wankel presents serious questions regarding seals, internal heat transfer, and combustion in small spaces.

3.2 Cooling

An important technological application for MECS is small-scale systems cooling. As mentioned previously, an advantage will be realized in areas where portability, compactness, weight, and point application are the driving considerations. Many specific cases that meet these criteria exist. For example, cooling protective suits used in chemical and biological “hot spots” represent one important application. Super zoning in residential buildings for heating and cooling is another. At the smallest length scales, electronic chips would benefit from onboard and integrated cooling mechanisms. The following is a brief overview of some of the work being done in this area.

3.2.1 Refrigeration and Cryocooling

Microtechnology can be employed to produce miniaturized refrigeration and cryocooling systems. Although process intensification is typically the route used for miniaturization, microtechnology can also be used for reducing the size of mechanical components that are necessary for operation. For example, to create a small-scale vapor compression refrigerator, the heat transfer components comprising the condenser and evaporator can be made in a micro-channel configuration for enhanced heat transfer rates. However, the mechanical compressor will

often be the determining factor for overall size and weight. Advanced compressor designs can be developed by replacing existing components with micro-machined ones, or entirely new types of compressors can be developed. In the former area, valve heads and drive mechanisms can be redesigned to take advantage of layered manufacturing techniques so that integrated components result. In the latter area, work has been under way focusing on electrostatically operated compressors micromachined out of silicon. Other approaches using thermopneumatic operation, electroactive polymers, and magnetic shape memory alloys have been pursued.

Cryocoolers can benefit from miniature components as well, but fundamental heat transfer issues must be considered. For example, counterflow heat exchangers or regenerators are necessary for cryocooler operation. This is due to thermal isolation requirements of the cold space from ambient temperature. The working fluid must pass relatively unimpeded (i.e., low pressure drop) through the heat exchanger while thermal energy is exchanged between the incoming and outgoing flows. To reduce the size of the overall cooler, each component (including the heat exchanger or regenerator) must be reduced in scale. However, this presents a heat transfer problem in the form of heat leakage into the cold space. As the size, and hence the length, is reduced, the temperature gradient along the heat exchanger increases, leading to enhanced heat transfer rates to the cold section. On further size reduction, a point is reached where the cooler load is entirely from this leakage and no useful heat lift takes place. Thus, there exists a limit, based on fundamental heat transfer principles, to which coolers can be reduced in size.

With these considerations, several of the techniques used at the macroscale for cryocooling have been investigated for miniaturization. For instance, both pulse tube and Stirling cycle devices have been studied for miniaturization. Conduction through the regenerators and heat exchangers limits the ultimate size of the system, but cryocoolers that are approximately a centimeter in length appear to be practical. Miniaturization of the cyclic compressors needed for cryocoolers is also an important need in this area. Typically, piston-based compressors are used; reducing their size involves all of the challenges as does the microengine area discussed earlier. Another possible approach to cryocooling uses a reverse Brayton cycle. Miniaturized turbo machinery could be effectively employed for this application. However, the performance of any Brayton cycle, whether

power producing or in cooling applications, is strongly influenced by the efficiency of the individual components. Thus, compressor and turbine elements in a reverse Brayton cycle machine must be relatively efficient for the cycle to be practical. Although this section has not covered solid-state coolers or refrigerators, much work has been under way for producing efficient small-scale devices (e.g., thermoelectric coolers). If high “figure-of-merit” materials become available for thermoelectric cooling, they will play an important role in miniaturized cooling applications.

3.2.2 Heat-Actuated Heat Pumps

Heat-actuated heat pumps are cooling systems requiring no electricity for operation. Rather, a heat source is used, avoiding the need for batteries. This consideration is critical for portable applications because on a weight basis, a stored liquid fuel has anywhere from 35 to 300 times the energy content of batteries, depending on the battery technology considered. Thus, although the simple solution of combining an electric motor, a battery package, and a vapor compression refrigerator may sound appealing, it actually suffers a severe volume and weight disadvantage when compared with a heat-actuated system. An efficient heat-actuated cooler, perhaps based on combustion of butane or a logistics fuel, can be especially applicable to cooling protective suits and vehicle interiors. In general, the coefficient of performance (COP) for a heat-actuated cooling system needs to be in the range of unity for effective systems to be fielded.

Two types of heat-activated cooling systems are considered viable for miniaturized systems. Both would rely on microtechnology for operation of critical components, and both would use combustion of a storable fuel to drive the process. The first is an absorption cycle heat pump with a working fluid of lithium bromide and water or ammonia and water. The key components of the cycle are shown schematically in Fig. 4. The feature of this device making it a possible choice for palm-sized cooling is the use of microtechnology in the absorber, desorber, and evaporator. As shown in the diagram, a small amount of electrical power is needed for pumping the two-component liquid to a higher pressure. However, more than 90% of the energy used in the device would come from combustion of the stored fuel.

Significant developmental challenges exist for miniature absorption coolers intended for protection suits. The absorber and desorber rely on thin,

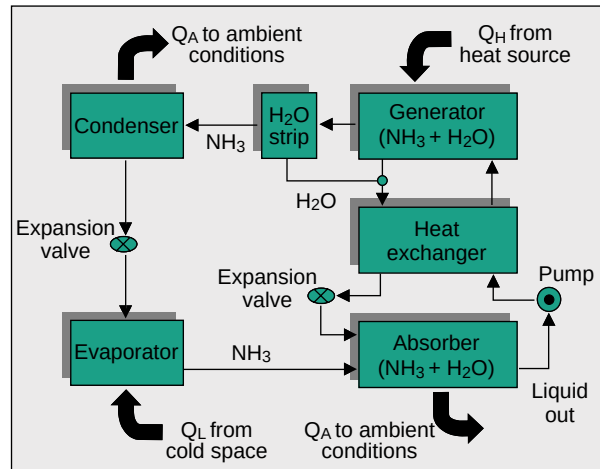


FIGURE 4 Schematic diagram of an ammonia/water absorption heat pump.

mechanically constrained membranes that mediate the separation of the two-component system. For practical use, orientation independence is required so that fieldable units can operate whether the wearer is standing upright or lying down. Finally, an integrated approach for the main components of the cooler, as well as for the balance of plant, must be developed so that economical fabrication is possible.

The second possible system for miniaturized heat pumps takes the form of an engine-driven vapor compression refrigerator. Microtechnology would be used throughout, but especially in the refrigerator components such as the compressor, evaporator, and condenser. The miniature engine for this type of heat pump has yet to be developed. Candidates are the aforementioned small-scale turbines and Wankel engines as well as other engine concepts using external combustion. High efficiency in a small package will be a key aspect to this heat pump concept. Furthermore, thermal management, including insulation around hot engine parts and energy recovery in the exhaust stream, will be critical if the engine-driven vapor compression cycle is to be realized.

4. MATERIALS, FABRICATION, AND COSTS

4.1 Criteria for Materials in the Energy Area

In fabricating microscale energy systems, a number of operating conditions must first be taken into account. Foremost are the operating temperatures throughout

the device. Often with mesoscale/microscale energy systems, two closely spaced parts of the same device have a large temperature difference between them; hence, a large gradient can be present. A primary constraint is to have the material withstand the maximum temperatures existing in the system, but the gradients and thermal expansion must also be tolerated. This is especially true if more than one material is used for construction where transitions must be made between material sets. Corrosion and oxidation are also important considerations when extended operating lifetime is required. Materials must be chosen for compatibility with the working fluid being used, especially at the temperature extremes of operation. The various materials comprising all wetted parts must not catalyze decomposition reactions at the fluid–surface interface—not even slightly—or else extended operation cannot be ensured. Furthermore, the materials must retain sufficient strength at the operating temperature and should have a low thermal conductivity. This latter criterion stems from the requirement of minimizing conductive heat loss throughout the device. Finally, cost, ease of machining, and material joining factors (e.g., brazing, diffusion bonding, soldering compatibility) must also be considered when selecting the materials of construction. With all of these criteria and constraints, the engineer must have a versatile material set to work with when designing and constructing MECS.

Other considerations come into play when selecting thermal insulation for isolating the elevated temperature (or cold) sections of the device from ambient conditions. Several potential material sets are available for this task, but operating temperature is one of the deciding factors in using a particular material or technique. For example, silica-based aerogel material has a very low effective thermal conductivity and can be used throughout the temperature range from cryogenic to elevated temperatures of up to 800°C. The aerogel can be either infiltrated with a gas or evacuated for very low thermal conductivity. However, radiation heat transfer in the infrared (IR) region must be effectively blocked at high and low temperatures. To accomplish this, carbon black is typically used as a dispersant. But because of oxidation of the carbon, this formulation is incompatible with air at temperatures above approximately 350°C. Other insulating systems, such as multifoil vacuum insulation and simple vacuum gaps, can be used. For long-term service, vacuum conditions must be maintained and might require gettering.

4.2 Types of Engineering Materials Available

Silicon is the material of choice for most MEMS due to (1) the feature size often being in the range of less than 20 μm requiring single crystalline material and (2) the dependence on fabrication techniques developed by the electronics industry for processing this material. However, it is a particularly poor choice for the larger thermally based devices that characterize MECS. The room temperature thermal conductivity of silicon is approximately 150 (W/mK). Use of this material leads to high rates of heat transfer across the short distances where temperature gradients occur. The thermal conductivity of silicon above 1000°C falls to below 25 W/mK, but the average value can still be high. Various metallic alloys and intermetallics, including common stainless steels, are potential fabrication materials for small-scale heat engines and coolers. Most stainless steels have a room temperature thermal conductivity between 11 and 15 W/mK. In contrast to silicon, the thermal conductivity of metals increases with increasing temperature. It is interesting to note that between approximately 1000 and 1200°C, the thermal conductivities of silicon and stainless steel are similar. However, heat loss through each material will depend on average thermal conductivities, and silicon will still have the higher average value. From a heat loss point of view, the best materials will probably turn out to be ceramics and glasses with amorphous structures. Fused silica is at the low end of the thermal conductivity range, with a room temperature value of 1.4 W/mK. This climbs to approximately 4 W/mK at 1200°C. It is also highly resistant to thermal shock due to its low thermal expansion coefficient. A class of material not known for elevated temperature use is polymers. However, polyimide (and other similar polymers) can be used up to approximately 400°C, and this could make these polymers attractive in certain applications.

In most MECS concepts, internal structures will be engineered to reduce internal thermal conduction, which implies thin sections. However, these same sections often contain a working fluid at elevated pressures and must be hermetic. Furthermore, they must support the stress due to the operation of the device and must not fail catastrophically. Single crystal silicon and other pure semiconductors (most notably silicon carbide) can be excellent structural material for microtechnology components. But they are brittle and unforgiving in a yield mode. The so-called engineering materials, including steels and

other metals, ceramics, and plastics, will probably be applied to the construction of larger microscale and mesoscale devices due to a wide selection of both properties and bonding techniques available. Material costs are reasonable for most of the engineering materials, and various forms such as thin sheets are commercially available. Another deciding factor, as discussed in the next subsection, is the availability of lower cost manufacturing techniques. In this respect, silicon requires a wide suite of expensive equipment for bulk and/or surface machining the material.

4.3 Microfabrication Techniques for Engineering Materials

Fabrication techniques developed for integrated circuit (IC) production have been refined to the extent of supporting a multi-billion-dollar industry. Chip manufacturing relies on silicon-based processing where micron-sized features are routinely used in production. MECS do not require the extremely small “line widths” needed for IC fabrication. Furthermore, in many energy applications, silicon is not the favored material, as discussed previously. Other fabrication techniques, such as LIGA, have been specifically developed for MEMS. Although many rely heavily on silicon processing, others can produce very small structures in metals electrodeposited on a surface or within a micromold. Again, for MECS applications, the feature size of these MEMS fabrication techniques is usually much smaller than what is needed.

Because MECS are fundamentally different from traditional ICs and MEMS, they require different materials and fabrication processes. One important fabrication method for microenergy applications is called microlamination or platelet technology (Fig. 5). Although its history dates back to the 1970s, when it was developed for liquid rocket engine injectors, it is now being pursued by a number of groups for fabricating MECS. The method is based on microlamination of metals, ceramics, and polymers. The process begins by surface machining, or cutting, a single lamina with a pattern containing the desired structure. The lamina is often a shim of a material having desirable mechanical and thermal properties important to the functioning of the final device. Once the pattern is cut, the shims are surface treated and precisely stacked in a prearranged order. The stack is then bonded together, forming a single block of material. For the platelet architecture to have utility, a machining method capable of fabricating structures

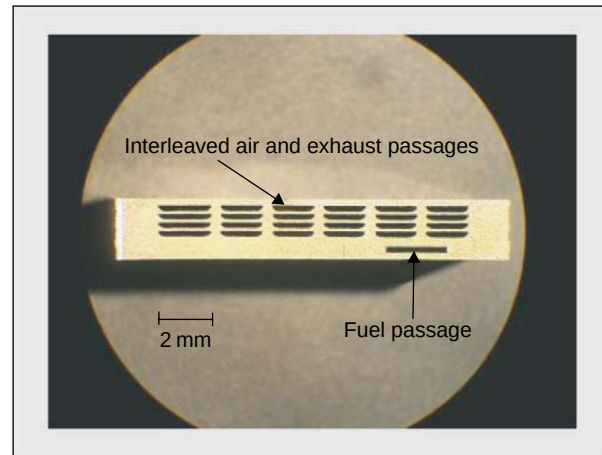


FIGURE 5 A three-stream counterflow heat exchanger fabricated using platelet technology. (Courtesy of Oregon State University.)

in the laminating material is needed. The method must be versatile, easy to use, and capable of rapid machining (with through-cuts and surface texturing) in a wide variety of materials. One of the most general techniques is laser numerically controlled micromachining. It is most useful for prototype runs and can be used on metals and polymers. Other useful techniques specific to only one class of material can also be used. For example, chemical etching through a photographically defined mask (photolithography) can be used on most metals. The process is commercially available and can be employed for high-volume production runs. Another machining technique applicable to most metals is wire-based electrodischarge machining (wire EDM). Current machines on the market are numerically controlled and have wire diameters as small as 50 μm . The cutting of metal platelets with this technique is primarily a two-dimensional operation and can be as precise as 5 μm . Note that conventional high-speed milling, with end mills as small as 100 μm , can be achieved with most modern equipment and can be applied to both metals and plastics. Although wire EDM and conventional milling with small tools have been described in the context of platelet fabrication, their precision and speed can also be applied to a wide range of small mechanical components needed for MECS.

The properties of ceramics (and of materials such as quartz) are desirable from an energy systems point of view, but this class of material is difficult to machine and form. Relatively few techniques are available for cutting the requisite structures in ceramics. However, work is progressing on various

additive techniques for making small components out of both plastics and ceramics. Some of these techniques go by the name stereolithography and often employ a laser to selectively fuse a particle bed of the starting material into a desired shape.

SEE ALSO THE FOLLOWING ARTICLES

Fuel Cells • Heat Transfer • Technology Innovation and Energy

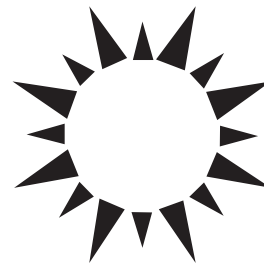
Further Reading

- Ameel, T. A., Warrington, R. O., Wegeng, R. S., and Drost, M. K. (1997). Miniaturization technologies applied to energy systems. *Energy Conversion Mgmt.* 38, 969–982.
- Ehrfeld, W., Hessel, V., and Lowe, H. (2000). “Microreactors: New Technology for Modern Chemistry.” Wiley–VCH, New York.
- Epstein, A. H., and Senturia, S. D. (1997). Macro power from micro machinery. *Science* 276, 1211.
- Fernandez-Pello, A. C. (2002). Micro-power generation using combustion: Issues and approaches. “The 29th International Symposium on Combustion.” The Combustion Institute, Supporo, Japan.
- Goemans, P. A. F. M. (1994). Microsystems and energy: The role of energy. In “Microsystem Technology: Exploring Opportunities” (G. K. Lebbink, Ed.), pp. 50–64. Samsom BedrijfsInformatie, Aphen aan den Rijn, Netherlands.
- Larminie, J., and Dicks, A. (2000). “Fuel Cell Systems Explained.” John Wiley, New York.
- Madou, M. (1997). “Fundamentals of Microfabrication.” CRC Press, Boca Raton, FL.
- Peterson, R. B. (2003). Miniature and microscale energy systems. In “Heat and Fluid Flow in Microscale and Nanoscale Structures” (M. Faghri and B. Sunden, Eds.). WIT Press, Southampton, UK.

Migration, Energy Costs of

CHARLES R. BLEM

Virginia Commonwealth University
Richmond, Virginia, United States



-
1. Importance of Migration
 2. Migration Patterns and Ranges
 3. Molecular Nature of Energy Stores
 4. Magnitude of Energy Stores
 5. Anatomical Storage Sites for Energy Reserves
 6. Patterns of Energy Storage
 7. Energetic Costs of Migration and Flight Ranges
 8. Energetic Benefits of Migration

Glossary

β -oxidation Metabolic process whereby fatty acids are oxidized to produce energy for locomotor activity.

fatty acid Major energy source for many migrants; part of triglycerides.

glycogen Carbohydrate formed by the union of two glucose molecules.

homeotherm Animals that are insulated and regulate their body temperature, typically with the assistance of physiological heat production by shivering.

hyperlipogenesis Excessive production of lipid stores brought about by hyperphagia and increased enzyme activity.

Krebs cycle The tricarboxylic acid cycle; chemical process whereby metabolic substrate provides energy for utilization and storage by the organism.

metabolizable energy Energy obtained from food; the basic amount of energy available for all activities.

passerine Birds belonging to the order Passeriformes; relatively small perching birds, including sparrows, thrushes, and swallows.

poikilotherm Animals that do not regulate their body temperature with precision; typically, they get heat from external sources.

triglyceride Molecular form of storage lipid; a glycerine backbone to which three fatty acid molecules are attached.

Migration is defined as the periodic, regular (usually annual) movement of animals between breeding sites

and wintering or feeding–maturation areas. Numerous species of animals migrate, but the majority making trips of significant distance include swimming (fish, sea turtles, sea snakes, cetaceans, and pinnipeds) or flying (insects, birds, and bats) species. The energetic cost of locomotion has been measured in numerous animals, large and small, as well as the energetic costs of locomotion (the latter in kJ/kg km).

1. IMPORTANCE OF MIGRATION

Because costs of transport of terrestrial migrants are high, and risks of predation severe, there are few species of small animals that migrate by walking or running, and all animals making terrestrial migrations of more than a few kilometers are relatively large mammals, such as wildebeest (*Connochaetes taurinus*), caribou (*Rangifer tarandus*), and bison (*Bos bison*).

In insects, migration occurs fairly frequently among species in the orders Lepidoptera (butterflies and moths) and Diptera (flies) and occasionally in Orthoptera (grasshoppers and locusts), Coleoptera (beetles), and Odonata (dragonflies). Many fish are migratory—so many so that migrants have been classified into three groups; anadromous, catadromous, and oceanic species. Anadromous fishes are those that are born in freshwater, mature in the ocean, and make spawning runs back in freshwater (salmon, herring, and lampreys). Catadromous fishes are those that are born in the ocean, then move up into freshwater habitats to mature, and finally make trips to spawning grounds back in the ocean. Oceanic species (sharks, tuna, and swordfish) migrate much like whales, following food, often with the assistance of currents such as the Gulf Stream. Like terrestrial mammals, migration is limited to larger fishes. Most small fish cannot swim rapidly and cannot store sufficient energy to travel long distances. Young eels are an exception. They are able to make much of their

migration by drifting with the current. Among birds, many migrants are members of the order Passeriformes (perching birds) or belong to the various groups of waterfowl (ducks, swans, cranes, storks, grebes, and loons). However, just about every order within the class Aves includes at least a few migratory species. For example, nearly all temperate zone swifts, hummingbirds, and goatsuckers are migratory and some of the pelagic seabirds (shearwaters, albatrosses, and jaegers) are remarkable travelers that wander over huge expanses of ocean. Among mammals, cetaceans (whales), pinnipeds (seals), and bats include many migratory species.

Because the risks of migration are great, it is axiomatic that migrants must derive significant selective advantages from changing location. The focus of this article is on the balance of costs and benefits that affect the ultimate reproductive success of migrants. Although many of the advantages are potentially related to reduced competition and lower predation rates, some of the justification discovered by investigators of the phenomenon has involved energy acquisition and use. Without doubt, natural selection favors animals that minimize energy expenditure during migration, either to increase their potential range over barren habitat or to reduce the amount of foraging necessary after shorter trips.

Although we often think of poikilotherms (cold-blooded animals) as lethargic organisms less active than homeotherms (warm-blooded), it is significant that migration is fairly common in both groups. There is a relatively large amount of information regarding the physiology and ecology of migration of insects and fishes, but little attention has been given to similar aspects of terrestrial migrants. Furthermore, the information regarding the ecophysiology of flying migrants, primarily birds, far exceeds all other sources and this article reflects that bias.

2. MIGRATION PATTERNS AND RANGES

Migration, strictly defined, can be as restricted as a trip of a few hundred meters or as much as thousands of kilometers. Some salamanders (e.g., mole salamanders and newts) and many species of frogs (particularly tree frogs) emerge from the soil and travel to temporary ponds to breed. They then return to their subterranean hiding places, where they may remain for the next 6–11 months. The distance involved may be only a few meters.

The most extensive migration may be that of the Arctic tern (*Sterna paradisaea*), which literally spends its lifetime in migration, traveling 40,000 km or more, its trip interrupted only by an annual period of reproduction. Note, however, that this trip is not nonstop and the tern feeds along the way, can drink seawater, and rests on the surface of the ocean. The migration of the blackpoll warbler (*Dendroica striata*) may be viewed as even more remarkable because it may fly more than 80 h nonstop from New England to Bermuda or the Caribbean Islands with no food or rest. However, the monarch butterfly (*Danaus plexippus*), which travels more than 3000 km, or dragonflies, which apparently fly across the Gulf of Mexico, are no less striking in their abilities to migrate.

Patterns of migration vary among different groups of migrants. Migrants may travel exclusively during the day (diurnal), at night (nocturnal), or up and down mountainsides (altitudinal). Some insects and mammals (e.g., lemmings) said to be migratory actually are not because their movements are emigrations from population centers to food sources. Their return to their site of birth is not predictable and often does not happen at all. Most migrants travel in groups that either are loosely formed (insects and small birds) or are in distinct flocks (waterfowl and cranes). Onset of migration is usually controlled by weather conditions. For example, individuals of migratory fishes typically gather in schools (e.g., salmon, eels, and shad) before setting out. In some fish species, the number of migrants passing a point pulses in correlation with environmental conditions such as water temperature. Migratory insects may incorporate several generations in one annual circuit (e.g., monarch butterfly), during which time adults survive long enough to get to the next stopover and lay eggs, which develop into the adults covering the next leg of the migration. Birds, particularly small passerine species, are the icons of migration. They annually depart from wintering grounds in spring and from breeding grounds in autumn at fairly precise dates, take similar routes from year to year, and make extensive preparations for the trips. Migratory birds and some insects (e.g., dragonflies) must cross great expanses in which there is no suitable habitat to land or refuel. Birds crossing the Gulf of Mexico number in the millions, and in the spring their passage (together with numerous insects also in migration) may be readily seen on radar screens.

Birds crossing more hospitable terrain typically alternate one or more days of flight followed by

longer stopover periods, during which energy for the next flight period is accumulated. Short flights require relatively small energy reserves, whereas long flights depend on large fuel loads. The total amount of energy reserve is a compromise among the effects of loading on energetics, risks of the trip, and necessary duration of the flight. Furthermore, a key factor in acquiring the birds' optimal fuel reserves and for their resulting migration speed is the rate of fuel deposition at stopover sites. In small birds, metabolizable energy and intake rates are very high, particularly 1 day after reaching stopover (feeding) areas. This may indicate an increase in the capacity of the digestive tract after many hours of fasting during flight.

Some fish have similar adaptations for migration. Sockeye salmon (*Oncorhynchus nerka*) accumulate large fat reserves that are largely consumed during swimming over long migratory trips to upstream breeding sites (as much as 1200 km). In the course of their migration, they may use 91–96% of their lipid (possibly all of the lipid that can be mobilized) and more than 50% of the protein content of their flesh.

The general timing of migration is under the control of circannian (endogenous annual) physiological rhythms more finely tuned by changes in photoperiod. Increasing day length stimulates spring migration in a number of species, and decreasing day length stimulates fall migration. This is especially evident in birds. Day length is the most accurate environmental indicator of time of the year, but many species begin subtle preparation for migration before the photoperiod changes greatly. Timing also varies with age of the migrant, body size, distance of the migration, and rate of accumulation of energy stores.

Many migratory species routinely return to the site of their birth with precision. This is called philopatry or site fidelity. Philopatry has long been known for birds, but it has also been recognized in sea turtles, bats, cetaceans, and other species. Returning to specific locations, especially if breeding has been successful, is of obvious importance and may be related to the ability of the organism to accumulate sufficient energy in its previous breeding attempt. However, such precise migration requires special skills. The navigation of long-range migrant birds depends on their striking sense of location, including their ability to detect magnetic fields. Some species correct their direction of movement by integrating information from star patterns with these fields. Wintering site fidelity is less well-known, but there are examples of species (passerine birds) that

appear to return to specific sites on the wintering grounds and this is especially true for nonpasserine birds, such as waterfowl and cranes, that travel in specific flyways and winter on the same refuges year after year.

No other single event in the life history of migratory organisms requires more energy than that expended in travel during this period. Indeed, within the span of a few days, some small passerine birds may spend a significant proportion of their total annual energy budget in spring and fall migration. The majority of this energy is expended in two activities: accumulation of energy stores for their journey and the locomotor costs of movement between starting points and ultimate destinations. Much of the evolution of migration centers on the amount and nature of the energy stores and on reduction of the costs of travel.

3. MOLECULAR NATURE OF ENERGY STORES

Several fuels have been shown to support the energetic demands of muscular activity of migrating animals. Among these are carbohydrates, lipids, and some amino acids. Not all animals use the same fuel and some may even use different substrates during different stages of a prolonged trip, particularly if feeding is not possible. Glycogen and trehalose, a disaccharide that hydrolyzes to glucose, may be used in metabolic pathways of some migrant Diptera (flies) and Lepidoptera (butterflies and moths). Other species typically store glycogen, which is transformed to glucose (a simple sugar) for use in the Krebs cycle. In birds, glucose and glycogen are used for muscular activity, but accumulation of lipid for migration has been observed widely; it has been demonstrated in at least 40 families of birds. Some insects and birds use large proportions of glycogen early in migration but begin to use lipid as glycogen stores are depleted. Protein can be used in cases of extreme exertion, but this usually is indicative of an animal near its limits of survival.

In many birds, some insects, and perhaps other animals, nonesterified fatty acids provide fuel for avian exercise, particularly as activity is prolonged. Declines in lipid reserves and downward shifts of respiratory quotients (carbon dioxide produced/oxygen consumed) during prolonged locomotion under controlled conditions in the laboratory support these observations. Premigratory fat deposits are generally

TABLE I

Energy Content of Common Fuels Used in Migration

Material	J/g
Carbohydrates	16.7–18.8
Proteins	~18.0
Lipids	37.7–39.7

composed of triglycerides containing a mixture of fatty acids of various chain lengths and degrees of saturation. The metabolic pathway in which these lipids are used as fuel in many insects, fishes, birds, and mammals is β -oxidation. In this process, fatty acids are broken down into two-carbon units, converted to acetyl-CoA, which then enters the Krebs cycle. A great deal of energy (38 ATP) may be generated for every two-carbon unit going through the biochemical pathway.

Lipid is superior to carbohydrate as an energy reserve because lipid contains more energy than carbohydrate (Table I), and storage of carbohydrate is typically accompanied by storage of water, which increases body mass but contributes no energy. The water may amount to approximately 3 g per gram of glycogen. In birds, the net result is that 1 g of lipid potentially provides more than 37 J, whereas 1 g of glycogen (including water) provides only approximately 5 J. It is obvious that lipid is superior to glucose or glycogen as a intrinsic fuel source for long-range migration, especially in flying animals.

Lipid depots are mostly in the form of triglycerides (triacylglycerols), which consist of three fatty acid molecules attached to a glycerol moiety. The triglyceride content of adipose tissue may exceed 80% in some species, particularly birds. The fatty acids can be released metabolically from the glycerol molecule either entirely or partially and are then carried to the mitochondria within muscle cells, where they are transformed to produce substrate for oxidative metabolism. They are converted to two-carbon (acetyl) fragments that are oxidized in the Krebs cycle, producing relatively large numbers of ATP molecules. The amount of energy produced depends on the carbon chain length of the fatty acid (Table II). Although fatty acids may vary from 10 to 24 carbon atoms (chain length), the most common fatty acids have chain lengths of 16 and 18 carbons. Triglyceride composition may vary seasonally, geographically, and with diet of the migrant. The most energy-rich forms of triglyceride are those with greatest chain length that are least saturated (more

TABLE II

Common Names, Melting Points, and Energy Content of Common Fatty Acids

Common name	c:d ^a	J/g	Melting point (°C)
Saturated fatty acids			
Capric	10:0	35.3	32
Lauric	12:0	36.8	45
Myristic	14:0	38.0	54
Palmitic	16:0	38.9	63
Stearic	18:0	39.7	70
Arachidic	20:0	40.5	76
Behenic	22:0	41.0	80
Lignoceric	24:0	41.4	88
Unsaturated fatty acids			
Palmitoleic	16:1	38.5	33
Oleic	18:1	39.3	14
Linoleic	18:2	39.7	-5
Linolenic	18:3	40.0	-11
Arachidonic	20:4	—	-50
Erucic	22:1	—	33
Nervonic	24:1	—	43

^aRatio of carbon atoms (c) to double bonds (d).

double bonds). Unfortunately, these have higher melting points and may be less easily mobilized for use. Migratory birds tend to have flight muscles that are largely “red” fibers with many mitochondria, a good blood supply, and the metabolic capacity for sustained exercise fueled by lipids.

4. MAGNITUDE OF ENERGY STORES

The amount of fuel stored for migratory trips is extensive in some species. For example, the blackpoll warbler, a bird that flies more than 80 h nonstop during fall migration, may add 10 g of lipid to its body mass of 11 g shortly before its fall migration. Hummingbirds that cross the Gulf of Mexico during migration may be even more obese. These birds arguably have the largest energy stores of any animal in the world, rivaled only by some mammals in their prehibernation period and some cetaceans and pinnipeds outside of migration periods. Insects and bats that migrate over fairly large distances also deposit substantial amounts of fat.

In general, the mass of energy stores varies with size of the migrant, the distance to be covered, and barriers to be crossed (Table III). For example, small

TABLE III

Lipid Reserves in Some Migrants

Species	Lipid (%) ^a
Short-range migrants	
<i>Chorizagrotis auxiliaris</i> (moth)	5–15
<i>Aphis fabae</i> (aphid)	31
Yellow-rumped warbler	55–84
Little brown bat	38
Long-range migrants	
Migratory locust	35–69
Monarch butterfly	43
Blackpoll warbler	342
Bobolink	273

^aPercentage of total dry mass.

birds have relatively larger premigratory energy stores than large birds, everything else being equal. Transoceanic migrants must have large fuel supplies, especially if they are not able to land on the water or forage en route. Accumulating such reserves may be done by any or all of the following: hyperphagia (excessive eating) in the period prior to departure, increased efficiency of nutrient utilization, reduction of energy expenditure in activities not related to preparation for migration, and selection of food items high in energy. In most species, only hyperphagia seems to be important. Some migrants, particularly small birds, can fatten significantly in a short period of time, perhaps 1 or several days. Premigratory hyperphagia is best known in birds, but relatively little evidence of the phenomenon exists for insect, bats, or other forms. In birds, hyperphagia results in hyperlipogenesis and lipid deposition may rapidly occur. Some investigators have not been convinced that all of the fattening occurs as a result of hyperphagia, but there is little evidence of increased efficiency of assimilation, decreased activity, or decreases in standard metabolism in the premigratory period. In fact, small birds demonstrate intensive nocturnal activity under caged conditions (Zugunruhe) during the premigratory period that is not shown at any other time. This paradoxical behavior would be wasteful of energy but probably represents a stereotyped behavior that replaces actual migration. There have been several observations of animals making dietary shifts to items containing more energy during the period just prior to migration.

Accumulation of sufficient reserves in a short period of time has a number of advantages. First, the

time of arrival of some birds on their breeding grounds may be affected by their ability to accumulate sufficient reserves on the wintering rounds (or v.v.). It appears that arrival dates are extremely important in gaining advantage on competitors, exploiting food resources that are only temporarily abundant, and avoiding inclement weather. Second, heavy fuel loads may increase wing loading in flying animals, thus making them vulnerable to predators, so rapid accumulation and use is adaptive.

5. ANATOMICAL STORAGE SITES FOR ENERGY RESERVES

Energy deposits are basically stored at two general locations within the migrant's body. Poikilothermic animals tend to store lipids in discrete fat bodies that are located within the body cavity. (Fat bodies in the thorax and abdomen are the major sites of storage in insects.) This is also generally true for short-distance migrants, such as frogs and some salamanders, but may vary widely in other poikilotherms. For example, in migratory fish, lipid storage sites are in connective tissue, mesenteries of intestines, skeletal muscle, liver, or the head, skin, and tail. Homeotherms also deposit lipids within the body cavity but usually store large amounts of fat at subcutaneous sites. Whales are obvious examples. Their subcutaneous fat deposits were the foundation of the whaling industry. The difference between the spatial patterns of poikilotherms (fish, amphibians, and reptiles) and homeotherm (birds and mammals) may be due to heat conservation by the latter through the increased insulation provided by the subcutaneous fat. In birds, fat deposition may be done in a fairly discrete sequence in which the lipid is placed at fairly specific sites, beginning in the body cavity and ending at several subcutaneous sites. For example, the white-crowned sparrow (*Zonotrichia albicollis*) deposits lipid in 15 recognizable regions. In birds, subcutaneous layers of fat associated with feather tracts appear first. Subsequent fattening is manifested as greater amounts of subcutaneous fat plus some abdominal storage. In the final stages of fattening, intraabdominal deposits become extreme. When fat is deposited at both abdominal and subcutaneous sites, it is usually used in the opposite sequence as its deposition (i.e., abdominal fat is used first and subcutaneous fat last). It is logical that lipid nearest sites of utilization would be easiest to mobilize, but there is evidence that the molecular nature of the fatty acids

may differ with the location of the depot. Long-chain, saturated fatty acids have higher melting points than short-chain, unsaturated ones and are presumed to be more easily mobilized. Subcutaneous sites are cooler than abdominal sites, so long-chain, saturated fatty acids may be less easily used.

6. PATTERNS OF ENERGY STORAGE

The storage of energy in the depot fats that fuel most of the migratory trip may play a dominant role in the species' behavior for short periods. Some animals spend the majority of their lives in transit in a way that requires little energy storage for travel (e.g., tuna and seabirds). Storing huge reserves by cetaceans may reflect food availability on opposite ends of the their migratory pathway more than the energy demands of migration. Most other species have life histories that involve annual sequences of breeding and wintering, punctuated with two annual migrations. This pattern usually involves periods of energy accumulation and storage immediately prior to migration, the magnitude of which may be remarkable. In migrants that cross large areas in which food is unavailable, fuel storage is more pronounced. For example, transoceanic migrant passerines deposit large amounts of fat; intracontinental migrants stop periodically, never fattening extensively. Swimming migrants (fish and cetaceans) may likewise pass through areas in which food is scarce and thus must store large amounts of fuel (typically lipid). Terrestrial, short-range migrants either make very short trips, after which they feed for a few days, or feed as they go, stopping only during part of the day. In both instances, they have relatively small energy depots in comparison with long-range migrants.

In some cases, the temporal pattern of energy storage differs between spring and autumn migration. The difference may be a response to availability of food resources but also may be due to the benefits of precise timing of migrants traveling to breeding sites. In some small birds, fattening in spring occurs during a shorter period of time than does lipid acquisition in the fall, and nonmigrants show little seasonal variation in fattening, especially if they live in tropical habitats (Fig. 1). Precision in spring is beneficial because the earliest arrivals tend to obtain better territories and are more prone to reproduce successfully.

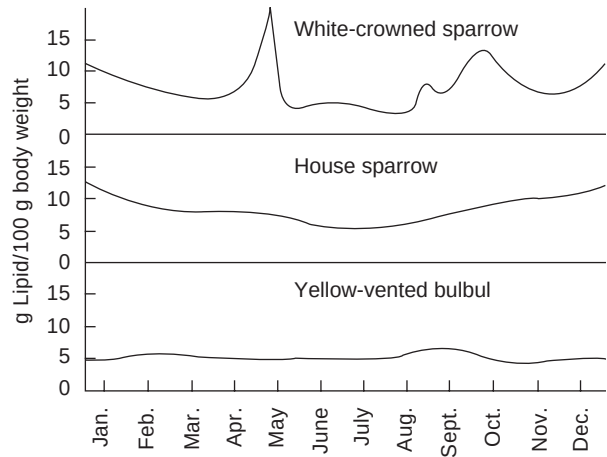


FIGURE 1 Annual cycles of lipid deposition in three passerine birds. The white-crowned sparrow (*Zonotrichia albicollis*) is a temperate zone migrant, the house sparrow (*Passer domesticus*) is a temperate zone nonmigrant, and the yellow-vented bulbul (*Pycnonotus goiavier*) is a tropical nonmigrant. From Blem (1990).

7. ENERGETIC COSTS OF MIGRATION AND FLIGHT RANGES

The energetic cost of migration, although striking in the short term, typically is not a great part of the annual budget of most animals. In long-range migrants such as some passerine birds, less than 1% of their annual energy is expended in migratory flight. In more leisurely migrants, this is similar, but the necessary energy can be collected as the migrant travels and thus the process is less demanding.

Long-range migrants typically have high rates of metabolism, great endurance, and a large metabolic scope (ratio of peak metabolic rate to standard or resting metabolic rate). Other factors affecting the ability of migrants to travel long distances include speed of travel and the mass of the migrant.

Migration is least costly for swimming organisms, more energetically expensive for flying organisms, and most costly for animals traveling terrestrially. Large animals have relatively low costs of transport and small animals have large costs (Table IV).

In flying migrants, flight range depends largely on the amount (and possibly chemical composition) of fuel stored prior to migration. The amount of fuel and the style of flight are a compromise among the costs of transportation of reserves, the risks of migration, and the nature of the migration (i.e., short-range intracontinental hops versus long-range transoceanic flights). However, energy demands of flight can be reduced by soaring, which is relatively inexpensive, and by flying with aid of favorable

TABLE IV
Costs of Transport of Some Species

Species	Mass (g)	Energy (J/kg km)
Mosquito	0.0082	0.033
Migratory locust	2.0	8.0
Hummingbird	4.0	16.0
Chaffinch	21.0	16.4
Laughing gull	310.0	6.1
Lemming ^a	61.0	164.0
Sockeye salmon	3.0	2.0
Sockeye salmon	1500.0	4.5

^aNot really migratory.

winds. Insects fly low within the boundary layers of shelter, use updrafts, or fly with the wind. Birds sometimes use soaring, flying information, or fly with the wind to reduce costs of migration. The monarch butterfly, however, has been observed flying into winds up to 12 km/h. In fact, the volume of migration of insects and birds is often a function of weather patterns. Migrant birds in eastern North America are usually associated with southerly winds in spring and north/northwest winds in autumn and often follow fronts. In some species of birds, migration only occurs over part of the range. In such species, the migratory birds may have longer wings, apparently reducing increased wing loading brought about by fuel loads. Empirical studies indicate that avian metabolic rates increase less rapidly with body mass than does the ability to store and carry lipid reserves, thus providing longer ranges for larger birds.

In birds and insects that cross large distances where refueling is difficult (deserts) or impossible (oceans), mass of reserve is important in several ways. Theoretical and empirical studies both suggest that there is an optimal flight velocity with regard to fuel use. Low speeds (near hovering) are expensive, as are very great speeds (Fig. 2). Furthermore, large flying objects travel less expensively at higher velocities than small ones at lower speeds. This results in greater flight ranges for large birds, even though their energy reserves are not proportionately greater (Fig. 3). Additionally, as travel progresses and body mass decreases as fuel is utilized, the optimal speed becomes lower. Some researchers have observed that migrants travel more slowly as they reach their destination. However, untested assumptions about physiology and aerodynamics may greatly affect estimated flight ranges. In the earliest studies of

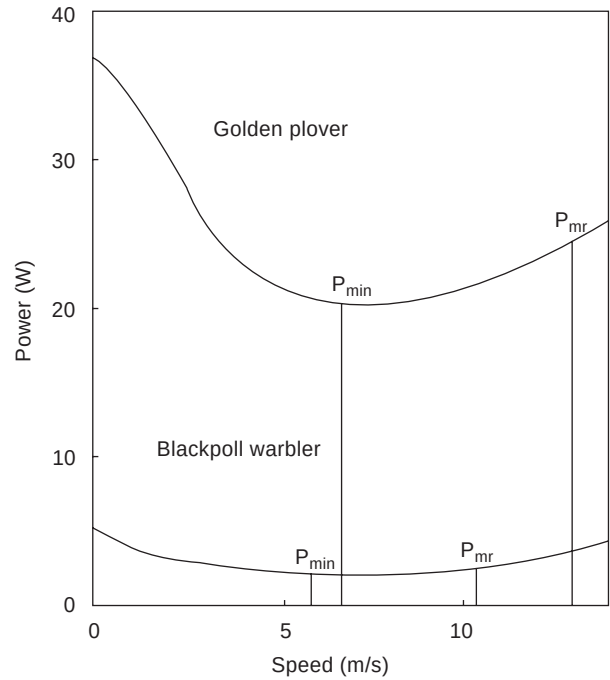


FIGURE 2 Theoretical power curves of two long-range migrants. Lines for minimum power required for flight extend to minimum energy costs ($P_{\min}$); lines for power producing maximum ranges extend to velocities that produce these maximum migratory ranges (P_{mr}). From Blem (1980).

ruby-throated hummingbird migration (*Archilochus colubris*) across the Gulf of Mexico, some investigators doubted that potential flight ranges of hummingbirds were not sufficient to complete the trip. Subsequent measurements of hummingbird flight metabolism indicated that a 4.5-g hummingbird (containing 2 g of fat) could fly nonstop for 26 h (at a cost of approximately 3 J/h). At 40 km/h, its range was estimated at 1050 km, and it is now believed that a trans-Gulf flight is possible by these birds and the available field data support this observation.

8. ENERGETIC BENEFITS OF MIGRATION

Migration is a risky activity. For example, during migration there is a high potential for mortality from predation, and there are predators that appear to seek out migrant insects, birds, and fish. Small birds that are exclusively diurnal throughout their lives migrate at night, ostensibly to avoid predation. Inclement weather may take a toll. There are numerous records of mass mortality caused by storms during passage over large bodies of water.

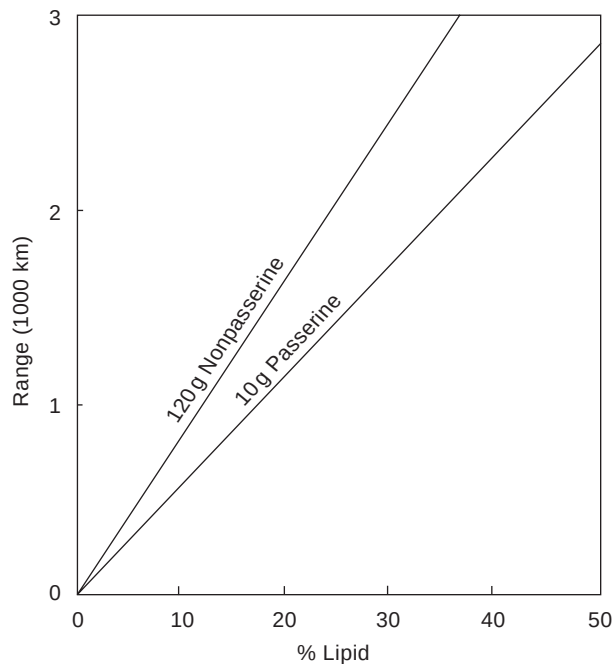


FIGURE 3 Nonstop flight ranges for two avian migrants as a function of their lipid reserves. These curves are often slightly to noticeably concave, which is not apparent here because of the narrow scale of the figure. Lipid content of the migrants is expressed as percentage of total wet body mass. From Blem (1980).

Rafts of dead birds, including kinglets, hummingbirds, and passenger pigeons (many years ago), and several species of insects have been found floating on seacoasts or along the shores of the Great Lakes following violent storms.

Starvation is a real possibility, particularly if the migrant flies or swims in opposing wind/water currents. At some locations, migrant birds are commonly seen in severely emaciated condition. They have little or no lipid, are greatly dehydrated, and are little more than “flying skeletons.” They may rapidly perish if they do not quickly find food and water. Finally, simple accidents occur at high rates. In recent years, accidental death from striking towers, buildings, and other man-made structures has become a major source of mortality, estimated to amount to millions of birds each year in North America alone.

There are several possible benefits of migration that offset the risks that accompany extensive travel, including avoidance of predation and parasitism on the breeding grounds and reduced competition at more distant breeding sites. Many migrants appear to obtain energetic benefits from migration because periodic, heterogeneous food resources become

available to them. In fact, some of these species literally never experience winter. They obtain the benefits of spring/summer throughout their lives. Some species (e.g., pinnipeds and cetaceans) may migrate only because of rich food sources at one or both ends of the journey. Longer day length on northern breeding grounds may provide energetic benefits for species whose foraging time would otherwise be limited. Tests of the hypothesis that birds or other migrants benefit energetically from the environment of the breeding area have provided both positive and neutral results. For example, the dickcissel (*Spiza americana*), a long-range neotropical migrant, attains a positive energy balance during the breeding season because of the longer photoperiod to which it is exposed in the north. American robins (*Turdus migratorius*) appear to obtain similar benefits. On the other hand, tests of field sparrows (*Spizella pusilla*) and tree sparrows (*Spizella arborea*) indicated no such benefits.

It is worth noting that reproductive output in some species increases with distance of migration. At distant breeding grounds, food may be more available as a result of the presence of fewer competitors. This not only provides obvious energetic benefits but also reduces risks of predation if the migrant needs to forage less during the breeding season. Foraging activity usually makes the animal more vulnerable to predation because its attention is directed elsewhere and it sometimes must move out of seclusion to find food sources. Long day length provides more time for foraging of diurnal animals (many insects and birds) and should provide energetic benefits, but these might be partially offset by relatively low environmental temperatures at northern breeding sites.

The energetic benefits of extended day length and higher ambient temperature at southern latitudes in winter may be a selective factor for fall migration in a variety of animals but are best understood in insects and birds. Furthermore, some species of birds appear to become distributed on their wintering grounds in a manner related to energetics. For example, in several species of small birds, sex ratios vary geographically, apparently because of differences in size and the metabolic requirements of the sexes. Males winter farther north and have larger body masses. Bergmann’s rule, an ecogeographic principle used to describe ecotypic variation in vertebrates, suggests that body size increases with latitude in such a way that it provides energetic benefits. Larger individuals may have greater endurance of periods of fasting caused by inclement weather.

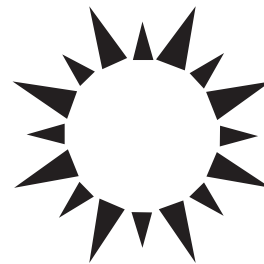
SEE ALSO THE FOLLOWING ARTICLES

Conversion of Energy: People and Animals • Food Capture, Energy Costs of • Heterotrophic Energy Flows • Human Energetics • Photosynthesis and Autotrophic Energy Flows • Reproduction, Energy Costs of

Further Reading

- Alerstam, T. (1990). "Bird Migration." Cambridge Univ. Press, Cambridge, UK.
- Berthold, P. (1996). "Control of Bird Migration." Chapman & Hall, London.
- Blem, C. R. (1976). Patterns of lipid storage and utilization in birds. *Am. Zool.* **16**, 671–684.
- Blem, C. R. (1980). The energetics of migration. In "Animal Migration, Orientation and Navigation" (S. A. Gauthreaux, Ed.), pp. 175–224. Academic Press, Orlando, FL.
- Blem, C. R. (1990). Avian energy storage. *Curr. Ornithol.* **7**, 59–113.
- Blem, C. R. (2000). Energy balance. In "Sturkie's Avian Physiology" (G. C. Whittow, Ed.), pp. 327–343. Academic Press, New York.
- Gwinner, E. (1990). "Bird Migration: The Physiology and Ecophysiology." Springer, New York.
- Klaassen, M. (1996). Metabolic constraints on long-distance migration in birds. *J. Exp. Biol.* **199**, 57–64.
- Lindstrom, A., and Alerstam, T. (1992). Optimal fat loads in migrating birds: A test of the time-minimization hypothesis. *Am. Nat.* **140**, 477–491.
- Pennycuik, C. J. (1975). Mechanics of flight. *Avian Biol.* **5**, 1–75.

Modeling Energy Markets and Climate Change Policy



HILLARD G. HUNTINGTON and JOHN P. WEYANT

Stanford University
Stanford, California, United States

-
1. Alternative Modeling Approaches
 2. The Key Role of Energy Prices
 3. Factors Influencing the Response to Climate Policies
 4. Conclusions

Glossary

autonomous energy efficiency improvement An exogenous technical change that is unrelated to energy prices or other variables predicted by the model.

foresight The ability of producers or consumers to anticipate future changes in market conditions. Perfect foresight means that they correctly anticipate the future (as predicted by the model); myopic foresight assumes that the conditions tomorrow will be like today's.

general equilibrium The joint solution to the market price for multiple energy and economic sectors.

induced technological change An endogenous technical change, usually resulting from price-induced behavior.

interindustry interactions The relationships between industries that reveal which sectors buy from and which sectors sell to each industry.

process analysis An engineering concept that refers to discrete technologies, each often requiring fixed input combinations.

production function A relationship that links inputs to output and specifies the rate at which each input can be substituted for each other input in response to shifts in input prices.

putty-clay capital stock malleability The assumption that original equipment cannot be modified once installed.

putty-putty capital stock malleability The assumption that both old and new capital can be reconfigured once installed, to fit the current price situation in each time period.

vintage Identifying equipment and its energy and other key characteristics by the year when it is introduced.

Various economic models of energy supply and demand have been used for global climate change policy analysis. Although these models are quite diverse in their structure and focus, all systems determine market-clearing equilibrium prices that balance production and consumption levels for different fuel types. On energy demand, these models (except for MARKAL-Macro) are “top-down” in that they seek to establish aggregate or sectoral relationships between consumption, prices, and economic activity, as opposed to the “bottom-up” assessments that focus on specific equipment and their energy use. This comparison focuses on the energy supply and demand adjustments that significantly influence the costs of various strategies for limiting carbon emissions, although the models also incorporate reduced-form representation of the climatic effects that determine the benefits of abatement.

1. ALTERNATIVE MODELING APPROACHES

A number of different economic models has been applied to the global climate change problem; the focus here is on 16 modeling systems by teams, most of whom have recently participated in model-comparison studies conducted by Stanford University's Energy Modeling Forum. The models are identified in Table I along with their principal investigators and sponsoring organization. Half of these teams are based in the United States and half outside of it. The model structures discussed herein appear to be most important for determining the effects of climate policies. Weyant has provided comprehensive discussions of the many major findings or insights from applying these models to this issue.

TABLE I

Models Analyzing Post-Kyoto Energy Modeling Forum Scenarios

Model	Principal investigators
ABARE–GTEM (Australian Bureau of Agriculture and Resources Economics–Global Trade and Environment Model)	B. Fisher, V. Tulpule, D. Kennedy, and S. Brown (ABARE)
AIM (Asian–Pacific Integrated Model)	T. Morita, M. Kainuma (NIES, Japan), and Y. Matsuoka (Kyoto University)
CETA (Carbon Emissions Trajectory Assessment)	S. Peck (Electric Policy Research Institute) and T. Teisberg (Teisberg Assoc.)
FUND (Framework for Uncertainty, Negotiation, and Distribution)	R. Tol (Vrije Universiteit, Amsterdam)
G-Cubed (Global General Equilibrium Growth)	W. McKibben (Australian National University), P. Wilcoxon (University of Texas), and R. Shackleton (U.S. Office of Management and Budget)
GRAPE (Global Relationship Assessment to Protect the Environment)	A. Kurosawa (Institute of Applied Energy and Research Institute of Innovative Technology for Earth, University of Tokyo)
IGEM (Intertemporal General Equilibrium Model)	D. Jorgenson (Harvard University), P. Wilcoxon (University of Texas), and R. Goettle, M. Sing Ho, and D. Slesnick (Dale W. Jorgenson Associates)
MARKAL-Macro	S. Morris (Brookhaven National Laboratory), A. Manne (Stanford University), and P. Tseng (U.S. Department of Energy)
MERGE 3.0 (Model for Evaluating Regional and Global Effects)	A. Manne (Stanford University) and R. Richels (Economic Policy Research Institute)
MIT-EPPA (Massachusetts Institute of Technology Emissions Projection and Policy Analysis)	H. Jacoby/J. Reiner (MIT) and I. Sue Wing (MIT)
MS–MRT (Multi-Sector–Multi-Region Trade)	D. Montgomery/P. Bernstein (Charles River Assoc.) and T. Rutherford (University of Colorado)
NEMS (National Energy Modeling System)	R. Earley, S. Holte, M. Hutzler, A. Kydes, R. Eynon, <i>et al.</i> (U.S. Energy Information Agency)
Oxford Econometrics Model	Adrian Cooper and John Walker (Oxford Econometrics)
RICE (Regional Integrated Climate and Economy)	W. Nordhaus and J. Boyer (Yale University)
SGM (Second Generation Model)	J. Edmonds, H. Pitcher, and R. Sands (Pacific Northwest National Lab)
WorldScan	A. Gielen/H. Timmer (Central Planning Bureau, Netherlands) and J. Bollen (Rijksinstituut voor Volksgezondheid Milieuhygiene, Netherlands)

Although each model has unique characteristics and has proved to be extremely valuable for studying certain types of issues, the structures of the models can be described in terms of five basic categories shown in Table II. Many of the models now employ combinations of traditional modeling paradigms. One category of models focuses on carbon as a key input to the economy. These models do not distinguish between energy types and therefore are similar to models of aggregate energy use. Trends toward less carbon-intensive fuels are incorporated into their projections for carbon. They consider the cost of reducing carbon emissions from an unconstrained baseline by using an aggregate cost function in each country/region. This approach uses a simple vintaging structure to incorporate the time lags in reducing carbon intensity in response to increases in the price of carbon. In these models, all industries are aggregated together, and gross domestic product

(GDP) is determined by an aggregate production function with capital, labor, and carbon inputs. These models generally omit interindustry interactions, include trade in carbon and carbon emissions rights but not in other goods and services, and assume full employment of capital and labor. The Regional Integrated Climate and Economy (RICE) and Framework for Uncertainty, Negotiation, and Distribution (FUND) models are examples of this category of models.

Another closely related category of models focuses heavily on the energy sector of the economy. These models consider the consumption and supplies of fossil fuels, renewable energy sources, and electric power generation technologies, as well as energy prices and transitions to future energy technologies. In general, they explicitly represent capital stock turnover and new technology introduction rate constraints in the energy industries, but take a more

TABLE II

Model Types

Economy model	Fuel supplies and demands by sector	Energy technology detail	Carbon coefficients
Aggregate production/cost function	—	CETA, MARKAL-Macro, MERGE 3.0, NEMS, GRAPE	FUND, RICE
Multisector general equilibrium	MIT-EPPA, WorldScan	ABARE-GTEM, AIM, MS-MRT, SGM	—
	G-Cubed, IGEN ^a	—	—
Multisector Macroeconometric	Oxford Econometrics Model	—	—

^aModels combine multisector general equilibrium with multisector macroeconometric approaches.

aggregated approach in representing the rest of the economy. In these models, all industries are aggregated together, and GDP is determined by an aggregate production function with capital, labor, and energy inputs. These models generally omit interindustry interactions and assume full employment of capital and labor. The Model for Evaluating Regional and Global Effects (MERGE 3.0), Carbon Emissions Trajectory Assessment (CETA), MARKAL-Macro, National Energy Modeling System (NEMS), and Global Relationship Assessment to Protect the Environment (GRAPE) paradigms are examples of this category of models. MERGE 3.0 and CETA have the same basic structure, with nine and four regions, respectively. GRAPE includes a somewhat broader set of technology options, including especially carbon sequestration technologies.

A third category of models is those that include multiple economic sectors within a general equilibrium framework. They focus on the interactions of the firms and consumers in various sectors and industries, allowing for interindustry interactions and international trade in nonenergy goods. In these models, adjustments in energy use result from changes in the prices of energy fuels produced by the energy industries included in the interindustry structure of the model (e.g., coal, oil, gas, and electricity). Explicit energy sector capital stock dynamics are generally omitted. These multisector general equilibrium models tend to ignore unemployment and financial market effects. The Intertemporal General Equilibrium Model (IGEM), the Massachusetts Institute of Technology Emissions Projection and Policy Analysis (MIT-EPPA) model, and the WorldScan model are examples of this type of model. The Global General Equilibrium Growth (G-Cubed) model does consider some unemployment and financial effects and is, therefore, a hybrid general equilibrium/macroconometric model. G-Cubed,

MIT-EPPA, and WorldScan all include trade in nonenergy goods.

A fourth basic class of models combines elements of the first two categories. They are multisector, multiregion economic models with explicit energy sector detail on capital stock turnover, energy efficiency, and fuel-switching possibilities. Examples of this type of hybrid model are the Asian-Pacific Integrated Model (AIM), Australian Bureau of Agriculture and Resources Economics-Global Trade Environment Model (ABARE-GTEM), Second Generation Model (SGM), and Multi-Sector-Multi-Region Trade (MS-MRT) model. These models include trade in nonenergy goods, with AIM including energy end-use detail, GTEM and MS-MRT including some energy supply detail, and the SGM considering five separate supply subsectors to the electric power industry.

By including unemployment, financial markets, international capital flows, and monetary policy, the Oxford model is the only model included here that is fundamentally macroeconomic in orientation. However, as shown in Table II, the G-Cubed and IGEN models do consider some unemployment and financial effects, as well as international capital flows.

2. THE KEY ROLE OF ENERGY PRICES

Economic models assign an important role to the price of energy in determining the economy's adjustment to climate change policies. The models compute the price of carbon that would be required to keep emissions levels controlled at some predetermined level. Although it is easiest to think about these scenarios as carbon taxes or permits (which may or may not be tradable), the additional carbon

costs also reveal important information about any program that seeks to reduce carbon emissions.

Most economic models solve a set of mathematical equations to obtain the prices of goods and services and key inputs, including different forms of energy. The simultaneous solution of these equations represents an equilibrium in which supply equals demand among consumers and producers. In this sense, these models generally determine fuel prices endogenously by searching for those prices that will balance the supply and demand for each fuel. This approach is very valuable for policy analysis of control strategies that will affect the supply and consumption of different fuels. In this framework, an energy price increase can be either the motivation for, or the result of, greenhouse gas (GHG) emissions reductions. For example, governments may impose emissions taxes to motivate GHG reductions. Emissions taxes increase the costs of fuels directly, and economies will adjust to reduce the use of those higher cost fuels, substituting goods and services that result in fewer GHG emissions. On the other hand, governments may cap the total amount of emissions, distribute or sell emissions "allowances," and let the market determine the price and distribution of these allowances. Such a "cap and trade" system will induce changes in prices that are difficult to predict. Because a cap would essentially restrict the supply of carbon-based fuels, GHG consumers would bid up the price until demand for such fuels no longer exceeded supply. In this way, the higher prices reduce emissions, but also allocate the GHGs to their highest value uses.

The effects of higher fossil fuel prices would diffuse throughout the economy. Prices of secondary energy sources, such as electricity and oil products, would rise as the higher primary fuel costs are passed through into electricity rates, fuel oil, and gasoline prices. Higher fuel costs would also increase operating costs in transportation, agriculture, and especially industry. Although energy costs make up only 2.2% of the total costs in U.S. industries, they constitute up to 25% of the total costs in the most energy-intensive sectors (e.g., iron and steel, aluminum, papermaking, and chemicals). Each industry's ability to pass these cost increases along to customers through higher product prices would depend on the strength of the demand for its products and on the severity of international competition. Because many of the major trading partners of the United States would also be implementing similar climate policies, it is likely that the energy cost increase would result in higher prices for a broad range of consumer products. Households could also be affected through increased heating,

transportation, and utility bills and, to a lesser degree, food bills and other costs of living.

A host of adjustments by producers and consumers in the economy would take place in parallel with the price increases, and, in fact, these substitutions would also serve to limit the extent of the price increases that would ultimately result. Higher energy costs would induce firms to accelerate the replacement of coal-based or obsolete plants with more energy-efficient or less carbon-intensive equipment. Utilities and their customers would seek alternatives to carbon-intensive coal-fired power plants, stimulating the market for hydro, nuclear, gas-fired, and renewable electricity sources. As coal prices rise relative to natural gas prices, modern gas-fired combined-cycle power plants would become even more competitive. Older, less efficient coal-fired plants would probably be retired from service or reserved for intermittent operations. Energy-intensive industries would also face a number of adjustment decisions: whether to retire obsolete facilities and concentrate production at more modern, low-cost facilities; whether to modify their furnaces to burn gas instead of coal; whether to generate their own electricity; whether to invest in a wide variety of energy-conserving process changes; whether to redesign products to save energy; and whether to alter their product mix. Ultimately, there would be an effective diminution in the value of the existing stock of plant and equipment because it is optimized for the set of input prices that prevailed when it was installed and would be suboptimal for the new price regime.

In the short run, consumers and producers would reduce their energy consumption by either consuming less energy services (for example, turning their thermostats down or driving their automobiles less) or producing less output. Consumers and producers may also, potentially, reduce energy use without reducing output by identifying energy efficiency measures previously believed to be uneconomic. In the intermediate time frame, there might be opportunities for fuel switching (or substitutions between other inputs) that would not involve substantial outlays for new equipment or infrastructure (for example, switching the fuel used in a multi-fuel-capable boiler from oil or coal to gas). In addition, consumers may be able to substitute goods that require less energy to produce (which would become relatively less expensive) for more energy-intensive ones (which would become relatively more expensive). In the long term, new technologies would be purchased that either use less GHG-intensive fuel or are more fuel efficient. In addition, new, less GHG-intensive technologies might become available over

time as a result of research and development expenditures or cumulative experience. The emergence of these new technologies might be related to the energy price increases, the base case trend of all other prices, or simply the passage of time.

Higher energy prices would lead to less energy use, and less energy use would decrease the productivity of capital and labor. These productivity changes would, in turn, generally result in a slow-down in the accumulation of capital equipment and infrastructure, and in lower wages for workers. Ultimately, even after all the adjustments have been made, consumers would have somewhat less income, which might cause them to adjust the amount of time they spend on work rather than leisure. The resulting adjustment in labor depends on two opposing effects. Workers would want to work more to make up for their loss in real income, but the lower wage would make working worth less relative to leisure. Replacing work with leisure would involve an additional change in welfare. Offsetting these welfare losses would be the benefits of reduced climate change and the benefit of making those responsible for GHG emissions pay for the damages they cause.

The complicated web of economic adjustments that would take place in response to rising prices of energy, or energy scarcity, makes the task of projecting the costs of GHG mitigation a challenge. Interpreting the results they produce is further complicated because different modeling systems emphasize different dimensions of the adjustment process. Also, different policymakers may be interested in different policy regimes, and in different impacts of climate change and climate change policies.

3. FACTORS INFLUENCING THE RESPONSE TO CLIMATE POLICIES

Baseline economic, technological, and political conditions, the opportunities to substitute away from fossil fuels, the nature of the capital stock turnover process, and the dynamics of technological progress are four very important factors that determine a model's response to climate change policies.

3.1 Baseline Conditions

One set of important issues is the base case emissions and climate impact scenarios, against which the costs and benefits of GHG mitigation policies are assessed. They are largely the product of assumptions that are

external to the analysis. Each GHG mitigation cost analysis relies on input assumptions in three areas:

- Population and economic activity.
- Energy resource availability and prices.
- Technology availability and costs.

Most of the researchers projecting the cost of reducing carbon emissions have relied on worldwide population growth projections made by others (e.g., the World Bank or the United Nations). These external projections are generally based on results from very simple demographic models. There is less uncertainty about projections for the developed countries, where population is expected to peak very soon, than for the developing countries, where population is typically assumed to peak somewhere around the middle of this century. Very few of the researchers analyzing GHG emissions reductions make their own projections of economic growth. Most rely on economic growth projections made by others, or on external assumptions about labor force participation and productivity growth.

Another key set of assumptions concerns the price and/or availability of energy resources. The prices of fossil fuels (oil, natural gas, and coal) are important because producers and consumers generally need to substitute away from these fuels when carbon emissions are restricted. Optimistic assumptions about natural gas availability and/or substitutability can make carbon emissions reductions easier to achieve in the short run. Natural gas plays an important role because its carbon emissions are about 60% of those from coal, and 80% of those from oil, per unit of energy consumed. In addition, the amount of unconventional oil and gas production that will ultimately be technically and economically feasible is highly uncertain. It depends on future economic incentives for oil and gas exploration and production, which could (absent climate policies) retard the development of carbon-free renewable and higher efficiency end-use energy technologies. How oil exporters would react to a climate policy that would reduce the demand for oil imports is another key dimension of the energy supply picture.

A final set of key assumptions includes those made about the costs and efficiencies of current and future energy-supply and energy-using technologies. These factors tend to be critical determinants of energy use in both the base case and control scenarios. Most analysts use a combination of statistical analysis of historical data on the demand for individual fuels and process analysis of individual technologies in use or under development, in order to represent trends in

learning by doing (past experience with the technology) and its effect on reducing costs as well as other characteristics of energy technologies. Particularly important, but difficult, is projecting technological progress within the energy sector. Jorgenson and Wilcoxon have attempted to estimate systematically and empirically future trends in energy productivity at a national level, but such efforts are rare. Typically, analysts take one of two approaches: (1) the costs and efficiencies of energy-using and energy-producing technologies are projected based on process analysis, and the characteristics of these technologies are extrapolated into the future, or (2) the trend in energy demand per unit of economic activity, independent of future price increases, is assumed. Some recent analyses have attempted to blend the two approaches. At some point, these two approaches tend to converge, because the end-use process analyst usually runs out of new technologies to predict. It is then assumed that the efficiency of the most efficient technologies for which there is an actual proposed design will continue to improve as time goes on.

Projections of the benefits of reductions in GHG emissions are also highly dependent on the base case scenario employed. The greater the base case damages (i.e., the damages that would occur in the absence of any new climate policies), the greater the benefits of a specific emissions target. The magnitude of the benefits from emissions reductions depends not only on the base case level of impacts, but also on where they occur, and on what sectors are being considered. In fact, a number of additional socio-economic inputs (e.g., income by economic class and region; infrastructure and institutional capability to adapt to changes) are required because they determine how well the affected populations can cope with any changes that occur. The task of projecting base case climate change impacts is particularly challenging because (1) most assessments project that serious impacts resulting from climate change will not begin for several decades and (2) most of the impacts are projected to occur in developing countries, where future conditions are highly uncertain. How well developing countries can cope with future climate change will depend largely on their rate of economic development.

3.2 Representation of Substitution Possibilities

As efforts are made to reduce GHG emissions, fossil fuel combustion and other GHG-generating activities

become more expensive. Producers adjust to these price increases by substituting inputs (i.e., switching to inputs that generate fewer GHG emissions in manufacturing any particular product) and by changing their product mix (i.e., producing different products that require less GHG emissions to make). The extent to which inputs can be shifted depends on the availability and cost of appropriate technologies as well as on the turnover rate of capital equipment and infrastructure. These two factors, as well as consumer preferences, determine an industry's ability to produce and sell alternative mixes of products. Increases in the costs of fossil fuels and products that depend on fossil fuel combustion will reduce consumers' real incomes. Consumers will simultaneously decide (1) the extent to which they wish to adjust their mix of purchases toward less carbon-intensive products and (2) how to adjust their mix of work and leisure time to compensate for the reduction in their real income.

3.2.1 Short-Term vs. Long-Term Substitution

If businesses and households have several decades to complete the substitution process, the current stocks of energy equipment and associated infrastructure do not constrain the substitutions that they may make. Businesses and households are limited primarily by the available technologies and by their own preferences regarding how much of each available product they would buy at the prevailing prices.

If climate policy is long term, the transition to a lower carbon energy system can be relatively smooth and the costs relatively moderate. To reach such an outcome, economic incentives should be designed to motivate producers and consumers to invest in more energy-efficient and less carbon-intensive equipment when their existing equipment has reached the end of its useful life. Useful life is an economic concept that compares the costs of operating existing equipment with the costs of purchasing and operating new equipment. A new and better computer may be purchased after 3 years, even though the old computer could be "useful" for 10 years, because the new one has superior cost and performance characteristics. Or an old car may be kept running because the performance advantage of the new car is not worth the cost.

Over shorter time spans, however, existing plant and equipment can significantly constrain the behavior of firms and households, adding transition costs to the long-run costs of GHG control policies. Policies implemented on this time scale (i.e., within 10 years) will lead to reductions in energy services

(e.g., industrial process heat and home heating and cooling), some easy fuel switching, and an increase in the purchase and use of available energy-efficient products and services. They will also influence the rate of retirement and replacement of existing equipment. Energy-producing and energy-using goods are relatively expensive and long-lived. Thus, it will generally take a substantial increase in energy prices to induce those who own such equipment to replace any of it before the end of its useful life.

The importance of capital stock dynamics creates a formidable challenge for the analytical community. Some data on the characteristics of the energy-producing and energy-using capital stock are available. It would be ideal to have information on the costs of operating and maintaining every piece of equipment currently in use. This would enable analysts to calculate all the trade-offs between retiring equipment early and using other strategies to achieve the specified targets. Unfortunately, the data that are available are generally aggregated across large classes of consumers and generally include all existing capacity without regard to when it was installed. An important exception is power plant data, which are very disaggregated and include the age of the equipment. However, even these data are generally not sufficient to ascertain precisely the point at which the carbon price incentives will influence the rate of replacement of plant and equipment. Limitations on data require the analyst to make a number of assumptions regarding the aggregation and interpretation of the available data.

3.2.2 Two Approaches to Representing Substitution Possibilities

In many models, technologies are represented with “production functions” that specify what combinations of inputs are needed to produce particular outputs. The production function specifies the rate at which each input can be substituted for each other input in response to shifts in input prices. As new capital investment occurs and older capital is retired, the technology mix within the model will change.

Two basic types of production functions may be specified: aggregate production functions and technology-by-technology production functions, also known as process analysis. Some models (e.g., G-Cubed, SGM, and EPPA; see Table I for model identification) use smooth and continuous aggregate production functions that allow incremental input substitutions as prices change, even if the resulting input configuration does not correspond to a known technology. These models do not represent individual

technologies. Such models often assume “nested” production functions. For example, at one level, substitutions are possible between energy, capital, and labor in producing final commodities; at a second level, substitutions are possible between electricity and fuel oil in producing energy; and, at a third level, substitutions are possible between coal and natural gas in producing electricity. Models employing continuous aggregate production functions do not account for individual technologies.

In contrast, other models (e.g., MARKAL-Macro and NEMS) draw from a menu of discrete technologies, each requiring fixed input combinations—i.e., each technology is essentially represented with its own production function. This approach is often referred to as “process analysis.” These combinations correspond to those employed in actual, or anticipated, technologies that the modeler specifies. The technology-rich MARKAL-Macro model specifies over 200 separate technologies. For discrete technology models, different technologies become cost-effective as input prices change. Modelers then assume that these technologies are selected and used to produce outputs. Process analysis represents capital stock turnover on a technology-by-technology basis. The data and analysis requirements for this type of model can be substantial.

A number of systems use a process analysis approach within the energy sector and an aggregate production approach for the remainder of the economy (e.g., MERGE, MARKAL-Macro). When using either approach, it is important to be able to distinguish between the causes of changes in the selections the models make among the existing technologies. Sometimes the technology choice changes because prices change, and sometimes it changes because new technologies become available.

Some models represent both individual energy supply technologies and individual energy consumption technologies, and do not represent the remainder of the economy explicitly. With these models, however, the analyst must either (1) assume that “end-use” energy demands (such as the demand for home heating and automotive transport) do not respond to changes in the prices of those services or (2) employ a complex statistical estimation technique (which requires some historical data on the cost of end-use energy equipment) to estimate the price responsiveness. The choice of production function depends, in part, on the time frame under consideration and the level of technological disaggregation. Short-term models intended to shed light on precise technology choices specify production functions for

large numbers of separate technologies. In contrast, models concerned with longer term effects can safely characterize technological trends using aggregate production functions. Many models blend the two approaches. Although they allow for smooth input substitution in determining new capital investment, they fix input proportions for all equipment installed in a certain year (sometimes called a “vintage”). Similarly, a model may have smooth production functions for conventional fuels, yet stipulate discrete technologies for a particular noncarbon fuel (e.g., EPPA).

3.2.3 Capital Stock Turnover and Malleability

In modeling capital investment in plant and equipment and turnover, each system must use assumptions about the flexibility the investor has in choosing technologies and in changing their characteristics after installation. Data availability and computational considerations limit the choice of modeling assumptions that can be employed. Fortunately, there are some simple formulations that seem to give plausible results in most circumstances.

In almost all models, it is assumed that in making decisions about new capital investment, the decision maker (firm, individual, or government entity) has complete flexibility (particularly in the mix of capital and energy inputs required) in choosing among available technologies before their purchase. The models differ, however, in their assumptions about how much the characteristics of the capital equipment can be changed after it has been installed. These adjustments may be desirable if changes in input prices occur, but retrofitting to a certain set of characteristics is generally more expensive than installing equipment with the same characteristics initially. On the other hand, technological improvements may reduce the costs of the retrofitting over time.

Most models make one of two polar assumptions about this process. To describe these assumptions, the metaphor of soft putty and hardened clay has proved useful (“putty” representing a flexible scenario and “clay” representing a hardened or inflexible scenario). In a putty–clay or putty–putty formulation, the first term refers to the assumption about the degree of flexibility in original capital investment, and the second term refers to the assumption about the degree of flexibility in modifying that capital after it is installed.

- In a putty–clay formulation, it is assumed that the original equipment cannot be modified once

installed. Putty–clay assumptions are more realistic in cases in which relative prices are changing rapidly. Here, new capital investments embody state-of-the-art technology and use input mixes that are appropriate for the price expectations that exist at the time of the investment. These characteristics then remain with that vintage until it is scrapped.

- In a putty–putty formulation, it is assumed that capital—old or new—can be reconfigured once installed to fit the current price situation in each time period. Under the so-called putty–putty assumption, the capital stock is a single entity that is neither broken down into separate vintages nor constrained to retain its initial technology and input mix. The term “putty–putty” is used to indicate that capital can be continuously reshaped both before and after investment has taken place. The inherited capital stock adjusts to changes in prices and technology as fully as brand new capital. In effect, the entire capital stock continually adapts itself to reflect current technologies and prices.

The precise details of the capital adjustment process differ from model to model. In some, there is a composite stock of old capital that reflects some average mix of inputs. In others, each vintage is identified and depreciated separately. In many models, the old capital stock cannot be altered. In others (e.g., NEMS), it can be retrofitted if doing so is more profitable than making brand new investments, or if it is required by regulation.

Modelers are just starting to experiment with various hybrids of the two formulations, i.e., putty–semiputty formulations, in which some retrofitting is allowed at some additional cost. One type of putty–semiputty specification allows plant and equipment to be retired before the end of its useful life if the operating cost of the old equipment is greater than the operating plus capital costs of replacement equipment. In this case, the remaining capital costs of the old equipment would have to be written off, so the changes in prices or new technologies would have to be quite significant for this to occur. Prices do rise to these levels in some models in Kyoto Protocol simulations in which the flexibility mechanisms are severely restricted.

3.3 Capital Stock Adjustment Process

Jacoby and Wing have described three characteristics of these models that are important in analyzing the time horizon for meeting the Kyoto targets: the time frame, the level of detail about capital stock and

production structure, and the specification of economic foresight. The first and most obvious attribute is the time interval over which a model solves its equations. If a model uses a 10-year time interval, the relatively long time period limits the model's ability to be used in analyzing phenomena occurring within a decade, such as the consequences of accepting a 2008–2012 Kyoto target after the year 2000. The results of such models may thus obscure important short-run dynamics of adjustment. The second important attribute of the models is the level of aggregation in the capital stock and the production structure. The level of aggregation affects how models represent the sources of rigidity in the production sectors of the economy. For example, the choice about whether to aggregate output and capital by sector or by technology determines the degree of substitution that is possible within the model's structure. Within a specific aggregate, substitutions are, by construction, assumed to be costless. Additional capital stock produces outputs using a combination of inputs that reflects (1) current and expected input prices and (2) the constraints and limits of existing technologies.

Models capture the aging of capital in different ways. In evaluating short-term adjustment to climate policies, the distinction between putty–putty and putty–clay specifications is critical. In the face of a stringent near-term policy, the putty–putty assumption may produce unrealistic results because this specification implies that large parts of the current capital stock can be transformed into more efficient and less carbon-intensive alternatives. However, for analysis of the long run, after fuel prices have settled at a new equilibrium level relative to other goods and services, the distinction is less important. In this postadjustment phase, the inherited capital stock will be increasingly fuel efficient and competitive under prevailing conditions, because those conditions will more closely match the conditions in place at the time the investments were made.

The third important characteristic of models of the capital stock turnover process is the way they treat foresight. Models may specify economic behavior as forward-looking or myopic. Forward-looking models assume that agents with perfect foresight find the path of emissions reductions that minimize discounted costs over the entire modeling horizon, choosing the timing and stringency of control measures so as to smooth optimally the costs of adjustment. In contrast, myopic models assume that economic agents seek to minimize the costs of policy on a period-by-period basis, and take little or no

action in advance of the onset of carbon constraints. Model results can be very sensitive to assumptions about investor foresight. Models that assume perfect foresight allow emissions targets to be met at lower costs because investment decisions are made in the full certainty that emissions limits will be set and achieved. Models that assume some degree of myopia generate higher costs because investors must scramble to alter the capital stock as the target period approaches, prematurely scrapping existing capital (e.g., coal-fired power stations) and quickly investing in less carbon-intensive alternatives.

Of the models reviewed here, a great majority assume perfect foresight, whereas only one is constrained to be myopic (EPPA). Some models (e.g., G-Cubed) allow alternative assumptions under different runs and/or can set expectations differently for different sectors. The NEMS and SGM models can allow industrial or utility investors to give greater consideration to future conditions than individual consumers do.

In practice, investors do not have perfect foresight, nor do they suffer from complete myopia. Although there is inevitable uncertainty regarding future economic conditions, policymakers can reduce uncertainties by making credible commitments to meet targets or to initiate market-based policies. Model results clearly demonstrate that the more convinced investors are that emissions targets will become binding, the less costly the transition to lower carbon emissions.

3.4 Technological Change

How these opportunities will change with time and with people's experience with new technologies also have important effects. Technological change can be thought of as increasing the amount of a product that can be produced from a given amount of inputs, or as expanding the universe of opportunities for substitution of inputs and products that were described in the last section. Technological change is discussed separately from input and product substitution here because the underlying determinants are somewhat different, because technological change is less well understood, and because of the opportunities for synergy between public support and private investment in stimulating new technology development.

In 1942, Schumpeter identified three distinct types of technological change that take place continually in modern economies: (1) invention of completely new ways of satisfying human needs and wants, or the creation of new needs not previously identified or

satisfied, (2) innovation, which takes place through continual improvement and refinement of existing ways of doing things, and (3) diffusion of new technologies throughout and across economies. These processes are all important for climate policy. It often takes decades for innovation and invention to pay off. Even diffusion may be difficult to accelerate over a decade, though, because it takes time to distribute information, analysis, and experience from one user to another.

New technologies can allow firms to produce a particular product using a mix of inputs not previously available, including, for example, less energy. In addition, new technologies can lead to new products. These new products compete with existing products, with further implications for carbon emissions reduction policies. If these new technologies and new products produce less carbon, then carbon emissions will be lower, fewer emissions reductions will be needed, and/or emissions reductions will be less expensive. Projecting how technological change might progress over time, both with and without climate policies, is challenging. The processes by which technological change occurs are very complex and the data required to estimate how these changes have been made in the past are generally not available. However, there are several ways economic models represent technological change.

3.4.1 Induced Technological Change

Inventions of productive technologies or processes are, by their very nature, hard to predict. However, past experience has shown that they can be revolutionary enough to justify large expenditures in basic research in strategic areas. Innovations could be of great help in lowering the costs of reducing GHG emissions. Thus, it would be worthwhile to find an appropriate combination of government interventions and private sector incentives that encourage innovation. Thus far, however, most of the policy debate on the influence of technological change on climate change policy has focused not on technology policy options, but rather on how restrictions on GHG emissions reduce aggregate mitigation costs over time. This latter effect has been labeled “induced technological change” (or ITC for short). ITC has to do with price-induced behavior—i.e., what private firms will do in response to higher prices. It does not incorporate what firms will do anyway in trying to become more competitive through investing in research and development, or what they would do in response to government sponsorship of research and development or other

direct government technology policies. There has been a good deal of discussion about the potential for induced technological change to substantially lower, and perhaps even eliminate, the costs of CO₂ abatement policies. These discussions have exposed very divergent views as to whether technological change can be induced at no cost, or at some cost.

Every ITC model must represent some incentive to induce technical change in one or more ways, such as the following examples:

1. The form of profits from innovations, as in the top-down models, which focus on the behavior of economic aggregates rather than the behavior of individual actors or the use of individual technologies.

2. At a more aggregate and abstract level, by means of cost-functions, research and development production functions, or empirical estimates. Similarly, the decision maker(s) considered may be either decentralized industries, representative firms, or a central planner.

3. By the inclusion of intrasectoral knowledge spillovers that are advances that individual firms within a sector cannot keep to themselves. For example, the level of investment may be determined by the rate of return the firm expects to earn on the research and development investment as compared with other available investment opportunities. However, the rate of innovation may far exceed that implied by the rate of return alone because other firms in the industry may be able to replicate the innovation.

4. By the dimension in which technological change is assumed to progress (i.e., new products or processes, substitution of inputs, or reorganization of production and distribution arrangements).

Some ITC models are based on empirical observations of past responses to energy price and policy changes. One advantage of this type of model is that different sectors may exhibit different rates of technological progress. However, only one model, IGEM, estimates all these parameters simultaneously because of the large amount of data necessary and the heavy computational burdens of such estimations. Another advantage is that this type of model implicitly takes into account real-world factors that are relevant to technological change and that are difficult to incorporate into conventional economic frameworks. Thus, this model relies on empirical observations of real events, not on a simplified representation of the phenomenon. All types and sources of short-term technical change are included. One disadvantage of this aggregation,

though, is that the approach may omit specific known technologies that are beginning to be introduced but that are not yet revealed in the available data. In addition, information about the underlying costs of research and development is lost. Also missing is explicit attention to how firms determine their research and development investments. Firms take into account both the cost of engaging in research and development and the expected benefits in terms of future profitability. Thus, models are unable to evaluate optimal policies with full consideration of the costs of research and development. Another disadvantage is that the model is as limited as the data set from which it is constructed. Only one historical path can be observed, and it is assumed that tomorrow's economy will respond to energy price changes in the same way as yesterday's economy. Thus, long-term technological change is beyond the feasible reach of this type of model. "Long-term" here refers to periods over which substantial technological development and major inventions may occur.

Nonetheless, empirical modeling of ITC may be valuable for short- to medium-term projections, or for estimating the short- to medium-term cost of policies on the economy. Empirical models may also be valuable in comparing or calibrating short-term projections from other types of ITC models. Also, the consideration of ITC helps clarify two key matters of debate: (1) whether prior studies (without ITC) have overstated the cost of achieving given emissions reduction targets and (2) the optimal size and timing of a carbon tax.

3.4.2 *Autonomous Energy Efficiency Improvement*

In contrast to the ITC models, many models include exogenous technical change. ("exogenous" can mean external to the model, or independent of price, or both.) A simple characterization of technological improvement, employed in many of the models, is a single scaling factor—the autonomous energy efficiency improvement (AEEI)—that makes aggregate energy use per unit of output decline over time, independent of any changes in energy prices. (Many modelers specify the AEEI as a percentage of gross domestic product growth, so that the value changes over time.) Although the definition of the AEEI varies from model to model, in all models it implicitly represents the effect of technological progress. In some models, it also represents one or both of two additional trends: (1) changes in the structure of the economy, resulting in a shift in the relative contribu-

tion of energy-intensive industry output to total economic output, and (2) an improvement in energy efficiency over time, reflecting the gradual removal of market barriers that prevent some energy consumers from choosing more efficient energy technologies.

Although the AEEI approach allows for energy improvements over time, it is limited in two respects. First, using the AEEI approach to represent technological change ignores price-induced technological progress (ITC). In reality, higher prices do spur greater innovation and more rapid diffusion of energy-saving technologies. Second, it is not clear what an appropriate rate for AEEI should be. This is important, especially for longer term projections, which are very sensitive to differences in assumed rates. More sophisticated specifications (often used in conjunction with an AEEI parameter) attempt to paint a more detailed picture of technological change by incorporating some degree of price sensitivity, distinguishing different sectors, and assessing changes to specific technologies.

3.4.3 *Learning by Doing*

In practice, much technological advance comes from learning by doing (LBD), which is the incremental improvement of processes through small modifications and adjustments. It is not until a technology is actually used that important lessons are learned that can be applied to its subsequent development. LBD is an integral part of the innovation process. Observations of past technological innovations show that initial installations are quite expensive, but that costs drop significantly the more the technology is used, and the more lessons are learned from using it. This type of learning may be the result of either exogenous or endogenous (induced) technological change.

The LBD approach does not reveal how learning occurs and whether the learning is associated with invention, innovation, or diffusion. Thus, it cannot evaluate which policies might be appropriate for increasing the learning associated with a technology. The approach also suffers from its inability to establish whether future cost reductions result from increased cumulative experience with the technology or whether they occur with the passage of time, which is closely associated with cumulative experience.

Although most models do not attempt to capture LBD, two models do mimic the process. MERGE assumes endogenous diffusion rates: the more investment there is in advanced technologies in the early years of the projection, the greater is the rate of

adoption in the later years. In the NEMS model, learning by doing is represented in the electricity generation sector, where the capital costs of particular types of new plants decline as more such plants are built.

4. CONCLUSIONS

As long as climate change policies are geared toward keeping a dynamic economy tied to some historical benchmark emissions level, projections of baseline economic and emissions conditions will dominate the costs of emissions control strategies. An equally important consideration is the adopted policy regime, such as the extent to which international emissions trading is permitted. In general, the more flexibility permitted as to where, when, and which GHG reductions may be used to satisfy a commitment, the smaller the economic impacts. In addition to these baseline and policy assumptions, the model structures also influence the cost estimates associated with climate change actions. Especially important are how each model's structure accounts for the rate and extent to which available inputs and products can be substituted for one another and the rate of improvement in the substitution possibilities over time (i.e., technological change). The representation of the substitution possibilities depends both on the available technologies and on how the retirement of existing equipment and the introduction of new technologies are represented. The more flexibility the model includes in the choice of technologies, retirement of old equipment, and introduction of new technologies, the lower the economic impacts of emissions reductions.

Technological change occurs when new technologies allow a particular good or service to be produced with fewer inputs, or when a new product is developed. Most models used to project GHG emissions and mitigation costs assume that technological change takes place steadily over time, but does not depend on changes in prices or the implementation of government policy options. Thus, different technologies are selected as prices change, but no new technologies are added to the menu. Analysts have recently started developing ways by which price-induced technological change and price-induced increases in the rate of diffusion of new technologies can be included.

The technological change that occurs over time, and that is included in most of the models, reduces

the costs of mitigating carbon emissions because it decreases the base case trajectory of GHG emissions. However, it is probably unrealistic to assume that changes in energy prices will not alter the course of technological progress. In the short run, price increases should encourage the diffusion of new technologies. In the intermediate term, they should lead to a more rapid rate of increase in the rate of improvement of existing technologies and earlier remodeling or replacement of other facilities and equipment. In the long run, they should stimulate the development of brand new technologies. Both kinds of changes should reduce the average rates of GHG emissions per unit of output.

SEE ALSO THE FOLLOWING ARTICLES

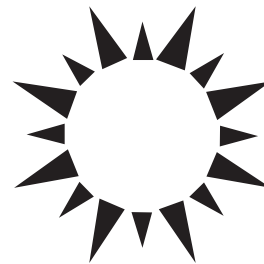
Bottom-Up Energy Modeling • Business Cycles and Energy Prices • Carbon Taxes and Climate Change • Clean Air Markets • Climate Change and Energy, Overview • Climate Change: Impact on the Demand for Energy • Greenhouse Gas Abatement: Controversies in Cost Assessment • Market-Based Instruments, Overview • Modeling Energy Supply and Demand: A Comparison of Approaches

Further Reading

- Jacoby, H. D., and Wing, I. S. (1999). Adjustment time, capital malleability and policy cost. In "The Costs of the Kyoto Protocol: A Multi-Model Evaluation" (J. P. Weyant, Ed.), pp. 73–92. International Association for Energy Economics, Cleveland, Ohio. [Special Issue of *The Energy Journal*.]
- Jorgenson, D. W., and Wilcoxon, P. J. (1993). Energy, the environment and economic growth. In "Handbook of Natural Resources and Energy Economics" (A. Kneese and J. Sweeney, Eds.), pp. 1267–1349. North-Holland Publ., Amsterdam.
- Manne, A. S., and Richels, R. G. (1992). "Buying Greenhouse Insurance: The Economic Costs of Carbon Dioxide Emission Limits." MIT Press, Cambridge, Massachusetts.
- Nordhaus, W. D. (1994). "Managing the Global Commons: The Economics of Climate Change." MIT Press, Cambridge, Massachusetts.
- Schumpeter, J. A. (1942). "Capitalism, Socialism, and Democracy." Harper & Brothers, New York and London.
- Toth, F. L., and Mwandosya, M. (2001). Decision-making frameworks. In "Climate Change 2001: Mitigation" (B. Metz, O. Davidson, R. Swart, and J. Pan, eds.), pp. 601–688. Cambridge University Press, Cambridge, U.K.
- Weyant, J. P. (ed.). (1996). Integrated assessment of climate change: An overview and comparison of approaches and results. In "Climate Change 1995—Volume 3: Economic and Social Dimensions of Climate Change" (J. P. Bruce, H.

- Lee, and E. F. Haites, eds), pp. 367–396. Cambridge University Press, Cambridge, U.K.
- Weyant, J. P. (ed.). (1999). “The Costs of the Kyoto Protocol: A Multi-Model Evaluation.” International Association for Energy Economics, Cleveland, Ohio. [Special Issue of *The Energy Journal*.]
- Weyant, J. P. (2000). “An Introduction to the Economics of Climate Change Policy.” Pew Center on Global Climate Change, Washington, D.C.
- Weyant, J. P., and Olavson, T. (1999). Issues in modeling induced technological change in energy, environmental, and climate policy. *Environ. Model. Assess.* 4, 67–85.

Modeling Energy Supply and Demand: A Comparison of Approaches



ALESSANDRO LANZA

Fondazione Eni Enrico Mattei, Milan, Italy

Eni S.p.A., Rome, Italy

FRANCESCO BOSELLO

Fondazione Eni Enrico Mattei

Milan, Italy

-
1. An Introduction and a Short Historical Overview
 2. Top-Down and Bottom-Up Models: Comparing Technical Features
 3. Top-Down and Bottom-Up Models: Comparing the Methodologies
 4. Top-Down and Bottom-Up Models: Possible Integration
 5. Concluding Comments

Glossary

constant elasticity of substitution (CES) technology Characterized by the fact that production factors are substitutable among each other at a given degree and that this substitution possibility is constant at any given production level.

Hicks-neutral technical change Refers to a technical improvement that uniformly reduces the input requirements associated with producing a given level of output.

Leontief technology Technology in which production is constrained by the input in the lowest availability; in this situation, increasing the availability of other production factors will not allow an increase in production.

panel data Composed of a repeated number of observations of the same group of individuals over time.

time-series data Composed of a repeated number of observations of a variable over time.

Walras's Law Often wrongly considered an equilibrium condition, a law stating simply that consumers fully expend their wealth.

After the first modeling efforts investigating the relationships between energy and economics emerged during the 1970s, two broad classes of

modeling approaches appeared: the economic or top-down models and the technical/engineering or bottom-up models. The top-down models adopted a general perspective and described the economic linkages between energy demand and supply and the rest of the economic system, with the main goal of analyzing energy or wider economic policies. The bottom-up models adopted a focused view of the energy sectors and explored the various technological options, with the main goal of highlighting low-cost energy production opportunities.

1. AN INTRODUCTION AND A SHORT HISTORICAL OVERVIEW

The first modeling efforts investigating the relationships between energy and economics date back to the 1970s. Since the beginning, two broad classes of modeling approaches appeared: the economic or top-down models and the technical/engineering or bottom-up models. The first approach, adopting a general perspective, described the economic linkages between energy demand and supply and the rest of the economic system, with the main goal of analyzing energy or wider economic policies. The second approach, adopting a focused view of the energy sectors, explored the various technological options, with the main goal of highlighting low-cost energy production opportunities. During the same period, natural scientists became interested in energy modeling as well. To assess the role of anthropogenic

greenhouse gas (GHG) emissions in determining global warming, they developed pioneering models consisting at that stage in simple extrapolations based on a restricted number of parameters where qualitative/quantitative experts' opinions were predominant with respect to direct estimation or calibration procedures. Simulation of energy and climate policies was well beyond the capacity of these first mathematical tools.

Interestingly, it was mainly environmental concerns that stimulated both bottom-up and top-down modeling efforts during the subsequent decade. In this respect, a particularly important event was the 1988 Toronto Conference on Climate that spurred a real boom in modelization. One of the outcomes of the conference was to establish a first preliminary limit for carbon dioxide (CO₂) emissions in the measure of a cut of 20% with respect to the 1988 level. This fostered immediate and great interest in the scientific and political communities to assess costs and benefits of such a target, its implications for the economic systems, and the feasible strategies for its accomplishment. Moreover, the need to analyse three different dimensions—energy, economics, and environment—encouraged a fruitful and unprecedented exchange of knowledge among disciplines.

Nowadays, bottom-up and top-down models are being used extensively to investigate climate change issues and the costs of the related policies. Moreover, it is a fact that even when these models are used to investigate energy policies and reforms not directly related to climate, they nevertheless take into account some environmental implications (e.g., GHG emissions are commonly considered a strategic variable to monitor anyway).

It could be said that the points in common between the two approaches end here. In fact, they have led to very different model properties and results. A striking example is again given by the case of determining the costs of climate change policies where top-down and bottom-up studies wound up producing opposite results. This induced two opposing views in the scientific community: one claiming the necessity to integrate the two methodologies so as to exploit the comparative advantages of both techniques and the other considering the two methodologies totally incompatible.

Recently, also due to the development of new flexible software packages and to the improvement in computers' computational capacity, the first view seemingly has prevailed, and a large number of "hybrid" models sharing top-down and bottom-up features is being developed. Still, the integration

techniques are far from reaching an uncontroversial standard and securing full validation from the scientific community; bridging the gap between the two approaches developed for different purposes and permitting different analyses designed to answer different questions remain a difficult and complex task. Yet despite efforts to merge characteristics in the hybrid models, and despite convergence between model categories, the distinction remains an important one that is essential to understanding the policy conclusions of influential studies and to avoiding misinterpretation of model results.

In light of this, the aim of this article is to describe the main technical and theoretical features of bottom-up and top-down modeling approaches, presenting their results, highlighting their respective strengths and weaknesses, and discussing the main issues that have arisen due to a possible integration.

In what follows, this article presents the technical characteristics of the two modeling approaches. It then compares and comments on the theoretical rationales shaping the models' use and results. Subsequently, it describes possible ways in which the two approaches could be and have been reconciled. The final section concludes the article.

2. TOP-DOWN AND BOTTOM-UP MODELS: COMPARING TECHNICAL FEATURES

2.1 Top-Down Modeling: Technical Features

Top-down models are considered economic, general equilibrium, or aggregated models; they aim to giving a comprehensive picture of the functioning of an economic system, including the relationship between energy markets and the rest of the economy, based on the behavior of representative and rational economic agents maximizing an objective function. The information used to build top-down models "comes from the past"; usually, historical data describing energy-economic interactions are used to shape present and future market behavior. With respect to this, two broad parameterization techniques can be identified: calibration (originating models in which the key parameters forging agents' behavior are based on the information stemming from a given point in time, i.e., the base year) and estimation (originating econometrically estimated models in which the parameterization is obtained extracting information from longer time periods, i.e.,

through time-series or panel econometric techniques). In fact, it is common for both approaches to coexist within a model, especially due to constraints imposed by data availability.

Traditionally the economic modeling literature has divided between general equilibrium models (GEMs), whether static or dynamic (on their turn, classifiable in recursive dynamic and growth models), and macroeconometric models. The view of GEMs is “Walrasian”; that is, optimizing producers and households demand and supply goods and factors, perfectly flexible prices adjust excess demand and supply in all markets that clear, and profit maximization under perfect competition and free market entrance guarantee zero profits and the optimal distribution of resources. Macroeconometric models are “neo-Keynesian” in the sense that the economic system is demand driven; moreover, perfect competition is abandoned, especially in energy and labor markets where market power and bargaining processes determine price settings.

Notwithstanding the great heterogeneity within the top-down model family, the explicit aim of comprehensiveness, the focus on agents’ decisions, and the idea that the past can describe the future consistently allow one to identify some important common features in the various modeling efforts.

2.1.1 Key Parameters and Variables

Usually, the most important parameters and variables constituting the main drivers of internal processes and results are the elasticity of energy demand in response to gross domestic product (GDP) or income changes; the elasticity of price substitution among capital, labor, and energy; the indicators of technical progress, usually assuming the form of non-price-induced decreases in energy use per unit of GDP (the so-called autonomous energy efficiency improvement [AEEI]) and/or of a more general autonomous factor-augmenting productivity (generally “Hicks-neutral” technical progress); the cost and availability of existing and future energy supplies; and the availability, conversion efficiencies, and costs of existing and future energy generation technologies, including “backstops” (discussed later). Together with the GDP growth rate, all of these elements determine future energy demand and the economic cost of a technological switch.

2.1.2 Disaggregation

It is a recognized fact that estimated econometric relationships among aggregated variables are generally more reliable than those among disaggregated

variables; as a consequence, the behavior of the models is more stable using such variables. Thus, it is common to adopt a high level of aggregation for top-down models when they are applied to long time frames. An additional constraint on the level of disaggregation attainable is imposed by computational burdens. This is true in particular for dynamic top-down models in which the process of intertemporal optimization used to find the equilibrium path for the economic variables taking into account strategic behavior among various agents (countries or sectors) requires the solution of huge nonlinear systems. In practice, although a static top-down model can reach a good degree of sectoral disaggregation (e.g., the 57 sectors of the GTAP model), a dynamic top-down model usually presents fewer than 10 to 15 sectors (e.g., the G-cubed model) or even no sectoral disaggregation (e.g., the RICE ’96 and RICE ’99 models).

2.1.3 Technology

Being concerned with the feedback between energy and other sectors, and between macroeconomic impacts of given policies (e.g., energy or climate change policies), on the national or even global scale, top-down models tend to propose minimal details of the energy-consuming side of the economy. More specifically, technology is usually represented by the shares of the purchase of a given input in intermediate consumption in the production function, as well as in labor, capital, and other inputs, and by the previously mentioned AEEI and/or the factor accounting for Hicks-neutral technical change. Depending on the functional form of the production function, these parameters describe the degree of substitutability among inputs and the evolution of their productivity. In other words, they shape economically a technology that finally determines energy (and other input) demands for the various energy types. This reflects the idea that it is mainly the characteristics of the specific energy type, such as its price, that govern demand developments. It is important to highlight that shares are usually constant and that parameters of technological progress are usually exogenous, being represented by some kind of time trend. This implies, on the one hand, that the approach to technology is inherently static, somehow “freezing” substitutability relationships among various inputs to what has already been observed (in a given year or a given “past”) and, on the other, that non-price-induced technological progress is not agents’ decision. In other words, agents can decide where to place themselves along the

production frontier, but they cannot determine the shape and shift of that frontier. More recently, the issue of endogenizing technological change was tackled by several researchers. Again the perspective is intrinsically economic. Put simply, the point here is to endow agents with an additional decision variable, usually the amount of investment in research and development (R&D), competing with the others, the stock of which determines, according to some functional form, the productivity and energy efficiency of the input bundle. In this case, AEEI and technical progress do depend on agents' decisions.

2.1.4 The Supply Side

Another key aspect that top-down modeling studies have to consider is the availability of primary inputs, in particular of energy sources and their cost. The majority of top-down models distinguish between carbon-emitting and non-carbon-emitting energy supplies. Usually, fossil fuels (coal, oil, and natural gas) provide energy and produce GHGs as by-products, whereas electricity use is considered carbon free. Some models explicitly consider the possibility of accessing alternative forms of renewable energy supplies with low or zero carbon content. This is done mainly following the so-called “back-stop approach,” a rather common technique. Alternative energy sources are already technologically feasible and available, but they are not economically viable due to their high costs. As time passes, extracting energy from traditional energy sources becomes more costly due to reserve exhaustion. Accordingly, at a given point in time, new energy sources and technologies become competitive and enter into the market, determining a total or partial shift away from carbon-emitting sources. Thus, the assumed cost and availability of backstop technologies are key determinants of the long-term marginal cost of supply in top-down models.

2.2 Bottom-Up Modeling: Technical Features

Bottom-up models are considered engineering, partial equilibrium, or disaggregated models. Their original aim was to find least cost opportunities to meet an exogenous demand for specific energy services. More recently, following environmental concerns, they have been used to devise least cost opportunities to achieve a given energy efficiency or carbon emission reduction. Bottom-up models are based on technological and economic data that describe in great detail past, present, and future

technological options that are used, or that can be used, for harnessing energy resources and convert them into energy services.

Bottom-up models can be grouped into two broad categories: spreadsheet models (which solve a simultaneous set of equations to describe the way in which a given set of technologies is or could be adopted) and simulation or optimization models (which simulate investment decisions endogenously). Again within the heterogeneity of bottom-up modeling, some common features can be devised.

2.2.1 Key Parameters and Variables

In a bottom-up model, the fundamental parameterization concerns the costs and initial values of installed capacities of technologies currently in use and of their alternatives, their residual lives, fuel and electricity costs, and the potential rates and limits of alternative technology penetration.

2.2.2 Disaggregation

As mentioned previously, a typical feature of bottom-up models is the high degree of detail or disaggregation in the representation of the energy sector. But note that disaggregation in bottom-up models has a different meaning with respect to top-down terminology. Most bottom-up models interconnect conversion and consumption of energy via energy carriers. Usually, energy carriers are disaggregated according to their involvement with primary supplies (e.g., mining, petroleum extraction), conversion and processing (e.g., power plants, refineries), and end use demand for energy services (e.g., boilers, automobiles, residential space conditioning). Demand for energy may, on its turn, be disaggregated by sector (e.g., residential, manufacturing, transportation) and by specific functions within a sector (e.g., air conditioning, heating, lighting).

2.2.3 Technology, Supply, and Demand

Bottom-up models capture technology in the engineering sense. This means, for instance, that technology A, with a given performance X determined by a whole set of parameters accessible at the direct cost $C(A)$, on its turn determined by a whole set of parameters, is compared with technology B with a given performance Y accessible at the direct cost $C(B)$. As an outcome, individual measures are ranked in order of increasing net cost to form “packets of options” represented as marginal cost or curves. These are called “conservation supply curves” (CSCs). More specifically, CSCs can be obtained through a three-step procedure: (1) energy-saving

potential and costs of all possible “packets” of options are evaluated, (2) the total “cost of conservation” curve is obtained as the lower convex bound of all conservation packets that may be justifiable at some energy price level, and (3) the CSC is obtained by plotting the slopes of each straight line segment of the total cost of conservation curve. Demand in these models is usually exogenous but can also be determined endogenously. The “typical” focus is on end use energy demand (for heating, lighting, ventilation etc.). This endorses the view that the evolution of energy demand is driven mainly by the various purposes for which energy is made use of. Household energy demands are typically specified by vintage models of a large number of end use technologies whose penetration rates follow a time profile with saturation levels. Penetration ratios could be specified to follow a

logistic function, and in some cases parameters of these logistic functions are estimated for each type of appliance.

All of the descriptions in this section are summarized in Table I and Fig. 1.

3. TOP-DOWN AND BOTTOM-UP MODELS: COMPARING THE METHODOLOGIES

It should not be surprising that the different technical characteristics of the two modeling approaches, peculiar to two different disciplines adopting different perspectives, determine a difference both in the kind of questions the specific models can address and in the qualitative and quantitative results they can produce.

TABLE I

Main Features of Top-Down and Bottom-Up Modeling Approaches

Criterion	Top-down	Bottom-up
Level of disaggregation	Usually low: 1–10 sectors or activities; can be high in some CGE static models	High: a wide range of energy end uses represented
Behavior representation	Comprehensive (general equilibrium approach = full economic feedback effects) but few energy-relevant details	Detailed (at end use level) but not comprehensive (partial equilibrium approach, “no rebounds”)
Representation of technologies	- Based on macro, input/output, or econometric analysis - Production functions determine substitution possibilities	- Based on engineering and cost data - Description of physical flows
Technological change	- Price and income effects - Usually exogenous technical progress (Hicks-neutral + AEEI); some models consider R&D and learning-by-doing-driven endogenous technical progress	- Assumptions on market shares or optimization - Projections of technological efficiency - Learning curves
Methodological approach	- Time-series/panel econometrics or calibration based on a single year - Economic growth estimated or exogenous	- Spreadsheet analysis (for descriptive reasons) - Simulation/optimization models
Efficiency gap	- No energy efficiency gap except in case of energy subsidies - Usually markets are fully competitive in CGE models; market power (oligopoly/monopolistic competition) in energy markets in Keynesian macroeconomic models	- Energy markets are not efficient (functioning inside the production frontier) - Potential for cost-effective energy savings
Assumptions about market barriers and hidden costs of new technologies	Costs of adopting new technologies are reflected in observed behavior	- Significant market barriers prevent adoption of new technologies - Hidden costs tend to be low
Transaction costs of removing market barriers and imperfections	High	Low

Source. Adapted from International Energy Agency (1998).

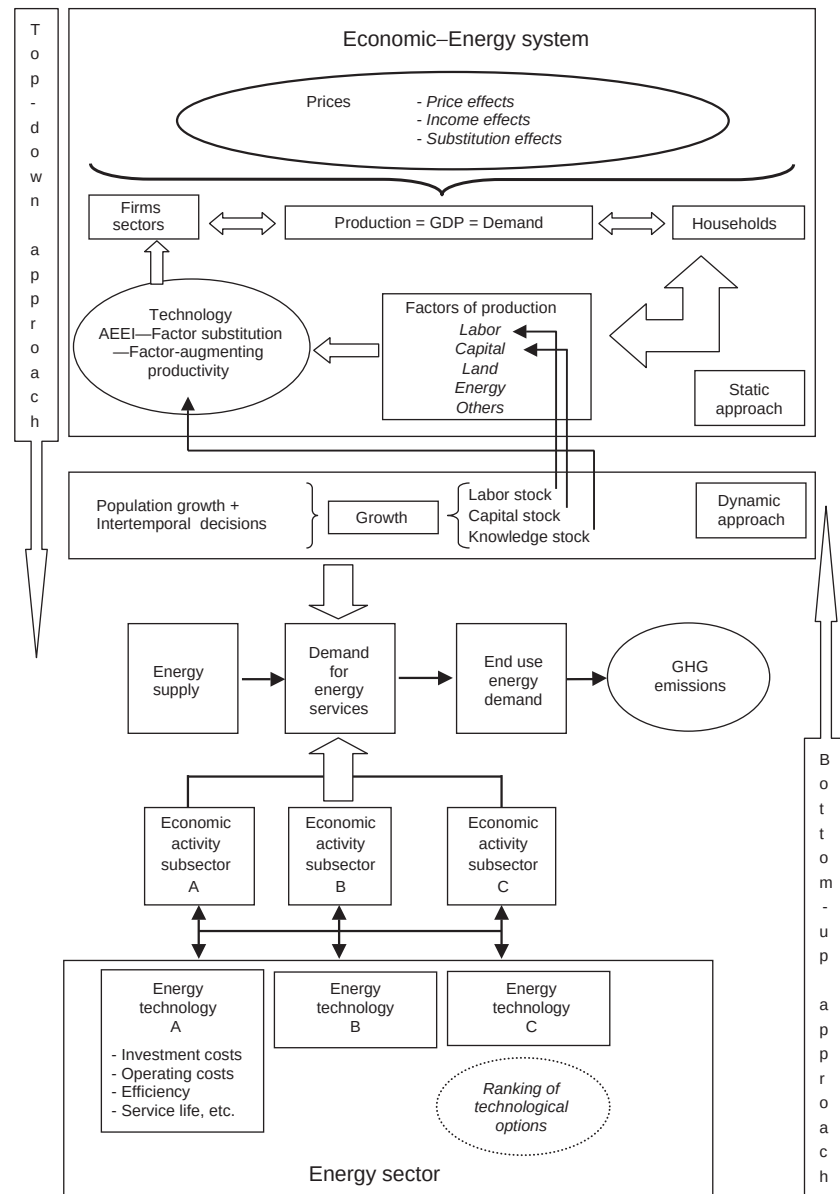


FIGURE 1 Schematic diagram illustrating the structural features of bottom-up and top-down modeling exercises. Top-down approaches are highlighted by the downward arrow, and bottom-up approaches are highlighted by the upward arrow.

A problem can arise when, using both tools to investigate the same issue with the explicit aim of providing policy advice, the two approaches end up with opposite answers. This is the specific case of climate change and climate change policies where both top-down and bottom-up models have been widely used to assess costs and benefits of various control options to decide how much, and according to which strategies (if any), to intervene. Unfortunately, top-down models generally conclude that mitigating global warming entails substantive costs

for the economic systems, whereas bottom-up models demonstrate that potential costs are negligible.

Understanding the reason for this discrepancy offers the opportunity to present and compare the different methodologies. Basically, a top-down model can perform two kinds of investigations that are strongly market oriented: the so-called “if-then” or simulation analysis and the optimization analysis. In the first case, the question answered is of the following type: “what are the economic consequences (in terms of international/intersectoral

trade, GDP, welfare, etc.) of a specific perturbation (e.g., a tax levied on the production or consumption of a given good, a change in factor productivity and endowment) imposed on a system?” or “what is the tax or subsidy required to accomplish a given policy target?” In the second case, the question answered is of the following type: “what is the path of a given control variable (e.g., the tax, the investment, the abatement rate) allowing a given target minimizing cost or maximizing welfare to be reached?” The perspective of top-down models is “wide” or “general”; that is, in describing the effect of the perturbation imposed on the system, they try to describe and measure all of the (macro)economic implications and feedback originated within the system. The technological part is necessarily simple, embodied in a small number of parameters.

Conversely, bottom-up models focus on problem solving and are solution oriented. They identify least energy and/or least cost strategies to providing energy services or to accomplish given GHG emissions reduction targets. Basically, the user of a bottom-up model can specify a given set of constraints (e.g., on technology or on demand but also derived from policy variables as emissions limits or taxes), and the model will produce the least cost “technological” solution that meets the provided set of constraints. The outcome of the procedure allows various future technology configurations, the paths of their adoption, and the optimal allocations of their investment flows to be described and compared. Bottom-up models are focused on the energy sector that they describe in great detail but leave the general macroeconomic picture outside the model.

This illustrates two important differentiations. First, top-down models can consider new technological options only marginally. This implies that top-down models, unlike bottom-up models, observe that an important part of the emissions reduction potential of the energy sector is missing. Second, the global perspective of top-down models takes into account the so-called rebound effects that usually escape from bottom-up analyses. Said simply, under an engineering perspective, the possibility of accessing a new “clean” technology at low cost directly implies a cheap decrease in the polluting potential of an economic sector. In a top-down framework, lower prices imply savings and/or an increase in demand for that technology. Both of these elements—an increase in demand and additional savings that can be expended somewhere—usually lead to an increase in energy consumption or consumption tout court

(“rebound”) that, on its turn, may lead to an increase or a lower decrease in emissions.

Another basic difference is more “philosophical”: top-down models adopt the perspective of markets led by rational agents. The direct consequence is that in a top-down world, new and low-cost technologies are hardly accessible because rational agents should have already exploited all of the best available options. Another consequence is that in top-down models, all physical and technological resources for supplying energy services are constrained to quantities available at the current time. A practical example is given by the previously mentioned treatment of backstop technologies. The possibility of obtaining energy from new sources is indeed accounted for, but notwithstanding technological progress, it is accounted for only at an increasing cost over time. Bottom-up models consider the future as “changeable”; they observe that technological progress can improve supply- and demand-side technologies, whose price can then decline. This is why bottom-up models usually highlight the potential for low-cost or even negative cost energy or emissions-saving opportunities.

In light of this, which approach is more reliable? Can the two approaches be reconciled or integrated? Regarding the first question, it must be recognized that both approaches present strengths and weaknesses. The strength of top-down models is their comprehensiveness; basically, these models point out that the energy sector cannot be considered in isolation from the whole economic system and, more important, that economic elements shaping and shaped by agents’ behavior are very important in determining the technological quality and performance of a production system. The main limitation is top-down models’ inherently “fixed” nature; they can offer valuable insights about the future so long as the relations obtained from the past remain unaltered and can be projected into the future. The main consequence is that a true analysis of impacts of technological innovation over time is precluded because these models implicitly neglect the fact that information, policy, or institutional changes can increase the availability of cost-effective technologies. Conversely, bottom-up models do highlight the possibility of low-cost options but suffer from their technology- and sector-specific focus. This neglects two important facts. On the one hand, even assuming that a technological option were available, it could be neither developed (due to, e.g., the huge investment necessary) nor diffused through the whole production system (due to, e.g., the existence of

transaction costs and other barriers imposed by investment planning). On the other hand, bottom-up models neglect rebounds or feedback that can counterbalance (sometimes dramatically) initial efficiency improvements.

These considerations could also shed some light on the issue of reconciling results. The basic point is to recognize that the two approaches are different. Additional confusion derives from the fact that sometimes the same terminology is used in the two fields but with different meanings. Top-down models are useful for describing macroeconomic (general equilibrium) effects and transmission channels related to taxes, various economic scenarios on energy and environment, and changes in the energy system. The subsequent quantification could be reliable so long as technical progress and cost-effective efficiency opportunities are limited and, more generally, the relationship among energy, economics, and the environment does not change over time. Conversely, bottom-up models are valuable tools for analyzing regulation and energy planning, restructuring of the energy supply sector, and introduction of technological standards and the technological development so as to quantify the aggregated development in energy efficiency, but they cannot be used generally, for example, to provide estimates of the costs of reducing GHGs on a global scale.

4. TOP-DOWN AND BOTTOM-UP MODELS: POSSIBLE INTEGRATION

As pointed out by some authors, it seems not only desirable but also natural to integrate the two approaches. Engineering data on technical potential and costs can offer valuable input to economic market investigations analyzing the effects of a given energy policy, whereas information about the economic intersectoral and international relationships described by top-down models can enrich the cost assessments of bottom-up research with the crucial dimensions of market effects.

Indeed, numerous attempts have been made to coherently and consistently exploit the benefits of both methodologies. One straightforward procedure is to use the results of bottom-up models as an input in macroeconomic models. This is the so-called soft link. An example derives from the MARKAL-MACRO model. MARKAL is a national system optimization model that determines the “best” energy technology solutions and provides marginal

abatement costs, energy prices, and rates of technical change as input for MACRO, a simple production and utility function for the economy.

A “harder” link among bottom-up and top-down rationales is obtained when bottom-up and top-down modules can work interdependently. An example derives from the Hybris model, developed by Jacobsen, in which the macroeconomic top-down model ADAM for Denmark is linked to three different bottom-up energy modules: energy supply, demand for heat, and demand for electricity. Basically, the bottom-up modules aggregate or disaggregate variables to fit the specification of the macro model. The model can run different scenarios for bottom-up and top-down initiatives separately and then compare them to scenarios with “combination of reduction” initiatives to highlight the interdependencies among them.

Another notable example is given by the AIM Asian-Pacific model documented by Fujino and colleagues. A set of bottom-up modules that can reproduce detailed processes of energy consumption, industrial production, land use changes, and waste management, as well as technology development and social demand changes, is linked to top-down modules to compute GHG emissions and their general equilibrium relationship with international trade and via an aggregation computable general equilibrium (CGE) interface to assess country and global sustainable development paths considering the feedback among energy, economics, and the environment.

As claimed by several authors, a soft link methodology suffers from the major limitation of failing an effective reconciliation between the logic of the two approaches. In particular, the separation between technological choices at the bottom-up level and economic choices at the top-down level is not observed in the reality where the two decisions are indeed taken simultaneously. Moreover, the top-down part of the model usually is built following the “neoclassical” investment theory that does not allow distinguishing between investment decisions related to the “quality” of the capital stock (e.g., its energy-saving potential) and those related to the “quantity” of the capital stock (e.g., its productive capacity). The consequence is that energy-saving investment is necessarily ill modeled.

This has led to the development of so-called hard link models that attempt a full fusion between top-down and bottom-up features. One possibility is to add a qualitative dimension to the capital stock. An interesting example derives from the WARM model for the European economy, developed by Carraro and

Galeotti, in which economic agents decide not only the quantity but also the quality of the total capital stock. This is done assuming the existence of two kinds of capital: a polluting one and an environmentally friendly one (with the latter allowing for energy savings). Each year, a new capital vintage enters into use, and its quality is determined by the relative weight of the “dirty” and “clean” components endogenously chosen by agents. This composition determines a quality index of the total capital stock that is crucial in determining energy demand responses to environmental and/or energy policies.

Alternatively, Bohringer proposed a hybrid approach that can be applied to a static CGE framework in which energy sectors are represented by bottom-up activity analysis and the other production sectors are characterized by top-down production functions. What is interesting is that the model is solved as a unique equilibrium. Although the technicality of the argument is beyond the scope of this article, this is intuitively possible by observing that equilibrium conditions stemming from the bottom-up representation of a sector are compatible with the equilibrium conditions implied by the standard Arrow–Debreu framework constituting the basis for CGE static top-down models when they are stated in their most general form. The treatment of this kind of problem requires the use of the so-called complementary format proposed by Cottle and Pang. The model is formulated as a nonlinear system of inequalities satisfying the usual Arrow–Debreu general equilibrium conditions (non-negativity of supply minus demand for every commodity, zero profit condition, and Walras’s Law). In top-down sectors, substitution is described by nested separable constant elasticity of substitution (CES) functions, whereas in bottom-up sectors, it is represented by a bundle of discrete Leontief technologies.

A possible integration procedure in a dynamic environment was instead proposed by Muller. The basic idea is to consider capital goods as heterogeneous, being characterized by both different production capacity and different energy efficiency. The representative firm chooses quality and quantity of new capital equipment simultaneously to minimize present and discounted future costs. Basically, this intertemporal optimization procedure is constrained by two laws of motion: one for capital stock and one for its quality. This last factor cannot be adjusted to the desired level immediately; rather, it can be improved only by replacing old inefficient vintages with new, more energy-efficient equipment. In this way, the model can take into account the double role

of investment. One is the top-down motivation of giving up one unit of consumption today in exchange for increasing productive capacity and consumption tomorrow, and the other is the bottom-up motivation of sustaining higher costs today by buying an efficient capital stock in exchange for gaining energy savings due to this higher efficiency tomorrow.

5. CONCLUDING COMMENTS

A model can be defined as a simplified but sufficiently representative picture of reality. Thus, the very first task of a model is to describe reality—or some aspects of it—that should appear clearer and more understandable with the imposed simplification of the model than without it. In the case of energy modeling, it is unquestionable that both top-down and bottom-up approaches contributed greatly to improving the knowledge of the respective economic and technological dynamics related to the energy system. Nevertheless the different foci of the two methodologies—one economic the other technological—usually leads them to different forecasts and prescriptions. Moreover, it must be recognized that, notwithstanding numerous promising attempts to bridge the gap between the two views, this difference still remains. Under a scientific perspective, this is not dramatic. The process of gaining knowledge takes time, and accordingly, a difference in results can be expected, and the consequent debate can constitute a major driver for scientific improvements. For scientists, it is a strong push to develop finer bottom-up and top-down methodologies and more appropriate integration techniques. On the contrary, under a policy perspective, the difference is problematic. On the one hand, it may lower the trust of policymakers regarding both approaches; on the other hand, it may leave policymakers without a sure framework for their decision making. This issue is very concrete insofar as although a full recognized and accomplished scientific reconciliation is missing, the choice between top-down and bottom-up modeling has to be made on political grounds. What science can do, while working on improvements, is present the results of the two approaches and state clearly under which conditions, and according to which hypotheses, these results could be expected to hold. Symmetrically, policymakers should make clear which vision of the world has endorsed their strategies, in any case bearing in mind that tackling energy policies adopting a bottom-up or top-down angle in isolation is misleading.

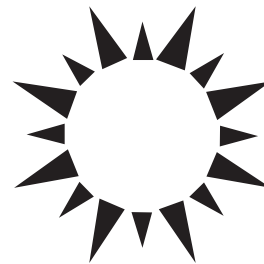
SEE ALSO THE FOLLOWING ARTICLES

Bottom-Up Energy Modeling • Decomposition Analysis Applied to Energy • Depletion and Valuation of Energy Resources • Economics of Energy Demand • Economics of Energy Supply • Fuzzy Logic Modeling of Energy Systems • Input–Output Analysis • Modeling Energy Markets and Climate Change Policy • Multicriteria Analysis of Energy • National Energy Modeling Systems

Further Reading

- Bohringer, C. (1998). The synthesis of bottom-up and top-down in energy policy modeling. *Energy Econ.* 20, 233–248.
- Carraro, C., and Galeotti, M. (1996). WARM: A European model for energy and environmental analysis. *Environ. Modeling Assess.* 1, 171–189.
- Fujino, J., Garg, A., *et al.* (2002). In “Climate Policy Assessment: Asia–Pacific Integrated Modeling” (M. Kainuma, Y. Matsuoka, and T. Morita, Eds.). Springer-Verlag, Tokyo.
- Hertel, T. (1998). “Global Trade Analysis: Modeling and Applications.” Cambridge University Press, Cambridge, MA.
- International Energy Agency. (1998). “The Energy Future: Energy Modeling and Climate Change Policy, Energy, and the Environment.” IEA, Paris.
- Jacobsen, H. K. (1998). Integrating the bottom-up and top-down approach to energy–economy modeling: The case of Denmark. *Energy Econ.* 20, 443–461.
- Manne, A., and Wene, C. O. (1994). MARKAL–MACRO: A linked model for energy–economy analysis. In “Advances in System Analysis: Modeling Energy-Related Emissions on a National and Global Level” (J. Hake, M. Kleemann *et al.*, Eds.). Konferenzen des Forschungszentrums, Jülich, Germany.
- McKibbin, W. J., and Wilcoxon, P. J. (1995). “The Theoretical and Empirical Structure of G-Cubed.” Brookings Institution, Washington, DC.
- Muller, T. (2000). “Integrating Bottom-Up and Top-Down Models for Energy Policy Analysis: A Dynamic Framework.” Centre Universitaire d’études, Université de Geneve, Switzerland.
- Nordhaus, W., and Yang, Z. (1996). A regional dynamic general equilibrium model of alternative climate change strategies. *Am. Econ. Rev.* 86, 726–741.
- Popp, D. (2000). Induced innovation and energy prices. *Am. Econ. Rev.* 92, 160–180.
- Wilson, D., and Swisher, J. (1993). Top-down versus bottom-up analyses of the cost of mitigating global warming. *Energy Policy* 21, 249–263.

Motor Vehicle Use, Social Costs of



MARK A. DELUCCHI

University of California, Davis
Davis, California, United States

-
1. Background
 2. Relevance of Analyses of the Social Cost of Motor Vehicle Use
 3. A Conceptual Framework
 4. Components of the Social Cost of Motor Vehicle Use
 5. Results of an Analysis
 6. Summary

Glossary

annualized (amortized) cost An initial or up-front payment for an asset converted to an economically equivalent stream of regular payments over the life of an asset; conversion depends on the interest rate.

average cost The total cost of a given total quantity divided by the total quantity (e.g., the total cost of 4 billion barrels of oil divided by 4 billion barrels) (cf. marginal cost).

bundled cost A cost, such as the cost of parking, that is not explicitly and separately priced, but rather is included in the cost of other items that together are explicitly priced as a “bundle.”

cost–benefit analysis A method of analysis in which the economic impacts of a proposed plan, policy, or project are identified, quantified, and valued in dollars for the purpose of comparing the total value of the negative impacts (costs) with the positive (benefits).

external cost The incidental cost of an economic transaction to persons who are not formally part of the transaction (see Section 4.5 for a more formal definition).

marginal cost The cost of an additional or incremental unit of consumption (e.g., the cost of the four-billionth barrel of oil) (cf. average cost). The marginal cost will exceed the average cost if the cost of producing each additional unit increases continually because of increasing scarcity of resources.

monetary cost A cost, such as the cost of gasoline, that is expressed in monetary terms (e.g., dollars) in market transactions. A nonmonetary cost is a cost, such as the cost of travel time, that is not expressed directly in money terms.

opportunity cost In economics, the opportunities or alternative resource uses given up as a result of a particular use of a resource.

optimal price A price on a good or service that incorporates all of the marginal costs to society of using the good or service.

private cost The opportunity cost of an action to a private individual, usually as part of a market transaction.

social cost The opportunity cost of an action to society as a whole; generally equal to the private cost plus the external cost.

The social cost of motor vehicle use—the all-inclusive economic cost to society of using motor vehicles—comprises explicitly priced private-sector costs, bundled private-sector costs, public-sector costs, external costs, and personal nonmonetary costs. Estimates of the social cost can be used for a variety of purposes, including analyses of efficient pricing of motor vehicle goods and services, cost–benefit analysis of motor vehicle projects and plans, and general policy analysis related to motor vehicle use. This article focuses on the classification, uses, and magnitude of estimates of the social cost of motor vehicle use.

1. BACKGROUND

Every year, American drivers spend hundreds of billions of dollars on highway transportation. They pay for vehicles, maintenance, repair, fuel, lubricants, tires, parts, insurance, parking, tolls, registration, fees, and other items. These expenditures buy Americans considerable personal mobility and economic productivity. But the use of motor vehicles costs society more than the hundreds of billions of dollars spent on explicitly priced transportation goods and services. There also are bundled costs, i.e., those goods and services that are not explicitly priced, but rather are bundled in the prices of

nontransportation goods and services. For example, “free” parking at a shopping mall is unpriced, but it is not costless; its cost is included—bundled—in the price of goods and services sold at the mall.

In addition to these priced or bundled private-sector costs, there are public-sector costs, of tens of billions of dollars per year, to build and maintain roads and to provide a wide range of services that partly support motor vehicle use. These costs include those for police protection, the judicial and legal systems, corrections, fire protection, environmental regulation, energy research and regulation, military protection of oil supplies, and more. Finally, beyond these monetary public and private-sector costs, are the nonmonetary costs of motor vehicle use, which, by definition, are not valued in dollars in normal market transactions. A wide variety of nonmonetary costs exist, including the health effects of air pollution, pain and suffering due to accidents, and travel time. Some of these nonmonetary costs, such as air pollution, are externalities, whereas others, such as travel time in uncongested conditions, are what may be called “personal nonmonetary costs.”

The social cost of motor vehicle use—the all-inclusive economic cost to society of using motor vehicles—is the sum of all of the costs just mentioned: explicitly priced private-sector costs, bundled private-sector costs, public-sector costs, external costs, and personal nonmonetary costs. These costs are listed in complete detail and classified more rigorously, in Table I. Over the past decade, a number of researchers have been doing detailed and comprehensive analyses of the social cost of motor vehicle use. In the following sections, the purpose of estimating the total social cost of motor vehicle use is explained, a conceptual framework and a cost classification are delineated, and recent cost estimates are presented and discussed.

2. RELEVANCE OF ANALYSES OF THE SOCIAL COST OF MOTOR VEHICLE USE

Researchers have performed social-cost analyses for a variety of reasons and have used them in a variety of ways to support a wide range of policy positions. Some researchers have used social-cost analyses to argue that motor vehicles and gasoline are terrifically underpriced, whereas others have used them to downplay the need for drastic policy intervention in the transportation sector. In any case, social-cost

analyses usually excite considerable interest, if only because nearly everyone uses motor vehicles. By itself, however, a total social-cost analysis does not determine whether motor vehicle use on balance is good or bad, or better or worse than some alternative, or whether it is wise to tax gasoline or restrict automobile use or encourage travel in trains. Rather, a social-cost analysis is but one of many pieces of information that might be useful to transportation policymakers. Specifically, a social-cost analysis can help in the following ways:

1. To establish efficient prices for transportation resources, such as roads or automobile emissions. A social-cost analysis can give some idea of the magnitude of the gap between current prices (which might be zero, as in the case of emissions) and theoretically optimal prices, and can inform discussions of the types of policies that might narrow the gaps and make people use transportation resources more efficiently. However, unless it is done with extraordinary specificity and explicitly with an eye to pricing, a social-cost analysis cannot tell us optimal prices for roads or emissions or anything else.

2. To evaluate all of the costs and benefits of alternative transportation investments in order to find the alternative that will provide the highest net benefit to society. A social-cost analysis, as is presented here, is conceptually one-half of the full social-cost-benefit analysis that must be performed in order to invest society’s resources efficiently. At a minimum, a broad but detailed social-cost analysis can be a source of data and methods for evaluations of specific projects.

3. To prioritize efforts to reduce the costs of transportation. A detailed comparison of costs can help decide how to fund research and development to improve the performance and reduce the costs of transportation. For example, when considering funding research into the sources, effects, and mitigation of pollution, it might be useful to know that emissions of road dust might be an order of magnitude more costly than are emissions of ozone precursors, which in turn might be an order of magnitude more costly than are emissions of toxic air pollutants.

The analysis and estimates are presented here with these relatively modest informational purposes in mind, not to promote any policies regarding motor vehicle use. There is no declamation, for example, about the correct price of gasoline or the correct level of investment in the highway infrastructure.

3. A CONCEPTUAL FRAMEWORK

In speaking of the social cost of motor vehicle use, what is meant is the annualized social cost of motor vehicle use. The annualized cost of motor vehicle use is equal to the sum of periodic or operating costs (such as fuel, vehicle maintenance, highway maintenance, salaries of police officers, travel time, noise, injuries from accidents, and disease from air pollution) plus the value of all capital (such as highways, parking lots, and residential garages; items that provide a stream of services), converted (annualized) into an equivalent stream of annual costs over the life of the capital. This annualization approach essentially is an investment analysis, or project evaluation, in which the project, in this case, is the entire motor vehicle and use system. Of course, it is awkward to treat the entire motor vehicle system—every car, every gallon of gasoline, every mile of highway—as a “project” up for evaluation, but endeavoring to generate data and methods that will be generally useful in analyses of pricing, investment, and research will not avoid the scale and its awkwardness.

3.1 What Counts as a Cost of Motor Vehicle Use or Infrastructure?

In economic analysis, “cost” means “opportunity cost.” The opportunity cost of action A is the opportunity that is foregone—what is given up, or used, or consumed—as a result of doing A. For some resource R to count as a cost of motor vehicle use, it must be true that a change in motor vehicle use will result in a change in use of R. Thus, gasoline is a cost of motor vehicle use because a change in motor vehicle use will result in a change in gasoline use, all else equal. But general spending on national health and education is not a cost of motor vehicle use because a change in motor vehicle use will not result in a change in the resources devoted to national health and education.

For the purposes of planning, evaluating, or pricing, it is of concern not only whether something is a cost of motor vehicle use, but, if it is a cost, exactly how it is related to motor vehicle use. For example, pollution is a direct, immediate cost of motor vehicle use: if motor vehicle use is changed a little, pollution is immediately changed a little. But defense expenditures in the Persian Gulf, if they are a cost of motor vehicle use at all, are an indirect, long-term, and tenuous cost (via congressional decisions about military needs, which might be informed by concerns about the security of oil supply for transportation). This sort of distinction is important

because the more tenuously linked costs are harder to estimate, often delayed considerably with respect to the causal changes in motor vehicle use, and often highly dependent on the specific characteristics and amount of the change in motor vehicle use.

3.2 Costs versus Benefits

In this discussion, the dollar social cost but not the dollar social benefit of motor vehicle use is considered. Of course, it is not forgotten that there are benefits of motor vehicle use (a charge occasionally leveled against social-cost analysts) and there is certainly no presumption here that the benefits somehow are less important than the costs of motor vehicle use. Rather, because there is no credible way to estimate all of the benefits of motor vehicle use, there is no attempt to do so here. The emphasis here, however, is that not only does motor vehicle use provide enormous social benefit, but that this benefit, if it could be expressed in dollars, almost certainly would greatly exceed the full social cost.

Nevertheless, because this is a cost analysis only, it is not possible to say much about net dollar benefits or cost–benefit ratios, or whether a particular transportation system is worthwhile, or better or worse than another system. For example, this analysis indicates that motor vehicle use might cost more than is realized (i.e., that the total social cost appreciably exceeds the commonly recognized private costs, such as for fuel, vehicles, and maintenance and repair). But even if this is so, it does not mean that motor vehicle use costs more than it is worth, or that there should be a preference for any transportation option that might have near-zero external costs, or even any transportation option that might have lower total social costs. To make these evaluations, the dollar value of all the benefits as well as the dollar value of all the costs must be estimated.

3.3 Average Cost as a Proxy for Marginal Cost

To be relevant, a social-cost estimate must apply to a potential (realistic) context or policy. Given this, the question might be asked: Is an estimate of the social cost of all motor vehicle use, which reveals what would be saved without a motor vehicle system and no motor vehicle use at all, of practical interest? Certainly an estimate of the total dollar cost of all motor vehicle use, by itself, is useful only for research purposes, not for evaluating any policy, because a policy requiring the elimination of motor vehicle use

TABLE I

Classification of the Social Costs of Motor Vehicle Use^a

Personal non-monetary costs	Private sector monetary costs			External costs of motor vehicle use		
	Goods and services produced and priced in the private sector (estimated net of producer surplus and taxes and fees)	Goods bundled in the private sector	Goods and services provided by government	Monetary externalities	Nonmonetary externalities	Nonmonetary impacts of the motor vehicle infrastructure ^b
Travel time (excluding travel delay imposed by others) that displaces unpaid activities	Costs usually included in GNP-type accounts: Annualized cost of the fleet (excluding vehicles replaced as a result of motor vehicle accidents) Cost of transactions for used cars	Annualized cost of nonresidential, offstreet parking included in the price of goods and services or offered as an employee benefit Annualized cost of residential, offstreet parking included in the price of housing	Annualized cost of public highways (including on-street parking and off-street private investment) Annualized cost of municipal and institutional off-street parking	Monetary costs of travel delay imposed by others (extra consumption of fuel, and foregone paid work) Accident costs not accounted for by the economically responsible party (property damage, medical, productivity, and legal and administrative costs)	Accidental pain, suffering, death, and lost nonmarket productivity, not accounted for by the economically responsible party Travel delay (imposed by other drivers) that displaces unpaid activities	Land-use damage: habitat, species loss due to highways, MV infrastructure The socially divisive effect of roads as physical barriers in communities
Accidental pain, suffering, death, and lost nonmarket productivity inflicted on oneself	Parts, supplies, maintenance, repair, cleaning, storage, renting, towing, etc. (excluding parts and services in the repair of vehicles damaged in accidents)	Annualized cost of roads provided or paid for by the private sector and recovered in the price of structures, goods, or services	Highway law enforcement and safety Regulation and control of MV air, water, and solid waste pollution MV and fuel technology R&D	Expected loss of GNP due to sudden changes in oil prices Price effect of using petroleum fuels for motor vehicles: increased payments to foreign countries	Air pollution: effects on human health, crops, materials, and visibility ^d Global warming due to fuel cycle emissions of greenhouse gases (U.S. damages only) Noise from motor vehicles Water pollution: effects of leaking	Esthetics of highways and vehicle and service establishments
Personal time spent working on MVs and garages, refueling MVs, and buying and disposing of MVs and parts	Motor fuel and lubricating oil (excluding cost of fuel use attributable to delay) Motor vehicle insurance (administrative and management costs)					

MV noise and air pollution inflicted on oneself	Priced private commercial and residential parking (excluding parking taxes) Costs not included in GNP-type accounts: Travel time (excluding travel delay imposed by others) that displaces paid work Overhead expenses of business and government fleets Private monetary costs of motor vehicle accidents (including user payments for cost of motor vehicle accidents inflicted on others, but excluding insurance administration costs)	for oil used in other sectors (not an external cost internationally) Monetary, non-public-sector costs of fires and net crimes ^c related to using or having MV goods, services, or infrastructure Police protection (excluding highway patrol); court and corrections system (net of cost of substitute crimes) Fire protection Motor-vehicle-related costs of other agencies Military expenditures related to the use of Persian Gulf oil by motor vehicles Annualized cost of the Strategic Petroleum Reserve	storage tanks, oil spills, urban runoff, road deicing Nonmonetary costs of fires and net crimes ^c related to using or having MV goods, services, or infrastructure Air pollution damages to ecosystems other than forests, costs of motor vehicle waste, vibration damages, fear of MVs and MV-related crime

^a Abbreviations: MV, motor vehicle; GNP, gross national product; R&D, research and development.

^b Although these are nonmonetary environmental and social costs of total motor vehicle use, they are not costs of marginal motor vehicle use, and hence technically are not externalities.

^c These should be classified not as external costs, within an economic framework, but rather as costs of illegal or immoral behavior, within a framework that encompasses more than just economic criteria. However, regardless of how these are classified, they in fact are related to using or having motor vehicle goods, services, or infrastructure.

^d The cost of crop loss, and some of the components of other costs of air pollution (e.g., the cost of medical treatment of sickness caused by motor vehicle air pollution), probably should be classified as monetary externalities.

is highly unlikely. However, an estimate of the cost of all motor vehicle use (i.e., as explained previously, an estimate of the annualized cost of the entire system) will be useful as a starting point to the extent that it accurately can be scaled down to an estimate of the cost of a more realistic "project size." That is, if the true cost of a proposal to increase the motor vehicle system and its use by 10% is approximately 10% of the cost of the entire motor vehicle system as estimated here, then the estimate of the cost of the entire system will be a useful starting point for evaluating the proposal.

Do social costs of motor vehicle use in fact scale linearly with output? In economic terms, the question can be stated as follows: Is the total cost function linear, so that marginal (or incremental) cost per unit is constant and equal to average cost per unit (where the average cost per unit is equal to the total cost divided by the total number of units)? In most cases, the answer is "not exactly," because most types of costs probably are not strictly a linear function of output. For example, it is known that the nonmarket costs of air pollution are a nonlinear function of motor vehicle pollution, and that congestion delay costs are a nonlinear function of motor vehicle travel. Still, even though most costs of motor vehicle use are not strictly a continuous linear function of motor vehicle use, down to the mile or gram or decibel or minute, in at least some scenarios of relatively large changes in motor vehicle use, the average cost ratio might be a serviceable approximation of the actual long-run marginal ratio of interest. At a minimum, some of the data and methods used in an analysis of total cost will be useful in an analysis of marginal cost.

3.4 Classification of Components of the Total Social Cost

In any social-cost analysis, the individual cost components, or cost items, should be identified and classified in consonance with how the cost estimates will be used. As discussed previously, estimates of the total social cost of motor vehicle use legitimately can be applied toward three ends: to ensure that motor vehicles are used efficiently, to evaluate transportation investments, and to prioritize efforts to reduce costs. Of these uses, only the first one, efficiency of use, comes with a set of widely accepted organizing principles. That is, when estimating costs in order to help policymakers improve the efficiency of the use of the transportation system, then costs should be organized into groups with similar characteristics with respect to this purpose of understanding the

economic efficiency of motor vehicle use (for example, costs that are priced and efficiently allocated, unpriced but efficiently allocated, or unpriced and inefficiently allocated). Therefore, in Table I, the costs of motor vehicle use are grouped with respect to efficiency of use. However, there is an additional criterion that has nothing to do with efficiency per se: whether a cost is monetary, which means "valued in dollars in real markets." (For example, gasoline and parking are monetary costs, but air pollution is nonmonetary cost.) The distinction between monetary and nonmonetary costs is important methodologically, because it is much harder to estimate the nonmonetary costs. These distinctions result in the six categories of Table I. These are reviewed briefly next.

4. COMPONENTS OF THE SOCIAL COST OF MOTOR VEHICLE USE

The discussions that follow are based on the classification set forth in Table I.

4.1 Column 1: Personal Nonmonetary Costs

Personal nonmonetary costs are those unpriced costs of motor vehicle use that a person self-imposes as a result of the decision to travel. The largest personal costs of motor vehicle use are personal travel time in uncongested conditions and the risk of getting into an accident that involves nobody else. With respect to economic efficiency, the particular issue in the case of personal nonmonetary costs is whether drivers fully recognize the personal nonmarket costs that they face. If a person does not correctly assess these costs (that is, if the true value to the user does not equal the true cost to the user), then the person will drive more or less than would be the case if he or she were fully informed and rational. For example, people who, on account of ignorance or poor judgment, underestimate their risk of falling asleep at the wheel, once in a while will make trips for which the real but underestimated risk-cost exceeds the value, and which, consequently, in principle, should not be made.

4.2 Column 2: Priced Private-Sector Motor Vehicle Goods and Services, Net of Producer Surplus and Taxes and Fees

Priced private-sector costs are related to goods and services that are explicitly priced in private markets:

motor vehicles, motor fuel, maintenance and repair, insurance, and so on. Because these goods and services are bought and sold by private individuals, they are the most familiar components of the total social cost. Also, a portion of these goods and services constitute the “transportation” subaccounts of the gross national product (GNP) accounting.

The social cost of motor vehicle goods and services supplied in private markets can be calculated by starting with known prices and quantities sold, but is not equal to prices multiplied by quantities. Rather, the social (economic) cost is the area under what is called the private supply curve: the dollar value of the resources that a private market allocates to supplying vehicles, fuel, parts, insurance, and so on. To estimate this area, two items must be subtracted from total price-times-quantity revenues, i.e., producer surplus, and taxes and fees. (It is appropriate to start with revenues because prices and quantities can be observed, but not supply curves.) The producer surplus is deducted because it is defined as revenue (“profit,” in lay terms) in excess of economic cost, and hence is a noncost wealth transfer from consumers to producers. Taxes and fees assessed on producers and consumers are deducted because they either are transfers from producers and consumers to government, or else are economically inefficient government charges for government services.

The deduction of producer surplus is not merely a theoretical aside: it bears directly on comparisons of alternatives. For example, in comparing the cost of motor fuel with the cost of alternative energy sources, it will not do to count the revenues received by oil producers as the cost, because a large portion of the revenues received by oil producers is extra “profit,” far in excess of costs and “normal” profit (which is why major oil-producing countries are so wealthy). It also is important to note that the prices and quantities are realized in private markets rarely if ever are optimal in economic terms, not only because of taxes and fees, which cause economic distortions, but because of imperfect competition (e.g., monopoly), standards and regulations that affect production and consumption, externalities (discussed later), and poor information. Put another way, there is no simple dichotomous world in which private-sector prices are perfect and can be left alone and all other prices (or nonprices) need to be fixed. Rather, there are a variety of market imperfections in every sector of the economy, including the most competitive, unregulated private sectors, and a corresponding range of issues pertaining to pricing, taxation, regulation, and so on. In some cases, there may be

as much concern about the market price of tires as there is about the best way to charge motor vehicle users for roads or about the complete lack of a price on motor vehicle emissions.

4.3 Column 3: Bundled Private-Sector Costs

Some very large costs of motor vehicle use are not explicitly priced as separate costs of motor vehicle use. Foremost among these are the cost of free, nonresidential parking, the cost of home garages, and the cost of local roads provided by private developers. However, all of these costs are included in the price of “packages” (such as houses and goods) that are explicitly priced. This bundling is not necessarily economically inefficient; in principle, a producer will bundle a cost and not price it separately if the administrative, operational, and customer (or employee) costs of collecting a separate price exceed the benefits. In a perfect market, it can be presumed that any observed bundling is economically efficient and that it would be economically undesirable to actually mandate unbundling. As concerns economic efficiency, then, the question is whether taxes or regulations (such as requirements that office buildings have a certain number of parking spaces, regardless of what the building owners think) and the like distort the decision to bundle and whether a supplier is correct in his assessment of the costs and benefits of bundling. To the extent that taxes and standards are distorting the market for parking, the ideal remedy is to eliminate the inefficient taxes and standards, not to force parking costs to be unbundled.

4.4 Column 4: Motor Vehicle Goods and Services Provided by the Public Sector

Government provides a wide range of infrastructure and services in support of motor vehicle use. The most costly of these is the capital of the highway infrastructure. Government costs are categorized separately because generally they either are not priced, or else are priced but not in the most efficient manner. Note that, whereas all government expenditures on highways and the highway patrol are a cost of motor vehicle use, only a portion of total government expenditures on local police, fire, corrections, jails, and so on is a cost of motor vehicle use. The portion of these expenditures that is a cost—a long-run or average cost, anyway—of motor

vehicle use has been estimated. This sort of allocation is valid for expenditures (such as for police protection) that arguably are economic costs of motor vehicle use, in the sense of “cost” as used previously. (For example, motor vehicle use results in motor vehicle-related crime, which consumes police-protection services, and motor vehicle fires, which consume fire-protection services.) Publicly provided goods and services that are not priced at all are arguably also monetary externalities, which are shown in column 5 of Table I and are discussed next. Those items that might be both a publicly provided good or service and a monetary externality are shown to straddle columns 4 and 5 in Table I.

4.5 Columns 5–7: Monetary and Nonmonetary Externalities

An externality is a cost or benefit that an action of person A imposes on person B, but that person A does not account for in his or her deliberations regarding the utility (usefulness) of the action. Environmental pollution, traffic delay, pain and suffering due to accidents, and the loss of GNP due to sudden changes in the price of oil are common examples of externalities. Some analysts distinguish “monetary externalities,” which are valued in dollars in markets, from “nonmonetary externalities,” which are not valued in any markets, even indirectly. Although monetary external costs are valued in markets, they are completely unpriced from the perspective of the responsible motor vehicle user and hence are externalities. The clearest example, shown in column 5 of Table I, is accident costs that are paid for in dollars by those who are not responsible for the accident. For example, vehicular repair costs inflicted by uninsured motorists are valued explicitly in dollars in private markets, but are unpriced from the perspective of the uninsured motorist responsible for the accident.

With respect to economic efficiency, the concern here, as with any external cost, is that the costs are not priced at all, and hence are associated with consumption that is larger than is socially desirable. The intuition behind this is straightforward: if people pay a price that is less than the full cost to society, they will consume more than they would if the price were raised to reflect the external costs. The largest monetary externalities (and hence the largest potential sources of economic inefficiency) are those resulting from motor vehicle accidents and congestion on the highways.

Most environmental damages are nonmonetary externalities. Environmental costs include those related to air pollution, global warming, water pollution, and noise due to motor vehicles. To estimate these costs, complex physical processes and biological responses must be modeled and the dollar value of the responses must then be estimated. (The valuation step, which often is quite difficult, is required for nonmonetary but not for monetary externalities; hence the distinction between monetary and nonmonetary costs.) By far the largest environmental externality of motor vehicle use is the cost of particulate air pollution. Interestingly, a typically overlooked and completely unregulated emissions source, particulate matter kicked up from the road bed by passing vehicles, may be one of the largest sources of pollution damages—much larger than damage from ozone.

5. RESULTS OF AN ANALYSIS

The results of a comprehensive analysis of most of the costs of Table I are summarized by aggregate cost category in Table II. Note that the aggregated totals are shown here in order to provide a sense of magnitudes, not because such aggregated totals are inherently useful. Indeed, as discussed next, care must be taken to avoid misusing estimates of the total social cost of motor vehicle use. Table II also shows two subtotals of interest: all monetary costs (those that are valued in dollars in markets) and costs that normally are included in GNP-type accounts of the economic value of transportation. Costs normally included in GNP-type accounts are of interest because they are the explicitly priced private-sector costs and hence represent what most people readily identify as “costs of motor vehicle use.” It can be seen that these are a small fraction of the total social costs of motor vehicle use.

An estimate of user payments for public highway and services is shown at the bottom of Table II. An important caveat regarding the use of this estimate is discussed in the next section.

5.1 Allocation of Costs to Individual Vehicle Categories

All of the costs shown in Table II pertain to all motor vehicles (autos, trucks, and buses). Although it can be interesting to estimate the cost of all motor vehicle use, it typically will be more useful to estimate the

TABLE II

Summary of the Social Costs of Motor Vehicle Use in the United States, 1990–1991

Cost item	Total cost (10 ⁹ \$)		Percentage of total	
	Low	High	Low	High
Personal nonmonetary costs of motor vehicle use	527	968	32	29
Motor vehicle goods and services produced and priced in the private sector (estimated net of producer surplus, taxes, and fees)	827	980	49	30
Motor vehicle goods and services bundled in the private sector	76	279	5	8
Motor vehicle infrastructure and services provided by the public sector ^a	131	247	8	7
Monetary externalities of motor vehicle use	43	104	3	3
Nonmonetary externalities of motor vehicle use ^b	68	730	4	22
Grand total social cost of highway transportation	1673	3308	100	100
Subtotal: costs of usually included in GNP-type accounts	559	674	33	20
(2) + (3) + (5) subtotal: monetary cost only	1077	1610	64	49
Item: payments by motor vehicle users for the use of public infrastructure and services	114	206		

^aIncludes items in Table I that straddle columns 4 and 5.^bIncludes motor vehicle noise and air pollution inflicted on oneself (from column 1 of Table I). Excludes nonmonetary impacts of the motor vehicle infrastructure (column 7 in Table I) and air pollution damages to ecosystems other than forests, costs of motor vehicle waste, vibration damages, and fear of motor vehicles and motor-vehicle-related crime (last item in column 6 of Table I).

cost of different classes of vehicles or of different fuel types, because analysts, policymakers, and regulators typically are interested in specific classes of vehicles and specific fuels, rather than all motor vehicles as a group. (For example, pollution regulations are set for individual classes of vehicles, not for all motor vehicles as a class.) For some cost items, such as the some of the costs of air pollution, analysts have estimated marginal costs as a function of vehicle type and other specific characteristics of motor vehicles and their use. For example, noise costs are estimated as a function of vehicle type, vehicle speed, distance from the road, the presence of sound barriers, housing values, and many other parameters. Air pollution costs are estimated as a function of vehicle characteristics, urban population, weather variables, and other parameters. Road maintenance and repair costs are estimated as a function of vehicle weight, traffic volume, and other factors.

If marginal cost functions are not available, total costs can be allocated to particular vehicle classes or uses on the basis of general cost-allocation factors. A cost-allocation factor shows the share of a particular vehicle class of some general measure of motor vehicle use. For example, it shows the share of light-duty gasoline autos of total vehicle miles of travel, or the share of heavy-duty diesel vehicles of total motor vehicle expenditures for maintenance

and repair. The use of these allocation factors is straightforward. For example, the heavy-duty diesel vehicle fraction of total vehicle ton-miles per axle, multiplied by any total motor vehicle cost that is a function of vehicle ton-miles per axle, yields the amount of that cost that is assignable to heavy-duty diesel vehicles. Thus, if total expenditures for highway repair are known, and it is believed that highway repair costs are related to ton-miles of travel per axle, the ton-mile/axle allocation factors can be used to allocate the total expenditures to individual vehicle classes.

5.2 How the Results of a Social-Cost Analysis Should Not Be Used

The legitimate uses of estimates were discussed in previous section. Caution against several common misuses of these estimates is necessary, however. First, the temptation to add up all of the imperfectly priced costs and to express the total per gallon of gasoline, as if the optimal strategy for remedying every inefficiency of motor vehicle use was simply to raise the tax on gasoline, should be resisted. The economically optimal strategy for dealing with imperfect prices is considerably more complex than this. In the first place, some sources of economic

inefficiency, such as imperfect competition and a distortionary income tax policy, are not externalities, and hence are not properly addressed by taxation at all. Taxation can be an appropriate remedy for external costs, but it is a particular kind of taxation that is appropriate (one in which the tax is equal to the actual marginal external cost), and it turns out that there is not a single external cost, with the possible exception of CO₂ emissions from vehicles, that in principle is best addressed by a gasoline tax. For example, an optimal air pollution tax would be a function of the amounts and types of emissions, the ambient conditions, and the size of the exposed population; it would not be proportional to gasoline consumption.

Second, it might be misleading to compare the total social cost of motor vehicle use with the GNP of the United States, because the GNP accounting is quite different from and generally more restricted than social-cost accounting. Most importantly, the GNP does not include nonmarket items. Third, the considerable uncertainty in most social-cost estimates should be represented properly. With regard to the estimates of Table II, the uncertainty actually is greater than is implied by the low and high ranges, even where the high is much higher than the low, because the estimates shown do not include every conceivable component or effect of every cost, and do not always accommodate the entire span of data or opinions in the literature. (Also, the costs "not estimated" in column 6 of Table I should not be assumed to be trivial.) Fourth, although Table II shows highway user tax and fee payments and the government-provided motor vehicle goods and services to which the user payments nominally apply, it may not be economically meaningful to compare the user payments with the government expenditures. Most emphatically, it is not true that any difference between user payments and government expenditures is a source of economic inefficiency that must be eliminated in order to have efficient resource use. This is because economic efficiency does not require that the government collect from users revenues sufficient to cover costs; rather, efficiency results from following certain pricing rules, which when applied to goods such as highways need not result in user revenues covering costs. An efficient pricing scheme for government goods and services would look nothing like the present tax and fee system. Consequently, a comparison of current user tax and fee payments with current government expenditures reveals little, if anything, quantitative about economic efficiency. However, such comparisons may be of

use in analyses of equity (questions of who pays for what), which is not addressed here.

Finally, estimates of *the total social cost* of motor vehicle use may be of use in the analysis of a particular policy or investment decision only if it is believed that the marginal costs associated with the particular policy or decision are reasonably close to average costs, which can be calculated easily from total costs. In this respect, average cost estimates derived from the results of Table II will be less and less applicable as times and places increasingly different from the United States in 1990 and 1991 (the basis of the estimates of Table II) are considered. However, even if total-cost estimates per se are irrelevant, the underlying data, methods, and concepts (which are presented in the sources listed in the reading list at the end of this article) might be useful in an analysis of specific pricing policies or investments.

6. SUMMARY

The social costs of motor vehicle use in the United States have been classified, discussed, and estimated. The analyses, which present some first-cut estimates of some of the costs, are meant to inform general decisions about pricing, investment, and research, providing a conceptual framework for analyzing social costs and developing analytical methods and data sources.

It should be clear that a social-cost analysis cannot describe precisely what should be done to improve the transportation system in the United States. There are several kinds of inefficiencies in the motor vehicle system, and hence several kinds of economically optimal measures. The magnitude of these inefficiencies is difficult to estimate accurately. Moreover, society cares at least as much about equity, opportunity, and justice as it does about efficiency. In sum, a total social-cost analysis contributes only modestly to one of several societal objectives regarding transportation.

SEE ALSO THE FOLLOWING ARTICLES

Bicycling • Cost-Benefit Analysis Applied to Energy • External Costs of Energy • Fuel Economy Initiatives: International Comparisons • Internal Combustion Engine Vehicles • Leisure, Energy Costs of •

*Lifestyles and Energy • Material Use in Automobiles
• Suburbanization and Energy • Transportation Fuel
Alternatives for Highway Vehicles*

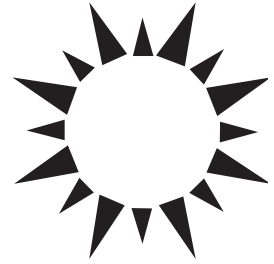
Further Reading

- Baumol, W. J., and Oates, W. E. (1988). "The Theory of Environmental Policy," 2nd ed. Cambridge University Press, New York.
- Delucchi, M. A. (1998). "The Annualized Social Cost of Motor-Vehicle Use, 1990–1991: Summary of Theory, Methods, Data, and Results." UCD-ITS-RR-96-3 (1). Institute of Transportation Studies, University of California, Davis, California.
- Friedrich, R., and Bickel, P. (eds.). (2001). "Environmental External Costs of Transport." Springer-Verlag, Stuttgart, Germany.
- Greene, D. L., Jones, D. W., and Delucchi, M. A. (eds.). (1997). "Measuring the Full Social Costs and Benefits of Transportation." Springer-Verlag, Heidelberg, Germany.
- Hohmeyer, O., Ottinger, R. L., and Rennings, K. (eds.). (1996). "Social Costs and Sustainability, Valuation and Implementation in the Energy and Transport Sector." Springer-Verlag, Berlin, Germany.
- Hohmeyer, O., Ottinger, R. L., and Rennings, K. (eds.). (1996). "Social Costs and Sustainability, Valuation and Implementation in the Energy and Transport Sector." Springer-Verlag, Berlin, Germany.
- Murphy, J. J., and Delucchi, M. A. (1998). A review of the literature on the social cost of motor-vehicle use in the united states. *J. Transport. Statistics* 1(1), 15–42.

Multicriteria Analysis of Energy

R. RAMANATHAN

Sultan Qaboos University
Muscat, Sultanate of Oman



-
1. Introduction
 2. Why Is Multicriteria Analysis Necessary for Energy Issues?
 3. MCDM: Terminologies, Classification, and Some General Characteristics
 4. An Overview of Some MCDM Methods and Their Energy Applications

Glossary

alternative One of several things or courses of action to be chosen by the decision maker. It is the most fundamental entity in a multicriteria decision-making model. Alternatives are also called solutions, especially when dealing with continuous variables, in the mathematical programming context.

attribute A surrogate measure of performance used to represent a criterion. Usually quantified, using some measurable unit, to identify the consequences arising from implementation of any particular decision alternative. Thus, whereas warmth is a criterion, temperature, measured in a suitable (say Celsius or Fahrenheit) scale, is an attribute.

criterion A tool allowing comparison of alternatives according to a particular significance axis or point of view.

decision maker (DM) The individual whose subjective opinions are considered in arriving at a solution for a decision problem.

decision problem A decision-making problem is characterized by the need to choose one or a few from among a number of alternatives.

group decision making (GDM) A situation in which the opinions of more than one decision maker need to be considered in arriving at a solution for a decision problem.

multiattribute decision making (MADM) Involves cases in which the set of alternatives is defined explicitly by a finite list from which one or a few alternatives should be chosen that reflect the DM's preference structure.

multicriteria decision making (MCDM) Making decisions in the face of multiple conflicting criteria. MCDM situations are also referred to herein as multicriteria analysis (MCA) decision situations.

multiobjective decision making (MODM) Involves cases in which the set of alternatives is defined implicitly by a mathematical programming structure with objective functions. Such alternatives are usually defined in terms of continuous variables, which results in an infinite number of alternatives. Multiobjective decision making is also referred to in the literature as multiobjective mathematical programming, multiobjective optimization, vector optimization, or simply multiobjective programming.

nondominated alternative An alternative that is clearly superior to others with respect to at least one criterion. The nondominated alternative is also called an efficient alternative, a noninferior alternative, or a Pareto-optimal alternative.

objective Usually represents the direction of improvement of the attributes; mainly used in mathematical programming problems. A maximizing objective refers to the case in which "more is better," whereas a minimizing objective refers to the case in which "less is better." For example, profit is an attribute, whereas maximizing profit is an objective.

preemptive priorities A situation in which it is assumed that a higher ranked objective is infinitely more important than any of the lower ranked objectives.

stakeholders People who affect and are affected by the decision under consideration.

The field of energy includes a number of issues that require analysis considering several criteria. For example, decisions on almost all energy projects require consideration of several, mostly conflicting, criteria related to economic development arising out of increased energy use, and environmental degradation arising out of increased energy production. Accordingly, multicriteria analysis has been applied to a number of energy-related problems.

1. INTRODUCTION

Multicriteria analysis (MCA), or multicriteria decision making (MCDM) as it is more often called, is a subfield of operations research. It is a special case of the so-called decision-making problems. A decision-making problem is characterized by the need to choose one or a few from among a number of alternatives. The person who is to choose the alternatives is normally called the decision maker (DM). The preferences of the DM will have to be considered in choosing the right alternative(s). In MCDM, the decision maker chooses the most preferred alternative(s) on the basis of two or more criteria.

2. WHY IS MULTICRITERIA ANALYSIS NECESSARY FOR ENERGY ISSUES?

Actually, single-objective programming can be considered as a specific case of MCA. Though some problems have been considered with a single criterion (which is captured by the single-objective function of the linear programming models), many of the issues often have impact on more than one criterion. When it can be safely assumed that only one criterion is overridingly important, linear programming approaches that use the single-objective function should be employed. However, if impacts should be considered in terms of several criteria, the MCA methods should be employed. Further, note that MCA is much more general, involving not only multiobjective programming models, but also non-programming-based models.

2.1 Need for Multiple Criteria

The issues of the energy sector that can be tackled by quantitative methods (including single-criterion linear programming and MCA methods) abound, from the location of power plants, to transmission and distribution systems, to the choice of demand-side management programs, to the choice of options that mitigate global warming. Financial criteria are the most dominant considerations usually evaluated in single-criterion analysis of energy issues. However, the growing importance of environmental impacts of decisions in the field of energy has increased the importance assigned to environment-related criteria such as air pollution (emission of sulfur dioxide, nitrogen oxides, etc.), water pollution, human

health, global warming, acid rain, oil spills, preservation of biodiversity, etc., increasing the need for multicriteria analysis of energy-related decision problems. Of course, other criteria, such as quality and reliability, are also relevant for specific problems.

Multicriteria analysis of energy problems encompasses technological, economic, social, environmental, risk, and financial criteria in making energy decisions. Multicriteria decisions require certain approaches that are not normally used in single-criterion decision-making situations. For example, it is necessary to carry out a trade-off analysis when distinctive characteristics in terms of cost, environmental emissions, human health, etc. are compared and weighed against each other. These comparisons equate to comparing apples and oranges, but are the necessary ingredient of any multicriteria analysis. The trade-off analysis is necessary only when the criteria employed are conflicting, i.e., when there is no alternative that performs well in terms of all the criteria. In some situations, such trade-offs may not be necessary. For example, in the power sector, there may be alternatives that have beneficial impacts on both environmental and economic objectives—most energy-efficient investments that are economically justifiable also bring about a reduction in emissions, and thus they score well in terms of both the criteria, i.e., environmental quality and economic efficiency. The two criteria are said to be not conflicting in this case. However, in most of the energy-related problems, alternatives are often evaluated in terms of conflicting criteria. For example, wind power plants score well in terms of environmental criteria but are more expensive than many other power generation options.

2.2 Group Decision Making

Another feature of multicriteria analysis is group decision making. Invariably, multiple criteria have to be considered to arrive at a decision, because different criteria are important to different people, who both affect and are affected by the decision under consideration. These people are normally referred to as the stakeholders. There may be different groups of stakeholders for a decision-making situation. For example, government and local authorities are normally stakeholders for most energy-related decisions because they would like to ensure that the resulting decision follows legal guidelines and is socially acceptable. A company, entrepreneurs, or potential investors are also stakeholders; here the greater interest is in the economic

and finance-related criteria. For a utility company, for example, the decision made about the problem of siting a power plant would have an economic impact. Environmentalists, public pressure groups (such as nongovernmental organizations and local media), and the local public, also stakeholders, may wish to ensure that the environment is not adversely affected. In some cases, opinions of experts and academia may also be important. Hence, involvement of all the stakeholders is important in arriving at a large-scale energy-related decision. And, the opinion of a stakeholder group should be synthesized from the opinions expressed by individual members of the group (such as individual local people, when the stakeholder group is the local public). MCA has been considered a valuable tool for facilitating early involvement of all stakeholders in the decision-making process, and enhances the fairness and transparency of the procedures for complex decision situations, including energy-related decision situations.

Ideally, the final decision of a stakeholder group should be arrived at by consensus. However, due to the conflicting nature of the criteria adopted by different people (or groups of people), especially in energy-related decision situations, it may not be possible to arrive at a consensus decision. In such cases, special procedures are required to estimate the most preferred solution for the group as a whole, from the opinions expressed by the group members. Most of the multicriteria methods provide such special procedures. Due to the need to consider the subjective opinions of a number of people, most of the MCA methods have been designed to be interactive.

3. MCDM: TERMINOLOGIES, CLASSIFICATION, AND SOME GENERAL CHARACTERISTICS

The field of multicriteria decision making has been succinctly defined as making decisions in the face of multiple conflicting criteria. Some special issues of journals have been devoted to the field of MCDM, including *Management Science* (Vol. 30, No. 1, 1984), *Interfaces* (Vol. 22, No. 6, 1992) (devoted to decision and risk analysis), and *Computers and Operations Research* (Vol. 19, No. 7, 1994). The *Journal of Multi-Criteria Decision Analysis*, since 1992, has published articles entirely devoted to MCDM. Special issues of the *Journal of the Opera-*

tional Research Society (April, 1982), and of *Interfaces* (November–December, 1991) provide a range of applications. Issue No. 2, Volume 133 (January 2001) of the *European Journal of Operational Research* is a special issue on goal programming.

3.1 MCDM Terminologies

Several terminologies are normally used when dealing with a decision problem that has multiple criteria. The terms “goals,” “objectives,” “criteria,” and “attributes” are commonly found in the MCDM literature and can be used with interchangeable ease. The general meaning of these words is similar in most cases.

Alternatives are the most fundamental entities in a MCDM model. They are normally compared with each other in terms of criteria. Identification of criteria for a particular problem is subjective, i.e., varies for each problem. Criteria are normally developed in a hierarchical fashion, starting from the broadest sense (usually called the goal of the problem) and refined into more and more precise sub-goals and sub-sub-goals. In general, some rules should be followed in identifying criteria for any decision problem. They have to be mutually exclusive or independent, collectively exhaustive, and should have operational clarity of definition.

The criteria of a decision problem are usually very general, abstract, and often ambiguous, and it can be impossible to associate criteria directly with alternatives. Attributes are objective and measurable features of the alternatives, and are sometimes used to describe criteria. Objectives, used mainly in mathematical programming problems, represent directions of improvement of the attributes. The term “criterion” is a general term comprising the concepts of attributes and objectives. It can represent either attribute or objective, depending on the nature of the problem. Due to this, MCDM is considered to encompass two distinct fields, namely multiattribute decision making and multiobjective decision making. These fields are discussed in the next section.

3.2 A Classification of Different MCDM Approaches

The broad area of MCDM can be divided into two general categories, multiattribute decision making (MADM) and multiobjective decision making (MODM). MADM involves cases in which the set of alternatives is defined explicitly by a finite list

from which one or a few alternatives should be chosen that reflect the decision maker's preference structure. MODM involves cases in which the set of alternatives is defined implicitly by a mathematical programming structure with objective functions. Such alternatives are usually defined in terms of continuous variables, which results in an infinite number of alternatives.

3.3 Multiobjective Decision Making—Some Basic Concepts

In this section, a comparison of a multiobjective decision making problem with a single-objective decision-making problem is used to introduce certain key concepts frequently used in MCDM literature. Note that many of the concepts introduced in this section are also applicable to MADM problems.

A single-objective decision-making (SODM) problem is generally written in the following form:

$$\begin{aligned} &\text{maximize/minimize } Z = CX \\ &\text{subject to } AX = b, X \geq 0, \end{aligned} \quad (1)$$

where Z is the objective function. The set of equations, $AX = b, X \geq 0$, represents the constraint set of the problem. Methods for solving single-objective mathematical programming problems have been studied extensively for the past 40 years. However, almost every important real-world problem involves more than one objective. A general multiobjective decision-making (MODM) problem has the following form:

$$\begin{aligned} &\text{maximize/minimize } Z_1 = C_1X \\ &\text{maximize/minimize } Z_2 = C_2X \\ &\text{maximize/minimize } Z_n = C_nX \\ &\quad \vdots \\ &\text{subject to } AX = b, X \geq 0, \end{aligned} \quad (2)$$

where the values $Z_i, i = 1, 2, \dots, n$, represent the n objective functions. The set of equations, $AX = b, X \geq 0$, represents the constraint set of the MODM problem. Thus the MODM problem is similar to a SODM problem, except that it has a stack of objective functions instead of only one.

Let us consider a simplified example involving choice of a coal technology for electric power

generation. Let the objectives be as follows:

Minimize Z_c = investment cost.

Maximize Z_G = savings of emissions of
global pollutants
(CO₂ here)

Maximize Z_L = savings of emissions of
local pollutants
(SO₂ here)

Subject to some constraints. (3)

Let us assume that we have some technologies (normally called alternatives of the decision problem), say pulverized fuel (PULV) technology, pressurized fluidized bed combustor (PFBC) technology, and integrated coal gasification combined cycle (IGCC) technology, that satisfy the constraints. This means that these are the three feasible technologies (or alternatives). The performance of the technologies in terms of the three criteria is shown in Table I. Note that the savings of emissions of CO₂ and SO₂ are lower and investment costs higher for PFBC technology compared to IGCC technology. Thus PFBC technology fares, on all the three objectives, worse than IGCC technology and hence it should not be considered any more. In MCDM terminology, PFBC technology is said to be inferior to (or dominated by) IGCC technology. Technologies PULV and IGCC are incomparable, as neither of them is at least as good as the other in terms of all the objectives. Although PULV is cheaper in terms of the investment cost compared to IGCC, it results in smaller savings of the environmental emissions. Hence, PULV and IGCC are said to be noninferior technologies and PFBC is called an inferior technology. (Please note that, in practice, comparisons of the

TABLE I

Some Characteristics of Coal Technologies for Power Generation in India^a

Technology	CO ₂ savings (g/kWh)	SO ₂ savings (g/kWh)	Investment cost (\$/kW)
Pulverized fuel (PULV)	220	3.30	1202
Pressurized fluidized bed combustor (PFBC)	240	3.32	1894
Integrated coal gasification combined cycle (IGCC)	590	3.75	1578

^aValues estimated by the author from miscellaneous sources.

different coal technologies will not be so simple. For example, PFBC technology may fare better than other technologies if some other criteria are considered.) Thus a noninferior alternative is one that is clearly superior with respect to at least one criterion. The noninferior alternative is also called an efficient alternative, nondominated alternative, or Pareto-optimal alternative. In any MODM exercise, the first task is to identify the noninferior options. Note that there may be more than one noninferior option for any multiobjective programming (MOP) problem.

Once the noninferior alternatives are identified, it is necessary to evaluate them so as to arrive at the best alternative. Because none of the noninferior alternatives will optimize (i.e., maximize/minimize) all the objective functions, it is necessary to identify those alternatives that best satisfy all the objective functions (in the opinion of the decision-maker). This has to be contrasted with single-objective optimization problems that are concerned with arriving at one or more optimal solutions. Thus, although the underlying philosophy in SODM problems is in optimizing solutions, MODM aims at identifying one or more satisfying solutions. One way of reducing the number of Pareto-optimal alternatives is by providing performance targets for each of the objectives. For example, if the DM specifies no interest in those technologies that cost more than \$1500/kW, then IGCC technology is automatically excluded from the DM preference set, and PULV technology now remains the only choice. This is a satisfying solution because it satisfies all the performance targets, but obviously it need not be an optimal solution. This is the basic idea of goal programming (discussed later). In general, identification of alternatives (technologies here) that are considered best by the decision maker requires additional information on the preference structure among the objectives, which is not necessary for SODM problems.

The preference structure can be ordinal or cardinal. This essentially involves identifying the "importance parameters" or "weights" associated with the objectives. Suppose it is possible for the DM to provide weights (w) to the three objectives as follows: $w_C = 5$, $w_S = 2$, $w_I = 8$, where the subscripts C, S, and I represent CO₂ savings, SO₂ savings, and investment cost, respectively. These weights may be specified exogenously before attempting to analyze the MOP problem. They imply the trade-offs between the objectives that the DM is willing to consider. It is necessary to have a trade-off because the objectives are conflicting: given the nondomi-

nated options, it is not possible to maximize one objective without getting reduction in the performance of other objective. The information as to how much of reduction in one objective can be tolerated to improve performance in another objective is called the trade-off information. Note that this trade-off information is subjective and depends on the particular DM for whom the MODM problem is being analyzed. The given structure means, in the opinion of the DM, that the objective Z_S (SO₂ savings) is the least important objective, a reduction in investment cost by \$1/kW is considered by the DM to be equivalent to an increase of $8/2 = 4$ g/kWh in SO₂ savings, an increase in CO₂ savings by 1 g/kWh is considered by the DM to be equivalent to an increase of 2.5 g/kWh in SO₂ savings, etc.

Given such cardinal weights, the MODM problem reduces to the SODM problem, with the single objective of minimizing $Z = \sum w_i Z_i$ subject to the constraints. In this case, IGCC technology has an aggregate score of about 15,582, and is preferable to PULV (which has an aggregate score of about 10,723). Thus, the MODM can be reduced to a SODM problem if such cardinal weights can be unambiguously specified. However, in practice, many DMs find it difficult to specify a precise set of weights. Imprecision occurs also because the different objectives have different performance measures or they are measured in incommensurable units. In situations when specification of cardinal weights is difficult, it may be easier to provide ordinal rankings or provide simple ranks of the objectives. For example, given the weight structure, even if the weights cannot be specified exactly, the ranking pattern may remain the same, i.e., Z_I is the most important of all the three objectives, followed in importance by Z_C , and Z_S is the least important objective, or

$$Z_I > Z_C > Z_S. \quad (4)$$

Here, the symbol $>$ stands for "is preferred to." Thus, $Z_I > Z_C$ means that Z_I is preferred to Z_C .

When it is assumed that no further trade-off is possible, this may be considered as a case of preemptive priorities, i.e., the higher ranked objective is assumed to be infinitely more important than any of the lower ranked objectives, or $w_I \gg w_C \gg w_S$. Preemptive priorities are more commonly denoted by P , i.e., $P_I > P_C > P_S$. Under these preemptive priorities, PULV technology is preferable to IGCC technology. In fact, PULV technology will be preferred even if we have a hypothetical technology with CO₂ savings of 500 g/kWh, SO₂ savings of

9 g/kWh, and investment cost of \$1203/kW. This is because, under the assumptions of the preceding preemptive priorities, a reduction of \$1/kW is worth infinitely more than any amount of CO₂ savings or SO₂ savings. This is the basic principle behind a version of goal programming, called preemptive goal programming.

3.4 Multiattribute Decision-Making Methods—Some Common Characteristics

In general, MADM methods have some common characteristics:

1. The persons or organizations that are the stakeholders of the decision problem being considered have to be identified. These people will act as the decision makers whose opinions will be elicited by the MCDM methods. This stage is equally applicable to MODM methods, especially while eliciting the preference information for choosing one or a few from the set of all nondominated alternatives (solutions).

2. A model of the decision problem has to be constructed. Typically, the model will consider the main goal of the problem, relevant criteria, attributes, and alternatives. For example, if it is desired to evaluate some selected technologies, it is important to identify the key criteria that distinguish the characteristics of technologies, and all the relevant technologies to be considered as alternatives in the model. The model is called a hierarchical model in some MCDM methods, because it represents a hierarchy of decisions—the goal of the exercise (say, selection of the most important energy technologies), the criteria (economic significance, social significance, etc.), and alternatives (the technologies to be evaluated).

3. Ideally speaking, criteria have to be mutually exclusive and collectively exhaustive. Collective exhaustivity is critical in some methods, such as in the analytic hierarchy process, because final rankings could change for the addition or deletion of alternatives at a later stage.

4. The relative importance of criteria, which are implicit trade-off data (see previous section) used in identification of the best alternative(s), has to be assessed. Although assessment of the relative importance of criteria forms an important step in many MCDM methods, some methods do not use this information (e.g., data envelopment analysis). The procedure for this assessment differs for each method and forms generally the single most important

distinguishing feature of each method. The procedures will be described in more detail in the next section for some important MCDM methods, along with some prominent applications of the methods in the field of energy. The assessment can be performed either by an individual, or by representatives of conflicting values acting separately, or by those representatives acting as a group. Again, some methods have special procedures to aggregate the opinions of individuals to form a group opinion, though in general the group members should be encouraged to arrive at a single unanimous assessment.

5. Assessment of alternatives with respect to different criteria.

6. Aggregation of performance of alternatives with respect to all criteria to provide the overall performance measures of alternatives. Simple additive or multiplicative aggregation is used in many methods, but can also be more sophisticated (e.g., data envelopment analysis uses linear programming). The overall performance measures could also be considered as the rankings of the alternatives.

Note that many of these characteristics are also valid for the MODM problems, especially when the preferences of the DM are elicited.

4. AN OVERVIEW OF SOME MCDM METHODS AND THEIR ENERGY APPLICATIONS

Several surveys are available that describe many of the MCDM methods suggested in the literature. In the following discussion, four selected MCDM methods that have been applied to a number of problems in the field of energy will be briefly described. Due to space limitations, additional methods cannot be discussed here, but are available in the references listed at the end of this article. Software packages are available for facilitating the implementation of most of the methods, but, again owing to space limitations, they will not be discussed here.

4.1 MADM Methods

4.1.1 Multiattribute Utility Theory

Utility measures the subjective “worth” of an outcome, even one that is not a monetary value. Traditionally, utility functions are defined for stochastic problems that involve uncertainty. In the case

of deterministic problems, the term “value functions” is more commonly used. The utility or value functions may be thought of as evaluative mechanisms that can be used to measure the value of a particular solution.

Utility functions are defined in terms of uncertainty and thus tie in the decision maker’s preferences under uncertainty, revealing the DM’s risk preference for an attribute. An uncertain situation faced by a decision maker can be considered similar to a lottery—the DM can earn \$ X with a probability p , and earn \$ Y with probability $(1-p)$. In these situations, a rational decision maker is expected to maximize the *expected utility*, given by $\$[pX + (1-p)Y]$. Utility functions are assessed by giving the DM a sequence of choice between alternatives or between alternative lotteries. The responses are used to generate functions.

Multiattribute utility theory (MAUT) consists of assessing and fitting utility functions and probabilities for each attribute, and then using the functions and probabilities to come up with rankings of alternatives. The utility functions for each attribute are aggregated to get the overall utility function. At least two methods of aggregation are used in MAUT, additive and multiplicative. The additive aggregation is

$$\begin{aligned} U(A) &= \sum_i w_i u_i(a_i) \\ \sum_i w_i &= 1, w_i \geq 0, \end{aligned} \quad (5)$$

where $0 \leq u_i(a_i) \leq 1$; $u_i(a_i)$ is the utility function describing preferences with respect to the attribute i , a_i represents the performance of the alternative A in terms of the attribute i , the scaling factors w_i define acceptable trade-offs between different attributes, and $U(A)$ represents the overall utility function of the alternative A when all the attributes are considered together. This form of additive aggregation is valid if and only if the decision maker’s preferences satisfy the mutual preferential independence. Suppose that there is a set of attributes, X . Let Y be a subset of X , and let Z be its complement, i.e., $Z = X - Y$. The subset Y is said to be preferentially independent of Z , if preferences relating to the attributes contained in Y do not depend on the level of attributes in Z .

Under certain assumptions, the multiplicative aggregation can be written as $U(A) = \prod_i u_i(a_i)$. This multiplicative utility function requires stronger assumptions such as the Utility independence. The utility functions may be used as objective functions for solving mathematical programming problems.

Utility theory has been criticized because of its “strict” assumptions, which are usually not empirically valid. Because of the strict assumptions, practical applications of MAUT are relatively difficult, though there are several practical successful applications of MAUT. This has led to some simplifications of the MAUT concepts. For example, the multiattribute value theory is a simplification of MAUT without uncertainty and risk. The simple multiattribute rating technique is another simplification that makes weaker assumptions while eliciting utilities. These are described in the next few subsections.

4.1.1.1 Multiattribute Value Theory Multiattribute value theory (MAVT) is a simplification of MAUT: MAVT does not seek to model the decision maker’s attitude to risk. As a result, MAVT is considered to be easier to implement. Value theory assumes that it is possible to represent a decision maker’s preferences in a defined context by a value function, $V()$, such that if alternative A is preferred to alternative B , then $V(A) > V(B)$. For this representation to be possible, the decision maker’s preferences should satisfy two properties: the transitivity property (if A is preferred to B , B is preferred to C , then A should be preferred to C) and comparability (given two alternatives A and B , the decision maker must be able to indicate whether A is preferred to B , or B is preferred to A , or there is no difference between the two).

Note that the value function is an ordinal function, i.e., it can be used only to rank alternatives. In contrast, utility function is cardinal, i.e., it can be used to measure the strength of preference among alternatives. Multiattribute value theory explicitly recognizes that the decision maker will use many attributes (criteria) when evaluating a set of alternatives. For each attribute i , a partial value function $v_i(a_i)$ describing preferences with respect to the attribute i is assessed by the decision maker, where a_i represents the performance of the alternative A in terms of the attribute i . Then, the overall value function $V(A)$ of the alternative when all the attributes are considered together is normally obtained using the additive form: $V(A) = \sum_i v_i(a_i)$. This is more generally expressed as follows:

$$\begin{aligned} V(A) &= \sum_i w_i v_i(a_i) \\ \sum_i w_i &= 1, w_i \geq 0, \end{aligned} \quad (6)$$

where $0 \leq v_i(a_i) \leq 1$. As mentioned above, the scaling factors w_i define acceptable trade-offs between

different attributes. Again, this additive value function is appropriate if and only if the decision maker's preferences satisfy the so-called mutual preferential independence discussed earlier.

4.1.1.2 Simple Multiattribute Rating Technique The simple multiattribute rating technique (SMART) follows the steps described in the previous section for modeling a decision problem. It uses the simple weighting technique for the assessment of importance of criteria, and for the assessment of alternatives with respect to criteria. To rate (i.e., assess the importance of) criteria, the DM is asked to start by assigning the least important criterion an importance of 10. Then, the DM has to consider the next-least-important criterion, and ask how much more important (if at all) is it compared to the least important criterion, and assign a number that reflects that ratio. This procedure is continued till all the criteria are assessed, checking each set of implied ratios as each new judgement is made. The DM will be given the opportunity to revise previous judgements to make them consistent. Once the numbers are assigned, the relative importance of criteria is obtained by summing the importance weights, and dividing each by the sum. Thus, the relative importance of the criterion $j(w_j)$ is the ratio of importance weight of this criterion to the sum. Note that $\sum_j w_j = 1$ by definition.

Alternatives are rated with respect to each criterion in a similar fashion. Though MAUT requires the development of complex utility functions for each criteria, SMART prefers to produce the rating using a more straightforward approach: the DM is asked to estimate the position of the alternative of a criterion on a scale of 0 to 100, where 0 is defined as the minimum plausible value and 100 is defined as the maximum plausible value. Once the two measures are available, the overall performance of an alternative i can be aggregated using the simple weighted average,

$$U_i = \sum_j w_j u_{ij}, \quad (7)$$

where U_i is the overall performance rating of alternative i , w_j is the relative importance of criterion j , and u_{ij} is the rating of the alternative i with respect to the criterion j . The alternative that has the maximum U_i is the most preferred alternative to achieve the goal of the decision problem. The values of U_i can be used to provide the overall rankings of the alternatives.

MAUT, or its simplified versions MAVT or SMART, has been used for several practical applications. Siting and licensing of nuclear power facilities is an interesting application of these methods in the field of energy. Siting of nuclear power facilities is an extremely complex task because there are many interest groups with their own set of multiple objectives. Five categories of objectives were considered in the study: environmental, human safety, consumer well being, economic, and national interest. Note that each category has one or more objectives. For example, the economic category comprises objectives such as maximize economic benefits to local community, maximize utility company profits, and improve balance of payments. Similarly, the consumer well-being category comprises two objectives: provide necessary power and minimize consumer power costs. There are many interest groups (stakeholders) for this problem. The power company is one of the interest groups. Objectives of interest to the company are maximizing company profits, satisfying consumer preferences for energy, minimizing the detrimental environmental impact of its facilities, and maximizing the net benefits of its facilities on the local communities. Hence, it is important to assess a utility function from the company that is a function of these four objectives. Eliciting the utility functions involves assessing the trade-off information among the four objectives. More detailed discussion of this application is available in the literature cited in the reading list at the end of this article.

Additive or multiplicative utility or value functions have been employed to handle multiple criteria in many applications in the field of energy: to integrate socioeconomic, environmental, and cost criteria in siting a power plant and routing its transmission lines; for the weighting of environmental and economic benefits in demand-side management programs; for the amalgamation of energy, controllability, flexibility, viability, and other related attributes of new resources in bidding procedures; and combining measures of air, water, land, visual, noise, and solid waste effects into an index of environmental impact for use in resource bidding and screening. Other energy-related applications of utility or value theory include air pollution control problems, transporting hazardous wastes, refueling a nuclear power plant, climate change analysis, the study of UK energy policy, selection of a portfolio for a solar energy project, selection of a technology for the disposition of surplus weapons-grade plutonium, performing "as low as reasonably practicable"

(ALARP) assessments in the nuclear industry, identifying appropriate technological alternatives to implement treatment of industrial solid residuals, and risk analysis in nuclear emergency management.

4.1.2 The Analytic Hierarchy Process

The analytic hierarchy process (AHP) is one of the most popular and widely employed multicriteria methods. In this technique, the processes of rating alternatives and aggregating to find the most relevant alternatives are integrated. The technique is employed for ranking a set of alternatives or for the selection of the best in a set of alternatives. The ranking/selection is done with respect to an overall goal, which is broken down into a set of criteria.

The application of the methodology consists of establishing the importance weights to be associated to the criteria in defining the overall goal. This is done by comparing the criteria pairwise. Let us consider two criteria, C_j and C_k . The DM is asked to express a graded comparative judgment about the pair in terms of the relative importance of C_j over C_k with respect to the goal. The comparative judgement is captured on a semantic scale (equally important/moderately more important/strongly important and so on) and is converted into a numerical integer value a_{jk} . The relative importance of C_k over C_j is defined as its reciprocal, i.e., $a_{kj} = 1/a_{jk}$. A reciprocal pairwise comparison matrix A is then formed using a_{jk} for all j and k . Note that $a_{jj} = 1$. It has been generally agreed that the weights of criteria can be estimated by finding the principal eigenvector w of the matrix A :

$$AW = \lambda_{\max} w. \quad (8)$$

When the vector w is normalized, it becomes the vector of priorities of the criteria with respect to the goal; $\lambda_{\max}$ is the largest eigenvalue of the matrix A and the corresponding eigenvector w contains only positive entries. The methodology also incorporates established procedures for checking the consistency of the judgments provided by the decision maker.

Using similar procedures, the weights of alternatives with respect to each criterion are computed. Then, the overall weights of alternatives are computed using the weighted summation

$$\begin{aligned} & \left(\begin{array}{c} \text{Overall weight} \\ \text{of alternative } i \end{array} \right) \\ &= \sum_j \left(\begin{array}{c} \text{Weight of alternative } i \text{ with} \\ \text{respect to } C_j \times \text{weight of } C_j \\ \text{with respect to the goal} \end{array} \right) \quad (9) \end{aligned}$$

The popularity of AHP stems from its simplicity, flexibility, intuitive appeal, and ability to mix quantitative and qualitative criteria in the same decision framework. Despite its popularity, several shortcomings of AHP have been reported in the literature, which have limited its applicability. However, several modifications have been suggested to the original AHP, such as the multiplicative AHP (MAHP), to overcome these limitations. Some of the prominent shortcomings of AHP include the following limitations:

- Rank reversal: The ranking of alternatives determined by the original AHP may be altered by the addition of another alternative for consideration. For example, when AHP is used for a technology selection problem, it is possible that the rankings of the technologies get reversed when a new technology is added to the list of technologies. One way to overcome this problem is to include all possible technologies and criteria at the beginning of the AHP exercise, and not to add or remove technologies while or after completing the exercise. Some variants, such as MAHP, do not suffer from this type of rank reversal.

- Number of comparisons: AHP uses redundant judgments for checking consistency, and this can exponentially increase the number of judgments to be elicited from DMs. For example, to compare eight alternatives on the basis of one criterion, 28 judgments are needed. If there are n criteria, then the total number of judgments for comparing alternatives on the basis of all these criteria will be $28n$. This is often a tiresome exercise for the decision maker. Some methods have been developed to reduce the number of judgments needed.

AHP has been applied to a variety of decision problems in the literature. Several detailed annotated bibliographies of AHP applications are available. In an illustrative application of AHP for the selection of greenhouse gas mitigation options, there are three criteria—namely, cost-effectiveness, feasibility, and other national benefits—to rank five demand-side management options (variable speed drives, good housekeeping, energy-efficient motors, compact fluorescent lamps, and cogeneration). Some other energy-related applications of AHP include solar energy utilization, energy resource allocation, integrated resource planning, climate change negotiations, environmental impact assessment, energy choices for the United States and for members of the Organization of Petroleum Exporting Countries (OPEC), choice of an appropriate of energy mix for

the United States, and, options for high-level nuclear waste management. In some of the applications, AHP has been combined with other MCA methods, such as goal programming and compromise programming. In an application of AHP for global negotiations on climate change, the method has been used to compare the preferences of nations of the world to decide about the extent of reduction of greenhouse gases. In another application in utility planning, AHP has been applied to evaluate the best bids from private parties using criteria such as flexibility in starting date, fuel diversity, reputation of the bidder, and environmental impacts during generation and transmission.

4.2 MODM Methods

As discussed earlier, most of the MODM problems start with the following mathematical programming problem:

$$\begin{aligned} &\text{maximize/minimize } Z_1 = C_1X \\ &\text{maximize/minimize } Z_2 = C_2X \\ &\text{maximize/minimize } Z_n = C_nX \\ &\quad \vdots \\ &\text{subject to } AX = b, \quad X \geq 0 \end{aligned} \quad (10)$$

where the values Z_i , $i = 1, 2, \dots, n$, represent the n objective functions.

4.2.1 Goal programming

Goal programming (GP), likely the oldest school of MCDM approaches, was developed in the 1950s as an extension of linear programming. In its simplest form, the method of GP assigns so-called aspiration levels (also called targets or goals) for the achievement of different objective functions, and minimizes the deviations of actual achievement from the aspiration levels. It is important to stress here that that the term “goal” has the connotation of target when used in goal programming, whereas in the general MCDM context, the term “goal” is taken to represent a generalization of all criteria.

Consider Problem (10). Suppose T_i is the aspiration level for the objective function Z_i . This means that the decision maker expects to achieve around T_i for the objective function Z_i as given by the approximate equation, $Z_i \approx T_i$, which can be further written as $Z_i + d_i^- - d_i^+ = T_i$, where d_i^- and d_i^+ are deviational variables measuring deviations of actual achievement below and above the aspiration level. They are usually called underachievement and overachievement deviational variables, respectively.

Because it is expected that the actual achievement will be as close to the aspiration levels as possible, the deviational variables are minimized in goal programming. Weights, reflecting the relative importance of the objective functions, can also be associated with the deviational variables during the minimization. Thus, one simple form of the goal programming objective function is given by the following relationship:

$$\min \sum_{i=1}^n (w_{im}d_i^- + w_{ip}d_i^+), \quad (11)$$

where the values w_i are the weight given to minimizing the selected deviational variable from the aspiration level T_i . This objective has to be minimized subject to the original constraints of Problem (1).

Here are two major approaches in goal programming: minimizing the weighted function of goals (discussed previously) and preemptive goal programming that avoids the weighted summation in the goal programming objective function. In preemptive GP, the weights used in the objective function are preemptive weights as defined in Section 3.4. Other mathematical forms, such as minimizing the maximum deviation, fractional goal programming, and nonlinear GP, are also available in the literature. GP has been criticized by several authors. For example, it can, in some circumstances, choose the dominated solution. The criticisms can be overcome by careful applications of the method. In one application, GP has been used to provide energy resource allocation for the city of Madras in India. The minimum energy requirements for cooking, water pumping, lighting, and using electrical appliances form the main constraints of the model. The objective functions pertain mainly to the needs for minimizing cost, maximizing efficiency, minimizing consumption of petroleum products, maximizing employment generation, maximizing the use of locally available resources, minimizing emissions of oxides of carbon, sulfur and nitrogen, and maximizing convenience and safety. The trade-off information about the relative importance of these objectives has been obtained using the AHP.

4.2.2 Compromise Programming

The multiple objective functions Z_i in Problem (10) are often conflicting. Compromise among the multiple conflicting objective functions is defined as deviation or “distance” from the ideal solution. The compromise programming approach minimizes the

distance from the ideal solution. The distance is defined as

$$d_p = \left\{ \sum_{i=1}^n w_i^p \left[\frac{Z_i(x^*) - Z_i(x)}{Z_i(x^*)} \right]^p \right\}^{1/p}, \quad (12)$$

where x is the vector of decision variables in the optimization exercise, the values $Z_i(x)$, $i = 1, 2, \dots, n$, are the n objective functions, $Z_i(x^*)$ represents the ideal solution [when the objective function $Z_i(x)$ is optimized separately without considering the other objective functions], w_i represents the weights to be provided to the objective function $Z_i(x)$, and the parameter p can range from 1 to ∞ .

Note that the deviations of all of the objective functions, $Z_i(x^*) - Z_i(x)$, are divided by the respective ideal values. This is done because the objectives are measured in different units and are not comparable in absolute terms. In the literature, p has been interpreted as a degree of noncompensation among the objectives. The case $p = 1$ refers to when the objectives are perfectly compensatory, i.e., an increase in one can compensate for a decrease in the value of the other. The case of $p = \infty$, on the other hand, corresponds to absolute noncompensation, i.e., the objectives are of completely different natures, and improvement in one does not compensate for worsening of the others. Thus, when compromise programming is used, one will have to minimize d_p subject to the original constraints, $AX = b$, $X \geq 0$.

Compromise programming has received some energy applications, especially related to utility planning and environmental economics. In an application, the method has been used to aid the utility planners to select the best bids for generation capacity addition. Bids from independent power producers have been evaluated using multiple objectives—minimizing cost, minimizing emissions, and maximizing qualitative measures such as flexibility, fuel diversity, reputation, etc., subject to constraints representing demands, capacity limits for individual generator and transmission projects, budget constraints, hydro energy availability, fuel production and transportation capacity limits, and minimum utilization of generation capacity. Note that the first two objectives are quantifiable whereas the remaining objectives are not directly quantifiable. Although the compromise programming model has used the first two objectives directly, other qualitative objectives have been handled using AHP.

4.3 Other Methods

The field of MCDM features many more techniques for solving MCA problems. Only a few were described in the previous section. Other techniques include ELECTRE, PROMETHEE, an aspiration-level interactive method, TOPSIS, the reference point approach, multiobjective linear programming, and data envelopment analysis. A technique named preference ratios through intervals in multiattribute evaluation (PRIME) is a relatively recent method that deals with incomplete information about the preferences of the decision makers. An approach involving measuring attractiveness by a categorical-based evaluation technique (MACBETH) was developed in Europe in the 1990s. It is an interactive approach for cardinal measurement of judgments about the degrees of attractiveness in decision processes. Other methods based on an outranking approach (the approach initially used in the ELECTRE methods) include the QUALIFLEX method, the ORESTE method, and the TACTIC method. There are more methods, such as the utilities additives (UTA) method, the ZAPROS method, and the NIMBUS method. Most of these methods are described in the literature mentioned in the list of suggestions for further reading.

SEE ALSO THE FOLLOWING ARTICLES

Bottom-Up Energy Modeling • Complex Systems and Energy • Depletion and Valuation of Energy Resources • Input–Output Analysis • Life Cycle Assessment and Energy Systems • Modeling Energy Markets and Climate Change Policy • Modeling Energy Supply and Demand: A Comparison of Approaches • Net Energy Analysis: Concepts and Methods

Further Reading

- Dyer, J. S., Fishburn, P. C., Steuer, R. E., Wallenius, J., and Zionts, S. (1992). Multiple criteria decision making, multiattribute utility theory: The next ten years. *Manage. Sci.* 38(5), 645–654.
- Edwards, W. (1977). How to use multiattribute utility measurement for social decision-making? *IEEE Trans. Syst. Man Cybernet* 7/5, 326–340.
- Gal, T., Stewart, T. J., and Hanne, T. (eds.). (1999). "Multicriteria Decision Making: Advances in MCDM Models, Algorithms, Theory, and Applications." Kluwer Academic Publ., Boston.
- Hobbs, B. F., and Meier, P. (2000). "Energy Decisions and the Environment: A Guide to the use of Multicriteria Methods." Kluwer Academic Publ., Boston.
- Keeney, R. L., and Raiffa, H. (1976). "Decisions with Multiple Objectives: Preferences and Value Tradeoffs." Wiley, New York. [Another edition (1993) available Cambridge University Press.]

- Roy, B. (1996). "Multicriteria Methodology for Decision Aiding." Kluwer Academic Publ., Dordrecht, The Netherlands.
- Saaty, T. L. (1980). "The Analytic Hierarchy Process: Planning, Priority Setting and Resource Allocation." McGraw-Hill, New York.
- Schniederjans, M. J. (1995). "Goal Programming: Methodology and Applications." Kluwer Academic Publ., Boston.
- Shi, Y., and Zeleny, M. (eds.). (2000). "New Frontiers of Decision Making for the Information Technology Era." World Scientific Publ. Co., Singapore.
- Zeleny, M. (1982). "Multiple Criteria Decision Making (McGraw-Hill Series in Quantitative Methods for Management)." McGraw-Hill, New York.

National Energy Modeling Systems



ANDY S. KYDES

U.S. Department of Energy
Washington, D.C., United States

AMIT KANUDIA

McGill University and KanORS Consulting Inc.
Montreal, Quebec, Canada

RICHARD LOULOU

McGill University and HALOA Inc.
Montreal, Quebec, Canada

-
1. Introduction
 2. Overview of NEMS
 3. Analysis of a 20% Non-hydroelectric Portfolio Standard
 4. Overview of MARKAL
 5. Example Application of the Canadian MARKAL model

Glossary

coefficient of performance The ratio of energy transferred out of a system (e.g., an air conditioner or heat pump transfers heat from inside a house to outside the house) to the energy input to the system.

competitive market Pertains to the degree that energy prices and supply are influenced or set by any particular group in the market; if energy prices and supplies are set by many market participants (no monopoly or monopsony power) and not by regulatory bodies, the market is said to be competitive.

corporate average fuel efficiency (CAFE) For new cars, a standard by which the average of all cars sold in the United States by a manufacturer must achieve 27.5 miles per gallon.

dynamic linear programming model A multiperiod linear programming model that is composed of multiple smaller linear program models (one for each time period) that are linked with intertemporal constraints; a dynamic linear programming model provides the optimal value of the objective over all time periods—decision making with perfect foresight.

efficiency The ratio of the energy output by an energy system to the energy content of the fuel input to the energy system.

exajoule (EJ) A unit of energy equal to 10^{18} joules; 1.0556 EJ is approximately 1 quadrillion Btu.

Gauss–Seidel convergence algorithm An iterative method for solving large linear or nonlinear models; in the National Energy Modeling System (NEMS), the algorithm finds where supply equals demand and the delivered prices for which this happens.

greenhouse gases Gases that have the property of absorbing or trapping energy in the earth's atmosphere; such gases include carbon dioxide, methane, and a number of manmade chemicals used in refrigeration and manufacturing, with the latter group being collectively known as “high global warming potential” gases.

intermittent technologies Pertain to technologies whose energy production is intermittent (e.g., wind, photovoltaics); for intermittent technologies, energy production depends on factors other than the availability of the technology (e.g., availability of wind and sunlight).

learning factor Pertains to the rate by which manufacturing costs decline for every doubling of capacity.

linear programming model A mathematical model with a linear objective function and linear constraints.

load shape Pertains to the time profile of energy consumption.

logit function A function form that often includes exponential functions in the denominator and numerator; it is typically used to determine market shares.

market diffusion Pertains to the adoption of a new technology or features into the market.

merit order dispatch Pertains to the order in which power plants are brought online and used to satisfy electricity demand; a merit order dispatch requires that plants be brought online in order of increasing operating plus fuel costs.

overnight capital cost Usually pertains to the costs quoted for electric power generation plants; overnight costs include all manufacturing costs, including owners' costs

and project contingency but not interest during construction.

reference energy system A schematic representation of energy flows in an energy system, from energy production, to intermediate conversions and transformations, to delivery at end use.

renewable portfolio standard Pertains to a potential policy requiring that a minimum percentage of all generation be provided from renewable energy sources such as wood, wind, other biomass, and geothermal.

Residential Energy Consumption Survey (RECS) A survey that was developed and is published periodically by the Energy Information Administration.

technological change Pertains to changes in the development, adoption, and characteristics of technologies, including innovations, efficiency improvements, cost reductions, and the addition of new qualitative features such as variable speed fans.

technological optimism Pertains to cost and performance (mis)estimates for technologies that have not been commercialized. Prior to commercialization, manufacturing costs tend to be underestimated and performance tends to be overestimated.

unit energy consumption (UEC) A surrogate term used to measure efficiency in a process that combines structures and energy conversion systems.

The development and introduction of new technologies to the market, relative costs and performance of technologies, physical lifetimes of installed equipment (which influence the natural turnover rate), and consumer preferences are key factors that determine the rate of market diffusion of technologies and how quickly energy use patterns, energy efficiency, and energy-related environmental emissions in a society can change. Estimates of the costs and benefits of new energy policies, such as those that may be implemented to constrain carbon dioxide or other greenhouse gas emissions, will depend on estimates of how rapidly advanced technologies are accepted in the market and how quickly existing end use equipment reaches the end of its useful life (i.e., wears out). Consequently, the representation of technological change, market diffusion, and other economic flexibility embodied in any energy-economy modeling system will be major determinants of the projected energy-economy adjustments or responses to policy/price changes. This article provides an overview of the National Energy Modeling System (NEMS), which was developed and used by the Energy Information Administration (EIA) to address energy-economic-environmental policy questions for the U.S. government—typically Congress and the executive branch of government. A

brief example of a recent NEMS application in assessing the potential impact of a 20%, non-hydroelectric, renewable portfolio standard (RPS) by 2020 on U.S. energy markets is also provided. The second national energy modeling system described in this article is the Canadian MARKAL model, one of the numerous variants of the market allocation (MARKAL) model, a dynamic linear programming formulation of the Canadian energy markets.

1. INTRODUCTION

The purpose of this article is to provide an overview of arguably two of the most prominent and influential national energy modeling systems used over the past decade or so: the National Energy Modeling System (NEMS) and the MARKAL model. NEMS was developed by the Energy Information Administration (EIA) during the early 1990s to analyze U.S. energy markets and is important because it is the primary analytical and projection tool used by the U.S. Department of Energy to analyze the energy, economic, and environmental impacts of proposed energy policies for the U.S. government. MARKAL is important because it has been used by more than 50 countries and 70 regional governments to analyze the impact of energy strategies and is supported by an international group of experts and the International Energy Agency (IEA).

The persistence of and refinements to these models over the past decade or so reflect both the increasing complexities of proposed energy-environmental policies and the recognition that energy policy has important impacts on an economy, even though energy's share of gross domestic product (GDP) may be less than 10%.

The next section of this article provides an overview of NEMS with more detailed descriptions of the residential and electricity generation markets to illustrate how two representative energy sectors are modeled in the system. This is followed by an example of the use of NEMS in policy analysis—a renewable portfolio study. The subsequent section provides an overview of MARKAL. The final section provides an example of policy analysis using the MARKAL model.

2. OVERVIEW OF NEMS

This section provides a brief overview of NEMS. A more extensive treatment of the NEMS model is

provided on the EIA's Web site (www.eia.doe.gov). NEMS is a large, technology-rich, regional, computer-based energy-economy model of U.S. energy markets for the midterm period through 2025.

In 1990, the secretary of the U.S. Department of Energy (DOE) directed that NEMS be developed. NEMS was designed and developed to support energy policy analysis and strategic planning, based on recommendations from the National Research Council (NRC) of the National Academy of Sciences.

Key features implemented in NEMS included (1) regional outputs of energy, economic, and environmental activity of the U.S. economy; (2) use of a modular modeling structure to facilitate and enable the model builders to work with particular aspects of the model independently; (3) integration of engineering and economic approaches to represent actual producer and consumer behavior; (4) use of a projection period spanning 20 to 25 years; and (5) incorporation of the broader energy analysis community and outside peer groups in the design and update of NEMS.

NEMS was completed at the end of 1993 and was used for the first time to develop the *Annual Energy Outlook 1994, with Projections to 2010*. More recently, NEMS has been extended to 2025 and further revised to address electricity restructuring and carbon mitigation issues and numerous other multipollutant policy studies.

NEMS is used to develop baseline projections that are published annually in the *Annual Energy Outlook (AEO)*. NEMS is also used by analysts to prepare special studies that are requested by the U.S. Congress, the DOE Office of Policy, other DOE offices, and other government agencies. In accordance with the requirement that the EIA remain policy neutral, the AEO projections assume that all existing legislation, regulations, and policies remain unchanged. Furthermore, these projections depend on additional uncertain assumptions, including the estimated size of the economically recoverable resource base of fossil fuels, changes in world energy supply and demand, the rate at which new energy technologies are developed, and the rate and extent of their adoption and penetration. Consequently, the AEO projections are not meant to be predictions about the future. This article describes an updated version of NEMS used for the AEO 2000.

2.1 Purpose of NEMS

The primary purpose of NEMS is to analyze the energy-related consequences for the United States of

alternative energy policies or pertinent economic or energy market influences. The policy questions of interest have determined the level of detail required within the structure of NEMS. For example, environmental issues relating to energy production and consumption have taken on new importance with the implementation of the Clean Air Act Amendments (CAAA) of 1990 and the proposed Kyoto Protocol on greenhouse gases (GHGs) in 1997. Accordingly, NEMS is designed to measure seven emissions (sulfur oxides [SO_x], nitrogen oxides [NO_x], carbon, carbon monoxide [CO], carbon dioxide [CO₂], volatile organic compounds, and mercury) released in the use of energy products to generate electricity and, in the case of carbon (or CO₂), to constrain national emissions using a pricing mechanism. The technology representation in NEMS is particularly helpful in the analysis of national carbon mitigation policies and utility sector SO_x, NO_x, and mercury mitigation policies due to its explicit representation of vintage (time-dependent) energy equipment and structures (e.g., building shells) and the careful tracking of vintage capital stock turnover rates. For similar reasons, NEMS contains sufficient detail in the transportation sector to project the use of alternative or reformulated fuels such as compressed natural gas, ethanol, and methanol. In addition to environmental concerns, NEMS is designed to account for existing and emerging government regulations (e.g., electricity restructuring), the potential for the development and use of new energy-related technologies, the increased use of renewable sources of energy (especially intermittent technologies), and the potential for demand-side management, conservation, and increases in the efficiency of energy use. These topics reflect the expected scope of current and future government policy.

The NEMS representation of energy markets focuses on four important interrelationships: (1) interactions among the energy supply, conversion, and consumption sectors; (2) interactions between the domestic energy system and the general domestic economy; (3) interactions between the U.S. energy system and world energy markets; and (4) interactions between current production and consumption decisions and expectations about the future.

2.1.1 Domestic Energy

System/Economy Interactions

The general level of economic activity in sectoral and regional detail has traditionally been used as an explanatory variable or "driver" for projections of energy consumption and prices. In reality, energy prices and other energy system activities themselves

influence the level of economic activity. NEMS is designed to capture this “feedback” between the domestic economy and the energy system. Thus, energy price changes are capable of affecting macroeconomic variables such as GDP.

2.1.2 Domestic and World Oil Market Interactions

The world oil price (WOP) is a key variable in domestic energy supply and demand decision making. As a result, WOP assumptions have been a key starting point in the development of energy system projections. In fact, the U.S. energy system itself exerts a significant influence on world oil markets, and this in turn influences the WOP—another example of a feedback effect. World energy market supply and demand are first specified outside of NEMS. Given this, NEMS models the interactions between the U.S. and world oil markets, with changes in U.S. oil markets affecting world supply and demand. As a result, domestic energy system projections and the WOP are made internally consistent.

2.1.3 Economic Decision Making over Time

The production and consumption of energy products today are influenced by past decisions to develop energy resources and acquire energy using capital. Similarly, the production and consumption of energy at a future time are influenced by decisions made today and in the past. Current investment decisions depend on expectations about future market circumstances. For example, the propensity to invest now to develop alternative energy sources is greater when it is expected that future energy prices will be higher. NEMS allows the application of different kinds of foresight assumptions to be applied differentially to its individual submodules. This feature allows the consequences of different kinds of planning horizons to be incorporated into NEMS projections.

2.2 System Design of NEMS

Like its predecessor models (e.g., the Intermediate Future Forecasting System [IFFS]), NEMS incorporates a market-based approach to energy analysis. NEMS balances the supply of and demand for energy for each fuel and consuming sector, taking into account the economic competition between energy sources.

NEMS is partitioned into a modular system, which is solved by applying the Gauss–Seidel convergence method with successive overrelaxation. The modules of NEMS represent each of the fuel supply markets, conversion sectors, and end use

consumption sectors and also include interactive macroeconomic and international modules. The primary flows between these modules are the delivered prices of energy and the energy quantities consumed by product, region, and sector but also include other information such as economic activity and technology characteristics. The delivered prices of fuels incorporate all of the activities necessary to produce, import, and transport fuels to the end users. Figure 1 provides a system overview of NEMS.

The integrating methodology controls the independent execution of the component modules. The modules are executed from the integrating module. To facilitate modularity, the components do not pass information to each other directly but instead communicate through a central data file. This modular design provides the capability of executing modules individually or to alternative modules, thereby allowing decentralized development of the system and independent substitute analysis and testing of individual modules. Furthermore, this modularity allows the flexibility of using the methodology and level of detail that is most appropriate to represent each energy sector.

A solution is achieved by equilibrating on the delivered prices and quantities of energy demanded, thereby ensuring an economic equilibrium of supply and demand in the consuming sectors. Each fuel supply, conversion, or end use demand module is called in sequence by the integrating module and solved, assuming that all other variables in the other energy markets are fixed. For example, when solving for the quantities of fuels demanded in the residential sector for an input set of energy product prices, all other sectors of the economy are held fixed. The modules are called iteratively until successive end use prices and quantities remain constant within a specified tolerance. This equilibration is achieved annually through the midterm period to 2025. Table I provides a summary of NEMS products and regional details. NEMS reflects market economics, industry structure, and energy policies and regulations that influence market behavior.

NEMS consists of four supply modules (oil and gas, natural gas transmission and distribution, coal, and renewable fuels), two conversion modules (electricity and petroleum refineries), four demand modules (residential, commercial, transportation, and industrial), one module to simulate energy–economy interactions (macroeconomic activity), one module to simulate world energy–domestic energy interactions (international energy activity), and one module to provide the mechanism for achieving a

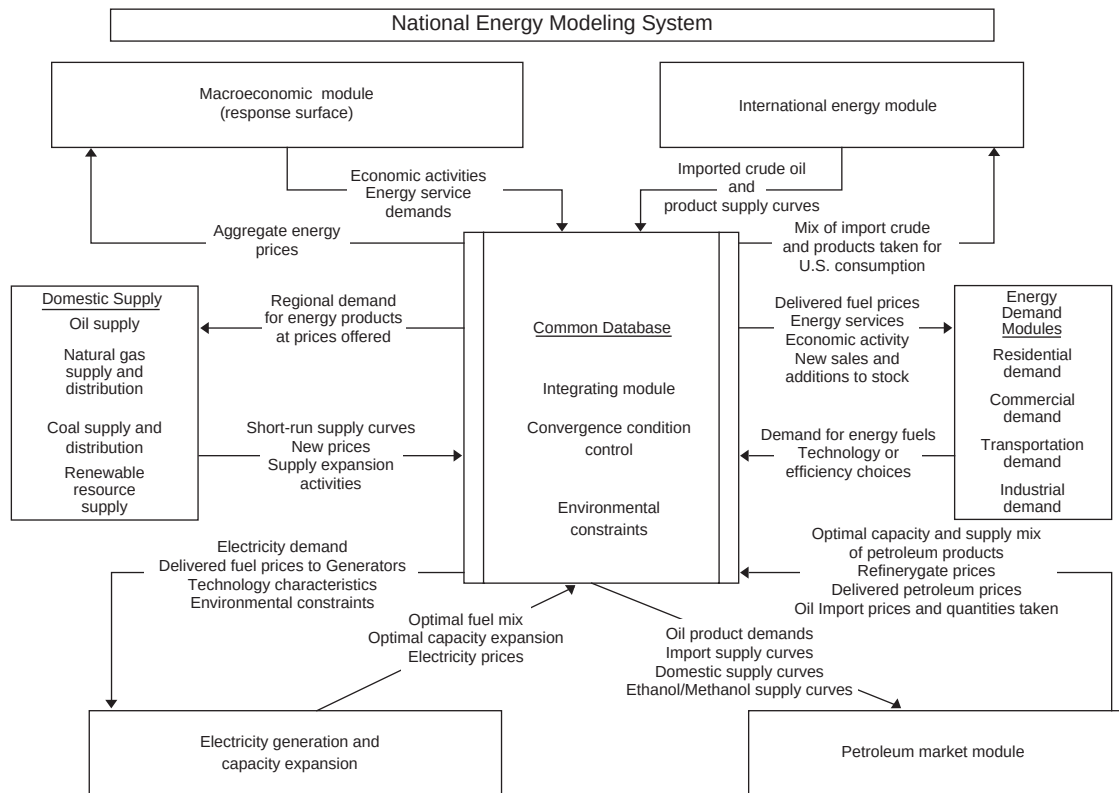


FIGURE 1 NEMS systems view.

general market equilibrium among all of the modules (the integrating module).

2.3 Foresight

All of the NEMS supply modules require assumptions of future energy prices and demands for energy so as to make capacity expansion decisions. Recognizing that there is valid evidence that decision making varies by sector, one implementation of foresight in NEMS allows each sector to use those foresight and behavioral assumptions deemed most appropriate for that sector. Two alternative decentralized options that have been implemented for foresight are myopic (i.e., assuming, within any forecast year, that the current prices will remain constant into the future for capacity expansion decisions) and extrapolative (i.e., assuming, within any forecast year, that expected prices, demands, and supplies are a function of historical trends and other critical assumptions regarding behavior). In the simplest case, the extrapolation may be a growth rate applied to the current forecast year.

Perfect foresight has been implemented using a recursive process where the previous run's prices are used as the expectations.

2.4 Emissions

Recognizing the importance of environmental issues associated with the use of energy, an environmental accounting capability has been incorporated within NEMS. Seven emissions are accounted for in the electricity generation sector: sulfur dioxide (SO_2), NO_x , CO, CO_2 , carbon, volatile organic compounds, and (most recently) mercury emissions from power plants. A subset of these emissions is computed for energy production activities and fuel combustion. In addition, NEMS is designed to represent all current environmental regulations (e.g., CAAA of 1990) as well as other mandated costs for controlling toxic substances.

NEMS also incorporates the capability of constraining systemwide CO_2 emissions as well as SO_2 emissions, NO_x , and mercury emissions in the utility market—important features for policy analysis. One option for achieving a prescribed level of systemwide CO_2 emissions is accomplished by raising the cost of producing emissions until the total system emissions are reduced to the level of the constraint. Another approach used in the utility market for SO_2 is to incorporate the constraint directly into a linear program. This is

TABLE I

Summary of NEMS Modeling Details

Energy system activity	Categories	Regions
Primary supply		
Oil	Conventional Enhanced Heavy oil Shale oil	Six lower 48 onshore Three lower 48 offshore Three Alaska
Gas	Conventional nonassociated Conventional associated Coalbed methane Three other types	Six lower 48 onshore Three lower 48 offshore Three Alaska
Natural gas transmission and distribution	Residential, commercial, industrial, utility firm versus interruptible Peak versus offpeak	12: nine census divisions, with further subdivisions for key transportation issues
Coal	Four sulfur categories Four thermal categories Deep and surface mining types	16 supply regions
Renewables	Hydropower, wind, geothermal, solar thermal, photovoltaic, municipal solid waste, biomass	Natural Environment Research Council/ Refining/Census
Conversion		
Electricity (including uranium for nuclear)	Utilities, independent power cogeneration	13 supply regions Nine census demand regions
Refining	Five crude categories 19 products	Three petroleum allocation for defense districts
Energy demand		
Residential	Eight end use services Three housing types	Nine census divisions
Commercial	Eight end use service 11 building types	Nine census divisions
Transportation	Six vehicle size categories 10 vehicle age categories	Nine census divisions
Industrial	35 industries Nine primary industries	Four census regions shared to nine census divisions

the method used by the electric market module to constrain SO₂ emissions as mandated by the CAAA of 1990.

2.5 Policy Analysis Capability with NEMS

The following list illustrates the main analytical capabilities available with NEMS:

- *Federal tax policies/fees*: impacts of energy tax policies on the economy and energy system; impacts of Btu or carbon taxes; revenue recycling options (how portions of taxes are recycled through the

economy: all to consumers, all to deficit reduction, or some other combination); federal reaction to rising unemployment (accommodating or not)

- *Emissions caps/restrictions*: carbon emissions cap (cap and trade impacts) from all sectors; national sulfur cap from electricity generation; NO_x cap from electricity generation; mercury emissions cap from electricity generation; other emissions caps

- *Effects of existing and proposed government laws and regulations related to energy production and use*: impacts of electricity restructuring on prices and treatment of stranded costs; impacts of renewable portfolio standards (RPSs); impacts of

increased use of renewable energy sources; potential savings from increased efficiency of energy use; changes in emission levels due to environmental policies; effects of appliance efficiency and building shell standards on energy consumption; impacts of fuel use restrictions (e.g., required use of oxygenated and reformulated gasoline, mandated use of alternative-fueled vehicles) on emissions, energy supply, prices, and economic growth; changes in natural gas prices and pipeline and import capacity in response to regulatory initiatives; impacts of automobile CAFE standards

- *Impacts of advanced/alternative technology menus/cost/performance:* potential impacts of new and advanced energy production, conversion, and consumption technologies; impact on technology learning of accelerated demonstration plants for utilities (government-funded/mandated capacity additions to accelerate learning and reduce capital and/or operating and maintenance [O&M] costs); impacts of new technologies on consumption and production patterns and emissions; impacts on the production of crude oil and natural gas resulting from improvements in exploration and production technologies

- *Oil market assessment:* responses of the energy and economic systems to changes in world oil market conditions as a result of changing levels of foreign production and demand in the developing countries.

2.6 NEMS Residential Sector

The residential demand module forecasts fuel consumption and energy equipment choices in the residential sector for 9 census divisions, 3 housing types (single-family, multifamily, and mobile homes), and 10 end uses of which 8 (space heating, space cooling, water heating, refrigerators, freezers, lighting, clothes dryers, and cooking) have equipment choices associated with them.

2.6.1 Important Interactions with NEMS

NEMS is initialized with information about the state of the U.S. energy system for a recent historical year. With respect to the residential demand module, this includes base year housing stock and retirement rates, appliance stocks and life expectancies, new appliances to be made available in current and future years with their costs and efficiencies, housing shell integrity (efficiency index), unit energy consumption per end use and household, and square footage per household and housing type. In addition, assumptions about population and associated demographics developed by the U.S. Department of Commerce are provided as inputs to the projection.

NEMS provides to the residential modules forecasts of residential energy product prices and new housing starts (Fig. 2). The housing stock submodule begins by adjusting the previous year's housing stock by adding new construction (new housing starts) and

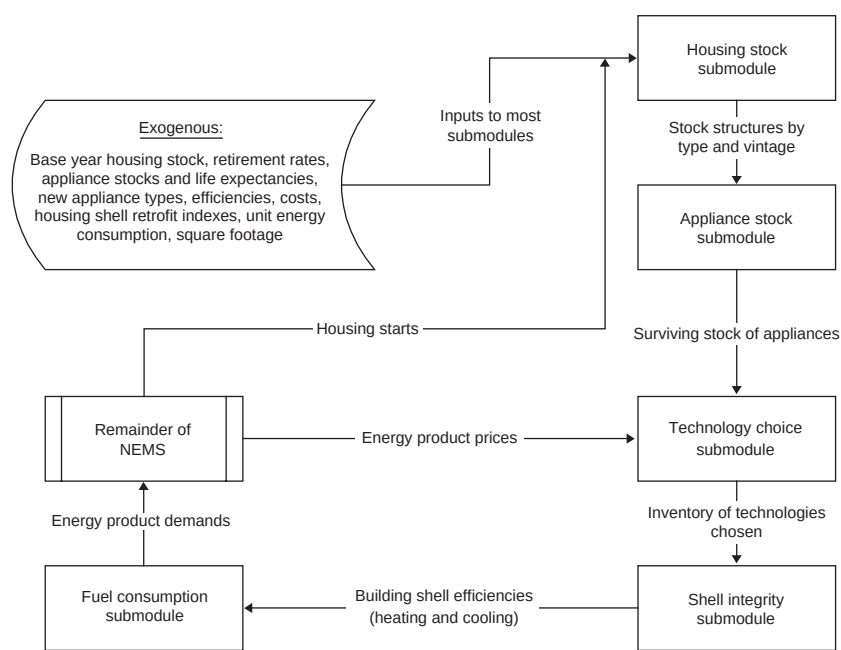


FIGURE 2 Residential demand module.

subtracting housing retirements. The new housing stock is allocated to each of the three housing types based on a combination of historical shares and income distribution and demographic changes. Control is then transferred to the appliance stock module, which (1) retires appliances in homes that were retired, (2) retires appliances that have reached the end of their useful lives, and (3) prepares the menu of technology choices with their associated cost and performance characteristics. Control is then transferred to the technology choice submodule. This submodule, using equipment cost and performance data with estimated consumer preference functions, selects appliances to replace retired appliances and appliances needed for new housing. Control is then passed to the shell integrity submodule, which adjusts the heating and cooling requirements of each building type based on an autonomous change to shell efficiency (improvements that are independent of price) and price-induced changes to building efficiency.

Once the appliances have been selected for each end use and end use energy intensities have been adjusted for changes to shell integrity and lifestyle changes such as increasing floor space per household, the energy consumption module computes energy consumption by fuel and census division, and this is passed back to NEMS for response. Figure 3 provides an overview of the residential demand module with an emphasis on the appliance and fuel choice components.

2.6.2 Equipment Characterizations and Choice Methodology

2.6.2.1 Technology Characterizations Residential sector energy technologies are characterized by (1) retail cost of equipment plus installation, (2) efficiency of the equipment, (3) dates available for purchase (efficiency standards may limit availability during future years), and (4) technological evolution, that is, either projected efficiency increases or projected cost declines. Cost reductions can be discrete

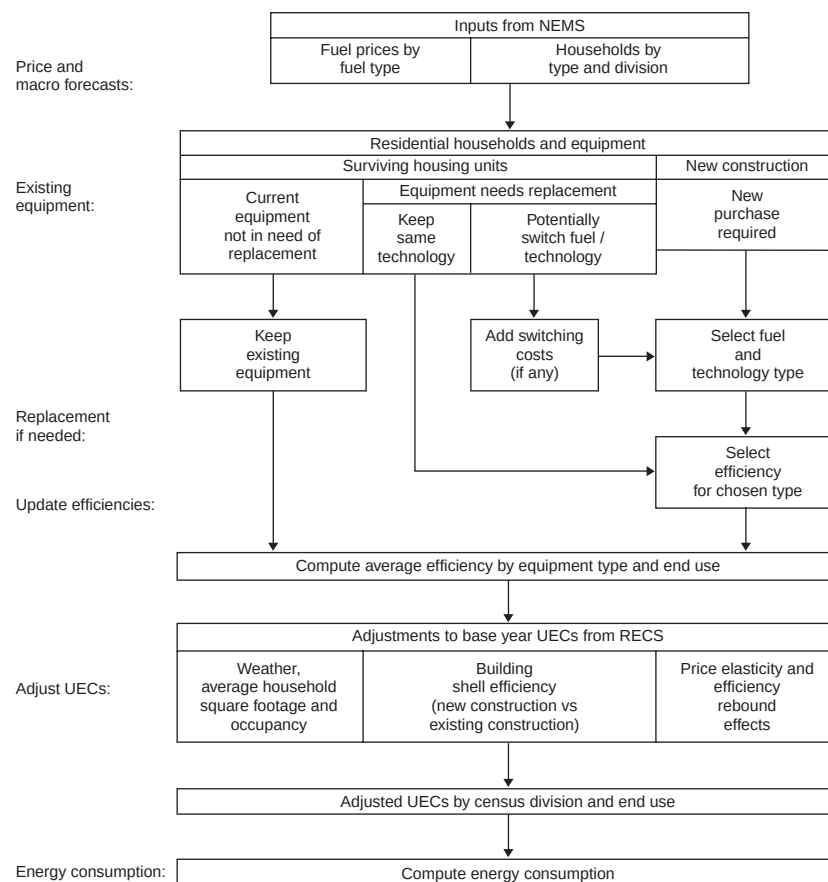


FIGURE 3 NEMS residential model overview. UECs, units energy consumption; RECS, Residential Energy Consumption Survey.

(e.g., a new characterization of a technology in a given year) or continuous (e.g., annually declining costs modeled for new heating technologies).

2.6.2.2 Market Penetration/Consumer Choice Algorithm The choice of equipment in the residential demand sector is made using a two-step selection process that incorporates the following important residential market attributes: (1) the inclination not to replace equipment until it fails (i.e., “if it isn’t broken, don’t fix it”), (2) the predisposition to stay with the same technology when equipment needs to be replaced, and (3) the incorporation of switching costs (“real and perceived”) when changing to a different fuel is contemplated.

Technology choice/penetration is a two-step process in the NEMS residential energy model. Equipment choices are segmented by new and replacement equipment. For equipment decisions in new homes or for replacement decisions that consider fuel switching as an option, the first step is to decide on what fuel and general equipment type will be used for each end use (e.g., natural gas furnace for space heating). The second step is to select the efficiency of the equipment (e.g., the furnace). For replacement decisions that do not consider fuel switching or technology switching, the only decision made is to select the efficiency of the technology. NEMS uses a logit function to represent the decisions made in each of these two steps.

In the residential sector, technologies compete on the basis of relative importance of installed cost and operating cost for each technology. Equipment choices are segmented by new and replacement equipment for each of nine census divisions and each of three building types.

2.6.2.3 Fuel–Technology Class Choice The following calculation is performed as the first step in determining the fuel–technology class combination in new homes and when considering fuel switching when replacing worn equipment. When fuel switching is considered for replacement decisions, a matrix of switching costs from one equipment/fuel type to another is added to the installed capital costs before computing life cycle costs. Life cycle costs are based on the implicit hurdle rate for each end use and technology combination using 7 years as the time horizon over which discounted costs are calculated. These hurdle rates combine not only financial aspects (the cost of money) but also nonfinancial aspects (institutional and physical obstacles and perceived risks). The implicit discount rates range from 15 to

more than 100%. The time horizon of 7 years reflects the fact that the average time between moves for homeowners in the United States is less than 7 years, and a longer time horizon for calculating discounted costs seems inappropriate for simulation purposes.

If LC_i is the life cycle cost of the representative technology for each fuel–technology class i , the share allocated to fuel–technology class i is given by the logit function

$$S_i = \frac{\exp(a_i + b_i LC_i)}{\sum \exp(a_j + b_j LC_j)}; \quad j = 1, n.$$

2.6.2.4 Efficiency–Technology Decision Each efficiency choice function is calibrated to historical shipment efficiency data; that is, the model replicates efficiency choices for historical years. The efficiency–technology decision is based on a three-parameter logit function of the form

$$S_i = \frac{\exp(a_i + b_i FC_i + c_i O\&M_i)}{\sum \exp(a_j + b_j FC_j + c_j O\&M_j)}; \quad j = 1, k_i,$$

where FC_i is the first cost of technology i , $O\&M_i$ is the O&M cost of technology i , and k_i is the number of technologies that use fuel i .

2.7 Electricity Market Module

The electricity market module (EMM) represents the capacity planning, generation, transmission and distribution, and pricing of electricity subject to delivered prices for coal, petroleum products, and natural gas; O&M costs for existing and new generation equipment; the costs of renewable fuels for generation; the capital costs of power generation investments; and electricity load shapes and demand. The submodules consist of capacity planning, fuel dispatching, finance and pricing, and load and demand-side management (Fig. 4). In addition, non-utility supply and electricity trade are represented in the fuel dispatching and capacity planning submodules. Nonutility generation from cogenerators and other facilities whose primary business is not electricity generation is represented in the demand and fuel supply modules. All other nonutility generation is represented in EMM. The generation of electricity is accounted for in 13 supply regions, whereas consumption is satisfied in the 9 census divisions.

Operating (dispatch) decisions are provided by the cost-minimizing mix of fuel and variable O&M costs, subject to environmental constraints. Capacity expansion is determined by the least cost mix of all

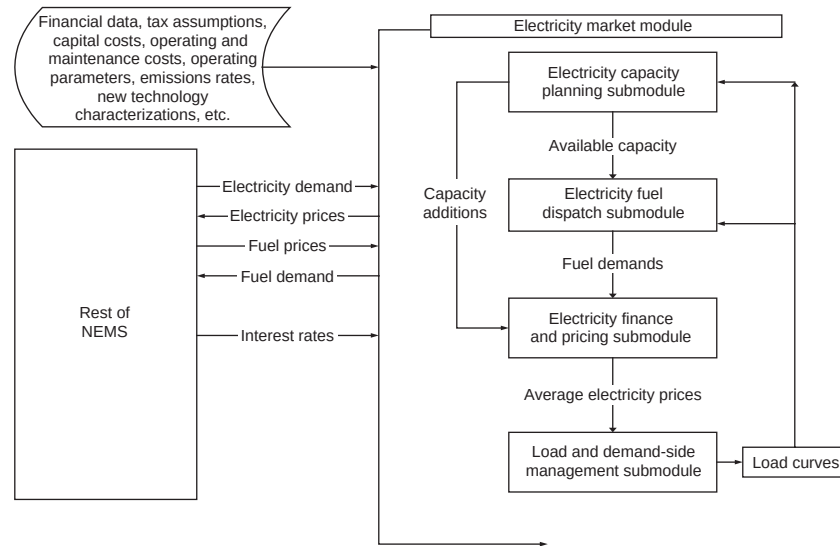


FIGURE 4 Electricity market module overview.

expected costs, including capital, O&M, and fuel, subject to meeting environmental restrictions and expected electricity demand. Construction of generating plants with long lead times is selected with planning horizons up to six periods into the future; the planning horizon can change with respect to the generating technology being considered. Electricity demand is represented by load curves, which vary by region, season, and time of day.

The solution to the submodules of EMM is simultaneous in that, directly or indirectly, the solution for each submodule depends on the solution to every other submodule. A solution sequence through the submodules can be viewed as follows:

- The load and demand-side management (LDSM) submodule processes electricity demand to construct load curves.
- The electricity capacity planning (ECP) submodule projects the construction of new utility and nonutility plants, the level of firm power trades, and the addition of scrubbers for environmental compliance.
- The electricity fuel dispatch (EFD) submodule dispatches the available generating units, both utility and nonutility, allowing surplus capacity in select regions to be dispatched for another region's needs (economy trade).
- The electricity finance and pricing (EFP) submodule calculates total revenue requirements for each operation and computes average and marginal cost-based electricity prices.

2.7.1 Load and Demand-Side Management Submodule

The LDSM submodule generates load curves representing the demand for electricity. The demand for electricity varies over the course of a day. Many different technologies and end uses, each requiring a different level of capacity for different lengths of time, are powered by electricity. For operational and planning analysis, an annual load duration curve, which represents the aggregated hourly demands, is constructed. Because demand varies by geographic area and time of year, the LDSM submodule generates load curves for each region and season.

2.7.2 Electricity Capacity Planning Submodule

The ECP submodule, a dynamic linear programming formulation of capacity expansion decision making, determines how best to meet expected growth in electricity demand given available resources, expected load shapes, expected demands and fuel prices, environmental constraints, and costs for utility and nonutility technologies. When new capacity is required to meet electricity demand, the timing of the demand increase, the expected use of the new capacity, the operating efficiencies, and the construction and operating costs of available technologies determine what technology is chosen.

The timing of the demand increase is important because the construction lead times of technologies differ. The ECP submodule looks up to six periods into the future when identifying new capacity needs. A multiperiod optimization is performed, whereby

capacity choices in each year are made by looking at several years in the future rather than at only a single year. Construction lead times also contribute to uncertainty about investment decisions. Technologies with long lead times are subject to greater financial risk. Compared with plants with shorter lead times, they are more sensitive to market changes in interest and inflation rates and are more vulnerable to uncertain demand projections that determine the need for new capacity. To capture these factors, the discount rate for each technology can be adjusted using risk premiums based on the construction lead time. The risk-adjusted discount rate results in the perception that a technology with a long lead time is less economically attractive than another technology with similar costs but a shorter lead time.

Uncertainty about investment costs for new technologies is captured in the ECP submodule using technological optimism and learning factors. The technological optimism factor reflects the inherent tendency to underestimate costs for new technologies.

The degree of technological optimism depends on the complexity of the engineering design and the stage of development. As development proceeds and more data become available, cost estimates become more accurate and the technological optimism factor declines. Learning factors represent reductions in capital costs due to “learning-by-doing.” For new technologies, cost reductions due to learning also account for international experience in building generating capacity.

The decrease in overnight capital costs due to learning depends on the stage of technological development. The costs of a “revolutionary” technology are assumed to decrease much faster than the costs of mature technologies, ranging from 10% for every doubling of capacity for advanced technologies to 1% for every doubling of capacity for mature technologies.

Capital costs for all new electricity generating technologies (fossil, nuclear, and renewable) decrease in response to foreign and domestic experience. Foreign units of new technologies are assumed to contribute to reductions in capital costs for units that are installed in the United States to the extent that (1) the technology characteristics are similar to those used in U.S. markets, (2) the design and construction firms and key personnel compete in the U.S. market, (3) the owning and operating firm competes actively in the United States, and (4) there exists relatively complete information about the status of the associated facilities. If the new foreign units do not satisfy one or more of these requirements, they are

given a reduced weight or are not included in the learning effects calculation.

Initially, investment decisions are determined in the ECP submodule using cost and performance characteristics that are represented as single-point estimates corresponding to the average (expected) cost. However, these parameters are also subject to uncertainty and are better represented by distributions. If the distributions of two or more options overlap, the option with the lowest average cost is not likely to capture the entire market. Therefore, ECP uses a market-sharing algorithm to adjust the initial solution and reallocate some of the capacity expansion decisions to technologies that are “competitive” but do not have the lowest average cost.

Fossil-fired steam plant retirements are calculated endogenously within the model. Fossil plants are retired if the market price of electricity is not sufficient to support continued operation. The expected revenues from these plants are compared with the annual going-forward costs, which are mainly fuel and O&M costs. A plant is retired if these costs exceed the revenues and the overall cost of electricity can be reduced by building replacement capacity.

Retirement decisions for nuclear capacity are also determined by the model. Four options for the operating license are considered. A unit (1) can be retired early (10 years prior to the end of the operation license), (2) can be retired when the license expires, or can be operated (3) an additional 10 years or (4) an additional 20 years by renewing the license. At each stage, the assumed aging-related expenditures due to capital additions, increased maintenance, and/or performance declines are compared with the cost of replacement capacity. A unit is retired if the aging costs, which are recovered over 10 years, exceed the cost of building new capacity.

The ECP submodule also determines whether to contract for unplanned firm power imports from Canada and from neighboring electricity supply regions. Imports from Canada are computed using supply curves developed from cost estimates for potential hydroelectric projects in Canada. Imports from neighboring electricity supply regions are computed in ECP based on the cost of the unit in the exporting region plus the additional cost of transmitting the power. Transmission costs are computed as a fraction of revenue. After building new capacity, the submodule passes total available capacity to the electricity fuel dispatch submodule and new capacity expenses to the electricity finance and pricing submodule.

2.7.3 Electricity Fuel Dispatch Submodule

Given available capacity, firm purchased power agreements, fuel prices, and load curves, the EFD submodule minimizes variable costs as it solves for generation facility use and economy power exchanges to satisfy demand during each time period and in each region. The submodule uses merit order dispatching; that is, utility, independent power producer, and small power producer plants are dispatched until demand is met in a sequence based on their operating costs, with least cost plants being operated first. Limits on emissions of SO₂ and NO_x (and mercury and CO₂ if such policies are requested) from generating units and the engineering characteristics of units serve as constraints. Coal-fired capacity can cofire with biomass to lower operating costs and/or emissions. During off-peak periods, the submodule institutes load following, which is the practice of running plants near their minimum operating levels rather than shutting them down and incurring shut-off and start-up costs. In addition, to account for scheduled and unscheduled maintenance, the capacity of each plant is derated (lowered) to the expected availability level. Finally, the operation of utility and nonutility plants for each region is simulated over six seasons to reflect the seasonal variation in electricity demand.

Interregional economy trade is also represented in the EFD submodule by allowing surplus generation in one region to satisfy electricity demand in an importing region, resulting in a cost savings. Economy trade with Canada is determined in a manner similar to that of interregional economy trade. Surplus Canadian energy is allowed to displace energy in an importing region if doing so results in a cost savings. After dispatching, fuel use is reported back to the fuel supply modules, and operating expenses and revenues from trade are reported to the EFP submodule.

2.7.4 Electricity Finance and Pricing Submodule

The costs of building capacity, buying power, and generating electricity are tallied in the EFP submodule, which simulates the cost-of-service method often used by state regulators to determine the price of electricity. Revenue requirements shared over sales by customer class yield the price of electricity for each class. Electricity prices are returned to the demand modules. In addition, the submodule generates detailed financial statements.

The EFP submodule also determines “competitive” prices for electricity generation. Unlike cost-of-service prices, which are based on average costs,

competitive prices are based on marginal costs. Marginal costs are primarily the operating costs of the most expensive plant required to meet demand. The competitive price also includes a “reliability price adjustment,” which represents the value consumers place on reliability of service when demands are high and available capacity is limited. Prices for transmission and distribution are assumed to remain regulated, so the delivered electricity price under competition is the sum of the marginal price of generation and the average price of transmission and distribution.

2.7.5 Emissions

The EMM tracks emission levels for SO₂, NO_x, and mercury. Facility development, retrofitting, and dispatch are constrained to comply with the pollution constraints of the CAAA of 1990 and other pollution constraints. An innovative feature of this legislation is a system of trading emissions allowances. The trading system allows a utility with a relatively low cost of compliance to sell its excess compliance (i.e., the degree to which its emissions per unit of power generated are below the maximum allowable levels) to utilities with a relatively high cost of compliance. The trading of emissions allowances does not change the national aggregate emissions level set by the CAAA, but it does tend to minimize the overall cost of compliance.

3. ANALYSIS OF A 20% NON-HYDROELECTRIC PORTFOLIO STANDARD

Concerns over the possibility that climate change may be caused by anthropogenic activities, particularly the combustion of fossil fuels, have raised interest in examining a series of policy options that may inhibit or reverse the growth of energy-related carbon emissions. Recently, a number of bills have been introduced by the U.S. Congress that would simultaneously reduce emissions of NO_x, SO₂, mercury, and CO₂ from power generators. Two of the more recent policy proposals include Senate bill S. 1766 and House bill H.R. 4. These analyses were developed using the AEO 2002 version of the NEMS. Other related analyses have been performed at the request of the House Committee on Government Reform and the Senate Committee on Environment and Public Works. These may be viewed or downloaded from the EIA Web site.

The policy considered in this article is a 20% non-hydroelectric RPS. A typical RPS requires that a share of the power sold must come from qualifying renewable facilities. Companies that generate power from qualifying renewable facilities are issued credits that they can hold for their own use or sell to others. To meet the RPS requirement, each individual electricity seller must hold credits—issued to qualifying renewable facilities or purchased from others—equal to the share required each year. For example, a supplier of 10 TWh of retail electricity sales during a year with a 10% RPS requirement would have to hold 1 TWh of renewable credits. In a competitive market, the price of renewable credits would increase to the level needed to meet the RPS requirement. The RPS provides a subsidy to renewable generators (from nuclear, coal, natural gas, oil, and hydroelectric generators) to make them competitive with other resource options while allowing the market to determine the most economical renewable options to develop.

3.1 Assumptions of the RPS

In the 20% RPS case, we assumed the *AEO 2002* reference case assumptions in all areas except the following:

- A non-hydroelectric RPS of 20% by 2020 was assumed, growing linearly from approximately 2.5% in 2003 to 20% by 2020. A renewable credit trading system was assumed. Sellers of electricity must hold renewable credits that are equal to the minimum renewable portfolio fraction of generation assumed for each year.
- Because a renewable portfolio policy implies greater public acceptance and, as a consequence, lower legal and other preparation costs, the RPS incorporates the assumption that wind system costs will be lower, and the maximum allowable generation from intermittent technologies (e.g., wind, photovoltaics [PV]) was raised from 15 to 20% of total generation in the RPS case. (Many electricity market experts assert that even 15% intermittent power generation is too high without backup due to reliability. As a rough rule of thumb, intermittent power cannot exceed the reserve margin of the power control center.)
- All existing state and federal programs or mandates for renewable power generation are assumed to adhere to existing legislation and are not included in the costs attributed to federal RPS programs.

3.2 Results

In 2003, 12 states had RPSs, targets, or other incentive programs. Without the current federal and state wind programs, grid-connected wind systems or other intermittent renewable generation technologies are not expected to be economical in the United States over the midterm through 2020. Niche markets are another matter because electricity options are often quite limited.

Approximately 4.8 GW of capacity is expected to result from state RPS programs, and 2.2 GW is expected to result from other state-sponsored stimulus programs, over the 2001–2020 time frame. However, because of these capacity additions, technological learning is expected to result in cost reductions for wind; the projected lower wind costs become competitive with conventional fossil-fueled technologies in some regions of the United States during the 2015–2020 period. Renewable capacity built under these programs reduces the incremental quantity of renewable generation needed to comply with the federal RPS program.

3.3 Fossil Fuel Use and Electricity Market Impacts of the RPS

The penetration of renewable generation technologies reduces the construction of the more efficient gas combined cycle and coal gasification generation technologies that would have been built in the reference case, thereby reducing the overall stock efficiency of fossil-fueled electricity generation plants. In 2020, the RPS case is projected to build approximately 70 GW less combined cycle and 26.8 GW fewer new advanced coal units. Renewable generation capacity is projected to increase by approximately 166 GW, much of which is intermittent wind capacity (~99 GW above the reference case) (Fig. 5).

Fossil fuel consumption in 2020 is expected to be approximately 5.6% lower in the RPS case than in the *AEO 2002* reference case—down from 120.9 to 114.2 EJ. For electricity generation in 2020, fossil fuel consumption is expected to be approximately 19% lower than in the reference case—down from 37.7 to 30.4 EJ.

The RPS electricity production is expected to be slightly lower in 2020 than in the reference case (~30 TWh or 0.6%) due to the higher electricity prices that result from the RPS. However, electricity generation from non-hydroelectric renewables is expected to increase from approximately 160 TWh

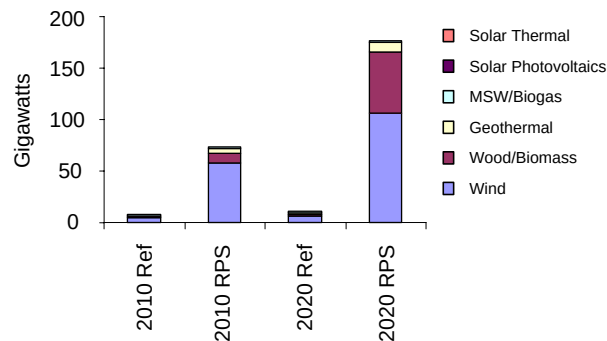


FIGURE 5 Renewable capacity additions: Reference versus RPS case. MSW, municipal solid waste; Ref, reference.

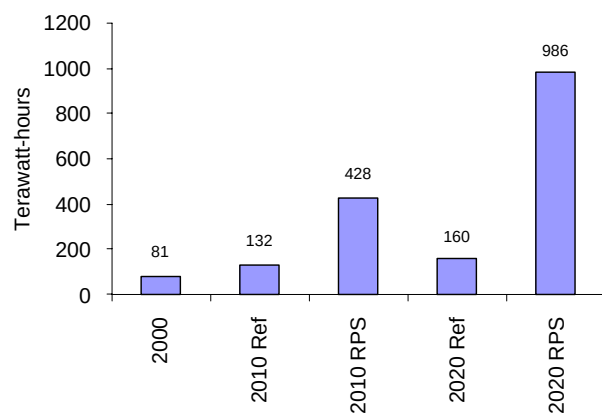


FIGURE 6 Non-hydroelectric renewable generation. Ref, reference.

in the reference case in 2020 to approximately 986 TWh in the RPS case (Fig. 6). The primary contributors to non-hydroelectric renewable generation in the RPS case in 2020 are expected to be biomass generation from both dedicated plants and cofiring with coal (~ 476 TWh), wind systems (~ 335 TWh), and geothermal (~ 82.5 TWh).

3.4 GHG Emissions

NEMS projects only CO₂ emissions on a systemwide basis for the United States. In 2020, carbon emissions are projected to be approximately 6% lower in the RPS case than in the AEO 2002 reference case (Fig. 7). However, carbon emissions from electricity generation are expected to be approximately 16.8% lower in the 20% RPS case than in the reference case despite meeting the 20% RPS target in 2020. Consequently, per capita carbon emissions are projected to decline by 0.4 metric tons ($\sim 5.5\%$).

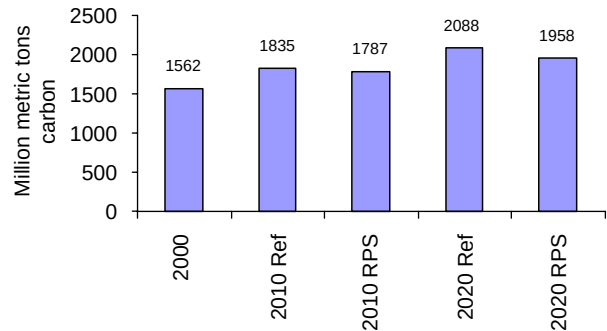


FIGURE 7 Carbon emissions: Reference versus RPS. Ref, reference.

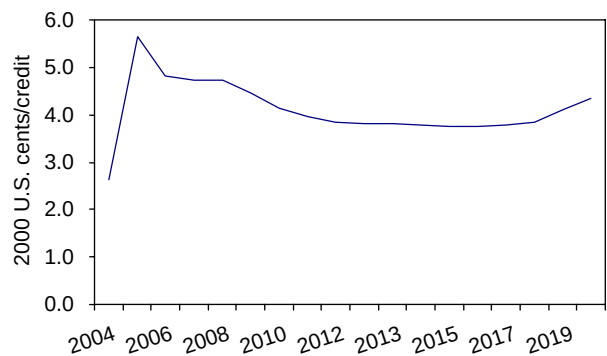


FIGURE 8 Renewable generation credit prices.

3.5 Costs of the 20% Renewable Policy

Imposition of an RPS on the U.S. generation markets is projected to have a relatively mild adverse affect on delivered electricity prices but at a significant cost to the electricity industry. Electricity prices in 2020 in the RPS case are projected to be approximately 3% higher than in the reference case as a result of the costs that were added to electricity prices to conform to the RPS. The reduction in natural gas prices in the RPS case in 2020 relative to the reference case mitigates against the price increases that would otherwise have occurred due to the renewable credit prices paid and the additional capital invested in renewable generation technologies. However, the cost to the electricity industry over the next 18 years ranges between \$38 billion and \$59 billion (in 2000 U.S. dollars), using a discount rate ranging from 0 to 10%.

As the industry ratchets up its use of renewables to 20% by 2020, significant issues emerge regarding the ability of the U.S. renewable resources to continue to expand as evidenced by the rising renewable credit prices and the slowing expansion rate of wind and geothermal capacity (Fig. 8). For

example, geothermal resources are projected to have used more than 50% of all remotely competitive sites, although an additional 10 GW of capacity is possible at significantly inferior and costlier sites. Wind expansion in 2020 is expected to begin to be limited by the amounts of cost competitive wind resources remaining in regions that are sufficiently close to demand and transmission centers to allow further expansion. Although wind capacity is projected to reach 109 GW by 2020 under a 20% RPS, expansion beyond 150 GW is likely to be very expensive. The rapid growth in biomass generation eventually is likely to be limited by the competition between agricultural uses and the generation of an energy feedstock for biomass power generators. Furthermore, dedicated biomass gasification plants must be built within a 50-mile radius of their resources to be cost-effective; otherwise, transportation costs are likely to make biomass generation very expensive (by raising renewable credit prices).

3.6 Conclusions

A 20% renewable portfolio standard for the United States is expected to increase total consumer costs of electricity by approximately 3%. Although this does not appear to be significant on a national level, the regional distributional price effects can be quite significant. For example, producers in regions rich in coal- or gas-based generation are likely to experience much larger revenue reductions than will those that are rich in renewable resources. The RPS is likely to significantly increase the costs to the power generation industry, from \$38 billion to \$59 billion, for the period ending in 2020.

Whether the benefits of a 20% RPS outweigh the costs is a matter of considerable policy debate within the United States. The answer clearly depends on how costs and benefits are perceived and whether they can be measured at all.

4. OVERVIEW OF MARKAL

4.1 Purpose of MARKAL

MARKAL stands for market allocation. The model's original purpose was to evaluate energy technologies in a comprehensive and integrated framework. Since its initial implementation during the early 1980s, MARKAL has undergone several significant enhancements that have extended its scope to the computation of a competitive, dynamic, energy

supply-demand equilibrium. The model represents all stages in the energy market—extraction, trade, conversion, and end use consumption—with each sector selecting from an array of technologies that produce and/or consume energy forms. The technologies, together with their input and output energy carriers, constitute a reference energy system (RES) of the regional energy system being modeled (local regions, countries, or provinces).

MARKAL's original capabilities as a tool for an economic assessment of technologies has been enhanced by the addition of new features to the model, including the ability to simulate endogenous technological learning, the inclusion of interregional trade variables, and the capability to represent major uncertainties.

MARKAL evaluates the economics of a technology in the context of the configuration of the entire energy system. This fact renders it difficult, if not impossible, to evaluate a single technology in isolation because its profitability depends on the presence or absence of competing technologies in all parts of the energy system.

MARKAL may be used in two distinct modes. In the first mode, a perfectly competitive energy market is simulated, where all economic agents (suppliers and producers) make their investment and operating decisions under perfect information and perfect foresight and minimize long-term energy system costs using a single discount rate. This mode is particularly useful for revealing the socially optimal set of technologies and for identifying those technologies that should be targeted to achieve the optimal technological mix over the long run. In the second mode, several market imperfections may be introduced to bring the model closer to a short- to medium-term forecasting tool. For instance, the assumptions of perfect information and of perfect foresight may be replaced by the introduction of uncertainty via stochastic programming or by using the model in a time-stepped fashion to simulate decision making with imperfect foresight. In addition, different discount rates may be assumed for different sectors (and even different technologies), thereby simulating the observed differences in behavior among economic agents. This forecasting mode is better suited to the short to medium term when simulating energy market agents under imperfect information.

4.2 MARKAL Model Structure

The MARKAL model minimizes energy system costs over a specific time horizon, subject to satisfying all of

the energy service demands and environmental or fuel use restrictions. This objective can also be viewed as minimizing the net social cost (or maximizing the social surplus) of the energy system, while satisfying a number of constraints, over the horizon.

We now proceed to flesh out several terms in the preceding description. To configure means to decide which technologies to invest in (and/or abandon) at different dates, which ones to operate (and at what level), how much of each energy form to produce during each period, how much of each energy form to buy and/or sell to other regions, and so on. In addition to time periods (which are usually 5 or 10 years long), MARKAL divides electricity demand into three seasons (winter, summer, and spring/fall) and two diurnal divisions (day and night). These time divisions result in six so-called time slices. These time slices are recognized only for technologies producing electricity or low-temperature heat, neither of which may be easily stored and so require a finer time disaggregation than do other energy forms. As a result, these two energy forms are disaggregated into time slices for each time period.

The energy system of a region includes energy supply sources such as mining, imports, and exports as well as processing, conversion of the various energy forms, and their transport (including inter-regional transport), distribution, and end use consumption by all sectors of the economy. These activities are represented in MARKAL by means of technologies, which consume and/or produce energy forms and/or materials (collectively termed commodities).

The end use technologies produce special kinds of commodities that are physical products or services for which demands are specified.

The set of technologies, demands, sources, and commodities determine the topology of the “reference energy system” of the region modeled. Figure 9 is a simplified representation of a typical MARKAL RES showing the five broad components usually recognized in each model instance: primary energy resources (SRC), energy conversion into electricity or low-temperature heat (CON), other energy processing (PRC), and energy end uses (DMD), with the right-most oval in the figure representing the demands (DM) for energy services and products.

Each “region” may be a country, a group of countries, or a province/state within a country. Multiregional MARKAL models have been created with a wide variety of geographical scopes, going from a relatively small community to the whole world divided into 15 regions. The regions are

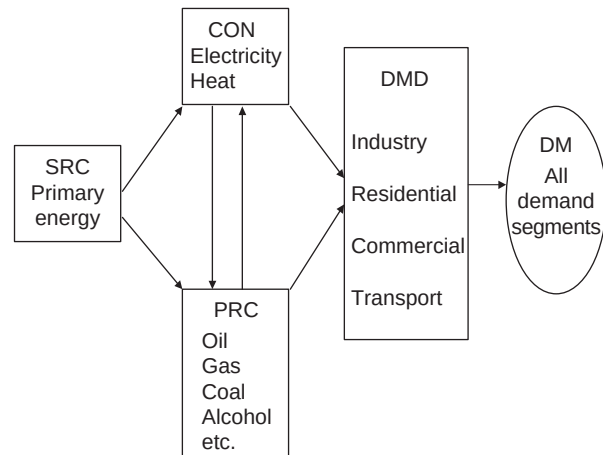


FIGURE 9 Simplified reference energy system. See text for abbreviations.

interconnected by technologies that transport energy forms (e.g., transmission lines, pipelines).

The “horizon” comprises at most nine periods, each having equal duration (usually 5 or 10 years, as chosen by the modeler). At least one of the initial periods is a historical period over which the model has no freedom and for which all of the quantities of interest are fixed to their historical values. This calibration to an initial period is one of the important tasks required for setting up a MARKAL model for a given region. The main variables that must be fixed are the capacity and operating levels of all technologies as well as extraction, exports, and imports for all energy forms and materials. Note that the initial period’s calibration also influences the model’s decisions over several future periods because the “profile of residual capacities” (i.e., capacities inherited from a historical period) is fully specified over the remaining lives of the technologies existing at the start of the model forecast horizon.

MARKAL minimizes the “total energy system cost,” which includes the following elements: investments, fixed and variable annual O&M costs, commodity import costs, minus export revenues, and demand losses incurred from reduced product and service demands. To correctly account for the remaining value of all equipment that is still operating at the end of the horizon, a salvage value (i.e., residual value) of all such equipments is computed and subtracted from the total cost. This is an important feature; without salvage value, the investment decisions made by the model would be severely distorted, especially toward the end of

the planning horizon. Optionally, the user may also specify taxes and subsidies on some commodities (e.g., an emission tax on some pollutant). Because the various costs are incurred at different times over the horizon, each cost is first discounted to a fixed year before being added to the total cost. The same or different discount rates may be used for the various sectors and regions.

The constraints imposed by MARKAL are many. The main ones are as follows:

- *Satisfaction of demands.* A reference demand scenario is provided by the user, specifying the reference demand trajectories in all subsectors. These demands are set for the reference case but may later be modified by MARKAL if some alternate scenario is such that it alters the prices of end use demands (demands can be price elastic in MARKAL). Note also that prices are computed endogenously by MARKAL for all commodities and end use demands.

- *Conservation of investments.* If the model decides to invest in a piece of equipment at a certain period, the capacity is increased accordingly for the life of the equipment. At the end of that life, the capacity is decreased by the same amount (unless the model decides to extend the life of the equipment by investing in a life extension technology). While computing the available capacity at some time period, the model takes into account the capacity resulting from all surviving (unretired) investments up to that period. Some of those investments may have been made prior to the initial period and remain in operating condition (embodied by the residual capacity of the equipment), whereas other investments may have been made by the model at or after the initial period. However, the model is not forced to use all of the available capacity.

- *Use of capacity.* At each period, the model may use some or all of the available capacity in that period times the availability factor of the technology. In some cases, the model may decide to use less than the available capacity at certain time slices or even throughout the whole period. In other words, some capacity may be inactive during some time periods. Of course, this will occur only if such a decision contributes to minimizing the overall cost. Optionally, there is a provision for the modeler to force specific technologies to use their capacity.

- *Electricity balance.* At each period, during each time slice, and in each region, electricity produced plus electricity imported (from other regions) must be at least as much as electricity consumed plus electricity exported (to other regions) plus grid

losses. A similar balance exists for low-temperature heat.

- *Balance for other energy forms.* This is the same as for electricity, but because time slice plays no role, there is only an annual balance equation.

- *Electricity peaking reserve constraint.* This constraint requires that at each time period and for each region, total available capacity exceeds the average load of the peaking time slice by a certain percentage. This percentage is called the peak reserve factor and is chosen to reflect the extra load at peak as well as uncertainties regarding electricity supply that may decrease capacity in an unpredictable way (e.g., water availability in a reservoir, unplanned equipment downtime). The peak reserve factor is typically set between 0.20 and 0.50. The peaking time slice is defined as the time slice when load is heaviest (it may be winter day in cold countries, summer day in warm countries, etc.).

- *Emissions constraint(s).* The user may impose on the whole system upper limits on emissions of one or more pollutants. The limits may be set for each time period separately to simulate a particular emissions profile (also called emission target).

- *Base load (electricity generation only).* The user may identify which technologies should be considered as base load technologies by MARKAL, that is, those whose operation must not fluctuate from day to night during a given season. The user may also specify what is the maximum fraction of night production that may be supplied from all base load technologies.

- *Seasonal availability factors (electricity sector only).* The user may specify seasonal and even day/night limitations on the use of the installed capacity of some equipment. This is especially needed when the operation of the equipment depends on the availability of a resource that cannot be stored, such as wind or sun, or that can be partially stored, such as water in a reservoir.

The Canadian incarnation of MARKAL has the following specific features:

- It is composed of 14 linked MARKAL modules, one for each Canadian province and territory plus a U.S. model. All 14 modules are linked by a large number of energy trade variables as well as permit trading variables when required.

- Each provincial module contains full technological descriptions of energy extraction, electricity production, oil and gas processing, coal extraction, industrial processes of all energy-intensive industries, transportation, and residential and commercial end

uses. A typical provincial module contains more than 2000 technologies.

- The demand segments number more than 50 in each provincial module.
- There are more than 500 energy trade links among the 13 Canadian jurisdictions and the U.S. module, encompassing detailed representation of the oil and gas pipelines, electricity interties, and coal and refined petroleum products trading.

4.3 Representation of Energy Market Interactions

MARKAL maximizes the long-term net energy-related social surplus (i.e., the sum of energy-related producers and consumers surpluses) of a state, province, country, or group of countries. Equivalently, the model computes a least cost partial equilibrium on energy markets (i.e., the prices and quantities such that all quantities produced are consumed at the computed market clearing prices). In MARKAL, all energy forms (quantities by fuel type) and their prices are endogenous (computed by the model) and result from the optimization. The demands for products and services are also endogenous but in a different fashion; their price elasticities are chosen by the user, and these choices in turn determine the demand levels.

Mathematically, MARKAL optimizes a linear objective function subject to various constraints. The economic significance of the result of the optimization is twofold. First, the primal solution consists of the investment levels, capacities, and operating levels for all technologies as well as the amounts of each energy form extracted, traded, produced, and consumed by each technology. Second, the dual solution provides shadow prices for all constraints. Three types of shadow prices are of particular practical interest: those of the energy balance constraints (which provide prices for all energy forms), those of the emissions constraints (which provide the price of each pollutant whose emissions are capped), and those of the demand constraints (which provide a price for each demand for a good or service).

4.4 Technology Representation

Each technology is explicitly represented by its technical and economic parameters as follows:

- Emission of each pollutant per unit of technology
- Annual availability factor of the technology
- Seasonal availability factors (electricity- and heat-producing technologies only)
- Technical life duration
- First period of availability
- Time profile of existing capacity (residual capacity at initial period)
- Investment cost per unit of capacity
- Fixed O&M cost per unit of capacity
- Variable O&M cost per unit of capacity
- Delivery cost per unit of each input into the technology
- Upper and/or lower bounds on capacity, on investment, and on operating level

Note that each parameter, except life duration and period of first availability, may be specified for each time period desired.

5. EXAMPLE APPLICATION OF THE CANADIAN MARKAL MODEL

In 1999, the Canadian government embarked on a systematic study of all aspects of climate change in Canada. This multistage process, named National Climate Change Implementation Process (NCCIP), directly involved a large number of stakeholders regrouped in 15 issue tables, charged with proposing a large array of measures and actions to reduce GHG emissions. The output from these issue tables was collected by the Analysis and Modeling Group (AMG), which was responsible for integrating the issue tables' proposals into a set of coherent strategies using several energy-economy models, including MARKAL. The MARKAL model was selected as one of two energy technology models to provide an integrated analysis of several paths to Kyoto, that is, to analyze the economic impacts on the Canadian economy of various ways of reducing GHG emissions. The results of the two energy technology models were then injected into the macroeconomic model TIM/RIM to complete the economic analysis with macroeconomic indicators.

The NCCIP was divided into two phases. During phase I, several paths to Kyoto were examined under fairly general economic instruments. During phase II, a restricted number of paths were examined more closely and the market instruments were made more precise.

5.1 Scenarios and Paths to Kyoto

5.1.1 Phase I of the AMG

Three scenarios were considered in addition to the baseline:

- *Canada acts alone (CA) scenario.* In this scenario, only Canada implements GHG reduction strategies, whereas the rest of the world does not. Therefore, the oil price remains the same as in the base case, and energy trade with the United States is negatively affected by Canada's GHG policies.

- *Kyoto-loose (KL) scenario.* Annex I countries (including the United States) implement Kyoto-like GHG reductions, and an international permit trading system is in place. The permit price is low (\$25/metric ton [tonne] CO₂). Canadian electricity and gas trade is positively affected, but oil trade is negatively affected.

- *Kyoto-tight (KT) scenario.* The situation is the same as in the KL scenario except that the permit price is high (\$50/tonne CO₂). Gas and electricity trade is positively affected, and oil trade is negatively affected but to different degrees than in the KL scenario.

A path describes the components of a GHG reduction strategy. Five paths were defined and simulated in the AMG phase I:

- *Path 0.* Force in all of the issue tables' measures without any imposed emissions target.
- *Path 1.* Impose a 1990–4.33% cap on each sector separately.
- *Path 2.* Impose a 1990–4.33% cap on the entire Canadian economy.
- *Path 3.* This is a variant of path 2 where the cap is imposed on approximately 90% of the Canadian emission sources, with the rest being treated via dedicated measures and actions.
- *Path 4.* This is a variant of path 3 where only approximately 40% of the GHG sources are covered by the cap, with the rest being treated via dedicated measures.

Path 0 is not really a path to Kyoto given that there is no guarantee that the desired target will be reached. The issue tables' measures are made available to the model but are not imposed in paths 1 to 4. In paths 1 to 4, only the sectors covered by permits could buy or sell permits from the international market. Path 2 is a 100% efficient path that achieves the target at minimum cost.

TABLE II

Cases Treated during Phase I

	Path 0	Path 1	Path 2	Path 3	Path 4
Scenario					
Canada acts alone	Y	Y	Y	Y	Y
Kyoto-tight			Y		Y
Kyoto-loose			Y		Y

Note. Y, yes.

The scenarios and paths potentially create 15 combinations. Of these, 9 were simulated during phase I plus the base case, as shown in Table II.

5.1.2 Phase II of the AMG

The CA scenario was dropped, and it was assumed that the Kyoto treaty was in effect and ratified by all industrialized countries except the United States.

Two carbon price scenarios were considered (high and low prices for permits). The paths considered (now called cases) were as follows:

- *Case 1.* This is similar to path 3 (maximum coverage of the economy). Permits were not allocated at all. Emitters must buy them at the prevailing carbon price.

- *Case 2a.* This is similar to path 4 (partial coverage of the economy). Permits were allocated gratis to covered industries in proportion to their emissions in 1990. The proportionality coefficient is calculated to achieve the Kyoto target.

- *Case 2.* This is similar to path 4 (partial coverage of the economy). Permits were allocated gratis but proportionally to the output of each industry in 2010.

- *Case 3.* This is similar to path 4 (partial coverage of the economy). Permits were allocated gratis according to a so-called triptych formula by sector and by province. The multiple-criteria formula was devised to favor development of some sectors and provinces while taking into account each sector/province's endowment in energy forms.

In addition to these, two sensitivity scenarios were simulated, both with case 2 only: one with an intermediate CO₂ price and the other with a supplementarity condition imposing a limit on the amounts of international permits that Canada was allowed to purchase. Table III shows the resulting 10 combinations that were simulated during phase II.

TABLE III
Cases Treated during Phase II

Carbon price	Case 1	Case 2	Case 2a	Case 3
High (\$50)	Y	Y	Y	Y
Intermediate (\$25)		Y		
Low (\$10)	Y	Y	Y	Y
Supplementarity constraint		Y		

Note. Y, yes.

5.2 Some Results from the Analysis

5.2.1 Main Conclusions for Phase I

5.2.1.1 CA Paths

- Path 0 did not achieve the Kyoto target (fell short by more than 40%), and its total cost was quite large.
- Path 1 nearly achieved the Kyoto target (short-fall of 10%), with a total cost 2.5 times that of path 2.
- Path 2 (efficient path) achieved the Kyoto target at a net present value cost equal to \$25 billion (the smallest of all paths).
- Path 3 (broad coverage) achieved the Kyoto target at a cost reasonably close to that of path 2. Because path 3 is more likely to be implementable in practice, it becomes a good candidate for further analysis and potential implementation.
- Path 4 (large final emitters coverage) achieved the Kyoto target at a cost closer to that of path 1 than to that of path 2. Therefore, narrow permit coverage is much less efficient than broad coverage.

5.2.1.2 KT and KL Paths When international permits are available, the overall Kyoto costs are significantly smaller than when Canada acts alone. In the KT path, Canada actually incurs negative costs when implementing the Kyoto GHG reductions. This is due to the increased energy trade with the United States (gas and electricity). In the KL path, overall costs for Canada are close to zero.

5.2.2 Main Conclusions for Phase II

- The overall cost for Canada of achieving the Kyoto target is very sensitive to the price of carbon rights, but it is much less sensitive to the case considered. For instance, for the high carbon price,

the net present value varies from +\$12 billion to +\$18 billion, depending on the case analyzed. For the low carbon price, the net present value stays in the range of −\$1 billion to −\$2 billion.

- However, the sectoral and provincial detailed costs vary quite a bit from case to case. Cases 1 and 2a show heavy burdens on industries, whereas cases 2 and 3 succeed in smoothing the costs better across sectors. This is due to the output-based allocation of permits adopted in these two cases.

- The imposition of a 50% limit on permit acquisition proves to be quite costly for Canada. In the low carbon price (case 2), Canadian costs jump from −\$2 billion to +\$5 billion.

- The triptych formula partially succeeds in alleviating the burden for the targeted provinces. It is inferred that some additional refinements of the formula would achieve a better repartition of the Kyoto burden than in case 2.

SEE ALSO THE FOLLOWING ARTICLES

Bottom-Up Energy Modeling • Computer Modeling of Renewable Power Systems • Decomposition Analysis Applied to Energy • Input–Output Analysis • Modeling Energy Markets and Climate Change Policy • Modeling Energy Supply and Demand: A Comparison of Approaches • Multicriteria Analysis of Energy • Net Energy Analysis: Concepts and Methods

Further Reading

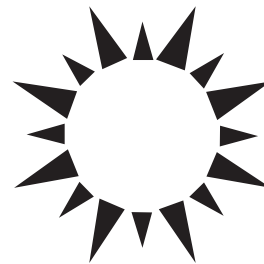
- Adams, D. M., Alig, R. J., Callaway, J. M., and McCarl, B. A. (1996). "The Forest and Agricultural Sector Optimization Model (FASOM): Model Structure and Policy Applications, USDA Forest Service Report PNW-RP-495." U.S. Department of Agriculture, Washington, DC.
- Edmonds, J. A., Pitcher, H. M., Barns, D., Baron, R., and Wise, M. A. (1995). Modeling future greenhouse gas emissions: the second generation model description. In "Modeling Global Change" (L. R. Klein and F. C. Lo, Eds.), pp. 295–340. United Nations University Press, Tokyo.
- Energy Information Administration (1994). "Annual Energy Outlook 1994, with Projections to 2010," DOE/EIA-0383. EIA, Washington, DC.
- Energy Information Administration (1998). "Impacts of the Kyoto Protocol on U.S. Energy Markets and Economic Activity," SR/OIAF/98-03. EIA, Washington, DC.
- Energy Information Administration (1999). "The Comprehensive Electricity Competition Act: A Comparison of Model Results," SR/OIAF/99-04. EIA, Washington, DC.
- Energy Information Administration (2001). "Analysis of Strategies for Reducing Multiple Emissions from Electric Power Plants:

- Sulfur Dioxide, Nitrogen Oxides, Carbon Dioxide, and Mercury and a Renewable Portfolio Standard,” SR/OIAF/2001-03. EIA, Washington, DC. www.eia.doe.gov/oiaf/servicerpt/epp/index.html.
- Energy Information Administration (2001). “Analysis of Strategies for Reducing Multiple Emissions from Power Plants with Advanced Technology Scenarios,” SR/OIAF/2001-05. EIA, Washington, DC. www.eia.doe.gov/oiaf/servicerpt/epp/index.html.
- Energy Information Administration (2001). “The National Energy Modeling System: An Overview,” DOE/EIA-0581. U.S. Department of Energy, Washington, DC.
- Gabriel, S., Kydes, A. S., and Whitman, P. (2001). The National Energy Modeling System: A large-scale energy-economic equilibrium model. *Operations Res.* 49(1).
- Hillier, F. S., and Lieberman, G. J. (1990). “Introduction to Operations Research.” McGraw-Hill, New York.
- Loulou, R., and Lavigne, D. (1996). MARKAL model with elastic demands: application to greenhouse emissions control. In “Operations Research and Environmental Management” (C. Carraro and A. Haurie, Eds.), pp. 201–220. Kluwer Academic, Dordrecht, Netherlands.
- Manne, A., Mendelsohn, R., and Richels, R. (1995). MERGE: A model for evaluating regional and global effects of GHG reduction policies. *Energy Policy* 3(1).
- Murphy, F. H. (1983). An overview of the Intermediate Future Forecasting System. In “Energy Modeling and Simulation” (A. S. Kydes, *et al.*, Eds.), pp. 67–73. North-Holland, Amsterdam, Netherlands.
- Murphy, F. H., Conti, J. J., Shaw, S. H., and Sanders, R. (1988). Modeling and forecasting energy markets with the intermediate future forecasting system. *Operations Res.* 36, 406–420.
- National Research Council (1992). “The National Energy Modeling System.” National Academy Press, Washington, DC.
- Nordhaus, W. D., and Yang, Z. (1996). A regional dynamic general equilibrium model of alternative climate change strategies. *Am. Econ. Rev.* 86, 741–765.

National Energy Policy: Brazil

SERGIO V. BAJAY

State University of Campinas
Campinas, São Paulo, Brazil



-
1. Organization of the Brazilian Energy Supply Industry
 2. The National Council for Energy Policy
 3. Fostering Energy Supply
 4. Energy Efficiency and Research and Development Programs
 5. Energy Prices and Social Issues
 6. Energy and the Environment
 7. An Integrated Approach

Glossary

cogeneration The simultaneous production of power (either electrical or mechanical) and useful heat (e.g., process steam) using a single fuel source.

energy service company A business that implements energy conservation measures for its customers and is paid by them part of the corresponding cost savings.

firm power Continually available power, or power that is available for a large, prespecified, part of the time.

“free” electricity or gas consumer A consumer free to choose a supplier of electricity or gas, in contrast to a “captive” consumer.

independent power producer An electrical power producer that is not a generation utility, i.e., does not have a concession contract and is not regulated. Independent producers compete, at their own risk, with other producers and, sometimes, with generation utilities in a power supply market.

indicative forward planning Prospective studies, carried out or contracted by government bodies, indicating possible expansion paths and providing guidance about future investment needs to interested agents.

Some of the current national energy policies in Brazil were adopted during the two terms of President Fernando Henrique Cardoso, from 1995 to 2002. Other policies that were set up earlier have evolved over time to their present form. Since the middle of 2003, under the government of President Luis Inácio Lula da Silva, who took office in January 2003, there

have been proposals to change some of these energy policies. Potential shifts that may impact the national energy policy of Brazil include institutional and managerial changes in the country’s energy supply industry; the fostering of some types of energy supply development and demand-side management programs; the formulation of energy price regulations; tariff making, with cross-subsidies for some large consumers; the granting of subsidies to the poor for fuel and electricity purchases; increasing the cross-links between energy and environmental policies; and integrating the approach to energy policymaking and forward planning. In this latter matter, the roles of the Ministry of Mines and Energy and the National Council for Energy Policy are of paramount importance.

1. ORGANIZATION OF THE BRAZILIAN ENERGY SUPPLY INDUSTRY

From the 1940s and through the 1950s and 1960s, the federal government of Brazil, with the help of the state governments, undertook the charge of assuring, through state-owned companies, the supply of most of the electricity, oil, and gas consumed in the country. A state monopoly for the production, importation, processing (with the exception of private refineries existing at that time), and transportation of oil and gas was defined by the 1953 mandate, Law No. 2004, and was granted to Petrobras, a federal-government-owned company created for the purpose. Distribution and retail trade of oil products were kept out of the monopoly, instead being shared between BR, a subsidiary of Petrobras, and large transnational oil supply companies such as Shell, Exxon, and Texaco. Some Brazilian states formed state-owned companies to distribute and trade initially town gas and later natural gas.

The Brazilian government created Eletrosul, Furnas, Chesf, and Eletronorte to generate and transmit electricity for, respectively, the southern, southeastern/midwestern, northeastern, and northern regions of the country, putting all of these entities under the control of a holding company called Eletrobras. Chesf also sells electricity for some very large consumers in the northeastern region, and Eletronorte, besides doing the same in the northern region, also provides distribution services in the capitals of the northern states. All of the state governments formed electrical energy distribution and retail supply companies; some of them, such as Cesp (in the state of São Paulo), Cemig (in the state of Minas Gerais), and Copel (in the state of Parana), also generate and transmit power. At a later stage, the federal government acquired control of two large electricity distribution utilities, Light and Escelsa, which supply areas located in the states of Rio de Janeiro and Espírito Santo, respectively.

Petrobras and Eletrobras, the latter with the cooperation of all of the large state-owned electricity supply utilities in the country, did the required forward (expansion) and operation planning exercises for the oil and gas and electrical power supply industries, respectively, and proposed the associated energy policies for the Ministry of Mines and Energy. This scheme, of having state-owned companies as the main operators in the Brazilian energy supply industry, involving both federal and state governments, succeeded up to the mid-1980s, when a number of policy positions cast shadows on the adequacy of the scheme for the future. The missteps included (1) the artificially low tariffs for electricity (mirroring most public service tariffs imposed by the federal government, in often vain efforts to control high inflation rates) and (2) the political misuse of electricity supply and gas distribution utilities (involving incompetent and often corrupt management and the initiation of construction of several plants, particularly electric power stations, primarily to reap political benefits to some politicians, but without the necessary funding to finish them on schedule), coupled with the desire of the federal government to have substantial and fast increases in domestic production of oil and gas. Discussions about what institutional changes should be made to correct the problems dragged along through several governments and lasted nearly a decade, up to the time when a deep financial crisis in the electricity supply industry required urgent action. President Fernando Henrique Cardoso, in the beginning of his first term in office, decided to sell all of the federally owned

electrical power utilities to private investors and also to exert political and economic pressure on state governments to do the same. At the same time, two amendments to the Brazilian Constitution were passed in congress, ending the legal monopoly of Petrobras and allowing the state governments to grant concessions to investor-owned gas distribution utilities and not just to state-owned ones, as was the case before this change. The control of Petrobras continued to be in the hands of the federal government, but Petrobras was expected to compete with private companies in the production, importation, processing, and transportation of oil and gas, allowing, according to the government's wishes, lower prices in the market and substantial increases in the domestic production of these commodities.

At the end of President Cardoso's second term (December 2002), some competition was achieved in domestic exploration; four bidding rounds were carried out to grant exploration and production (EP) licenses, but no licenses have been granted for production of oil and gas (no large findings have occurred outside of Petrobras' EP areas) and few licenses were granted for importation of natural gas. No competition has occurred in the processing and transportation of oil and gas. Several transnational companies operating in this industry have preferred so far to set up joint ventures with Petrobras, rather than to challenge a competitor with such market power as Petrobras still has in Brazil.

Privatization in the electricity supply industry has occurred to a much more limited extent than was planned initially by the government. Around 70% of the distribution capacity was actually privatized but less than 30% of the generation capacity went to private ownership (the large-scale generation and transmission utilities Furnas, Chesf, Eletronorte, Cemig, and Copel continue to be state owned). This partial failure of President Cardoso's government plans was caused by strong political opposition to the privatization of these utilities, not just from opposition parties but also from the government's own rank and file, particularly after the electricity supply shortage of 2001. Rolling blackouts were avoided due to a power rationing program, in effect from June 2001 through March 2002; also, several short-construction-time generating plants were built, to provide reserve capacity, and some generation and transmission facilities were brought online ahead of schedule.

Big changes, however, were made in the Brazilian electrical power supply industry. A regulated third-party access system was mandated for both transmission and distribution networks. An independent

regulatory agency (ANEEL), a national system operator (ONS), and a wholesale market (MAE) were created; because of legal disputes among some utilities, the latter did not settle the short-term transactions (spot market) up to the end of 2002. The distribution and trade activities of the distribution utilities now have separate accounting systems, and for some utilities, the generation and transmission businesses were split into different companies during the privatization process. Some new agents in the electricity market were created, such as the independent power producers (IPPs), the “free” consumers (who, as opposed to the traditional “captive” consumers, can choose their electricity suppliers), and the pure traders (who do not own any generation, transmission, or distribution assets).

In the oil and gas supply industry, a negotiated third-party access scheme was defined by Law No. 9478, which detailed, in 1997, the new “rules of the game,” to promote competition in the industry. The same law created the National Petroleum Agency (ANP), an independent agency that regulates the whole oil supply chain and the upstream activities of the natural gas supply chain.

The opening up of the Brazilian energy supply industry to private investors, in order to redirect public investments to other areas and to introduce competition in the industry, in line with what is happening in several other countries, was the main energy policy of President Fernando Henrique Cardoso. As a result of this policy, there are now both private and state-owned large companies in both main branches of the industry (oil/gas and electricity supply). Electricity tariffs, rising much faster than the country’s inflation rates, represent a big problem facing the new federal administration. This has been exacerbated in the past few years by growing marginal costs (particularly for generation), by clauses in concession contracts linking the annual tariff updates to an inflation index (which has been overvaluing the strong devaluation of the local currency, the real, since 1999), and by the perception among private investors in the industry of a high-risk business environment, which, in the short term, either increases profit expectations or decreases investments.

In order to solve this problem, by making some changes in the current institutional model of the Brazilian electric power supply industry, the government of President Luis Inácio Lula da Silva intends to negotiate with the utilities the inflation index issue and to pursue programs to reduce the perceptions of high risk. The main changes proposed by the Ministry of Mines and Energy in July 2003 and, after dis-

cussions with the interested parties, reaffirmed with minor modifications in December 2003, are as follows:

1. The electricity supply market will be divided into two parts, one comprising the free consumers and the other comprising the captive consumers. Free consumers choose their suppliers among independent power producers, or traders, and buy from them their energy requirements, through freely negotiated bilateral contracts; the captive consumers deal with a distribution utility, through a pool managed by a new entity, the Administrator of Electricity Contracts (ACEE), which will replace the current electricity wholesale market (MAE) organization.

2. The tasks of ACEE will be management of long-term bilateral contracts among generators and distribution utilities and settlement of contractual differences for all market agents.

3. A new state-owned company, the energy research company (EPE) will be created to carry out the long-term (20 years ahead) and medium-term (10 years ahead) expansion planning exercises for the Ministry of Mines and Energy (MME); the resulting plans will be publicly discussed and eventually modified before final approval and implementation by the Ministry.

4. The plan for 10 years ahead will define the hydropower plant projects, the predefined energy and capacity generation blocks for thermal power plants, the regional constraints, and the transmission lines that should be auctioned by MME (no longer ANEEL), in addition to the required commissioning dates, to meet the forecasted demand of the pool consumers.

5. The bidding process referred to previously will allow proposals, by interested agents, of alternative projects to fulfill the energy supply or transmission needs as outlined in the plan. The proposal requiring the least revenue during the concession period will be the winning bid.

6. Preference will be given to public service generation utilities, instead of independent power producers, to supply the pool. Such utilities will sign a concession contract with MME and will have their firm power shared among all distribution utilities of the national interconnected grid, through compulsory long-term bilateral power purchase contracts.

7. Meeting the forecasted demand of the distribution utilities for the next 5 years should be fully assured through these long-term power purchase contracts. Special contractual arrangements are proposed for additional power purchases, in association with nonpredicted demand requirements.

8. EPE, ONS, ACEE, and ANEEL will help MME to oversee the supply conditions over the next 5 years, proposing corrective measures whenever necessary, which includes setting up generation reserve margins.

These changes were brought to the Brazilian Congress in December 2003, in the form of proposals of two new laws, to be enacted in the first half of 2004.

President Silva has emphasized that no further privatizations in the electricity supply industry will take place during his presidency. The partial deverticalization process carried out in this industry under the previous administration will continue in the new administration; vertically integrated utilities will form separate companies to carry out generation, distribution, and trade activities with free consumers, whereas separate accounting systems will suffice in the case of generation and transmission.

2. THE NATIONAL COUNCIL FOR ENERGY POLICY

The National Council for Energy Policy (CNPE) was created in 1997 by Law No. 9478, but was not actually installed until October 2000. According to a report issued in 2001 by a commission formed by the government to search for the causes of that year's electricity supply shortage, earlier activation of the council would have helped to avert or, at least, to minimize the effects of the shortage.

The National Council of Energy Policy consists of ten members, seven ministers of state bureaus (Mines and Energy, Planning, Economy, Environment, Industry and Trade, Civil House, and Science and Technology), one representative of the state governments, one representative of the universities, and one citizen expert on energy policy issues; the President of the Republic appoints the latter two members. The CNPE is headed by the Minister of Mines and Energy, who forwards proposals of energy policy resolutions to the President of the Republic; once approved, the proposals have the power of a presidential decree. Thus, CNPE is the most important forum for setting energy policies in the country. At the end of President Cardoso's second term, three technical committees were lending support for the council activities: one committee addressed the activities of the electrical power sector, another dealt with fuel supply chains, and the third focused on activities concerning required changes in the institutional model of the Brazilian electrical power supply industry.

3. FOSTERING ENERGY SUPPLY

Most of the national energy policies aiming to foster various forms of energy supply in Brazil were conceived in the 1970s. Policy development focused on medium- and large-scale hydroelectricity plants, coal-fired power plants, nuclear power stations, large-scale petroleum and gas production from off-shore wells located in deep waters, and fuel alcohol production from sugarcane. Policies to boost the generation of electricity from gas-fired power plants, small hydropower stations, wind power, and biomass resulted from decisions made during President Cardoso's government. All of these policies are briefly reviewed in the following sections.

3.1 Hydroelectricity

Brazil has a large hydroelectric potential, i.e., 258,420 MW, of which just 23.9% corresponded to plants in operation in 2002 and 4.3% represented hydropower stations under construction at that time. Since the middle of the 20th century, particularly after the 1960s, the federal and state governments have made large-scale efforts to tap this valuable and comparatively cheap resource, building the plants themselves through state-owned utilities or, more recently, providing credit facilities for private investors, through the National Bank for Economic and Social Development (BNDES).

Of the total installed capacity of electrical power plants in Brazil as of December 2001, the share of hydropower stations was 82.25%; the corresponding figure for public supply plants at that time was 84.92%. During President Cardoso's government, there was a policy guideline establishing that investments in new hydroelectric power plants should be carried out preferentially by private entrepreneurs, with possible minority participation of state-owned utilities in the case of projects of strategic interest for the federal administration. Spokesmen from the new government have declared recently that state-owned utilities will have more opportunity to invest in hydroelectricity than they had in the previous administration.

3.2 Coal-Fired Power Plants

Brazil's recoverable coal reserves as of December 2001 were estimated at 32.4 billion tons, the largest coal reserves in Latin America; the mines are located in the southern states of Rio Grande do Sul, Santa

Catarina, and Parana. The coal's high content of ash and, in most of the mines, sulfur severely limits the use of Brazilian coal in the iron and steel industries, and the remoteness of the mines necessitates great transport distances; in the year 2001, 99.9% of the coal consumed in Brazil's iron and steel plants was imported. The lack of appropriate railway networks in the mining regions also adds a further difficulty to the transportation problem. The Brazilian coal-mining industry has therefore always depended on the construction of new coal-fired power plants to survive. However, these plants have never been competitive with hydropower stations in Brazil, and have thus required subsidies to be built and operated.

Under the old rules of the Brazilian electricity supply industry, state-owned utilities had been building coal-fired power stations in the southern states of the country, close to the mines, for strategic reasons (diversification of the fuel mix for power generation and, as a result of an industrial policy, aiming to increase the domestic production of components for such plants); a fund (CCC) created by an electricity surcharge was formed to subsidize the operation of these plants when required, i.e., during years and seasons of low inflows to the hydro plant reservoirs. A minimum capacity factor for the plants, however, has been fixed, because of minimum annual consumption levels specified in the coal supply contracts, required to keep the mines running.

The new rules of the game in the Brazilian electricity supply industry, aiming to foster industry competition, cast shadows on the future of the coal mining industry in the country; the CCC fund, for instance, will be downsized from 2003 to 2005 and eliminated in 2006, according to 1997 legislation (Law No. 9648). Law No. 10,438, passed by the Brazilian Congress on April 26, 2002, however, opened a new door for the coal producers; the resources of a new fund (CDE), created by this law for the electricity supply industry, can, among other uses, be employed to finance both old stations (for operating expenses, replacing the cash flows from the CCC fund) and new coal-fired power stations. The amount of the CDE fund to be made available for such purposes will be defined on a regular basis by the National Council for Energy Policy.

3.3 Nuclear Power Stations

President Cardoso's administration created Eletro-nuclear, a subsidiary of Eletrobras, to assume responsibility for the nuclear plants in Brazil. Brazil

has two operational nuclear power plants, Angra-1 (675 MW) and Angra-2 (1.3 GW), both located at the town of Angra dos Reis, in the state of Rio de Janeiro. Angra-1 was bought from the U.S. company Westinghouse in 1969. The Angra-2 plant came online in 2000, 23 years and \$10 billion (U.S. dollars) after construction began; it was the single outcome, in terms of power plant building, of a very ambitious nuclear power agreement signed between the Brazilian and German governments in 1975, which envisaged the construction, by Siemens of Germany, of eight nuclear power stations in Brazil; the agreement also specified the transfer of German technology related to fuel cycle activities (mining, processing, fuel enrichment, fuel element manufacturing, and reprocessing), and the joint development of a new uranium enrichment process.

The construction of a second nuclear plant (Angra-3, with an installed capacity of 1.3 GW), included in the Brazilian/German agreement, was started in 1981, involving foundation works and the acquisition of German equipment. Due to budget cuts and to some opposition from environmental groups, the construction was stalled and the equipment for the plant has been mothballed. However, the electricity supply crisis of 2001 bolstered interest in bringing the Angra-3 plant into service. Those in favor focus on the need of the country to diversify its sources of power generation and to take advantage of its substantial uranium reserves (the world sixth largest: 309,370 t of U_3O_8 as of December 2001); furthermore, there is the fact that about \$750 million (U.S. dollars) has already been spent on the plant, including the purchase of about 60% of the required equipment. These resources will be lost if the project is abandoned and Eletronuclear will be unable to develop sufficient scale to become competitive. On the other hand, those against Angra-3 point out that the project will require around an additional \$1.7 billion (U.S. dollars) and will take at least 5 years to be completed; it is also emphasized that the population still views nuclear energy with suspicion, because issues surrounding safety and the final disposal of the radioactive residues have not yet been resolved.

CNPE authorized Eletronuclear in 2001 to carry out the necessary economic and environmental feasibility studies (Resolution No. 05, approved in December 2001). In August 2002, the Council voted in favor of Eletronuclear resuming the construction of Angra-3 after the necessary environmental licenses have been granted, if the new government does not decide to halt the process (a CNPE meeting was

scheduled for May 2003 via Resolution No. 8, of September 17, 2002; major outlays in the project will occur only after that meeting). This decision was postponed for 2004 by the new federal administration, in part because of current surplus power supplies in MAE and in part because of urgent cash problems at Eletronuclear, requiring a tariff revision by ANEEL.

Apart from Angra-3, no other new nuclear plants are planned for Brazil. Such plants may be built in the future, but only if safer and less expensive new technologies, still at the research and development stage, succeed on an international scale.

3.4 Oil and Gas Production from Offshore Deepwater Fields

Brazil has the second largest proved oil reserves in South America (after Venezuela), at 1.35 billion m³, or 8.48 billion barrels, as of December 2001; 88% of the total is in offshore basins and 80% is at depths above 400 m. The natural gas proved reserves as of December 2001 stood at 219.84 billion m³, or 7.76 trillion ft³, the fifth largest in South America behind Venezuela, Argentina, Bolivia, and Peru; 60% of the total is in offshore basins and 40% is at depths above 400 m.

The production of both petroleum and natural gas has been rising steadily in Brazil since the early 1990s, reaching, in 2001, 75.22 million m³, or 1.3 barrels per day (bbl/d) of petroleum, which met 79.3% of the consumption in that year, and 14.04 billion m³, or 1.36 billion ft³ per day of gas. Imports of gas from Bolivia started in 1999 and imports from Argentina started in the next year, totaling 4.61 billion m³, or 0.45 billion ft³ per day in 2001. The offshore Campos Basin, north of Rio de Janeiro, is the country's most prolific production area for both oil and gas, containing around 80% of the national reserves. The Santos Basin also holds large gas fields. Brazil's oil imports come mostly from Venezuela and Argentina.

As was the case with many other national oil companies, Petrobras initially concentrated on building up its downstream infrastructure, particularly from 1965 to 1974. In the wake of the first oil price shock, in the middle 1970s, the Brazilian government ordered the management of Petrobras to implement three new policies aiming to decrease the effects of the oil price rises on the national balance of payments: (1) international expansion of the company in upstream activities, through a subsidiary, Petrobras International (Braspetro); (2) signature of

service contracts, with a risk clause, with private oil companies in regions not yet under exploration; and (3) an increase in the national production of oil and gas through exploitation of offshore, mainly deepwater, fields, which make up most of the Brazilian reserves. The first two strategies failed, but the last one has been highly successful. Petrobras' accomplishments in deepwater production have been internationally acknowledged and the Campos Basin's success at proving giant reserves at great depths has attracted attention worldwide. Many companies have been encouraged to come to Brazil to participate in ANP's promoted bidding rounds, in order to develop upstream exploration, some of them without partnering with Petrobras.

In 1986, Petrobras began the first of the Procap programs (the Petrobras technological development program on deepwater production systems, or Procap 1000). The main objective of this program was to improve the company's expertise in oil and gas production in water as deep as 1000 m. It also consolidated Petrobras' production concept based on floating production systems. Petrobras' Procap 2000, launched in 1993, emphasized the development of technologies aimed at reducing investment and operational costs as well as improving efficiency and extending the working life of equipment at water depths of 1000–2000 m. Procap 3000, implemented in 2000, goes even further, seeking to develop technologies that will make oil and gas production in ultradeep waters, below 2000 m, technically and economically feasible. In 2000, the U.S. Geological Survey published new estimates of global oil reserves suggesting that Brazil might still have some 47 billion barrels of undiscovered oil, almost all in offshore fields, with about 35% in the Campos Basin.

Apart from the second half of the 1980s, when Petrobras' management became involved in a long battle with the Ministry of Economy, which decided to cut back the company's expenditures and investments, the rest of the time the federal administration has supported the company's effort to boost oil and gas production from offshore deepwater fields, striving in the medium term for self-sufficiency in oil production. This is likely to continue with the new administration.

3.5 Fuel Alcohol from Sugarcane

Since 1975, with the creation of the National Alcohol Program (Proalcool) by Federal Government Decree No. 76,593, Brazil has produced anhydrous alcohol from sugarcane; this alcohol is blended with

gasoline in Otto cycle car engines in proportions of up to 25%. With the second phase of Proalcool, which started in 1979 (Federal Government Decree No. 83,700), hydrated alcohol has also been produced for use in Otto cycle engines modified to run on 100% ethanol, or neat alcohol.

Currently, Brazil is the world's largest producer of sugarcane, with crops often yielding over 300 million tonnes of crushed cane per harvest season. Prior to Proalcool, the Brazilian share was less than 15% of worldwide production. During the 1970s, many ethanol distilleries were installed in the country, either as new plants or as distilleries annexed to existing sugar mills. The main alcohol-producing states are São Paulo (contributing over two-thirds of the total), Rio de Janeiro, Alagoas, and Pernambuco. Since Proalcool was created, two main products have been obtained from sugarcane: sugar and fuel ethanol. The former has been an important component of the basket of commodities exported by the country since the time when Brazil was a colony of Portugal; in contrast, exports of fuel ethanol have been sporadic and have faced many protectionist barriers abroad.

The production rate of fuel ethanol has varied according to the relative prices of both sugar, particularly in the export markets, and alcohol, which are the main factors affecting production, besides climatic and environmental variables. When sugar prices are high, the production of alcohol decreases, and vice versa. Up to the beginning of the current decade, the price of fuel ethanol was fixed by the government, tracking with the controlled price of gasoline; now both prices are determined by market forces, although they will eventually be subject to government pressures on Petrobras and on the alcohol producers when price increases are considered too high.

The main objective of Proalcool, rapid growth of the alcohol industry, in conjunction with subsidies that increased alcohol production capacity in the 1970s and early 1980s, has facilitated the building of a large alcohol industry. Within the industry, there is still considerable need for increasing energy efficiency and reducing production costs; government policies for this industry have so far failed to address the important issue of cost-effectiveness. The Proalcool program was discontinued in the early 1990s during President Collor de Mello's term in office. The federal government, however, continues to foster the production of fuel ethanol by maintaining the requirement of a mandatory blend of anhydrous alcohol with gasoline. The blend formulation set by

the government is between 18 and 25% ethanol, depending on the annual production of ethanol, which, as already pointed out, is strongly affected by sugar prices abroad.

In support of the fuel ethanol program, there are discounts on federal taxes applied to alcohol-fueled cars (the IPI tax) and to fuel ethanol (the recently created CIDE tax). The share of alcohol-fueled cars in total sales of new cars dropped from 96% in 1985 to 1.15% in 2001, after a low of 0.07% in 1997. At the end of 2001, there was an aging fleet of neat alcohol-fueled cars, estimated at about 2.5 million vehicles. New policies have been considered by the federal government to boost the production of hydrated ethanol once more. The most important measures that have been envisaged are the compulsory addition of ethanol to diesel oil in buses and lorries and the establishment of government "green fleets" that will run on neat ethanol. None of these measures, however, has yet been adopted, and the outlook for hydrated ethanol production in Brazil is not promising. On the other hand, the prospects for future growth in the production of anhydrous alcohol to blend with gasoline are bright, not only because of the environmental benefits of such blends, in terms of reduction of air pollution, particularly in large cities, but also because of the good prospects for "flexible fuel" vehicles. These vehicles employ electronic fuel management technologies that allow use of any blend of anhydrous alcohol with ethanol. Some flexible fuel models are already available in the Brazilian market. Because the prices for these vehicles are higher than the prices for the low-ethanol-blend counterparts, the government is evaluating the adoption of financial incentives to help boost sales.

A rapid hydrolysis process to produce ethanol from sugarcane bagasse (crushed cane) is being developed in the state of São Paulo. A demonstration plant should be operating soon. If this technology proves economically feasible, it will allow an increase of around 30% in alcohol production with the use of 50% of the currently available sugarcane waste (tops and leaves, or "barbojo"), without any additional contribution from sugarcane plantations. There have been significant improvements in the productivity of both sugarcane agriculture and the ethanol-based industrial sector. These gains have been due to a combination of factors, including (1) introduction of new and improved sugarcane varieties, (2) better economies of scale from larger and more efficient new plants, and (3) technological improvements and energy conservation measures in old plants. However,

there is still room for further cost reductions. Finding better uses for the sugar and alcohol by-products, such as sugarcane bagasse, barbojo, and vinasse (organic wastewater), is certainly an excellent route to improving the economic performance of these plants. The prospects to increase the current generation of surplus electricity in cogeneration plants located in sugar mills and alcohol distilleries are also promising.

3.6 Gas-Fired Thermal Power Plants

Brazil has a large natural gas pipeline network to transport the gas produced in the Campos and Santos basins to the cities of Rio de Janeiro, São Paulo, and Belo Horizonte. There is also a long pipeline for collecting and transporting the output of gas fields located in the northeastern region of the country to the local capitals and industrial areas; there are plans to interconnect the two pipeline systems. There are also smaller transportation networks to receive the production from the offshore Espirito Santo Basin and from the onshore field of Urucu; the former delivers gas to Vitoria, the capital of the state of Espirito Santo, and to industrial areas in the northern part of that state, and should be connected soon to the Campos/Santos network. Commissioning of the onshore field of Urucu awaits completion of pipelines that will supply Manaus and Porto Velho, the capitals of the states of Amazonas and Rondonia, respectively. All of these gas pipelines are owned by Petrobras.

There are two international gas pipeline connections in the country. The first pipeline to connect Brazil to foreign gas sources was the Bolivia-to-Brazil pipeline, tapping Bolivia's Rio Grande sources and servicing the states of Mato Grosso do Sul, São Paulo, Parana, Santa Catarina, and Rio Grande do Sul; this pipeline came onstream in July 1999. In the Bolivian part of the pipeline, there is a diversion to supply a power plant and other consumers in Cuiaba, the capital of the state of Mato Grosso. Partners in the Brazilian section of the pipeline include Petrobras, which is the major shareholder, Enron, Shell, and BBPP Holdings. The second international pipeline links the city of Parana, in Argentina, to Uruguiana, in the state of Rio Grande do Sul, Brazil, where it supplies gas to a 600-MW power plant. Transportadora de Gas del Mercosur is the pipeline's operator. Service began in July 2000. An extension of the pipeline, which will connect Uruguiana to Porto Alegre, the capital of the State of Rio Grande do Sul, to service a new power plant in Porto Alegre, was planned but has been post-

poned. Additional Argentina-Brazil pipelines are in various stages of planning, although recent natural gas discoveries in Bolivia and Brazil could discourage the development of these projects. It is also possible that a second Bolivia-Brazil pipeline will be built.

The primary motivation behind the projects of most of the recently built or planned pipelines has been the hope for a fast buildup of natural gas demand in Brazil, in conjunction with expectations that there will be construction of a large number of gas-fired thermal power plants. The source of these expectations was the belief of President Cardoso's government that the private investors, under the new rules of the Brazilian electrical power supply industry, would prefer to build efficient, combined cycle, gas-fired power plants, as has been the case in many developed and developing countries, instead of new hydro plants, as has been the Brazilian practice in past decades. Some initial uncertainties among the potential investors caused the Brazilian government to step in; in September 1999, the Gas-Fired Thermal Power Plants Priority Plan, or simply PPT, was announced. The first version of the PPT identified 15 projects, totaling 12 GW, expected to be online by 2003. Specific regulations were established for these projects, such as a specific value for the upper limit pass-through that the electricity distribution companies are allowed to pass on to their "captive" consumers' tariffs. To reassure investors concerned about fluctuations in gas prices, which were in U.S. dollars and were indexed to a basket of fuel oils, the government, through Petrobras, set price ceilings on 20-year fuel supply contracts. There was just one gas price, revised quarterly, regardless of the power plants' location. In addition, the national development bank, BNDES, offered a special loan program.

Uncertainty among the investors remained, however, such that none of the 15 projects got underway. The variation of gas prices with fuel oil prices, the exchange rate, the quarterly price revisions, and the lack of synchronism between the revisions of electricity and gas prices generated investor anxiety. So Petrobras was required to offer an alternative solution with a blended gas price indexed to the U.S. Producer Price Index (All Commodities) (PPI), revised annually. In April 2000, the federal government issued a revised version of the PPT, with the new price option and, in response to political pressure from local politicians and state governors, increased the number of projects from 15 to 51, all over the country. Apart from the ambitious and unrealistic number of proposed plants, the rapid devaluation of the Brazilian currency, the real,

against the U.S. dollar created further difficulties, given that the gas price was set in dollars.

The electricity shortage of 2001 forced the government to launch its Emergency Thermal Power Plant Program, the last version of the PPT program, improving the conditions for all the project developers with gas already contracted or coming on-stream before June 2003 (this was later extended to December 2004), up to a maximum volume of 40 million m³/day. For these plants, MME/MF Order No. 176, on June 1, 2001, set a new gas price formula, valid for 12 years. The timetable for tariff revisions was rescheduled to bring gas and electricity into line. Petrobras will assume the exchange rate risk for 1 year before passing it on to the power plants at the time of their tariff revision. The annual revision of gas prices considers the Brazilian inflation index IGPM, with a weighting of 20%, and the exchange rate plus the PPI, with a weighting of 80%. Prices will be renegotiated every 3 years and the gas supply contracts are transferable. A further benefit to the plants under the new program, established by Law No. 10,312 on November 27, 2001, was the elimination of PIS/PASEP and COFINS, two federal taxes on the gross revenue accruing from the gas sales to such plants. According to estimates made by the Ministry of Mines and Energy in October 2002, based on the regular follow-ups carried out by the Ministry and considering plants in operation, undergoing trial runs, and under construction, at several stages and contracting levels, 19 gas-fired thermal power plants are likely to come online by 2004, with a total installed capacity of 7157.6 MW, under the umbrella of the last version of the PPT.

When the gas supply contract for the Uruguaiana power plant was signed, during the early stages of the Brazilian electric power supply industry reform, new gas-fired thermal power plants were competitive with new hydropower stations, according to calculations made using the prevailing cheap gas price negotiated for the contract and the reference unit costs for both types of plants, in Brazilian reals. The sharp devaluation of the real since 1999 and high oil prices, however, changed this picture, against the gas-fired thermal power plants. The Brazilian government believed that, in the medium term, such plants would regain their competitiveness. This reasoning was based on (1) decreasing thermal power unit costs accruing from competition among gas suppliers and from the growth in industrial uses for the gas, creating the necessary conditions for the development of a secondary gas market, which, in turn, would allow more flexible “take or pay” and “ship or

pay” contractual requirements, and (2) increasing hydropower unit costs arising from plants located farther from the main load centers. Thus the government of President Cardoso decided to subsidize the cost of transporting gas in the country. Law No. 10,604, enacted on December 17, 2002, allows such a subsidy up to R\$500,000,000 per year, using the CIDE tax as a resource. With this subsidy, with the lower prices for the commodity expected to accrue from negotiations with the Bolivian government and producers, and with the substantially increased medium-term local production made possible by recent discoveries of large fields in the Santos and Campos basins, the government expects to reduce by \$0.50 (U.S. dollars) per million British thermal units or more the price of the gas, which should make the gas-fired power stations competitive again in Brazil.

A fundamental issue is the fact that, in Brazil, in contrast to most other countries, the opportunity cost of natural gas for power generation in public supply plants is determined by hydro generation in new plants. Thus, indexing the price of such gas to the prices of a basket of fuel oils, as is traditional in the oil and gas industry, is meaningless, in economic terms, in Brazil. The new federal administration is less enthusiastic than the previous one about large expansion plans involving gas-fired thermal power stations. Regarding possible new energy policies to boost gas demand in the medium term, incentives may be given for other gas uses, particularly for cogeneration plants in the industrial and services sectors.

3.7 Generation of Electricity from Small Hydropower Plants, Biomass, and Wind Power

Power generation units employing renewable sources of energy (e.g., small hydropower plants, wind power, solar energy, and biomass) and cogeneration plants have received financial incentives in some countries. During the 1970s and part of the 1980s, the major reason was that they represented indigenous sources of energy, reducing the dependence on foreign sources. More recently, with globalization and the formation of economic blocks of countries, this argument has lost much of its early appeal. However, the potential of these generating units to create environmental benefits is being realized. The financial incentives can be orthodox, such as tax relief and attractive credit terms, or heterodox, such as (1) compulsory purchases by utilities of the power

generated by these plants at avoided costs; (2) purchase, for the public grid, of energy blocks through bidding restricted to some types of these plants; or (3) granting of purchase tariffs above the market rate for the energy generated in these plants (to be paid for by all consumers, or on a voluntary basis, depending on the willingness of consumers to pay more for “green” energy).

The sugar and alcohol, and paper and pulp sectors are the Brazilian industrial branches that rely most heavily on self-production to meet their electricity needs. They use mostly cogeneration plants burning industrial residues from biomass, e.g., sugarcane bagasse, firewood, and black liquor (a mixture of chemicals and dissolved wood materials). The new rules of the Brazilian electricity supply industry tend to encourage greater use of such industrial residues for process steam and power generation in cogeneration units, together with forestry residues and sugarcane waste (*barbojo*), since recent developments in harvesting machinery design and new collection practices are reducing the cost of the latter waste fuels for power generation, particularly if gasification is involved.

During the period 1996–1997, the Brazilian Minister of Mines and Energy discussed with interested parties a possible federal government decree that would oblige utilities to buy surplus power from cogenerators, up to a certain share of their market growth, more or less along the same lines of the American Public Utilities Regulatory Policies Act (PURPA) legislation, during its first phase. The project was badly designed and the proposed measures came up against the main directives dealing with the opening up of the Brazilian power sector. As a consequence, many utilities rallied against the project and succeeded in aborting it. Bearing in mind the American experience related to the application of the PURPA legislation, the board of directors of Brazil’s regulatory agency for the electrical power industry, ANEEL, defined, through Resolution No. 021, on January 20, 2000, the minimum share of thermal energy production and the minimum total efficiency requirements that a cogeneration plant owner should meet to become a “qualified cogenerator.” The creation of this qualification process aimed to set up procedures allowing the selection of eligible cogeneration units to receive incentives still to be defined. (Some of the requirements of ANEEL Resolution No. 021 should be revised, however, because they are either too slack regarding some combinations of technologies and fuels, or they are too strict regarding other combina-

tions.) The first of the incentives was specified in December 2000 through MME Order No. 551, which included qualified cogeneration plants, using any kind of fuel available in the country, to be commissioned up to December 2003 (this was later extended to December 2004) in the PPT program, having rights to all program benefits.

For quite a while, Brazil’s most important development bank, Banco Nacional de Desenvolvimento Econômico e Social (BNDES), has been offering some credit facilities for the building of cogeneration units and electrical power plants using nonconventional renewable energy sources. Although the conditions specified by BNDES are more favorable than what usually can be found on the Brazilian credit market, they are worse than those in the international market, particularly because of the long-enduring prevailing high interest rates in the country. Thus, this credit line was little used before 2001.

Before 2002, the owners of small hydropower stations (up to 30 MW) were the only renewable power producers to enjoy “heterodox” financial incentives in Brazil. In 1998, Law No. 9648 granted these producers access to any consumer with a contracted demand higher than 0.5 MW and relieved them from the payment of half the value of transmission grid use fees. The current minimum power demand limit, which defines “free” consumers, is 3 MW. Law No. 10,438, enacted on April 26, 2002, created the Incentive Program to Generate Electricity from Alternative Sources (Proinfa), comprising wind power, biomass, and small hydropower plants, to be implemented in two stages. Associated with this program, the law defined a new kind of agent in the Brazilian electrical power supply industry—the autonomous independent producer, whose business cannot be controlled or associated with any electricity generation, transmission, or distribution utility. Producers that do not meet this requirement can participate in the program, provided their share in the contracts does not exceed 25% (50% for wind power producers, in the first stage of the program), and no autonomous producer is precluded because of the requirements. Equipment manufacturers can be autonomous independent producers if at least 50% of the value of the equipment involved in the program is produced in Brazil.

Proinfa will hold public solicitations for each kind of power source. Priority will be given first to plants that have already obtained the Installation Environmental License (LI) and then to those holding a Preliminary Environmental License (LP). If more capacity is offered, satisfying the conditions above,

than the capacity scheduled to be contracted, the plants with the shortest remaining environmental license periods will be chosen. In the first stage of the program, 3300 MW, equally distributed among small hydropower plants, wind power stations, and biomass-fueled thermal power stations, will be installed up to the year 2006. Eletrobras will provide long-term contracts to purchase the energy produced by these plants, paying the so-called “economic value” associated with each technology, which should correspond at least to 80% of the average electricity tariff in the country. The cost of these acquisitions as well as the administrative cost of Eletrobras to manage this scheme will be shared among all categories of consumers in the National Interlinked System, proportional to measured individual consumption.

After completion of the first stage, a second stage will continue up to 2022, during which the generation from the plants should meet 15% of the annual load growth and, considering the results of the first stage, 10% of the electricity consumption in the country. Throughout 15-year-long contracts, Eletrobras will again buy the output of these plants, equally among the three technologies if there is enough supply. The purchase will, as before, be preceded by public calls and there will be the same selection criteria as used in Proinfa’s first stage, but the price paid will be equal to the weighted average unit cost of new hydroelectric plants with an installed capacity above 30 MW and new gas-fired thermal power stations. The expenses associated with this purchase will again be shared among all consumers proportional to their measured consumption. The difference between the generation cost of each technology and the average unit cost will be paid straight to the producers, using the resources of a new fund (CDE) created by Law No. 10,438. ANEEL is responsible for overseeing the whole process, using, for this purpose, the Renewable Energy Certificates issued by the generators. The CDE fund consists of monies from the annual fees paid by the electrical power supply industry investors to the government for the right to use public goods, the revenues collected by ANEEL from the application of fines, and a new annual fee paid by all agents who sell electricity to consumers.

The creation of Proinfa by Law No. 10,438, inspired by successful legislation in Germany and Denmark, is a landmark approach to foster the generation of electricity from distributed renewable energy sources in Brazil. It has, however, some drawbacks, which should be addressed in future legislation. The first problem is the fact that the market share targets set for the generation of the

renewable energy sources have no relationship to environmental policy targets, to the amount of these resources available in Brazil at reasonable costs, to the indigenous manufacturing capacity the government wishes to foster, or to supplementary power sources, e.g., new thermal power plants, required to complement the generation of random energy sources such as wind and hydropower. Also, there was no study of the impact of Proinfa targets on electricity tariffs, which is a major issue in a country with so many poor consumers. These targets should be reviewed in the future. Apart from eventually decreasing “economic values” for each technology, set by the Ministry of Mines and Energy, the program provides no further incentives to increase the cost-effectiveness of these sources; making the Renewable Energy Certificates tradable would be an important step forward in this direction.

In December 2003, the new federal administration put forward a proposal that would limit the annual addition of plants generating electricity from distributed renewable energy sources in the second stage of Proinfa, through specific auctions for such sources, to amounts which would not cause increases in the new pool prices (discussed in Section 1) above 0.5% in any single year and 5% on a cumulative basis. According to the same proposal, from January 2005 onward, the generators interested in participating in the auctions will have to prove that at least 60% of their equipment and services will be produced in Brazil; this share will increase to 90% in 2007.

4. ENERGY EFFICIENCY AND RESEARCH AND DEVELOPMENT PROGRAMS

Several energy efficiency programs have been sponsored by the federal government, as well as by the governments of some states (São Paulo, Bahia, Minas Gerais, and Rio Grande do Sul), since the 1970s. Of the national programs implemented in previous decades and still in operation, the most important ones are Procel, Conpet, and the mandatory energy efficiency programs run by the electricity distribution utilities and overseen by ANEEL.

The Ministry of Mines and Energy and the Ministry of Industry and Trade created, through the MME/MIC Order No. 1877, on December 30, 1985, the Program to Reduce the Waste of Electrical Energy (Procel), to be managed by Eletrobras. A

presidential decree of July 1991 gave the coordination of the program to the Ministry of Mines and Energy and increased its scope, which included electricity consumption labels for household appliances and electrical motors; electricity consumption audits in small and medium-size industrial and commercial premises; financing of research and development initiatives (particularly at the Eletrobras research center, Cepel), directed to the manufacturing of more efficient electrical appliances and motors; support for new legislation and regulation in the country concerning energy efficiency; support of projects directed to reduce losses in the generation, transmission, and distribution of electricity; setting up information systems and marketing activities on energy efficiency; and running educational and training programs aiming to bolster a culture of energy conservation. The Procel program has impacted households, the commercial and industrial sectors, public services such as illumination and water supply, and efficient management of electricity consumption in public buildings. Procel has gone through ups and downs. Since the mandatory energy efficiency programs run by the electricity distribution utilities were set up in the late 1990s, Procel's role has been downgraded; its activities have been redirected to support ANEEL in the evaluation of the utilities' programs.

A 1991 presidential decree created the National Program to Rationalize the Use of Oil Products and Natural Gas (Conpet), to be coordinated by the Ministry of Mines and Energy and to be operated by Petrobras. "Conpet in the School" is its main institutional project. In the transportation sector, Conpet has two successful projects, SIGA-BEM and ECONOMIZAR. SIGA-BEM is a partnership with BR, Petrobras' subsidiary company for the distribution and retail trade of oil products, directed to advise truck drivers in BR's filling stations about how to reduce the consumption of diesel oil in their vehicles. ECONOMIZAR is a partnership with the National Confederation of Transportation, which, through mobile units, provides assistance to garages and service shops of load/passenger transportation companies in the search of the optimal management of their diesel oil stocks, aiming to reduce specific fuel consumption and pollutant emissions of the serviced fleets. Most of Conpet's initiatives in the industrial sector have been directed to energy efficiency improvements in Petrobras' refineries, including a wider use of efficient cogeneration units. In the residential/commercial sectors, Conpet is responsible for pioneering the labeling of liquefied

petroleum gas (LPG) consumption of stoves. So far, Conpet has been managing with less financial resources and a shorter scope of action compared to Procel, but has shown a much more stable performance in running its projects.

Since 1998, the concession contracts of the electricity distribution utilities have contained a clause that requires them to apply at least 1% of their annual income to energy efficiency and research and development programs, with at least 0.25% going to demand-side management programs and at least 0.1% going to research and development activities. ANEEL regulates these programs and oversees their results with the help of Procel's staff and experts from some state regulatory agencies that have contracts with ANEEL. ANEEL has set boundary conditions for the range of activities covered by these programs, which, in essence, has been similar to many of those developed earlier by Eletrobras in Procel. Law No. 9991, enacted in July 2000, rules that the electricity distribution utilities should apply annually at least 0.75% (0.5% up to December 2005) of their net operational income to research and development projects and at least 0.25% (0.5% up to December 2005) to energy efficiency programs on the demand side. Electricity generation utilities, independent power producers, and distribution utilities are also required by this law to spend at least 1% of their net operational income on research and development programs. Half of all these research and development resources will be managed by ANEEL, and the other half will be channeled to the Electricity R&D Fund, created by Law No. 9991, to be managed by the Ministry of Science and Technology. The new federal administration intends to redirect half of the resources managed by ANEEL to partially fund EPE's activities.

Petrobras has been sponsoring research and development activities related to the production chain of oil and natural gas since the early days of the company, particularly at its research center, Cenpes. The Petroleum National Agency has, since its installation in 1998, been regulating research and development programs in this field, with resources coming from the royalties paid for by oil and gas exploration and production concessions (Presidential Decree No. 2851, November 30, 1998, which created the National Fund for Scientific and Technological Development for the oil and gas industry) and by the concession contract of Petrobras (1% of the company's gross revenue accruing from production activities). The Ministry of Mines and Energy also funds several applied research projects in

various universities, involving alternative fuels and/or technologies, particularly in the Amazon region.

In terms of energy policy, the most important incentive in recent years to the search for a higher level of energy efficiency in Brazil was the enactment of Law No. 10.295, on October 17, 2001; this law allows the government to set up maximum levels of specific energy consumption, or minimum levels of energy efficiency, for energy-consuming equipment produced in Brazil or imported, after public hearings involving the interested parties. This law also mandates the government to promote energy efficiency measures in buildings. A permanent committee (CGIEE), with members from several ministries, was formed to set goals and to elaborate proposals for the public hearings (Decree No. 4059, December 19, 2001). The first type of equipment to fall under the mandate for minimum energy efficiency levels is the three-phase induction squirrel-cage rotor electrical motor (Decree No. 4508, December 11, 2002).

The mandatory energy efficiency programs run by the electricity distribution utilities, and the electricity supply shortage of 2001, boosted the market of the Energy Services Companies (ESCOs). The main barrier found for further development of this market is the financing of ESCOs, which has been addressed by the Ministry of Mines and Energy through proposals for the opening of new credit lines in state-owned banks and the setting up of a fund with resources provided by Eletrobras and/or BNDES. The certification of ESCOs and the technical qualification of energy efficiency projects are other measures being considered by the Ministry to push forward this market. The promotion of a greater use of high-efficiency electrical motors and household appliances, via credit facilities and tax reductions, and a gradual integration of the national programs directed to energy efficiency improvements are two other important policy measures that have been pursued recently by the MME.

5. ENERGY PRICES AND SOCIAL ISSUES

The price of oil products (gasoline, diesel oil, fuel oil, LPG, naphtha, aviation kerosene, and lubricating oil) in Brazil has been set by market conditions, without any regulation, since January 2002, as mandated by two laws (No. 9478/97 and No. 9990/00). The price of LPG was subsidized before January 2002 because a large number of poor people in the country use this fuel for cooking purposes. Resolution CNPE No. 4, on December 5, 2001, followed by Decree No. 4102

(January 24, 2002), and Law No. 10,453 (May 13, 2002), substituted the cross-subsidy then existing for all consumers of LPG with government grants just for the low-income consumers, registered in government aid-to-the-poor programs, using the resources of the CIDE tax. During the same meeting, on December 5, 2001, CNPE decided that the Ministry of Mines and Energy and the Ministry of Economy should keep regulating the prices of the natural gas produced in Brazil after December 2001, because, in practice, Petrobras had retained its monopolistic position. This control should remain in place until real competition in this market materializes. As defined at the end of the last government by the Ministry of Mines and Energy, this "real" competition will be obvious when there are at least three suppliers, none of which hold a market share larger than 75%. The state governments regulate the gas price for the consumers and there will be no competition in the downstream part of the gas chain in the short to medium term unless a general understanding is achieved among the federal and state governments or the Constitution is changed. There are currently far fewer subsidies available for the sugarcane producers and fuel alcohol manufacturers than in the past, but they still exist, using resources provided by the CIDE tax (Law No. 10,453, December 13, 2002), particularly for the poor northeastern region.

Frustrating the expectations of President Cardoso's government, so far few eligible "captive" electricity consumers have opted to become "free" consumers, i.e., to choose their supplier and negotiate the price to be paid for energy consumption. The main reason is that there are cross-subsidies for energy-intensive large-scale consumers, based on past regulations, discouraging the move from regulated to free arrangements. Recognizing this distortion, but concerned about likely losses in export revenues and job losses in the short term if large tariff increases are imposed on the energy-intensive industrial branches, the previous federal administration decided to spread these increases over 4 years, at 25% per year (Decree No. 4562, December 31, 2002). The current administration has thus decided to increase the transition period to 5 years, with just a 10% increase in 2003, provided the interested energy-intensive industrial consumer invests in generation expansion (Decree No. 4667, April 4, 2003). The government also aims to convince this type of industrial consumer to buy at favorable prices, through medium-term contracts (duration of 1 to 2 years), about 2000 MW, on average, out of a total

of 7500 MW, on average, available in the generation system in the first months of 2003, because of the decrease in demand caused by the current recession. The current recession was triggered by the electricity shortage of 2001, which also speeded up the construction and commissioning of several new power plants.

There have been cross-subsidies to low-income electricity consumers in Brazil for many years. Law No. 10,438 broadened the definition of the low-income consumer, particularly in the poor northern and northeastern regions of the country. In order to avoid sharp tariff increases for certain utilities, especially in those regions, the government also decided to create direct subsidies to low-income electricity consumers using resources provided by eventual extra income earned by the federally owned generation utilities. This extra income could come from trading in the public biddings, defined by Law No. 10,438; it was also decided, if necessary, during 2002 and 2003, to use resources from the RGR fund, managed by Eletrobras.

Expensive thermal energy, generated by engines fueled by diesel oil or, for the larger machines, fuel oil, supplies isolated networks, located mostly in the northern region of the country. This generation is subsidized by all consumers in the national integrated network through a fund known as CCC-Isol, which, according to Law No. 10,438, should last up to 2022.

Between 10 and 15% of the Brazilian population, comprising from 4 to 5 million households, mainly in remote, rural areas, has no access to electricity supply. To improve this situation, the Federal government set up two rural electrification programs in the 1990s, "Luz no Campo" and "Prodeem." Luz no Campo, managed by Eletrobras, has the goal of electrifying 1 million rural properties, basically through grid extensions, using resources from the RGR fund. Up to the beginning of 2003, 541,115 properties were electrified by the program. Prodeem, on the other hand, was conceived to meet the basic social demands, in terms of electricity supply, of isolated small rural communities through local generation, with preference to the use of renewable energy sources. The program, managed by the Ministry of Mines and Energy, has been employing photovoltaic panels as generating units in most cases. Law No. 10,438 determines that the payments for the use of public goods, and fines applied by ANEEL, which contribute to the CDE fund, should be applied with preference to rural electrification programs. The law also states that ANEEL should assign rural

electrification targets to each distribution utility. The Agency board of directors hopes to have all Brazilian households electrified by 2016.

The new federal administration report on changes in the institutional model of the Brazilian electric power supply industry, issued in December 2003, proposes that the subsidies for power plants consuming distributed renewable energy sources should be borne by (small) tariff increases; thus the government hints that all of the CDE resources should be channeled with priority to rural electrification programs and for granting subsidies to poor electricity consumers.

6. ENERGY AND THE ENVIRONMENT

Brazil has an advanced body of legislation concerning the environment. Enforcement of this legislation, however, has had failures, mainly due to short budgets of the regulatory bodies, at both federal and state government levels; the situation has improved since the electricity shortage of 2001.

In terms of regulatory tools, the Brazilian environmental legislation uses "command and control" measures such as environmental licenses, pollutant emissions limits, and establishment of zones where certain activities are restricted due to potential environmental damages. The regulation leave little room for market-driven measures involving economic incentives and for negotiated agreements among the regulatory bodies and the agents they regulate, as is happening now in some countries. Environmental policies, planning, and regulation activities are decentralized in Brazil, involving not only federal and state government bodies, but also municipal ones. The same kind of decentralization was established by Law No. 9433, of August 1, 1997, for water resources. This law created a new agent, the Hydrographic Basin Committee, made up by representatives of municipalities, who are responsible for elaborating a Hydrographic Basin Plan and for defining the priorities of water usage in the basin; needless to say this committee is very important to the interests and activities of the energy supply industry.

There has been little connection so far between environmental and energy policies in Brazil. The energy supply shortage of 2001 brought together the work carried out by the Ministry of Mines and Energy and that of the Ministry of Environment, but essentially only on particular projects and mostly to speed up environmental licensing procedures. A joint

agenda for the electrical power sector set by the two ministries in 2002 is expected to enlarge the scope of joint activities, including formal exchanges between technical committees of CNPE and CONAMA, the National Council for the Environment, and a more proactive treatment of environment issues in the electricity supply industry's 10-year forward planning.

7. AN INTEGRATED APPROACH

Energy policies in Brazil have been formulated in the past mainly by the federal government. Separate policies have been developed for each energy supply industry branch (oil and gas, electricity, coal, nuclear, etc.), and these have had little or no relation to other public policies. This has been changing slowly in recent years. The installation of CNPE, which includes the seven most important ministers of state, was a big step forward toward the integration of energy policies with other public policies in Brazil. The development of long-term (20 years ahead) integrated prospective studies each year by the Ministry of Mines and Energy for CNPE, since 2001, for the energy sector as a whole, using alternative development scenarios that take into account the current and possible new economic, technological, and environmental policies, has provided a consistent technical background for such integration. There is, however, a big challenge still to be faced, which is to engage the state and municipal governments, similar to what has happened in the environmental and water resources areas, in order to decentralize to some extent the policymaking process, under the direction of CNPE and with the technical support and supervision of MME. Two important measures in this direction were taken in the second half of 2002 at MME, involving the integration of the National Energy Balance with the state balances, in

terms of methodology and data basis employed, and the start of integrated resources planning studies in four Brazilian hydrographic basins.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Development and Energy, Overview • European Union Energy Policy • Geopolitics of Energy • National Energy Policy: China • National Energy Policy: India • National Energy Policy: Japan • National Energy Policy: United States • Nationalism and Oil

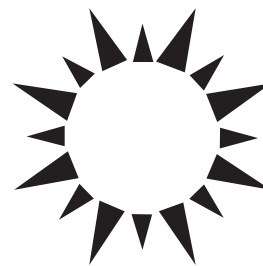
Further Reading

- Bajay, S. V., Carvalho, E. B., and Ferreira, A. L. (2000). Energy from biomass in Brazil. In "Industrial Uses of Biomass Energy—The Example of Brazil" (F. Rosillo-Calle, S. V. Bajay, and H. Rothman, Eds.), pp. 27–52. Taylor & Francis Inc., New York.
- Comitê Coordenador de Planejamento da Expansão dos Sistemas Elétricos (2002). "Plano Decenal de Expansão 2003–2012—Sumário Executivo." Secretaria de Energia, Ministério de Minas e Energia, Brasília, DF.
- Kelman, J., Ventura Filho, A., Bajay, S. V., Penna, J. C., and Haddad, C. L. S. (2001). "Relatório." Comissão de Análise do Sistema Hidrotérmico de Energia Elétrica (criada por Decreto do Presidente da República, em 22 de Maio de 2001), Brasília, DF.
- Martins, A. R. S., Alveal, C., Santos, E. M., La Rovere, E. L., Haddad, J., Lisbôa, M. L. V., Correia, P. R. S., Schaeffer, R., Aguiar, S. C., and Bajay, S. V. (1999). "Eficiência Energética—Integrando Usos e Reduzindo Desperdícios." ANEEL/ANP, Brasília, DF.
- Ministério de Minas e Energia. (2003). "Modelo Institucional do Setor Elétrico." Ministério de Minas e Energia, Brasília, DF.
- Santos, E. M. (2001). "The Brazil Oil and Gas Sector—Outlook and Opportunities." CWC Publishing Ltd., London.
- Secretary of Energy. (2003). "National Energy Balance." Ministry of Mines and Energy, Brasília, DF.

National Energy Policy: China

MARK D. LEVINE and JONATHAN E. SINTON

Lawrence Berkeley National Laboratory
Berkeley, California, United States



1. Energy Policy under Central Planning, 1949–1979
2. Energy Policy under the Transition to a Market Economy, 1979–Present
3. Outlook for China's Energy Policy

Glossary

exajoule (EJ) A measure of energy; an exajoule is 10^{18} joules, equivalent to 0.9478 quads (10^{15} British thermal units).

Five-Year Plan (FYP) China's Five-Year Plans are developed by the State Development Planning Commission and are used to guide overall socioeconomic development policy.

State Development Reform Commission (SDRC) The comprehensive national agency in China that coordinates long-term planning of economic and social development; formerly the State Planning Commission, and then, until 2003, the State Development Planning Commission.

State Economic and Trade Commission (SETC) The agency in China that, until 2003, coordinated day-to-day government regulation and economic activities.

In China, the national government has a strong role in the energy sector. Under the central planning system, from the establishment of the People's Republic in 1949 to the initial economic reforms in 1979 championed by Deng Xiaoping, the government directly controlled extraction, generation, transport, and allocation of fossil fuels and electricity. During the transition to a market-oriented economy, the government has gradually, though not always consistently, withdrawn. It lifted controls on energy prices, created energy corporations, and assumed a regulatory role. Like China's other governmental structures, the institutional apparatus for energy policy has been frequently reorganized, with a single ministry responsible for energy in some

periods and specialized agencies for each energy type in other periods. In the first three decades of the People's Republic, policy was concerned mainly with increasing supplies of coal, oil, natural gas, and electricity. In the early 1980s, after proposing ambitious economic development goals that would have outstripped conceivable growth in energy supply, China also adopted strong energy efficiency policies. There have been significant efforts to provide the two-thirds of the population that lives in rural areas with better energy services. In recent years, China has made greater efforts to integrate energy supply and efficiency policy with environmental protection and other social goals, in addition to bold economic development goals. The government faces energy security concerns as the dependence on imported oil and natural gas rises.

1. ENERGY POLICY UNDER CENTRAL PLANNING, 1949–1979

1.1 Energy for Industrialization

From 1949 until the 1970s, the Soviet system provided the model for industrial development in China. Investment in energy supply became the highest priority of national investment. As in the Soviet Union, there were large subsidies for extraction of natural resources. Subsidies for energy under the Chinese regime were intended to support expansion of heavy industry and to make energy available and affordable to all citizens. Energy and other natural resources were priced well below their cost of production. The result was very high growth in energy supply. China experienced rapid growth in commercial energy output, from a minuscule 0.7 EJ in 1949 to 18.9 EJ in 1979, an average annual growth rate of 11.8%. Large quantities of biomass fuels were used in rural areas, approximately 3 EJ in 1949 and currently about 6 EJ. The economy grew rapidly, also

from a very small base. Energy supply grew even faster; energy elasticity (percent change in energy divided by percent change in GDP) from 1953 to 1979 was 1.26. Energy supply commanded a larger share of total capital investment than any other industrial sector. From 1970 to 1979, one-third of capital investment in industry was allocated to energy. From 1953 to 1979, 33% of energy investment went to coal, 24% went to oil and natural gas, and 43% went to electricity. No attention was paid to the environmental consequences of energy development. Large numbers of miners were killed in coal mines, and mines degraded surface water and groundwater and produced scars on the countryside. Emissions to the air from coal burning were uncontrolled, and solid waste was not treated before disposal.

Because of large subsidies provided to energy supply, lack of incentives to control its use, and absence of economic and environmental discipline, in 1978 Deng Xiaoping inherited a set of energy institutions that were deeply troubled. Most of them were state owned, and virtually all of them would have been bankrupt but for government subsidies. They were inefficient, overstaffed, lacking in modern technology, and without effective management structures and with few incentives to create them.

1.2 Energy Institutions

Under the Soviet-style bureaucratic structure, a web of constantly changing agencies managed planning, production, and distribution of energy. The State Planning Commission (SPC) included energy investment, production, and distribution in its Five-Year Plans (FYPs). The State Economic Commission coordinated and monitored implementation of annual and short-term plans. Day-to-day energy production was directed from a ministerial level. At the beginning of the 1950s, one body, the Ministry of Fuels and Power (MFP), managed production of all types of energy. In 1955, the State Council replaced the MFP with individual ministries for coal, petroleum, and electric power. During the 1960s, the Cultural Revolution reshaped China's institutional structures. Prevailing ideology favored institutional simplification, so in 1970 the Ministries of Coal, Petroleum, and Chemical Industry were merged into a Ministry of Fuels and Chemical Industries. Similarly, the Ministries of Electric Power and of Ministry of Water Resources Utilization were combined. Management of state-run enterprises in the energy sector was transferred to local governments. Poor performance of the energy sector eventually led

the central government to resume control of larger enterprises and to reestablish separate ministries for coal and petroleum. Throughout the central planning period, ministries concentrated on short-term planning, and the SPC's FYPs provided only rough guides for long-term direction. Implementation relied on administrative decisions for production and allocation, and each year's targets were based on incremental changes from the previous year's targets.

2. ENERGY POLICY UNDER THE TRANSITION TO A MARKET ECONOMY, 1979–PRESENT

2.1 Energy Supply and Efficiency for Rapid Economic Growth

In 1979, Deng Xiaoping announced that China would quadruple its per capita gross domestic product between 1980 and 2000. This goal had significant implications for energy policy. If energy use continued to grow 26% faster than the economy, as it had over the previous 30 years, total energy use in 2000 would have reached 106 EJ. Actual consumption in 2000 was 38 EJ, while per capita GDP more than quadrupled (Fig. 1; Table I). For comparison, the United States in 2000 used about 117 EJ of energy. Had China continued its previous energy path after 1979, it would have suffered intolerable environmental insults. Hence, one of the

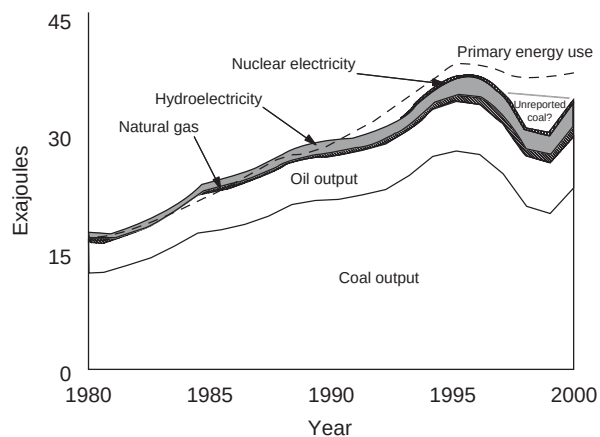


FIGURE 1 Primary energy output and use, 1980–2001. According to official statistics, China's energy output and consumption, mainly coal, rose from 1980 to the mid-1990s before falling temporarily. The difference between total energy output and use in most years represents net exports or imports of oil, except in the period from 1997 to 2000, when most of the difference is probably due to unreported output from small mines.

TABLE I
China's Energy Balances, 1949–2000^a

Energy balance	Exajoules used in year										
	1949	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000
Total primary energy supply	0.69	2.15	8.69	5.51	9.08	14.29	18.04	22.75	28.17	37.82	33.98
Coal	0.67	2.06	8.31	4.85	7.41	10.09	13.0	18.3	22.6	28.5	20.9
Oil	0.01	0.04	0.22	0.47	1.28	3.23	4.4	5.2	5.8	6.3	6.8
Natural gas	0.00	0.00	0.04	0.04	0.11	0.34	0.6	0.5	0.6	0.7	1.1
Hydroelectricity	0.01	0.05	0.12	0.15	0.28	0.63	0.7	1.1	1.5	2.1	2.6
Nuclear electricity										0.1	0.2
Net imports and stock changes							(0.64)	(2.33)	(2.28)	0.15	2.39
Total primary energy use		2.05	8.83	5.54	8.56	13.33	17.67	22.48	28.93	38.45	38.19
End use ^b											
Agriculture							1.02	1.19	1.42	1.61	1.70
Industry							11.50	14.46	18.89	26.62	24.95
Transport							0.85	1.09	1.33	1.72	2.91
Services							0.69	0.95	1.38	1.92	2.53
Households							2.81	3.90	4.63	4.61	4.37
Losses							0.81	0.91	1.29	2.03	1.84
Balance							0.38	0.27	(0.76)	(0.62)	(4.21)

^aData from the National Bureau of Statistics for various years; *China Statistical Yearbook* (China Statistics Press, Beijing, China).

^bIn Chinese energy accounts, electricity is converted according to the average amount of primary energy consumed in power plants to generate electricity.

most important issues to explore in China's energy policy post-1979 is how the country was able to cut its energy elasticity from 1.26 to a figure less than 0.5.

A second major issue regards the choice of energy forms most suitable to the country's economic development. This has meant a continuation of the emphasis on electricity development, which fostered production of the most valuable and versatile energy form, but limited investment capital to the coal, oil, and gas sectors. A third major issue was protecting the environment, because the greatest environmental impacts on air, water, and land come from energy supply and use. Although energy elasticity fell, energy supply and demand still grew rapidly, because China sustained a 20-year period of extraordinary economic growth (roughly 10% per year by official figures and about 8% according to some independent economic analysts).

Finally, the structure of energy markets was a critical issue in development of the energy system after 1979. It became clear that energy prices needed to reflect costs, which they did not under the central planning system. Energy shortages in the 1980s and

nonproductive investments led the Chinese to strive for reforms that encouraged a stronger role for markets for the energy system and for other sectors.

2.2 Transformation of Energy Institutions

Reforms of China's political economy have touched all sectors. Responsibility for overall planning has remained the domain of the State Council, the State Development and Reform Commission (SDRC), and, until its disbanding in 2003, the State Economic and Trade Commission (SETC), but ministries have been repeatedly reshuffled. Authority for energy sector activities has been spun off to large state-owned corporations, which retain many of the same personnel but have more freedom to allocate investments and manage production. The state no longer has the resources to direct the economy through its agencies and now seeks to influence development through institutions that use other means of control. Institutions currently in place may have more of the gloss than the substance of a market-oriented system, but they reflect the irreversible shift to a market

economy. Moreover, local agencies have become larger players.

Since the early 1980s, four major reorganizations occurred: in 1981–1983, 1985–1988, 1993–1994, and 1998–2001. The first reorganization focused on the oil sector and split the Ministry of Petroleum Industry into the China National Petroleum Corporation (CNPC), the China National Petrochemical Corporation (Sinopec), which were responsible, respectively, for upstream exploration and production and downstream refining and marketing. The China National Offshore Oil Corporation (CNOOC) was created in 1979 to manage offshore oil development.

The second wave of reorganizations, in 1985–1988, encompassed all energy subsectors. In place of central ministries with direct responsibility for investment and production, large state-owned companies were formed. A Ministry of Energy (MOE) was established as a coordinating body under the SPC, although it was active only in electricity. A State Energy Investment Corporation (SEIC) was formed and given responsibility for the central government's major investments in the energy sector. In the 1980s, China also incorporated energy conservation into the institutional structure. Energy conservation offices were established at central and provincial levels in planning and production agencies, and over 200 energy conservation technology service centers were set up nationally.

In 1993, MOE was disbanded and once again replaced with separate ministries for coal and electricity. SETC was established and given responsibility for coordinating short-term activities in energy supply. In 1994, control over state investment was unified, and SEIC, along with other state investment companies, was subsumed by the new State Development Bank. The China Energy Conservation Investment Corporation, formed in 1988, was the only exception and remained independent.

A major reorganization in 1998 transformed most remaining industrial line ministries, including those for coal and electricity, into departments of the SETC. Simultaneously, staffs were cut by half. The result was even greater independence for energy corporations and there was concentration of authority over energy supply policy, with SDPC responsible for long-term guidance and SETC responsible for short-term direction. To introduce competition in the oil industry, wells and refinery assets were split geographically between Sinopec and CNPC, which thus became vertically integrated oil companies. Further reorganization of SETC in 2001 absorbed

separate departments responsible for energy into a single department with authority for all industrial activity. In 2003, SETC was broken up and its functions spun off to the renamed State Development and Reform Commission and to other agencies. Possible future institutional reforms under consideration include establishment of a utilities commission to set rules for the electric power sector, and reestablishment of a Ministry of Energy.

Nongovernmental organizations (NGOs) have been absent in most areas of Chinese policymaking, but the China Energy Research Society (CERS) has been influential in charting the course from central planning of the energy system to reliance on markets.

2.3 Coal

Coal is China's major fuel, providing an average of 73% of total energy consumption from 1980 to 2000. For many reasons, including its abundance in China, coal received limited policy attention and less capital investment than did any other energy form besides natural gas. Figure 2 makes clear that, after 1984, investment in oil and gas outstripped coal and that electric power investment "took off." By 1997, investment in electricity was 50% more than for oil and gas, and 3.5 times that for coal. This policy

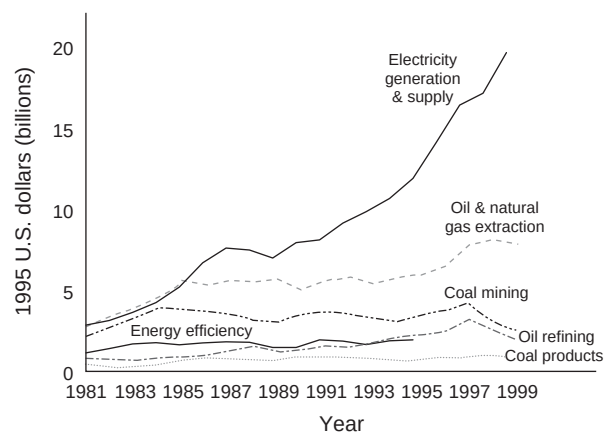


FIGURE 2 Energy supply and efficiency investment, 1981–1999. Investment in power generation and transmission has grown steadily since the early 1980s, and by the late 1990s made up three-fifths of all investment in the energy sector. Investment in other energy supply sectors remained flat until the late 1990s; during that period, investment in oil and natural gas extraction rose, while that in oil refining and coal mining fell, as capacities sufficed to meet demand. China has maintained a steady, even growing, level of investment in energy efficiency projects, in later years exceeding investment in coal mining. These data cover only state-owned and -invested enterprises, which account for the vast majority of energy investment in China.

essentially starved coal for capital, even more than previously. As a result, almost all expansion of coal output from 1980 through 1995 was from small collective and private mines, as shown in Fig. 3. This created a major new source of coal for China at virtually no investment from the central government. Unfortunately, it was an unsustainable policy. The government exerted little oversight over these small mines, which typically produce low-quality coal. Many of them are unsafe, contributing disproportionately to high accident and death rates. They cause serious environmental problems, polluting underground water sources, causing subsidence, and other harm. In most cases, the coal is used locally because it is not of high enough quality to be sold in regional markets, increasing local air pollution as well as pollution in homes (most rural and many urban households burn coal directly for cooking and heating).

Although the policy of encouraging the expansion of coal output from these small mines was clearly not desirable in the long term, there was a clear rationale in the 1980s. China faced a substantial shortage of coal; capital for energy was devoted to forms of energy that would best serve a modernizing economy,

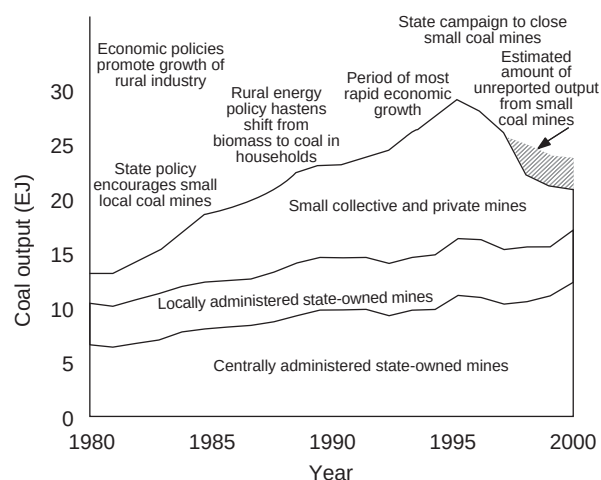


FIGURE 3 Coal production by type of mine, 1980–2001. National policy beginning in the early 1980s promoted small “township and village” mines. By 1995, these often tiny mines produced nearly half of China’s coal, but at a tremendous cost in human life, wasted resources, and environmental pollution. Officially reported coal production peaked in 1996 at 1374 million metric tons. Small-mine output was already falling when, in 1998, the central government began a vigorous campaign to close small mines. Reported output from small mines fell by half within 2 years, but actual production from small mines may have been up to one-third higher than reported output, because many small mines continued to operate clandestinely.

and energy sectors were already consuming a large portion of total industrial investment, up from 37% in 1980 to 59% in 1997. Thus, China faced a serious dilemma in allocating investment. It followed a path that promoted economic development, but contributed to the increasingly serious environmental problems. By the late 1990s, China had not only overcome the energy shortages of the 1980s, but, because of declining economic growth, was in a position to reduce its coal consumption. This situation led to a campaign by the central government to close the small coal mines on a massive scale. As shown in Fig. 3, coal output from small collective and private mines declined from more than 600 Mt of coal (45% of all coal produced) to 200 Mt from 1996 to 2001, according to official figures. In fact, as shown by the upper right-hand shaded area in Fig. 3, it is unlikely that all of the private and collective coal mines were closed. These prohibited mines may have produced as much as 150 Mt of coal. Even so, the reduction of coal use from these mines was enormous.

China’s coal industry is not appreciably stronger than it was some years ago. It is still undercapitalized, has poor environmental and safety records (about 6000 coal miners die annually in accidents), and has generally weak management, despite the country’s heavy reliance on coal. Closing small mines has improved the environmental situation, but either alternatives to coal must be found or the coal industry must be modernized and pollution from both production and consumption reduced.

2.4 Oil and Natural Gas

In the late 1950s, the huge Daqing oil field was discovered in northeastern China, followed a few years later by the large Shengli field in Shandong. Development of these and other fields began in earnest in the early 1970s and oil production expanded rapidly. There was a widespread belief, in China and elsewhere, that the country would become the world’s next major source of oil, with oil resources similar to those in the Middle East. The early promise of vast oil resources was not fulfilled. No large fields have been found in eastern China since the 1970s. Untapped oil resources in China are in harsh environments, i.e., in bitterly cold northwestern China and in offshore fields. Both sources are already being exploited. The need for advanced technology, experience, and capital to develop oil in such climates has meant that major projects are joint ventures with international oil companies, and are expensive.

The history of oil imports and exports tells an important story for China. As a result of the early oil discoveries and policy restrictions on nonindustrial oil use, China was an exporter for many years. The peak of its exports, at 1.6 EJ (36 Mt oil) occurred in 1985. Export earnings from energy that year were \$7.1 billion (U.S. dollars), or 25% of export earnings for the entire economy. Net oil exports dropped to zero by 1993 and have since increased as dramatically as exports once did. By 2000, net imports were 3.3 EJ (76 Mt oil). With the automobile industry in its infancy in China, but growing very rapidly, dramatic increases in oil demand are expected. This oil will come from overseas, presenting China with a growing security issue.

China's interest in natural gas has grown, and a major exploration program has been initiated. Natural gas is still a very small part of total energy supply: 1.2 EJ (30 billion cubic meters) in 2001, or 3.4% of total energy production. However, production has been growing at an average of over 10% per year since the early 1990s. China is building a major pipeline from the northwest to Shanghai, to be completed in the middle of the first decade of the 2000s. It is not yet clear how the gas will be utilized, because the price of delivered gas will be very expensive. Natural gas has the clear advantage of over coal and oil of reducing pollutant emissions, but unless Shanghai decides to impose high externality costs on coal burning, or forbids the use of coal in various applications, development of the gas market may be constrained.

The major barrier to greater use of natural gas is its high cost in China. In the absence of large gas fields near the eastern population centers, gas will have to be brought in by pipeline from the far western region, imported from Russia or Central Asia, or imported as liquefied natural gas (LNG). These alternatives are costly and pose perceived geopolitical risks. Coal has been used in some industrial processes in highly inefficient ways (e.g., as a feedstock for fertilizer production) because of the lack of natural gas. Substitution of natural gas for coal would be a natural development. A second large potential market is as a substitute for electricity (e.g., in household water heaters), especially as direct coal burning is phased out. In small-scale applications, for which pollution control is prohibitively expensive, natural gas or oil is often the only viable alternative to coal. However, given China's ambitious plans for natural gas development, promoting gas for power generation is an obvious way to spur long-term growth of the fuel. Many markets will depend on

China's policies to control emissions from coal combustion and the relative incremental costs of emissions controls and of natural gas.

Oil imports will continue to increase until an affordable substitute for liquid fuels from petroleum is found and widely disseminated. The drivers of future growth will be the rapidly expanding automotive and petrochemicals sectors. Without new large discoveries, China will become more dependent on imports, and has taken steps to mitigate dependency through investments in producing fields overseas. Predicting the future of natural gas is difficult. The key factors will be the rate of discovery of natural gas fields, their size and location relative to demand centers (thus determining the cost), and the development of suitable markets for the natural gas, which likely will remain expensive. After the current efforts to create a natural gas market, investment in continued exploration and development will depend on the assurance of markets for the gas, which will be closely related to environmental policy.

2.5 Electricity

From 1980 to 1998, electricity supply (and demand) grew at 8.3% per year, about the same rate as revised estimates of gross domestic product (GDP) growth. Over 80% of electricity is now generated by fossil-fired power plants, with the remainder coming mainly from hydropower, some from nuclear, and a tiny fraction from other sources. As already noted, the focus of China's energy policy has been electricity development. Investments in electrification have been highly beneficial to the nation in several ways. First, electric power is essential for a modernizing economy, because it provides the highest quality energy suitable for many uses that fuels cannot meet. Second, access to electricity in rural areas, including those suffering great poverty, provides a critical input into the economic and social development of the region. Moreover, it is easier to reduce pollutant emissions from large sources such as power plants, as compared to small ones such as household stoves, and a growing share of China's coal is used in power plants.

On the other hand, the policy of investing heavily in electricity has, as noted earlier, starved other supply sectors of funds. Two resulting problems are the poor quality of marketed coal and the lack of exploration for natural gas. Such problems could be overcome by making markets for international investments more open and transparent, a process that has proceeded slowly since the 1980s. Because China has long viewed energy as key to national

security, it has been cautious about relying on foreign energy supplies or encouraging “too much” foreign investment, although the definition of “too much” has never been made explicit and has apparently varied over time. Recently, the government has permitted companies outside of China to have a controlling interest in energy projects, subject to careful review.

China has the second largest hydropower resources after Russia, but the biggest dam sites are far inland, distant from coastal demand centers. The 1960s and 1970s saw preferential development of hydropower, but as electricity demand grew exponentially in the 1980s, China turned to coal. Large dams continued to be built, particularly with international development assistance, but more slowly than fossil plants, due to high costs of transmission, major problems encountered in earlier projects, and increasing environmental sensitivity. Since the late 1980s, most attention given to hydropower went to the giant Three Gorges project, but concerns about pollution from coal-fired generation and a desire to develop western provinces are helping to drive construction of other hydropower projects. As in most countries, large dams are planned as multipurpose projects, with water supply and flood-control objectives being particularly important. Pumped-storage facilities are also being built in coastal regions where power shortages during peak demand periods have sometimes been severe. Current policy aims to promote delivery of electricity from hydropower and coal from poorer western provinces to the wealthier coastal regions. This is intended to relieve projected shortfalls in generation capacity in the East, and to provide income to the West. Significant issues remain in funding investment in generation and transmission and in restructuring the regulatory system to allow large interregional transfers.

At least five factors affect the efficiency of China’s fossil-fired power plants. Two factors—the high quality of large, new state-of-the-art power plants and the relatively high percentage of cogeneration (i.e., use of heat as well as power) in the power system—have contributed to improved efficiency. Two other factors—the small size and outmoded technology of many existing power plants and the relatively low quality of coal burned—have resulted in many inefficient plants. A final factor, the lack of natural gas as a supply source of electricity, has precluded the use of advanced combustion turbines that can produce electricity at high efficiency and with low levels of pollution (as well as favor

cogeneration in many cases). In order to improve the technical and economic performance of power generation, the government has long had policies to close the least efficient power plants and to prohibit construction of new small power plants unless they are cogenerators. Over the past two decades, the average size of China’s power plants has grown, and the average efficiency of fossil fuel power plants has improved.

The Chinese have long had the goal of developing nuclear power on a large scale, but growth has been slower than planned, largely because of high costs compared with coal. Nonetheless, as of 1993, China had installed eight nuclear generating units totaling 5980 MW, with another three units totaling 2510 MW under construction. China has sought to develop the capability to construct nuclear power on its own, but almost all construction so far has had active overseas involvement.

Hydropower has always been an important part of Chinese electricity supply. The 18,000-MW Three Gorges Dam will be the world’s largest hydropower project, nine times the size of the Hoover Dam in the United States. It has been controversial from its inception because of environmental impacts, involving the relocation of 1.2 million people and the loss of historical sites. The reservoir is already being filled; the first turbine is expected to go online in 2004 and the project will be completed by 2009.

China has an active nonhydro renewable electricity program. There are large resources of geothermal energy, but because the high-temperature geothermal resources are in the West, far from end users, they contribute little to total electricity supply. Development of wind power is rapid, although from a low base. There are presently 28 wind farms totaling 403 MW. Two more, one near Shanghai and the other outside of Beijing, are committed for construction. There is discussion of policies to promote wind power, in the hopes that overcoming barriers will render wind a cost-competitive electricity source. Because China has many good-to-excellent wind regimes, wind could become competitive with coal-fired power, depending on environmental regulations.

The government is now restructuring the power system and planning to create regulatory commissions at the national and provincial levels, thus separating power generation from direct government oversight of performance and prices. Generation and transmission will also be separated. The government expects to introduce competition into the power generation market, although there may be only a

small number of companies owning large numbers of generators initially. The government will maintain authority over transmission lines and has ambitious plans for creating a national grid among its six regional and six provincial grids.

Some pressing issues regarding future electricity generation can be addressed by the following questions:

- Will China create meaningful environmental standards for new and existing power plants? When?
- Will China strive to replace coal in a significant portion of new power plants? What roles will natural gas, wind, and biomass play?
- Will the Three Gorges Dam work as planned, or will it run into significant operational and/or environmental problems?
- Will restructuring encourage improved performance of the electricity system?
- Will the bulk of domestic energy investment continue to be allocated to electricity? What will be the role of foreign investment?

2.6 Rural Energy

China's top priorities for rural energy development have been to meet basic fuel needs for cooking and heating, to provide electricity to as many households as possible, and to supply energy for rural industry. This last function has been accomplished mainly through allowing the development of local coal mines and grid extension. Most policy measures have been directed at fulfilling household energy needs. The vast majority of the rural population has at least minimal access to electricity. In remote areas in southwestern China, this has been accomplished through construction of small hydropower projects. Many of the "microhydro" projects of the 1960s and 1970s were poorly designed, but more recent projects have been more successful, and China has become a model for other developing countries. A growing number of remote communities are supplied by small wind generators. Still, as of 2000, about 16 million rural residents (i.e., slightly more people than the population of Florida) had no access to electricity. Most options for supplying those who remain without electricity are beyond local means, so the continued efforts to supply regions without electricity will depend on the central government's limited funds.

Throughout history, China's farmers have periodically faced serious fuel shortages. Various programs over the years have aimed to resolve those shortages.

In the 1970s, for example, a campaign resulted in the construction of tens of thousands of small biogas digesters, intended to turn crop and livestock waste into fuel gas. Virtually all of these facilities failed, due to poor design, inadequate materials, and insufficient operating expertise. A new program, instituted in the late 1990s, emphasizes larger scales and better designs, and shows more promise. Direct combustion remains the major way biomass fuels are used. In the early 1980s, in response to rural fuel shortages, the government organized a large-scale National Improved Stove Program (NISP) to provide rural households with more efficient biomass stoves for cooking and heating, and later also improved coal stoves. By the early 1990s, the pressure on biomass supplies had eased in most areas, in part due to increased incomes and availability of coal and electricity. From the mid-1990s onward, support for the stove industry was replaced with extension services and certification systems to standardize stoves. The development and dissemination of improved stoves is now left to market actors. The government claims that by 1998, 185 million of China's 236 million rural households had improved stoves. Although many rural households now use fossil fuels, especially coal and some liquefied petroleum gas (LPG), and ownership of electrical appliances is increasing, nearly all rural households use at least some biomass. Solar water heaters and large plastic-covered greenhouses have become common features in rural areas. Passive solar design has been incorporated into some new housing, and solar cookers have been distributed in some western areas where fuel is short and sunlight is plentiful.

2.7 Energy Efficiency

The achievements of China's energy efficiency policy have been remarkable. As noted earlier, since 1980, energy has grown half as fast as GDP (Fig. 4). Virtually all other developing countries have seen energy demand grow at least as fast as GDP. In many ways, China serves as the "existence proof" that energy can grow at a substantially lower rate than that of the economy over a considerable period of time.

2.7.1 Energy Efficiency under the Planning System

China achieved results in the large-scale promotion of energy efficiency beginning with instruments of the planned economy. Energy efficiency became a national priority in 1980, after a confidential study by

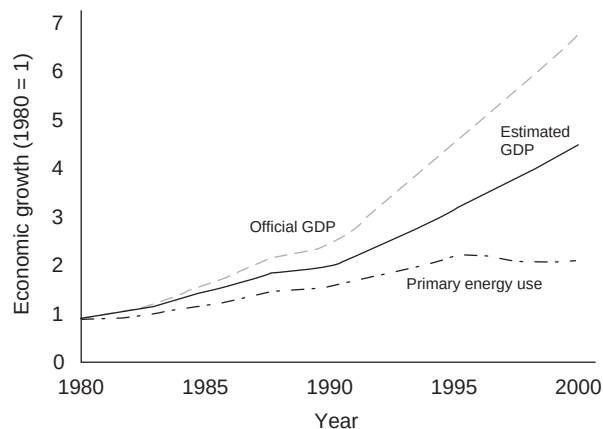


FIGURE 4 Official and estimated gross domestic product (GDP) and primary energy use, 1980–2001. Since 1980, according to official statistics, China's economy grew nearly sevenfold. Many economists believe official statistics exaggerate actual growth. Although experts differ on the correction to apply, a reduction of two percentage points in annual growth rates is a typical estimate. Even with revisions, China's economy in 2001 was 4.6 times larger than in 1980. Over the same period, primary energy use rose to 2.2 times its previous level. Unlike most developing countries, in which energy use rises faster than economic output, China has experienced growth in energy use only one-third to one-half as fast as economic growth.

a group of academics that proposed major changes in China's energy policy; among the proposals was establishment of a national efficiency initiative. In the Sixth FYP (1981–1985), China proposed and implemented a massive effort to promote energy efficiency. In 1981, investment in energy efficiency by the government was 9.5% of total energy investment (Fig. 2), or 1.5 billion yuan (about \$830 million U.S. dollars in then-current exchange rates). Investment in energy efficiency, including cogeneration, grew to 3.2 billion yuan by 1985.

No other country in the world at that time, or since, has had a national program that allocated energy investment to energy efficiency, or one that was implemented so rapidly. China created a bureau of Comprehensive Energy Savings and Resource Utilization within the SPC, which was at that time China's top executive agency. This bureau undertook a wide range of reforms, some of which were carried out by a newly created (China) Energy Conservation Investment Corporation. Along with its regional offices, the bureau evaluated proposals from throughout the nation and chose the projects to be funded. The energy efficiency investments were only the most important of many measures put in place during the very innovative period of the early and middle 1980s. Other innovations that remained in place until the late 1990s concerned energy manage-

ment, financial incentives, technical service centers, and research and development. Large factories were required to have energy managers to monitor energy use. Annual quotas for energy use were set and evolved into regulations relating to energy use of various types of equipment (e.g., motors and boilers) and manufacturing processes. Factories that performed better than the quota or exceeded the standards were given financial rewards. Many inefficient factories were shut down. Demonstration projects illustrated performance of efficient technologies.

Initially, energy conservation investments were financed by the national government. Later, the cost was shared with provincial governments. By the early 1990s, the investments had turned into low-interest loans. Other incentives included reduced taxes on energy-efficient products, subsidies to develop new energy-efficient technologies, and monetary rewards to energy-efficient enterprises. As mentioned previously, China created a network of over 200 energy conservation service centers that employed nearly 5000 staff at their peak; the staff performed energy efficiency feasibility studies, participated in investment projects, and trained energy managers and technicians. China instituted a national energy conservation week (in November), during which the national and local press spotlight energy efficiency and efficiency is included in school curricula.

China began an energy conservation research and development program in 1981 to complement its shorter term programs. The areas receiving the greatest attention were efficiency of coal combustion (e.g., improved briquettes and cookstoves), electricity end-use technologies (e.g., fans, pumps, and motor controls), boilers and furnaces for steel production, heat exchangers and waste heat recovery systems, transport technologies, fertilizer production processes, and industrial-process controls. Few of these programs have been formally evaluated. However, the reduction of the energy intensity of the Chinese economy from 1980 to the present is strongly suggestive of the overall success of the programs, which made more efficient equipment and practices widely known and available to the enterprises that were expanding production capacity in those years.

2.7.2 The Energy Conservation Law

In 1998, the national Energy Conservation Law came into force, codifying the country's approach to promoting energy efficiency under a more market-oriented economic system. Implementing provisions

are still being formulated to address some of the challenges:

- Creation of new fiscal and other incentives.
- Reassessment of existing energy price regulations.
- Establishment and enforcement of mandatory efficiency standards for common energy-using equipment and buildings.
- Integration of energy efficiency with environmental protection efforts.
- Improving efficiency of the rapidly growing rural industrial sector.
- Retaining and transforming the former system of efficiency centers.
- Institutional coordination in implementation, management, and supervision of energy conservation law.

2.8 Energy Pricing

Energy pricing remains heavily influenced by the central government, which has only slowly relinquished control. Under the planning system, state-dictated prices accompanied planned allocations to users. Controls on energy prices determined which energy enterprises and subsectors would make profits and suffer losses. Low coal prices, for example, consigned coal mines to a constant struggle to make ends meet, while price differentials between crude oil and oil products ensured that refineries became wealthy.

Energy-price reform began in the early 1980s and proceeded tentatively, because price policy was—and remains—politically sensitive. A portion of the energy products was allowed to be sold at prices higher than the in-plan prices, resulting in a multi-track pricing system. Today, most energy prices in China track international market levels. China's accession to the World Trade Organization in 2001 represents a step toward full integration of China's energy markets with international markets.

Price liberalization began with coal. In the early 1980s, the government allowed limited free markets for coal and permitted state-owned mines to sell small amounts at market prices, and then allowed a multitrack system. In 1993, the price of nearly all coal was liberalized and prices rose rapidly. State-owned enterprises, which had previously been required to purchase from state-owned mines, were able to buy lower cost (and lower quality) coal from small local mines. In the mid-1990s, coal prices leveled off and then fell as coal demand fell. Price

controls remain in place for some users, particularly fertilizer plants that use coal feedstock.

Two distinct markets for coal have emerged, with two different pricing systems. Large mines and consumers, like power plants and steel mills, negotiate supply contracts at national semiannual coal-marketing conferences. The product is of relatively high quality, is shipped long distances, and is expensive. Local markets match small non-state-owned mines with consumers, often in the same area, and provide generally cheaper, lower quality coal. Prices vary significantly by region; in coastal areas, far from the major coal mines, domestic coal can be expensive by world standards, and small amounts of coal are imported.

Oil prices have followed a similar path of deregulation. Central control of wellhead prices contributed to slow growth in oil production during the 1980s, even though a two-tiered pricing system was introduced in 1982. In 1988 the government increased plan prices, but by a very small amount compared to market levels, and by 1990 the cost of production had surpassed the average in-plan well-head price. Distortions in pricing of upstream and downstream products allowed refineries to profit at the expense of oil fields. Local governments built many small, inefficient refining facilities to capture rents created by this system. At the same time, oil exploration and extraction became much less attractive, resulting in a slowdown in oil field development.

In 1994, the two-tiered pricing system was replaced by a single set of central price controls on crude oil and oil products that were adjusted periodically with reference to international prices. Faced with fixed ex-refinery and retail prices up to twice the level of import prices, oil consumers sought alternatives to the high-priced oil. By 1998, smuggling had become rampant, and a new import and pricing regime—linking Chinese domestic prices for the first time to the Singapore market—was announced along with a crackdown on smuggling. Consequently, oil imports dropped in 1998 (before rising again; Fig. 5) and smuggling activity has declined. In 2001, the basis for the pricing regime was broadened to refer to Rotterdam and New York prices as well as Singapore prices. As yet, China has few domestic price-discovery mechanisms and continues to rely on references from international markets.

Natural gas prices, particularly the large portion of gas provided to fertilizer plants, remain tightly controlled by the government, which has tried to make investment in natural gas more attractive by adjusting domestic prices closer to international levels.

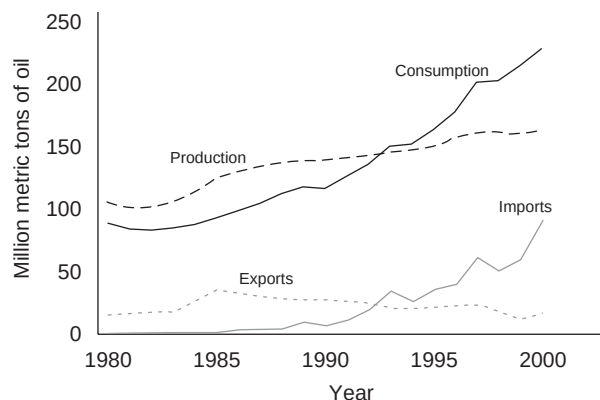


FIGURE 5 Oil production, consumption, and international trade, 1980–2000. In 1993, China went from being a net exporter to a net importer of oil.

However, gas pricing varies widely from city to city, and in the absence of a national pipeline network, a true market for natural gas does not yet exist.

Pricing policy in the electricity sector has been motivated by attempts to ease the serious electricity supply shortages that plagued much of China until the late 1980s. To encourage electricity production, the central government established a multitiered and diversified price system for the sector in 1985. Prices for enterprise-owned power plants (China's version of independent power producers) and small hydro-power stations were set higher than plan prices for state-run utilities, and state-owned utilities were allowed to sell above-quota generation at higher prices. As shortages persisted in the late 1980s, the government allowed tariffs to be set on the basis of total costs plus a profit margin to encourage capacity expansion. In 1987, peak-load and seasonal pricing were introduced in some areas. In 1988, state-owned enterprises were levied an added electricity consumption tax for national electric power development, in addition to other national and local taxes on electricity, e.g., the national tax for financing the Three Gorges Dam project. Electricity price classifications, as well as rules and regulations for price settings, remain confusing, defeating any incentives for users to conserve. Prices vary greatly depending on region and class of user, as well as between urban and rural areas. The impact on prices of the current wave of power-sector reforms remains unclear.

2.9 Energy and the Environment

All across China, and particularly in the wealthy coastal provinces, cities and towns are becoming stricter in enforcing limits to pollutant emissions.

When local administrations are supportive, environmental protection bureaus can levy significant emissions fees and fines, mandate process changes, and apply more drastic measures. Forceful application of environmental regulations could change industrial energy demand significantly. The regulation of sulfur dioxide emissions in China's legislatively defined "acid rain control zones" may, for instance, result in greater use of washed coal and installation of flue-gas desulfurization (FGD) equipment at power plants. Coal washing would provide a higher heat-content product that would burn more efficiently, reducing demand for coal, all else being equal. FGD, on the other hand, requires a great deal of a power plant's output, raising demand for coal inputs to power generation. Requiring urban factories to move or to replace equipment often results in the use of newer, generally larger, cleaner, and more efficient equipment. Sustained support for environmental policies is a factor in the continuing decline of coal use by households. In households, coal is being replaced by LPG, natural gas, town gas, and electricity.

3. OUTLOOK FOR CHINA'S ENERGY POLICY

China's demand for energy services will grow to meet the demands of its expanding economy and urbanizing population. Widely circulated baseline scenarios of primary energy demand in China to 2020 forecast total energy use to be between 70 and 90 EJ in 2020, compared to actual consumption in 2001 of 39 EJ (Fig. 6). The country faces numerous challenges in meeting this demand in a way that is economically, socially, and environmentally sound. It will be crucial for China to continue its progress in structuring markets, establishing institutions of corporate governance, revamping the finance sector, and finding other means to fulfill the social-welfare functions that formerly were provided by enterprises. These are broad tasks with multiple goals, but they will affect energy supply and use at least as much as policies aimed directly at energy.

Challenges for energy-supply policy abound. In the long term, commitment to developing renewable energy will have a tremendous impact on fuel structure, but for now conventional energy will be central. For coal, ensuring that large mines become financially viable remains a key task. How China finds ways to improve coal supply to meet environmental goals, such as reducing acid precipitation, will affect how coal is used. The debate, driven by

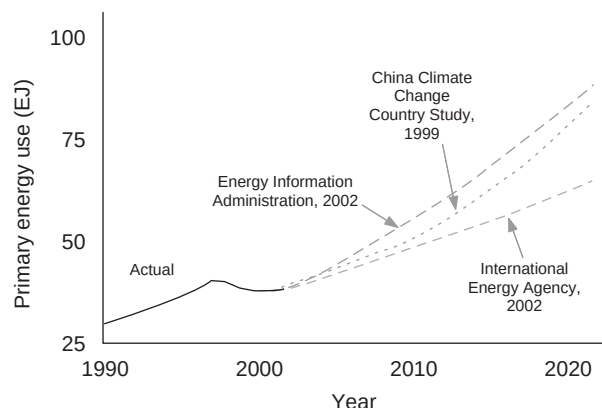


FIGURE 6 Major baseline scenarios of future fossil energy consumption in China to 2020. Widely circulated baseline scenarios of primary energy demand in China to 2020 forecast total energy use to be between 70 and 90 EJ in 2020, compared to actual consumption in 2001 of 39 EJ. The scenarios were adjusted to account for the discrepancies between forecast and actual energy use 2000; trends are the same as in original forecasts.

financial and security concerns, over whether China should rely more heavily on domestic coal resources instead of imported oil, also bears watching, though the potential for greater coal use will hinge again on coal quality, which, in turn, will depend on water availability in the arid coal-mining regions. Greater reliance on coal will signal higher growth in energy use and carbon emissions, because coal emits more carbon dioxide per unit of useful energy. In the long term, China's reliance on coal will be determined not by resources, but by climate change. If, as seems likely, significant reductions of carbon dioxide emissions from fossil fuel use in all countries are required, then coal use will be limited, unless an inexpensive way can be found to permanently sequester large amounts of carbon dioxide.

Because accession to the World Trade Organization has left China with fewer tools to restrict oil imports, policies that affect demand will be crucial. Transport policies that affect mode choices (e.g., road vs. rail; private vehicle vs. public transportation) and fuel-efficiency standards for vehicles, and policies affecting demand for petrochemicals, will influence China's oil imports. Natural gas is such a desirable fuel for so many reasons that consumption will likely be limited by supply. Policies affecting international trade in natural gas, support for construction of pipelines and distribution networks, pricing policies, and regulatory development all become relevant.

For electricity, the main question is how regulatory reform of utilities will proceed. Issues include

how government will reduce its role in the management and operation of utilities, how generation will be separated from transmission and distribution, and how markets for electricity will be transformed and regulated. How China treats the activities of foreign participants in the sector will also be important, as well as national policy regarding development of nuclear power and nonhydro renewables. Changes in these areas will affect what types and scale of new generating units are built, system efficiency, environmental performance, and what kinds of demand-side electricity efficiency programs can be feasibly deployed in China. Energy efficiency policies are unlikely to bring about great change on their own, but they will help to create attractive opportunities for energy suppliers and users to raise efficiency—and slow growth in energy use—when the economic and institutional environments permit.

Within all these areas, there are substantial opportunities for international assistance and cooperation. On the supply side, efforts to help ease China's participation in international energy markets will be important to ensuring that China has access to adequate supplies of oil and natural gas. International experience is a valuable guide to establishment of China's national gas network. For the other oil-importing countries, a key challenge will be to find ways to accommodate China's needs for access to energy from the Middle East and Central Asia. Although China still has some claim to status as a developing nation, it is increasingly a heterogeneous country, with aspects that are highly developed and others that are much less developed. Consequently, the kind of multi- and bilateral assistance that would be most valuable is different from that usually provided to developing countries. As the country's financial system evolves, direct grants and loans will become less important than efforts to develop strong domestic financial institutions and commercial links to international capital markets. Promoting development and transfer of efficient technologies could be particularly valuable, including joint precommercial research and development. Even though most exchanges affecting how particular actors obtain and use energy will occur in the commercial realm, cooperation at the policy level remains important. If the guiding philosophy behind assistance is one of helping a partner build capabilities that will serve mutual interests, rather than one of providing a handout to a poor neighbor, then the chances will be much greater that strong, peaceful, trusting, collaborative relationships will develop, relationships that

will allow countries to work together to solve global challenges.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • European Union Energy Policy • National Energy Policy: Brazil • National Energy Policy: India • National Energy Policy: Japan • National Energy Policy: United States • Rural Energy in China

Further Reading

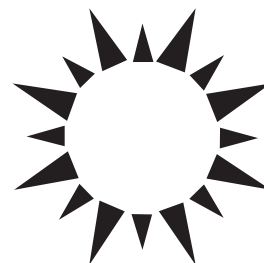
- Andrews-Speed, P. (2001). China's energy policy in transition: Pressures and constraints. *J. Energy Lit.* VII(2), 3–34.
- Deng, K. Y., Gu, S. H., and Liu, W. Q. (1996). Rural energy development in China. *Energy Sustain. Dev.* III(3), 31–36.
- Horii, N., and Gu, S. H. (eds.). (2001). "Transformation of China's Energy Industries in Market Transition and Its Prospects." Institute of Developing Economies, Japan External Trade Organization, Chiba, Japan.
- International Energy Agency (IEA). (2000). "China's Worldwide Quest for Energy Security." Organization for Economic Cooperation and Development, Paris.
- Lieberthal, K., and Oksenberg, M. (1988). "Policy Making in China: Leaders, Structures, and Processes." Princeton Univ. Press, Princeton, New Jersey.
- Lu, Y. Z. (1993). "Fueling One Billion: An Insider's Story of Chinese Energy Policy Development." The Washington Institute, Washington, D.C.
- McElroy, M. B., Nielsen, C. P., and Lydon, P. (eds.). (1998). "Energizing China: Reconciling Environmental Protection and Economic Growth." Harvard Univ. Press, Cambridge, Massachusetts.
- National Bureau of Statistics. (2002). "China Statistical Yearbook 2002." China Statistics Press, Beijing, China.
- Sinton, J. E., Levine, M. D., and Wang, Q. Y. (1998). Energy efficiency in China: Accomplishments and challenges. *Energy Policy* 26(11), 813–829.
- Smil, V. (1988). "Energy in China's Modernization: Advances and Limitations." M. E. Sharpe, Armonk, New York.
- Thomson, E. (2002). "China's Coal Industry: An Economic History." Routledge Curzon Press, London.
- World Bank. (1997). "Clear Water, Blue Skies: China's Environment in the New Century." The World Bank, Washington, D.C.
- Yang, F. Q., et al. (1995). "A Review of China's Energy Policy." Report No. LBL-35336. Lawrence Berkeley National Laboratory, Berkeley, California.
- Zhao, J. M. (2001). "Reform of China's Energy Institutions and Policies: Historical Evolution and Current Challenges." Energy Technology Innovation Project, John F. Kennedy School of Government, Harvard University. Harvard Univ. Press, Cambridge, Massachusetts.

National Energy Policy: India

R. K. PACHAURI and PREETY BHANDARI

The Energy and Resources Institute (TERI)

New Delhi, India



-
1. Evolution of the National Energy Policy in India
 2. Energy for All
 3. Ensuring Security of Energy Supplies
 4. Improving the Efficiency of the Energy System
 5. Reducing the Negative Environmental Impacts
 6. Imperatives for the National Energy Policy
 7. Implications for National Energy Policy

Glossary

biogas A gas composed of methane and carbon dioxide; produced from the anaerobic decomposition of organic material in landfills, biogas fuel provides a medium level of energy (British thermal units); also called biomass gas.

chulhas Traditional cooking devices that use biomass as a fuel.

coal bed methane (CBM) An environmentally friendly, clean fuel with properties similar to those of natural gas. Most of CBM is in an adsorbed state on the micropores on the surface of coal.

energy intensity The amount of energy required by an economy to produce one unit of national product.

fuel cell A type of cell capable of generating an electrical current by converting the chemical energy of a fuel directly into electrical energy. Fuel cells differ from conventional electrical cells in that the active materials, such as fuel and oxygen, are not contained within the cell, but are supplied from outside.

gas hydrates Solid, crystalline, waxlike substances composed of water, methane, and usually a small amount of other gases, with the gases being trapped in the interstices of a water and ice lattice. Gas hydrates form beneath the permafrost and on the ocean floor under conditions of moderately high pressure and at temperatures near the freezing point of water.

reserve replacement ratio The amount of natural resource added in a unit of time. It is calculated as the gross addition of reserves minus the production during the time frame, taken as a ratio of existing reserves.

solar thermal collector A device designed to receive solar radiation and convert it to thermal energy. Normally, a

solar thermal collector includes a frame, glazing, and an absorber, together with appropriate insulation. The heat collected by the solar thermal collector may be used immediately or stored for later use. Solar thermal collectors are used for space heating, for domestic hot water heating, and for heating swimming pools, hot tubs, or spas.

In India, several committees appointed by the government to review energy policy from time to time have emphasized the need for a nationally coordinated approach to energy policy formulation. As a result, sporadic efforts have been made to bring together various departments or ministries dealing with energy for the purpose of arriving at a unified and integrated approach. However, this has not met with much success and initial efforts have been abandoned in every case. Currently, the National Development Council (NDC) functions as an umbrella organization to approve of each Five-Year Plan as prepared by the Planning Commission. The development of each successive Five-Year Plan involves a significant effort in consultation between the Planning Commission and several ministries; the plan receives the final stamp of approval from the NDC. The aggregation of component plans, however, is not an effective substitute for a properly integrated comprehensive national energy policy. For instance, several countries worldwide have a single energy ministry that serves the requirement of integration; with the tradition established in India, such an arrangement has neither been attempted explicitly nor has it presented great appeal to political decision makers. With rapid growth of the economy and increase in demand for energy, issues of energy security, efficiency of the entire energy system, and the effects of energy production, conversion, and consumption on the environment require policies that optimize some of these variables on an integrated basis, rather than by fuel or by specific source of energy. There is, therefore, a growing

imperative for India to articulate a forward-looking integrated energy policy.

1. EVOLUTION OF THE NATIONAL ENERGY POLICY IN INDIA

The energy sector in India is administered at the apex level by four ministries and one department, representing the five supply sectors, namely, coal, oil and gas, power, renewable energy, and atomic energy. Coordination and integration are achieved within the Government of India (GoI) through the Cabinet and the Planning Commission. The respective ministries do coordination and integration with state governments, with the National Development Council providing a forum at the highest level.

This structure has evolved over time and has also led to a vacuum for coordinated vision and action on a sustained basis. However, this has not always been the position. A Fuel Policy Committee that had been set up in 1974 took an integrated view of the energy requirement of the country and made recommendations for the entire energy sector. With a rapidly changing international oil scenario, a Working Group on Energy Policy was set up in 1977. Its report, submitted in 1979 (after the beginning of the second oil shock), expressed that there could be no guarantee of steady oil supply to any importing country. It emphasized that although efforts were required to conserve resources of commercial fossil fuels through efficient production and utilization, it was essential to switch to renewable resources such as agricultural and animal wastes and to new and renewable sources such as energy from the sun and wind. An important recommendation of the Working Group was that energy planning and policy formulation should be done on a full-time basis and an energy ministry, or, alternatively, an energy commission, should be in place to deal with all these matters.

In 1982, the Ministry of Energy, comprising the Departments of Coal and Power, was expanded to include the Departments of Petroleum and Nonconventional Energy Sources. Thus, except for atomic energy, all other energy departments in the GoI had been brought under one ministry. In addition, in 1983, the GoI established an Advisory Board on Energy. The functions of the board included formulating an integrated energy policy covering commercial and noncommercial sources of energy. Its functions were advisory in nature and it was to submit its reports directly to the Prime Minister

(Table I). Since then, this structure has become more decentralized, with all the energy subsectors falling under separate ministries or departments. It is only at the time of the formulation of Five-Year Plans (FYPs) that the sector is looked at as a whole. A brief review of the various pronouncements related to the energy sector in the FYPs is summarized in Table II.

TABLE I
Evolution of National Energy Policy

Major milestone	Impact/function
Fuel Policy Committee (1974)	Looked at energy sector as a whole
Working Group on Energy Policy (1977)	Report submitted in the aftermath of the second oil shock expressed apprehensions regarding security of oil supplies and thus recommended conservation of energy and encouragement of renewable energy
Expansion of Ministry of Energy to include Departments of Petroleum and Nonconventional Energy Sources (1982)	Except for atomic energy, all other forms of energy were brought under one ministry
Advisory Board on Energy (1983)	Set up to formulate an integrated energy policy covering commercial and noncommercial energy resources
Exploration rounds	The exploration rounds to attract the private sector were introduced in 1979, but not much headway was made till 1995
Committee on Integrated Coal Policy (1996)	Recommended adoption of coal conservation measures, inviting private capital in the sector, deregulating the coal and coal product prices, and setting up a regulatory body, among other things
Common Minimum National Action Plan for Power (1996)	Initiated the reforms in the power sector with trifurcation of state electricity boards, setting up of state and union level regulatory commissions, and rationalization of tariffs
New Exploration and Licensing Policy (NELP) (1997)	NELP was introduced in 1997 with modified terms and conditions; three rounds have been held with moderate success
Administered Pricing Mechanism (APM) dismantling (1997)	Phase-wise dismantling of APM started in 1997; APM was completely abolished early in 2003

Broadly speaking, the overall objective of a National Energy Policy (NEP) should be to meet the energy needs of all segments of population in the most efficient and cost-effective manner while ensuring

TABLE II
Five-Year Plans^a

Plan	Goal
First FYP (1951–1956)	Set out the objectives of national planning with immediate focus on agricultural and industrial development and attendant development of irrigation and power
Second FYP (1956–1961)	Emphasized the long-term perspective while planning for short-term horizons
Third FYP (1961–1966)	Elaborated the need for a comprehensive review of the extent of available information and data on India's natural resources
Sixth FYP (1980–1985)	Aimed at a decentralized approach to increasing availability of renewables to meet rural energy requirements; thrust was on management of oil demand, conservation, and renewable resources; contained the first oblique reference to the environmental impact of energy use
Seventh FYP (1985–1990)	Recognized that energy planning involves not only potential increase in indigenous availability of energy, but also better utilization, because the trends in India's commercial energy consumption show high rates of growth and because of the growing share of oil dependency
Eighth FYP (1992/ 1993–1996/ 1997)	The environmental dimension was given due significance only in this plan, which emphasized the long-term need for promotion of technologies of production and consumption of energy that are environmentally benign and cost-efficient
Ninth (1997–2002) and Tenth FYP (2002–2007)	These plans placed special emphasis on the need to introduce reforms in the energy sector in order to improve efficiency in the sector and to enhance infrastructure development in view of the impending growth in the sector in the long term. Due importance has also been given to the pricing issues in the sector and the need to make prices indicative of the market. Concern was also raised regarding the direct and indirect subsidies present in the sector and the need to streamline them and make the process more focused. Energy conservation is to be promoted in the plan periods mainly through improvements in the industrial sector

^aData are from each plan, as specified by the Indian government.

ing long-term sustainability. This objective entails the following major tasks:

- Providing clean and affordable energy to all.
- Ensuring security of the energy supply.
- Improving the efficiency of the energy system.
- Reducing the adverse environmental impacts of energy use.

The following discussions review the Indian experience in light of these objectives.

2. ENERGY FOR ALL

2.1 Present Scenario in Energy Consumption

Total primary energy consumption in India has increased nearly fivefold in the three decades from 1970 to 2001. The total primary energy consumption stood at 437.7 million tonnes of oil equivalent (MTOE) in 2001/2002 (GoI 2002). The share of commercial energy in total primary energy consumed rose to more than 68% in 2001, up from 28% in 1950 and, as the economy develops, this share is expected to rise even further. However, the per capita energy consumption in India is still 479 kilograms of oil equivalent (kgOE), which indicates a tremendous scope for even higher energy consumption in the coming years. Additionally, the 5.4% annual growth rate of energy consumption achieved by India, over the period 1970–2001, is the 14th highest in the world.

Figure 1 shows the comparative position of India with respect to per capita income and per capita energy consumption. Countries with high per capita income are also characterized by high per

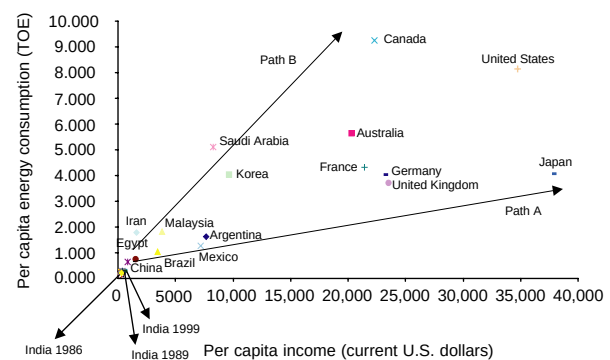


FIGURE 1 Comparison of per capita energy consumption and income in India and other countries. The gross domestic product and population data are from the World Bank; energy consumption data are from *BP Statistical Review of World Energy* (2002).

capita energy consumption. This is not to suggest that India should be looking toward energy consumption levels of the West, as it marches on its path of development, because that would simply be unsustainable. In fact, it does indicate that there is tremendous scope for increase in energy consumption, admitting, though, that the link between energy and economy will gradually become weak as the economy matures and moves more toward the services sector. This is shown in Table III, which gives the trend in elasticity of consumption of various fuels with reference to the gross domestic product (GDP) over time. From Fig. 1, it is evident that Japan is more efficient than the United States because its energy consumption is much lower than that of the United States for the same level of per capita income. Accordingly, developing countries could aim to follow path A for development in order to build a sustainable economy. Development along path B will result in a high, potentially fatal dependence on energy for countries like India, which are dependent on imports for meeting their energy needs. Figure 2 shows the path that India has taken for the past two decades. As shown, since 1990, the per capita energy consumption in India has not risen commensurately with its per capita income, which indicates that it is possible to reduce energy intensity and/or improve energy efficiency in the economy while still enabling economic development.

The elasticity of consumption of various fuels with respect to the GDP has declined over the years, with current commercial energy elasticity lower than unity; this means that a 1% increase in the GDP will increase the commercial energy consumption by less than one. However, underprivileged sections of

TABLE III

Elasticity of Consumption of Commercial Energy with Respect to GDP^a

Period	Coal	Oil	Electricity	Total commercial energy
1953/1954–1960/1961	1.10	2.14	3.02	1.37
1960/1961–1970/1971	0.53	2.31	3.04	1.16
1970/1971–1980/1981	0.98	1.83	2.06	1.37
1980/1981–1990/1991	0.31	1.12	1.57	0.82

^aData from Sengupta (1993).

Indian society do not have access to efficient sources of energy. The World Bank estimates indicate that, of the 2 billion people who have no access to modern energy sources such as electricity and liquefied petroleum gas (LPG), about 700 million live in India. Within India, there is a big rural–urban divide in terms of access to modern energy sources (Figs. 3 and 4). Access to modern fuels can be improved by short-term government measures such as subsidies as was done by the introduction of Public Distribution Systems to make available kerosene to the underprivileged section. However, these measures have not proved to be sustainable in the long term. State interventions in the form of programs for poverty alleviation are important to enable continued use of modern resources even after subsidies are phased out.

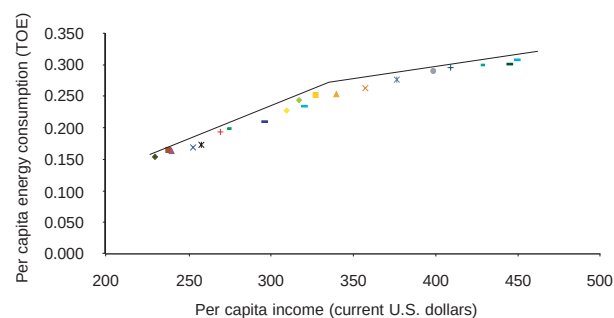


FIGURE 2 Trend in Indian per capita income and energy consumption (countries are represented by symbols as in Fig. 1). The gross domestic product data are from *Economic Survey 2001/2002*; population data are from *Provisional Census (2000/2001)*; energy consumption data are from *BP Statistical Review of World Energy (2002)*.

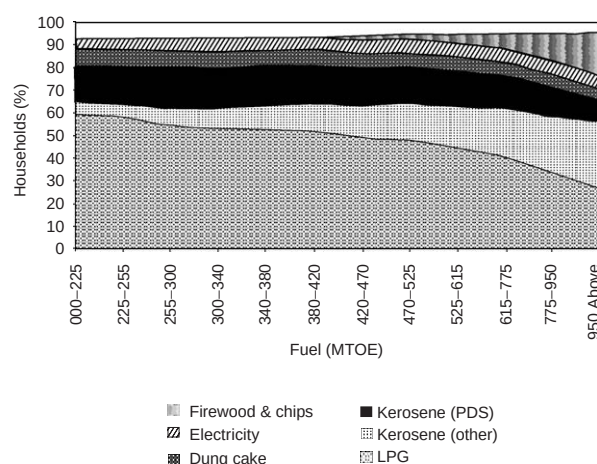


FIGURE 3 Fuels used by rural households (percentage). PDS kerosene is kerosene sold through the “fair price” shops under the Public Distribution System. Data are for 2001, from the National Sample Survey Organization.

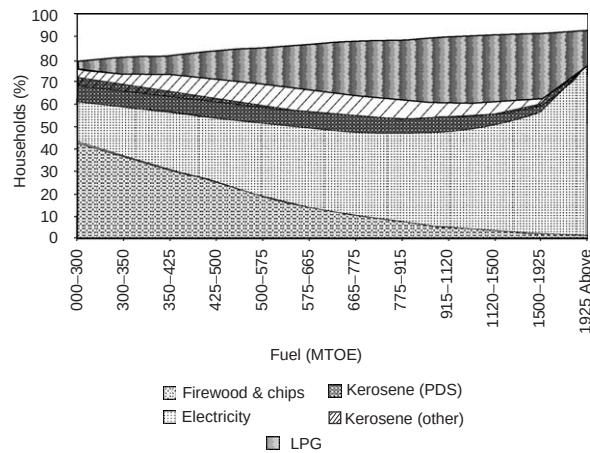


FIGURE 4 Fuels used by urban households (percentage). PDS kerosene is subsidized kerosene sold through the “fair price” shops under the Public Distribution System. Data are for 2001, from the National Sample Survey Organization.

Figs. 3 and 4 show that the percentage of electricity usage among all classes of urban households is quite high, as is the consumption of LPG. However, the dominant fuel among the rural households is firewood, which is both inefficient and leads to indoor air pollution. Thus, promoting fuels such as electricity and LPG would not only promote efficiency but would also help curb respiratory problems among rural women.

2.2 Energy–Economy Linkage

Energy is the key driver of many sectors of the economy, namely, agriculture, industry, commerce, services, and domestic. With advances in technology and the growing need for higher productivity, most of these sectors are also becoming more energy intensive. This is especially true in developing countries, which show a strong energy–economy linkage in the form of high-energy intensity/elasticity of the economy. The developed countries have been able to weaken this linkage due to development of more efficient processes that can do the same amount of work with less energy, or more work with the same amount of energy. However, developing countries—most of them highly industrialized—are still highly dependent on energy.

The picture is not different for India. Though India now boasts of a big services sector, accounting for 49% of its GDP, the energy intensity of Indian economy is still high. Figure 5, which shows the trends in the GDP and the primary energy consumption in India since 1970, indicates a strong linkage

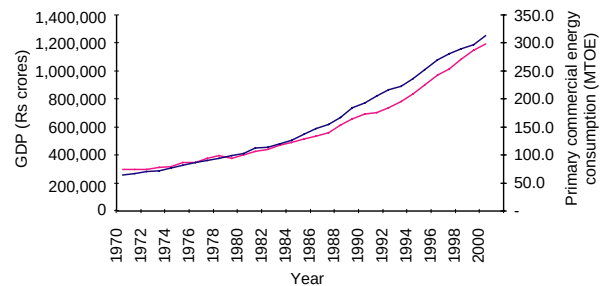


FIGURE 5 Comparative trends in gross domestic product (GDP) and total commercial energy consumption. The GDP is expressed in crores (=10 million) of rupees (Rs). The gross domestic product data are from *Economic Survey 2002/2003*; energy consumption data are from *BP Statistical Review of World Energy (2003)*.

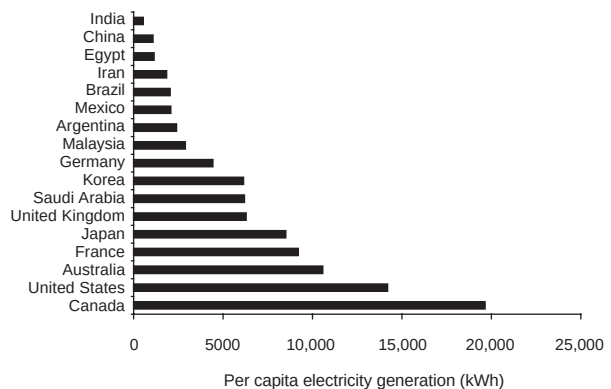


FIGURE 6 Comparison among countries with respect to per capita electricity generation (in kWh, kilowatt-hours). Electricity data are from the International Energy Agency; population figures are from the World Bank.

between energy and economy. However, the years since 1985 have witnessed a gradual distancing between the two curves that is indicative of a growing share of the services sector, which is less energy intensive as compared to the industrial sector. Per capita electricity generation and oil consumption levels in India are among the lowest in the world, lower even than those of its neighbors, as shown in Figs. 6 and 7.

3. ENSURING SECURITY OF ENERGY SUPPLIES

The issue of ensuring affordable supplies of energy in various required forms is central to India’s socio-economic development. As already mentioned, the link between energy and economy in India is still strong and any disruption in energy supplies is bound

to have negative impact on the economy. The important issue is to assess both the vulnerability of the economy to an external supply shock and the measures that can be taken to cushion the economy against such a shock.

With the entry of the private sector, market forces in the energy sector have added a new dimension to the whole issue of energy security. Market forces will ensure that, during crises, demand and supply will always balance. Prices will indicate the relative

shortage and surplus in the sector. Thus, the impact of external shock will be felt in prices rather than in demand and supply.

3.1 India's Fuel Mix

India's energy mix has not undergone much change in the past 30 years. Coal still remains the dominant fuel; 55% of the total primary energy is accounted for by coal, down from 58% in 1970, though the consumption of coal has grown by 5.21% annually. The share of oil has increased marginally from 30% in 1970 to 31% in 2001, though consumption has grown at 5.5% annually since 1970. Gas has witnessed major gain, increasing by a share of 8%, which is up from virtually nothing in 1970, with gas consumption growing at 13.73% annually. However, because of domestic discoveries, it is only since the 1980s that gas has really seen major improvement. The share of hydropower has declined, although it displayed an annual growth rate of 2.38% over the past 30 years. This is primarily due to high growth rates achieved by oil and gas (Table IV).

The high percentage of oil consumption in the economy, along with high oil elasticity, points toward increasing dependence of Indian economy on oil as a fuel. This dependence becomes more

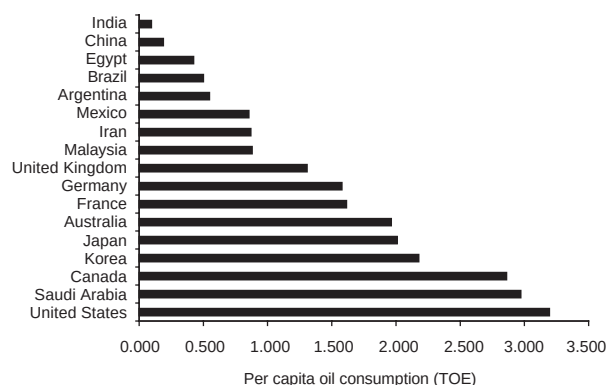


FIGURE 7 Comparison among countries with respect to per capita oil consumption (in TOE, tonnes of oil equivalent). Oil consumption data are from *BP Statistical Review of World Energy* (2003); population figures are from the World Bank.

TABLE IV

India's Fuel Mix^a

Fuel	Mix per year (in MTOE)					
	1953/1954	1960/1961	1970/1971	1980/1981	1990/1991	2001/2002
Commercial primary energy						
Coal	23.62	35.64	36.48	56.96	94.68	133.89
Lignite	—	0.01	0.81	1.23	3.34	6.52
Crude oil	0.19	0.46	7.01	10.79	33.92	32.03
Natural gas	—	—	0.6	1.41	11.73	26.72
Hydropower	0.24	0.67	2.17	4	6.16	6.37
Nuclear power	—	—	0.63	0.78	1.6	5.15
Wind power	—	—	—	—	—	0.14
Total	24.05	36.78	47.7	75.17	151.43	210.82
Net imports (+)	2.2	6.04	12.66	24.63	31.69	87.85
Stock changes (—)	0.24	2.87	0.69	3.8	5.37	—
International bunkers (—)	0.53	0.5	0.24	0.21	0.14	—
Total commercial energy supply	25.48	39.45	59.43	95.79	177.61	298.67
Noncommercial primary energy supply	64.13	74.38	86.72	108.48	122.07	139.02
Total primary energy supply	89.61	113.83	146.15	204.27	299.68	437.69

^aData from the Tenth Five-Year Plan, developed in 2002 by the Planning Commission.

threatening if it is seen in conjunction with the meager quantity of domestic crude oil production and, consequently, the high level of dependence on imported oil. However, the economy's dependence on oil is much more than is suggested by the share of 31%. This is because of the heavy dependence of the transport sector on oil. The avenues for fuel substitution in the transport sector are limited, notwithstanding recent endeavors to use compressed natural gas (CNG) in lieu of gas oil and gasoline. Initiatives to increase use of solar-powered vehicles and fuel cell vehicles are still many years from impacting the share of oil use by transport sector. Certain reform strategies to improve energy efficiency, however, can be implemented with immediate effect in the transport sector:

- Measures to improve the efficiency of oil use in vehicles (this would also improve the already degraded urban environment).
- Taxes on fuel consumption to signal importance of fuel conservation.
- Strict adherence to vehicular fuel-efficiency parameters.
- Policy interventions to encourage faster turnover of old vehicles.
- Incentives to attract foreign players into the market to create competition, promoting the latest and most efficient technologies.
- Improvements in highways and internal city and town roads and promotion of foreign direct investment (FDI).
- Strengthening of public transport systems to reduce the use of personal vehicles.
- Policy options and improvements in railways to promote cargo movement and to reduce the pressure on roadways.
- Policy options and rapid work on the megaproject to link rivers across the country, to reduce the load on road transportation.

The issue of energy security assumes even greater importance when future energy demand and supply scenarios are examined.

3.2 Future Energy Demand

The preceding discussion highlighted the low per capita income and high per capita energy consumption that characterize the Indian economy. These trends indicate the possibility of even larger energy consumption in the coming years. The *International Energy Outlook 2003*, published by the U.S. Department of Energy, projects India's energy consumption

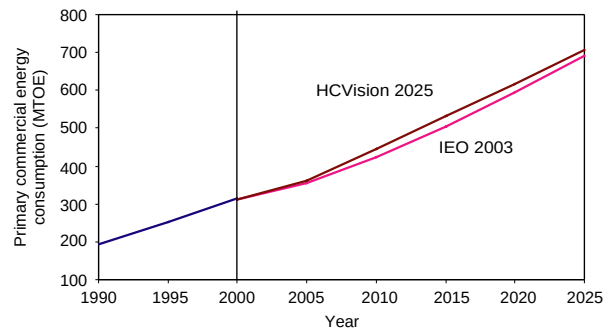


FIGURE 8 Past and projected energy consumption trend. MTOE, Million tonnes of oil equivalent. Data are from *Hydrocarbon Vision 2025* (HCVision) and *International Energy Outlook 2003* (IEO); additional data are from *TERI Energy Data Directory and Yearbook* for years 1999–2000.

at 690 MTOE by 2025, up from 322 MTOE in 2001, and an average annual percentage change of 3.2% during the period 2001–2025. Based on the population estimates given by the World Bank, the per capita energy consumption works out to be 529 kgOE, which is 10% higher than the current level. The energy intensity at this level of energy consumption, assuming a GDP growth rate of 6%, works out to be 13.67 kgOE per 1000 rupees, which indicates a decline of 43% over the 22-year period (2003–2025). This level of consumption will increase India's share in total world primary energy consumption to 4.27%, up from the current level of 3.1%.

Hydrocarbon Vision 2025, the long-term energy policy document published by the GoI, also projects energy demand and fuel mix for the period 2002–2025. It pegs India's total primary energy consumption at 616 MTOE in 2020, of which 238 MTOE will be derived from oil, representing 38% of the total; gas will contribute 16%; coal, 38%; hydro, 6%; and the rest will be met by nuclear power. These projections are not very different from the *International Energy Outlook 2003* projections, as shown in Fig. 8. The resultant change in fuel mix is shown in Fig. 9. The share of coal is projected to decrease, from 55% in 2001 to 38% in 2020 (Fig. 9), whereas that of natural gas is projected to double to 16%. The share of oil is projected to rise to 38%. This fuel mix indicates a gradual shift toward cleaner fuels (natural gas and oil) in the wake of rising environmental concerns. On the other hand, the projection also indicates a rise in dependence of the economy on imports, because domestic reserves of oil and gas cannot support such high consumption levels under the most optimistic scenario.

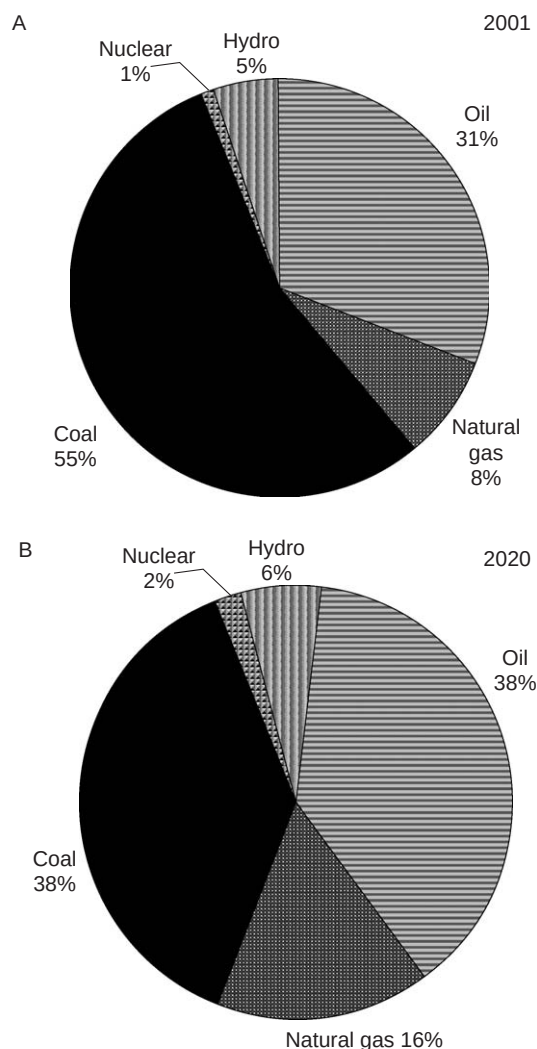


FIGURE 9 Current (A) and projected (B) commercial energy mix. Data are from *Hydrocarbon Vision 2025*.

3.3 Present Resource Availability and Future Outlook

Given that India is a huge consumer of energy, supply options are very limited. Already, India imports around 74% of its crude oil requirement and, though India has vast reserves of coal, coal imports are on the rise due to the low quality of native coal and the faulty taxation structure that makes coal imports cheaper for some consumers located near the coast. Development in the nuclear energy sector has increased in recent years, but it is unlikely that this sector will play a role in the development of Indian economy comparable to the role it played in France. India has a huge potential for hydropower, especially

TABLE V
Coal, Oil, and Natural Gas Reserves^a

Fuel	Reserves	Current production	Reserve replacement ratio
Crude oil (million tonnes)	644.8	32.45	0.59
Natural gas (billion cubic meters)	647.5	26.57	0.33
Coal (million tonnes)	213,905.5	322.2	8.46

^aData from TEDDY 2002/03, see TERI (2003).

small/micro hydro/electric (hydel) power projects, but these opportunities are concentrated in a few states. This makes it imperative to improve the efficiency of the transmission system so that surplus power in one state can be reliably transferred to a deficit state. To achieve this, reforms in the power sector are essential. Table V shows the level of reserves in coal, oil, and natural gas in the country along with the reserve replacement ratio. The reserve replacement ratio for crude oil, though positive, is still low and the same goes for gas. India's position in coal seems to be comfortable but, as is discussed later, there are some environmental concerns.

Efforts such as the New Exploration Licensing Policy (NELP) are being made to increase the indigenous production of crude oil, but the prospects for discovery of another oil field equivalent to the Bombay High field (discovered in the late 1970s) remain remote. Availability of gas in the form of liquefied natural gas (LNG) is slated to improve in the coming years, but there are many policy issues that remain to be sorted out before gas can be delivered cheaply in the country. Several pipeline projects are also planned, but political risks seem insurmountable. These issues are discussed in detail in the following sections. The locations of various coal, lignite, oil, and gas fields are shown in Fig. 10.

3.3.1 Crude Oil Supply Outlook

It has been almost two decades since Bombay High was discovered. Given the fact that consumption is around 100 million tonnes (MT) of petroleum products every year and production is only 33 MT, four more Bombay Highs would be required for India to become self-reliant in oil at present. As oil consumption rises in the future, discovering even more domestic reserves becomes imperative with



FIGURE 10 Crude oil, natural gas, and coal/lignite reserves in India. Data are from The Energy and Resources Institute.

respect to security in both oil supplies and prices. This appears far-fetched, given the lackluster response to the NELP by international oil companies, which were expected to bring in both technology and capital. Though the Oil and Natural Gas Corporation Ltd. (ONGCL) is planning to double its reserves in 20 years, the oil demand by then would also have more than doubled, leaving the import dependency unchanged. The Planning Commission has projected domestic production of only 45 MT by 2011/2012 against the projected demand for petroleum products ranging from 187 to 160 MT. This indicates high import dependency with all its repercussions on energy security and the foreign exchange reserves. The low prospect of discoveries in Indian sedimentary basins, due to poor exploration methodology, is an indication of the need for diversification of energy sources and application of effective and efficient

exploration techniques to discover new oil and gas basins in India.

3.3.2 Electricity Supply Outlook

Given the high positive elasticity of electric power with respect to the GDP, reliable electricity supply is a necessary ingredient for economic development. The current installed capacity in India is 105,083 megawatts (MW). However, there is a 10.5% energy shortage and a 15.9% peak shortage. The high transmission and distribution (T/D) losses mean that even the peak demand of 78,841 MW, which is 26,242 MW less than the installed capacity, is not being met. To meet the electricity requirement, the National Thermal Power Corporation (NTPC) is planning to add 20,000 MW of additional capacity by 2020; but these plans are largely dependent on the availability of LNG at reasonable prices. The Central

Electricity Authority (CEA) projects an addition of 47,000 MW during the period 2002–2007, which seems highly unlikely, given that India has been able to add only about 20,000 MW per plan period previously. The private sector was expected to play an important role in improving the power situation in the country, but the slow pace of reforms has discouraged investors, and many independent power producers (IPPs) have thus abandoned their plans for new plants. As indicated previously, to realize the full potential of hydropower, substantial investments in T/D are required.

The government has targeted a nuclear-based capacity of 20,335 MW by 2020. Considering the fact that current nuclear-based capacity is only 2720 MW, this also seems a bit ambitious. This is not only because of high initial investment costs, but also due to many other issues concerning nuclear power, such as long development period, reactor safety, waste disposal, and the continuance of government support. Thus, in the present scenario, nuclear energy is not expected to play the same important role that it played in France. The important aspect that emerges is that interventions are required to expedite the pace of reforms, attract IPPs back into the energy market in India, and create investment opportunities in T/D of power grids to tap the high prospect of hydropower.

3.3.3 Gas Supply Outlook

Gas is slated to play an important role in India's energy mix. As per *Hydrocarbon Vision 2025*, the share of gas in total energy consumption is projected to be 16%, up from 5% at present. However, projections for indigenous gas production imply that much of this gas is likely to be imported, primarily in the form of LNG. The Planning Commission has projected domestic production of 105 million standard cubic meters per day (MSCMD) by 2011/2012; the demand projected by *Hydrocarbon Vision 2025* ranges from 216 to 313 MSCMD in the year 2011/2012.

The recent discovery of 10.5 trillion cubic feet (TCF) of natural gas by Reliance in the Krishna–Godavari Basin is expected to boost the current gas production by 50% in 2–3 years. This discovery, in the NELP 1 block, has led to an improved outlook on indigenous gas production, and many plans for importing natural gas are being reviewed. However, this gas is concentrated in southeastern India and the major demand centers are in northern and western India. Whether this gas will flow to these demand centers remains to be seen, but in the present

scenario, the dependence on imported gas will continue. Several proposed LNG projects will serve to fill this gap, but progress is slow due to various issues concerning the reforms in the oil and gas sector in the country.

Several pipeline projects are also planned to reduce the demand–supply mismatch in the country, including the Indo–Iran pipeline, the Indo–Oman pipeline, and the Indo–Bangladesh pipeline. The security concerns surrounding the Indo–Iran pipeline are given great weight by the policymakers, rendering this project, too, a nonstarter. The Indo–Oman pipeline was officially declared dead when the feasibility report admitted the lack of technology for deepwater pipelines. Bangladesh has recently hinted at limited gas exports to India, but the issue is still under debate.

The GoI has also framed policies for the development of nonconventional gas resources such as coal-bed methane (CBM) and gas hydrates. India has 850 billion cubic meters (BCM) of assessed CBM resources and 6150 trillion cubic meters (TCM) of assessed gas hydrate reserves. Private sector players such as Reliance and Essar have evinced interest in CBM, and the National Gas Hydrates Program is underway to develop gas hydrates. All these initiatives to strengthen the nonconventional gas resources are steps in the right direction and will boost India's energy security.

4. IMPROVING THE EFFICIENCY OF THE ENERGY SYSTEM

Figure 11 shows energy intensity (kilograms of oil equivalent per 1000 rupees) and energy consumption

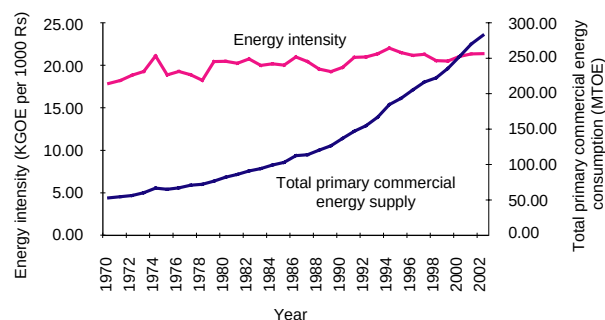


FIGURE 11 Trend in commercial energy intensity and total commercial energy consumption (in kgOE, kilograms of oil equivalent); Rs, rupees; MTOE, million tonnes of oil equivalent. Total commercial energy consumption data are from *BP Statistical Review of World Energy (2003)*; gross domestic product data are from *Economic Survey 2002/2003*.

over several decades. The rate of growth of commercial energy intensity of the economy has been slower than that of the total commercial energy consumption. This decline is due to the high growth in the services sector and/or improvements in the energy efficiency of the economy.

Sectoral trends in energy intensity also yield important insights into the country's energy efficiency. The energy intensity of the industrial sector has declined since the 1980s, primarily due to a decline in the oil intensity of the industry, although intensity of gas usage has increased. Presently, gas demand in the country is constrained due to deficiency in supply, even though gas is more efficient than oil in industrial applications. Energy intensity of the agriculture sector has risen, primarily due to a huge rise in electrical energy intensity of the sector. This trend is also expected to continue as Indian agriculture tries to rely more on man-made irrigation options than on monsoons. Also, new-age farming processes require measured irrigation at the right time to yield desirable results. This will further necessitate the use of pump sets in irrigation.

Energy intensity of transport (excluding railways) rose till 1989/1990, when it reached its peak, and declined thereafter. This is probably due to the introduction of more fuel-efficient vehicles following the liberalization of the economy in 1990/1991. Improving efficiency in the sector would entail introducing market-based reforms. Some progress

has been made along these lines (a detailed analysis of the status of reforms to date is discussed later). According to estimates of the United Nations Development Program (Table VI), there is still room for improvement in energy efficiency in various sectors of the economy.

5. REDUCING THE NEGATIVE ENVIRONMENTAL IMPACTS

5.1 Effects of Conventional Energy on the Environment

Being a developing country, India does not have any commitments for reduction of CO₂ (carbon dioxide) emissions, though India has ratified the United Nations Framework Convention on Climate Change (UNFCCC) and is party to several international environmental treaties (e.g., the Montreal Protocol). However, the noncommitment on emissions should not lead India to a pursue development that is unsustainable and has adverse impacts on the environment. Through the concept of sustainable development, a long-term vision that can bind various policy objectives of each subsector of the economy should be developed. Reducing environmental distress should undoubtedly be among the main policy objectives of an integrated energy policy.

India's energy-related carbon emissions have grown ninefold over the past four decades. With 162 MT of carbon released from consumption and burning of fossil fuels in 2001, India ranked fourth in the world. Coal burned in electricity generation accounts for 64% of all carbon emissions from coal. Technology improvements and diversification of fuels for the power sector are an utmost priority. In a business-as-usual scenario, according to *International Energy Outlook 2003* projections, CO₂ emissions from the power sector will rise to 261 MT of carbon equivalent by 2020. The average annual percentage change in CO₂ emissions from 1990 to 2020 will be 2.2%. India should gear up now, because it may be called on to cut emissions in the second commitment period (2012–2018). Given the long lives of power plants, these issues have to be kept in mind while planning national electricity capacity. Due emphasis should be given to research and development to implement effective and efficient technology for future emissions reductions.

India's carbon intensity is also high, primarily due to low efficiency of coal-based generating plants. Taxation policies that make domestic coal competi-

TABLE VI

Potential for Improvement in Energy Efficiency^a

Sector	Energy efficiency (%)
Industry	
Iron and steel	15
Cement	17
Pulp and paper	20–25
Textile	23
Aluminium	15–20
Household	
Lighting	10–70
Refrigeration	25
Air-conditioning	10
Agriculture	
Pump sets	25–55
Transportation	
Cars	7.5–10.0
Trains (diesel)	5–10
Trains (electric)	5–10

^aData from United Nations Development Program (2000).

tive with cleaner imported coal are also responsible for the high carbon intensity of the economy. An integrated energy plan would consider the effects of such energy policies on the already degraded environment of the country. In sum, analysis easily identifies the key areas—decentralized grass-roots initiatives and production and usage of renewable energy—for policy interventions for sustainable development and for reducing adverse environmental impacts.

5.2 Potential for Renewable Energy

The share of hydrocarbons in India's commercial energy mix is slated to rise to 54% in 2020 from the current level of 38% (see Fig. 9). Given the paucity of hydrocarbons in India, it is important to develop a renewable energy base that would reduce import dependence. Table VII shows the estimated potential for renewable energy in the country. Comparison of the potential with achievement paints a bleak picture. Although India ranks first in the world in utilization levels of solar cookers and biomass gasifiers, and can claim decent utilization levels in biogas plants and cookstoves, in solar photovoltaic energy, and in wind power, the potential to exploit these resources still remains unfulfilled. One major deterrent to the

utilization of renewable energy is the high utilization cost of these technologies. However, these costs are slated to decline in the future, which will make the technologies economically competitive. Solar photovoltaic (PV) energy is an important alternative for power generation, and specific intervention is required to promote its use. The annual global market for PV energy is estimated to be around 200 MW and has grown at a rate of 15% over the past 15 years. In India, the annual PV production volumes are of the order of 10 MW, making India one of the largest markets and manufacturers of PV modules in the world after the United States, Japan, and the European Union.

Very few power-generating renewable energy sources are currently connected to the grid in India. As energy from renewables becomes more and more competitive, this aspect will automatically cease. Renewable energy can be used not only to generate power, but also for cooking, space heating, water desalination, etc. All these uses suggest that rural India represents a vast market for renewable energy; with provision of energy becoming more and more decentralized, such options will be explored more thoroughly. In fact, the Ministry of Nonconventional Energy Sources has targeted 5750 remote villages for electrification through the use of renewable energy by 2012.

TABLE VII

Potential and Achievement in Renewable Energy^a

Energy source	Potential	Achievement ^b
Biogas plants	120 lakh ^c	32.75 lakh
Improved chulhas ^d	1200 lakh	338 lakh
Wind	45 000 MW	1507 MW
Small hydro	15 000 MW	1423 MW
Biomass power/ cogeneration	19 500 MW	358 MW
Biomass gasification		42.8 MW
Solar photovoltaic	20 MW per km ²	82 MW per km ²
Waste-to-energy fuels	1700 MW	17.1 MW
Solar water heating	1400-lakh m ² collector area	6-lakh km ² collector area

^aData from Ministry of Nonconventional Energy Sources (2002).

^bAs of 31 December 2001.

^cLakh is a term denoting 100,000; it is used here to indicate efficiency.

^dA chulha is a fuel-burning stove; chulhas are made of many materials and have shapes and components that make them more or less efficient.

6. IMPERATIVES FOR THE NATIONAL ENERGY POLICY

Energy demand in India is projected to increase in the coming decades, as discussed earlier. Though efforts are being made to ensure reliable supply of energy at reasonable prices, entry of market forces due to the globalization and privatization process will play an important role in the process. Ensuring the reliable supply of energy, in required forms and at reasonable rates, has become one of the priorities of governments the world over, and this holds true especially for energy-deficient countries such as India. However, market forces are increasingly encroaching on this traditional domain of state influence. This essentially means that though demand and supply will always match, prices will reflect the shortage or surplus in the economy. Hence, energy shortages are likely to translate into higher prices. Because there is a strong energy–economy linkage, this will slow down the pace of development and depress the economy. Thus, the

National Energy Policy should aim first to weaken the energy–economy linkage, by improving the efficiency of energy use, and then to ensure that dependence on any one source is reduced effectively, thereby reducing the vulnerability of the economy to external shocks. The various policy issues that are critical to ensuring these aims are discussed in the following sections.

6.1 Role of Market Forces

Traditionally, energy has been one of the main domains of state influence in which the private sector had been excluded from effective participation. This is evident in developing countries, in which scarce resources are utilized in optimal ways in the interest of the entire nation. At the same time, the developed world has moved toward market forces. In the recent past, market forces have also begun to make their presence felt in increasing numbers of developing countries (India, Brazil, Mexico, and several East Asian countries). It is being gradually realized that the market is the best allocator of resources and that the government should at best be a facilitator and regulator that sends signals in the form of incentives to the market, to orient market objectives toward national aims. In India, entry of market forces in each sector of the economy was planned and allowed when reforms were launched in 1990/1991.

The entry of private players in the energy sector means that prices will become important tools for signaling shortage or surplus in the market. In the controlled era, quantitative restrictions played a major role in balancing demand and supply. This, however, created a lopsided consumption pattern and efficiency was put on the back burner. Industries, assured of fixed returns, paid no attention to the efficiency of the process, which increased the energy intensity of the economy. Wasteful spending by consumers who could afford it led to shortages in the market, which then forced the government to resort to quantitative restrictions in order to supply energy to those who could not afford it. Meanwhile, the import bill for the country continued rising, given the importance of oil in the energy mix.

With prices now becoming important tools, such tendencies are likely to be curbed. This is already showing in the improving energy intensity of the economy, which is becoming more and more competitive. For example, the prices of petroleum products have been decontrolled and are now set by oil companies fortnightly, based on international

price of crude. This gives the consumer strong signals about the real value of these products and, hence, is likely to have an impact on the consumption pattern and fuel economy of vehicles. Such a rise of market forces, however, raises the question of regulation of the market by the government. Markets are nowhere near perfect; to prevent market abuse, regulation has to play an important role. Thus, regulating the energy market is one of the major tasks for the NEP.

Pricing policies adopted by the government so far have resulted in the largest state-run power-generating company in the country, the National Thermal Power Corporation, which holds the decision to use gas for power plants (Anta, Auraiya, Kawas, and Kayamkulam) in abeyance due to uncertainty over final delivered gas price. In the absence of an integrated LNG policy, the tax status of LNG remains uncertain and, hence, the cost of delivered LNG differs from state to state. On the other hand, an increasing number of studies are proving that gas will be competitive with coal, at least in the western region of the country. This uncertainty, in turn, affects the investment in the infrastructure sector of the country, which is crucial for development.

6.2 Status of Reforms

Reforms were launched in India almost a decade ago. So far, the scorecard has been mixed. In some sectors, reforms have seen greater consumer choice, availability of latest technologies, and reduction in prices due to increased competition. However, reforms in the energy sector have not been adequately effective, because this sector requires an integrated approach involving all the sectors that are directly or indirectly linked with it.

6.2.1 *The Energy Conservation Act (2001)*

The Energy Conservation Act (2001) was enacted in September 2001 to deal with all matters related to the efficient use of energy and energy conservation. Under the act, the Bureau of Energy Efficiency was set up to discharge the following functions:

- To look into the energy consumption norms for each energy-intensive industry and encourage proper labeling of energy consumption indicators on every electrical appliance.
- To provide guidelines for energy conservation building codes.

- To take measures to create awareness and disseminate information for efficient use and conservation of energy.
- To strengthen consultancy services in the field of energy conservation, develop testing and certification procedures, and promote testing facilities for certification and energy consumption labeling of equipment and appliances.
- To provide financing for certain measures taken by consumers to enhance efficiency in energy consumption.
- To maintain a list of accredited energy auditors to carry out energy audits of industries and recommend measures for improving efficiency.

The responsibility of the central and the state governments under the Energy Conservation Act is to ensure effective and efficient implementation of the suggestions given by the Bureau of Energy Efficiency, for which a system of penalties and incentives has been devised. On the whole, the enactment signifies the importance that the government has accorded to efficiency improvements in all sectors.

6.2.2 Oil and Gas Sector Reforms

In 1997, the government had established a timetable for abolishing the Administered Pricing Mechanism (APM) by 2002 and for deregulating the entire oil and gas sector. Though the APM has been abolished, it is not certain that the oil and gas sector has been freed of government diktats. Though the pricing of sensitive fuels (motor fuels, high-speed diesel fuel, LPG, and superior kerosene oil) is supposed to be market determined, prices are still set after consulting the government.

In order to prevent exploitation of consumers in the deregulated scenario, the Petroleum Regulatory Board Bill was introduced to ensure that each marketing company displays the maximum retail prices for the notified petroleum products and take steps in accordance with regulations to prevent profiteering by the entities. However, the bill has not been passed and has been referred back to the Parliamentary Standing Committee. Therefore, the regulator is still not in place and gas prices are not yet completely linked to international parity. The ONGCL still negotiates with the refineries for an internationally competitive crude price.

LNG importers have been demanding granting of infrastructure status for LNG that would make it eligible for a central sales tax of only 4%. Because the sales tax rates vary widely from state to state,

LNG is likely to be costlier than coal, which jeopardizes investments being made by companies in building LNG terminals. To deal with this, the government has introduced the draft Integrated LNG Policy, which focuses on the various tax sops given to the project developers to make LNG viable in the country. It is expected that the delivered LNG will be 22% cheaper under the new policy and will bring it on par with 100% fuel oil parity price of the domestic gas. The government has also announced a new Petroleum Product Pipeline Policy on a common carrier principle. The new guidelines for grant of right of use (ROU) on land do not contemplate any restrictions or conditions for grant of ROU for crude oil pipelines. As per the guidelines, the investors will have complete freedom with respect to the pipelines originating from refineries or meant for captive use by companies for which ROU will be unconditional.

The delay in the sell off of the Hindustan Petroleum Corporation Ltd. and Bharat Petroleum Corporation Ltd. has not gone in the government's favor and has discouraged foreign investors. So far, NELP rounds have been only moderately successful in attracting small international oil companies, but have failed to generate interest among the very large-size oil companies. Whether this is due to low prospects of Indian sedimentary basins or due to faulty policy has to be examined. However, the recent discovery in the Krishna–Godavari basin should improve the prospect of Indian sedimentary basins in the international market. Subsequent to this discovery, the government decided to include more deepwater areas in the fourth round of NELP.

6.2.3 Power Sector Reforms

Mounting power shortages and critical financial condition of state electricity boards (SEBs) have rendered the boards unable to add significantly to the power generation capacity, prompting the national government, in 1991, to encourage private-sector participation, with an objective of mobilizing additional resources for the sector. The milestones in power sector reforms are briefly summarized in the Table VIII. However, power sector reforms are also plagued with multiple problems. Even though the government has provided incentives to the reforming state in the form of higher allocations for the sector, this has not yielded the expected result, i.e., restructuring of the SEBs. Trifurcation of the SEBs into generation, transmission, and distribution companies has been initiated in only a few states, with no visible benefits yet evident. In view of high electricity

TABLE VIII

Milestones in Power Sector Reforms^a

The Ministry of Power formulated a Common Minimum National Action Plan (MNAP) in 1996. The MNAP introduced wide-ranging reforms in the power sector, touching every aspect from generation to transmission and distribution and the state of finances of state electricity boards. The agenda for reforms included the following decisions:

1. Each state/Union Territory shall set up an independent State Electricity Regulatory Commission (SERC), which shall have the tariff fixation, licensing, and planning powers. The Indian Electricity Act of 1920 and The Electricity [supply] Act of 1948 were amended to enable these changes
2. A Central Electricity Regulatory Commission (CERC) was set up to regulate the bulk tariffs for central generating power plant and transmission utilities. Licensing and planning will come under CERC when the Union government gives notice.
3. The action plan also aimed to rationalize retail tariffs, which were to be set by SERC. Norms for minimum tariffs and cross-subsidization were also to serve to guide SERC in the tariff fixation process
4. Private-sector participation was allowed in the distribution of electricity, with initial coverage limited to few areas
5. The action plan also envisaged greater autonomy for state electricity boards, which were to be restructured and corporatized and run on a commercial basis

^aThe progress on all these fronts has been slow. SERC has not been established in all states, and tariffs for many sectors are still not reasonable. Major issues, such as allowing Independent Power Producers (IPPs) to sell power directly to bulk consumers, are still unresolved and so is the sorry state of finances of the state electricity boards. With the status of LNG being uncertain, many many gas-based power plants have either been postponed or are being run on naphtha. The National Thermal Power Corporation is still uncertain about the status of its gas-based power plants at Anta, Auraiya, Kawayas, Gandhar, and Kayamkulam. The transmission and distribution losses are still very high in a majority of states, which is inhibiting improvements of the financial positions of state electricity boards. The procedure for getting permission for setting up a generating station is still lengthy and opaque. All these factors have caused many foreign IPPs to either abandon or postpone their plans for setting up power stations in India.

generating requirements projected for the future, and to promote private-sector investments, the government recently passed the Electricity Bill (2003), which seeks to create a liberal framework of development for the power sector by distancing the government from regulation. It replaces three existing laws, the Indian Electricity Act of 1910, the Electricity (Supply) Act of 1948; and the Electricity Regulatory Commissions Act of 1998. The Electricity Bill seeks to promote competition in the electricity sector in India by decoupling generation, transmission, distribution, and supply of electricity.

The bill also envisages preparation of a national electricity policy (including tariff) for the development of the power system based on optimal utilization of natural resources. In consonance with this policy, the Central Electricity Authority will prepare the National Electricity Plan once every 5 years. Perhaps the most important feature of the bill is the provision of open access to the transmission and distribution infrastructure in the country. Under the new bill, the generator and the consumer can individually negotiate the power purchase and use the common-access T/D system to meet the goals. The commissions aim to reduce cross-subsidization in the system and reward efficiency in performance. Thus, the Electricity Bill (2003) maintains the trend in electricity reforms witnessed the world over by exposing generation and the supply side of the market to competition, but placing T/D sections under an incentive regulation.

6.2.4 Coal Sector Reforms

A Committee on Integrated Coal Policy was initiated by the Planning Commission in 1996 to address problems in the coal sector. The proposals of this committee included the following major recommendations:

- Adoption of coal conservation measures.
- Augmentation of domestic coal production by inviting private capital.
- Integration of the exploratory efforts of coal and lignite.
- Deregulation of coal prices.
- Creation of a regulatory body.
- Establishment of more pithead thermal power plants.
- Augmentation of the railways and port infrastructure facilities.
- Acceleration of project clearance procedures.

The recent move by the government toward reforming the coal sector is exemplified in the constitution of Expenditure Reforms Commission (ERC). However, to date, deregulation of prices and the distribution of coal of all grades are the only recommendations that have been implemented by the government.

6.3 Subsidies

Subsidies form an issue of major concern, not only for the energy sector, but also for the entire economy. Problems due to subsidies arise when long-term gains are sacrificed for short-term gains, which are guided

by political will and have major influence on policy-makers. It has been common experience that subsidies are not withdrawn until long after they have stopped serving the purpose for which they were instituted. Subsidies doled out by the government to one sector often have a lasting impact on some other sector.

The International Energy Agency (IEA) estimated in 1999 that energy savings from removal of subsidies amount to 7.2% of total primary energy supply, and that CO₂ emissions would be cut by 14%. The fiscal savings would amount to \$8.6 billion. Eliminating all subsidies may not be feasible for the Indian political system, but rationalization of the plethora of subsidy schemes is definitely of prime importance and the government has taken important steps toward removing price controls on oil and coal and lowering of subsidies generally. Coal prices were decontrolled in 2000, and there are no longer any direct subsidies to coal production or consumption. Delivered coal prices, nonetheless, remain below market levels due to continuing subsidies on rail transportation. In April 2002, the government completed the dismantling of the APM for oil products and natural gas, and the removal of all subsidies, except for those for kerosene and LPG used by households. The Indian electricity sector is heavily subsidized. In 2000/2001, the average rate of subsidy expressed as a proportion of the estimated full cost of electricity supply was 93% for farmers and 58% for households. Industrial and commercial customers and the railways pay above-cost prices.

6.3.1 Effects of Subsidies on Sector Development

One prominent example of cross-subsidization impact, i.e., hurting one sector of the economy at the expense of another, is the imposition of a high electricity tariff on the Indian Railways. The Working Group on Energy Policy in 1980 had recommended electrification of 1000 km of route of railway track every year because of low costs and high efficiency of electrified engines. It is estimated that the cost of energy per 1000 gross tonne km for a diesel engine is 78 rupees, compared to 55 rupees for an electric engine. The cost of operation and maintenance per 1000 gross tonne km for a diesel-based engine is 26.5 rupees, whereas that for an electrified engine is 17.5 rupees. However, because the railways subsidize the residential and agricultural sectors and pay a high electricity tariff, only 550 km of track has been converted per year since 1980. This type of cross-subsidy, apart from exerting a toll on railway finances, undermines system efficiency and

clogs up the railway network. More important, it makes the railways dependent on crude oil for operation, a situation not desirable from either the economic or the security point of view.

6.3.2 Timely Review of Subsidies Essential for Effectiveness

LPG subsidies were instituted in order to encourage consumers to use LPG as cooking fuel. The penetration of LPG in urban centers is now near saturation, yet subsidies are still in place. Efforts by the government to eliminate them have not been successful, even though the subsidy monies could be better used elsewhere. This serves as another example to highlight the impact that a policy framed in one sector may have on other sectors. The National Energy Policy should recognize linkages between sectors and account for them in the policy framework.

7. IMPLICATIONS FOR NATIONAL ENERGY POLICY

Energy consumption in India is bound to grow manifoldly as India aims for sustained economic growth; growth will primarily be driven by increasing commercial energy consumption, given the positive and greater-than-unity energy elasticity of the economy. Usage of commercial energy sources, particularly oil and gas and power, will remain crucial for development. Therefore, policy initiatives to ensure flow of investments in these sectors become imperative. Domestic energy supply is likely to fall short of indigenous demand and, hence, reliance on imports will be unavoidable. To meet the growing demand for energy in a sustainable manner, India needs to look at all possible ways of bridging the energy gap. To achieve this, India has to take a long-term view of economic imperatives and move away from crisis management to providing strategic direction for energy development.

Before embarking on an integrated energy policy, policymakers should study the effect of existing energy policies on the already degraded environment and initiate effective research and development for alternatives to the conventional form of energy use. A long-term integrated NEP should keep in view both macro- and microlevel implications. Various recent technological innovations are giving rise to a competitive fuel market. For example, in the new-age power plants, gas is more efficient than coal. There-

fore, with improved LNG supplies, prospects of using gas for power generation are also improving, which will lead to gas providing tough competition to coal. Hence, for balanced energy consumption, linkages between different ministries need to be strengthened; in view of the fuel choice facing the power sector today, pricing policies adopted by the government will play an important role in influencing the decisions made by power generators.

Environmental concerns can no longer be ignored and policies should take into account the likely impact on the environment, and every attempt should be made to reduce negative impacts. Reforms in the energy sector are also important, both to attract private sector investment and to improve the energy efficiency of the economy.

To sum up, achieving energy efficiency and sustainable development requires the following essential approaches:

- A cabinet committee under the chairmanship of the Prime Minister (for effective coordination and monitoring between ministries).
- Immediate and effective reforms in various energy sectors and subsectors, representing both the supply side and the demand side, with an integrated approach.
- Policy coordination across energy ministries and other ministries for an integrated energy policy.
- Effective linkages with foreign policy.
- Revival of the National Environmental Council and its promotion as the forum for discussion and debate on energy–environment policies for the country and the changes that need to be brought about from time to time.

Overall, it can be said that energy has to be seen as part of overall development strategy. Interventions in the energy sector provide only short-term benefits unless they are combined with matching interventions in other sectors that lead to the spread of economic opportunities among the poor.

Acknowledgments

The authors appreciate the assistance of Ms. Shobha Mishra.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • European Union Energy Policy • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: Japan • National Energy Policy: United States • Rural Energy in India

Further Reading

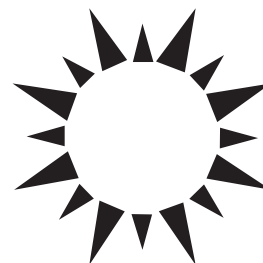
- Energy Information Administration (EIA). (2003). "International Energy Outlook 2003." EIA, Washington, D.C.
- International Energy Agency (IEA). (2002). "World Energy Outlook 2002." IEA, Paris.
- Ministry of Law, Justice, and Company Affairs (Legislative Department). (2001). "The Energy Conservation Act, 2001." Ministry of Law, Justice and Company Affairs, Government of India, New Delhi.
- Ministry of Nonconventional Energy Sources. (2002). "Annual Report 2001–2002." Ministry of Nonconventional Energy Sources, Government of India, New Delhi.
- Ministry of Petroleum & Natural Gas. (2002). "Indian Petroleum and Natural Gas Statistics, 2000–2001." Ministry of Petroleum & Natural Gas, Government of India, New Delhi.
- Ministry of Petroleum & Natural Gas. (2002). "The Petroleum Regulatory Board Bill, 2002." Ministry of Petroleum & Natural Gas, Government of India, New Delhi.
- Ministry of Power. (2003). "The Electricity Act, 2003." Ministry of Power, Government of India, New Delhi.
- Planning Commission. (1999). "Hydrocarbon Vision 2025." Planning Commission, Government of India, New Delhi.
- Sengupta, R. (1993). "Energy Modelling for India; Towards a Policy for Commercial Energy." Planning Commission, Government of India, New Delhi.
- The Energy and Resources Institute (TERI). (1998). "Green India 2047; Looking Back to Think Ahead." TERI, New Delhi.
- The Energy and Resources Institute (TERI). (2001). "Green India 2047; Direction, Innovations, and Strategies for Harnessing Actions for Sustainable Development." TERI, New Delhi.
- The Energy and Resources Institute (TERI). (2002). "Defining an Integrated Energy Strategy for India." TERI, New Delhi.
- The Energy and Resources Institute (TERI). (2003). "TERI Energy Data Directory and Yearbook (TEDDY), 2002/03." TERI, New Delhi.
- United Nations Development Program (UNDP). (2000). "World Energy Assessment." UNDP, New York.
- World Bank (2000). "World Development Indicators, 2000." The World Bank, Washington, D.C.
- World Bank Group. (2001). "Indoor Air Pollution. Energy and Health for the Poor 4." The World Bank, Washington, D.C.

National Energy Policy: Japan

PAUL J. SCALISE

Consultant

Tokyo, Japan



-
1. Energy Resources
 2. Energy Deregulation
 3. Social and Environmental Impacts
 4. Conclusion

Glossary

deregulation The time period over which countries/states open up to full retail competition of a preferred energy supplier, down to the household level.

distribution network The low-voltage electricity network that runs from electricity substations to the end user. Distribution is effectively a monopoly activity.

economies of scale A cost function exhibiting “natural monopoly” properties; economies of scale are present if the marginal costs of production of a single-product firm are less than the average costs of production over the relevant range of output. Put differently, economies of scale are said to exist over the relevant range of output should unit costs decline with the volume of production (on a kilowatt-hour basis.)

independent power producer (IPP) A producer of electricity whose plant is not affiliated with a local utility company; independent plants operate in a competitive, unregulated environment.

kilowatt-hour (kWh) A standard unit of electric consumption corresponding to usage of 1000 W for 1 hour. A 100-W light bulb burning for 10 hours consumes 1 kWh.

tariff A public schedule detailing utility rates, rules, service territory, and terms of service; tariffs are filed for official approval with a regulatory agency.

total primary energy supply (TPES) This is made up of production + net exports – international marine bunkers $\pm$ stock changes.

The objectives of Japan’s national energy policy today are defined by the Ministry of Economy, Trade, and Industry (METI) as the “four Es”: energy security, environmental protection, economic efficiency, and

economic growth. Although each goal is separately crucial to the long-term national good, implementation and compatibility of these four objectives have been in conflict for much of the time since World War II. The second largest economy in the world, Japan in the year 2000 represented 0.28% of the world’s landmass, 2.1% of the world’s population, and 15.3% of the world’s gross domestic product. Such statistics form the backdrop to a densely populated industrial society that heavily depends on stable supplies of energy to survive. Yet, the country suffers from a lack of indigenous natural resources, especially fossil fuels such as oil, coal, and natural gas—resources that are crucial in the maintenance of infrastructure and rising aggregate consumption. As such, Japan imported nearly 84% of the total amount of its primary energy supply in fiscal year 2000. This rate is the highest of any major industrial country (save Italy.) The problems and prospects of Japan’s changing economy and how this relates to its energy policy are discussed in this article.

1. ENERGY RESOURCES

Historically, geographical and commodity vulnerabilities have led Japan to a “dependence-based” energy strategy in which Middle Eastern oil-producing countries have been afforded special considerations. This strategy became critical during the oil shock in 1973, when the price of Brent crude oil rose substantially, thereby forcing Japan to take special measures not always compatible with the broader goals of its U.S. ally. Thus, reevaluation of Japan’s domestic energy policy resulted in further energy diversification and, in particular, a major nuclear construction program.

With the end of the Cold War and advent of the economically stagnant 1990s, Japan’s energy policy has moved away from one of pure “dependency” and

“economic aid in exchange for oil” within highly regulated energy markets to a policy of gradual deregulation, market efficiency, and alignment with the United States to obtain regional security in the maintenance of its oil supply flow. Judging by the country’s position in energy consumption among the Group of Seven (G7) countries (Table I), Japan ranks second in terms of total primary energy consumption as well as in oil and coal, third in nuclear and hydroelectric energy, and fifth in natural gas. Moreover, the country relies heavily on overseas sources for its total amount of primary energy supplied (84%) in fiscal year 2000. This percentage is the highest of any major industrial country. Consequently, Japan accounts for a very large proportion of the world’s energy imports, ranking second after the United States in total imports of primary energy, first in coal and natural gas, and second in crude oil and oil products. This situation suggests that Japan is both a driver and a vulnerable target of material fluctuations in the price of fossil fuels on the world market.

1.1 Domestic Energy Supply and Demand: A History

The process of energy development in Japan’s postwar economy can be divided into five stages: postwar reconstruction (1945–1955), the high-growth period (1955–1973) lasting until the 1973 oil crisis, the stable growth period (1973–1985) and its subsequent adjustments, followed by the incipient “bubble” growth period (1985–1991) that ultimately

led to the country’s current stage, the so-called restructuring period (1991–present). These five periods of development are different in their respective patterns and correspond to changes in Japan’s energy supply and demand.

1.1.1 Postwar Reconstruction (1945–1955)

The postwar reconstruction period was characterized by its industrial reorganization, development, and modernization in the aftermath of Japan’s defeat in World War II. In an attempt to meet the challenge of securing a stable supply of energy and saving foreign exchange, the Japanese government adopted “a priority production system” in 1946 that targeted the domestic extraction of coal and increased development of hydroelectric power. As shown in Fig. 1, coal continued to play a central role in Japan’s primary energy supply until it was overtaken by oil in 1963. Concurrently, the shift toward greater use of electricity as an alternative to coal coincided with the establishment of nine regional, but privately owned, electric power companies in 1951.

1.1.2 Rapid Growth (1955–1973)

Throughout a stage of rapid growth from the mid-1950s to the early 1970s, Japan’s real gross domestic product (GDP) growth averaged 9.2% per annum. One important factor in the development of this performance was the increasing supply of imported fossil fuels, such as inexpensive crude oil. In conjunction with this “oil revolution,” the country structurally shifted its emphasis away from agricultural services to heavy and chemical industries; the

TABLE I
Energy Consumption, 2000: Japan’s Rank among G7 Countries^a

Rank	TPES (%) ^b	Oil (%)	Nuclear (%)	Hydro (%)	Coal (%)	Natural gas (%)
1st	United States (56.4)	United States (56)	United States (42.9)	Canada (43.1)	United States (66.9)	United States (58.2)
2nd	Japan (12.9)	Japan (13.9)	France (22.3)	United States (29.9)	Japan (11.6)	United Kingdom (9.3)
3rd	Germany (8.3)	Germany (6.9)	Japan (17.3)	Japan (10.5)	Germany (10.0)	Canada (8.0)
4th	France (6.3)	Canada (6.1)	Germany (9.1)	France (8.0)	United Kingdom (4.4)	Germany (7.7)
5th	Canada (6.2)	Italy (5.9)	United Kingdom (4.6)	Italy (5.3)	Canada (3.8)	Japan (6.9)
6th	United Kingdom (5.7)	United Kingdom (5.8)	Canada (3.9)	Germany (2.6)	France (1.8)	Italy (6.2)
7th	Italy (4.2)	France (5.4)	Italy (0)	United Kingdom (0.6)	Italy (1.6)	France (3.8)

^aData from International Energy Agency (2002); calculations by the author.

^bTPES, Total primary energy supply.

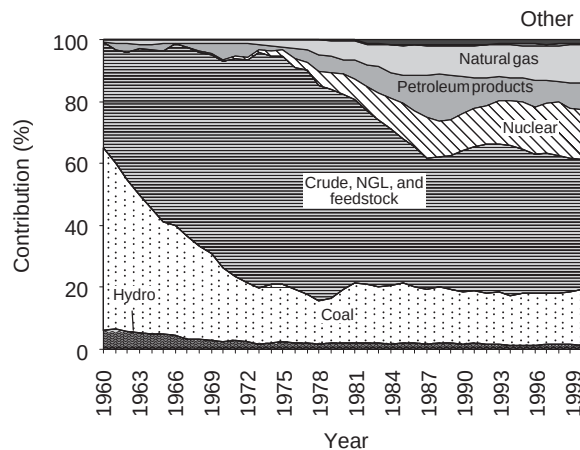


FIGURE 1 Long-term trends in primary energy supply by fuel. NGL, Natural gas liquids. Data from International Energy Agency (2002); calculations by the author.

government's primary objective was to support an "export-led" strategy of economic development in which oil consumption for domestic use increased materially. The share of oil as a percentage of the total primary energy supply increased from 17.6% in 1955 to 55.2% in 1965, while coal declined from 47.2% to 36.5% during the same period (Fig. 1). This trend continued as oil's share peaked at 76.4% in 1973, the year of the first world oil crisis. Oil replaced coal as the country's leading source of energy in 1963, and in that year thermal power replaced hydroelectric power as a percentage of total electric energy supplied.

1.1.3 The First Oil Shock and Stable Growth (1973–1986)

By 1973, Japan's energy self-sufficiency rate reached its nadir (10.6%). The country's dependence on imported oil (99.8%) suggested that any sudden fluctuation in world prices would cause serious discontinuities within the Japanese economy. The outbreak of the first oil crisis in 1973 dispelled the notion that cheap, readily available supplies of imported oil would always prevail. Crude oil prices increased fourfold, and Japan was compelled to rethink its national energy strategy. The second oil crisis in 1978/1979 reinforced this need. Industries began to adopt energy efficiency measures, and Japan's industrial framework shifted again from energy-intensive industries (e.g., basic materials) to less energy-demanding industries (e.g., assembly and processing). Consequently, energy consumption only grew 1.5% per annum from fiscal 1973 to

1979, and decelerated to 0.5% per annum from fiscal 1979 to 1985.

With the onset of the two oil crises, the energy policy of the government prioritized national energy security, and this was implemented in a few ways. First, the government sought to minimize its risk by increasing oil stockpiles and diversifying its imported oil sources. The second policy aim was to reduce dependence on oil by adopting a "diversified fuel-mix strategy." Alternative energy sources such as coal, liquefied natural gas (LNG), and nuclear energy were emphasized (Fig. 1).

1.1.4 Japan's "Bubble" Growth Era (1985–1991)

The Japanese economy witnessed a material turnaround in both performance and consumer activity during the so-called bubble years; this was a period characterized by rapidly rising real estate and stock market prices in highly speculative markets. In 1986, the material decline in crude oil prices, brought about by a reduction in the crude target prices of the Organization of Petroleum Exporting Countries (OPEC) for the first time since the 1973 oil crisis, led to a surplus on the world market. In turn, Japan's economy became a major beneficiary of the development as the compound annual growth of the primary energy supply from 1986 to 1991 increased by 3.9%, or 4.5% in industry, 4.3% in the household/service sector, and 5.3% in transportation. Japan, once again, was seeing an increase in energy consumption.

Concurrently, export-led companies attempted to compensate for the economic shocks imposed on the Japanese market of the previous decade, by rationalization efforts, cost-cutting, and increased exports. As the trade surplus with the United States increased, international pressure to shift Japan away from an "export-led growth strategy" to a "consumer-led growth strategy" intensified. In April 1986, the Advisory Group on Economic Structural Adjustment for International Harmony, a panel appointed by the prime minister and headed by former Bank of Japan Governor Maekawa Haruo, issued a set of recommendations otherwise known as the *Maekawa Report*. [The Japanese practice of putting family names of Japanese people first is followed in this article (for example, Ono Yoko, rather than Yoko Ono). Western names retain the standard form of first and last names (for example, Dwight Eisenhower).] This report set the stage for a series of proposals to restructure the Japanese economy with long-term ramifications for many domestic sectors, including energy. The concrete steps proposed

centered first on the expansion of domestic demand, focusing on areas such as industry restructuring, technology development, and changes to the quality of housing.

1.1.5 Structural Adjustment (1991–Present)

Since 1991, after the bursting of its bubble economy, Japan has faced unprecedented and difficult questions. Electricity demand growth, crippled by stagnant GDP elasticities, has fallen to barely flat levels, and incumbent energy industries have been forced to accept new regulatory enhancements to ensure greater market efficiencies under the banners of internationalization, globalization, and liberalization.

Throughout the decade, the government fundamentally shifted its policy from energy security to an overall reconsideration of its national agenda. Economic efficiency as an economic policy driver became another pillar to the METI's growing list of goals. Although deregulation incrementally lowered prices in energy industries, including oil, electricity, and gas, it also undermined profitability, forcing industrial consolidation and corporate efficiency measures.

1.2 Energy Consumption by Sector

Changes in demand can also offer valuable insights into the nature of energy growth. Since the 1973 oil shock, industrialized nations have made concerted efforts to conserve energy (Fig. 2). On an index with 1973 set at 100, Japan led the industrialized nations in energy conservation; between 1973 and 1987, the total consumption level as a percentage of GDP fell

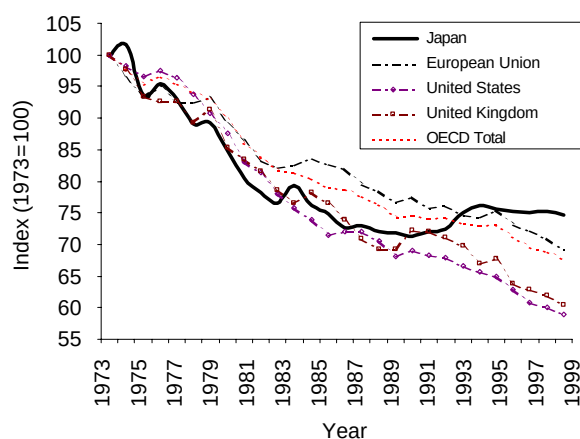


FIGURE 2 Index of total primary energy supply per unit of gross domestic product. OECD, Organization for Economic Cooperation and Development. Data from International Energy Agency (2002); calculations by the author.

(energy efficiency improved) at a compound annual rate of 2.3%, which was higher than that of any other Organization for Economic Cooperation and Development (OECD) country. As mentioned, such efficiency gains were due, in part, to a shift in orientation from heavy, energy-intensive industries, such as iron and steel, cement, nonferrous metals, and petrochemicals, to machine-based and high-technology industries, such as automobiles, appliances, consumer electronics, and industrial ceramics, and a further shift from secondary to tertiary industries with increasing emphasis on service trades. This shift can be expected to continue, but at a slower pace than in the past. Japan is under pressure to increase its domestic consumption to a level more in line with its trade competitors. This international pressure, in conjunction with growth in road transport, can be expected to increase energy intensity.

It is important to note, however, that the industrial sector has been the only contributor to major energy efficiency improvements. As Fig. 2 indicates, by the 1990s, the total primary energy supply per unit of GDP materially shifted; energy consumption levels within the household/service, commercial, freight transportation, and passenger transportation sectors markedly increased, thus outweighing previous efficiency gains in the industrial sector. An international comparison of per-household energy consumption reveals that, adjusting for heating figures related to atmospheric temperature differences among countries, Japanese data values are still quite low, suggesting that the current growth in energy consumption is the direct result of income elasticity among affluent economies. In 1973, for example, there was only one car for every two families in Japan. In 2000, the number of cars on the road reached over 52 million versus only 45 million households. Moreover, in 1970, fewer than 10 out of every 100 Japanese homes had such things as videocassette recorders, microwave ovens, and air conditioners, and only 43 out of every 100 homes had color televisions. By 2000, the number per 100 households had risen to 84 for videocassette recorders, 95 for microwave ovens, and 217 for air conditioners, and color televisions topped the list at 231.

2. ENERGY DEREGULATION

Deregulation of the Japanese economy began in the mid-1980s with the introduction of the so-called *Maekawa Report*. In the 1990s, deregulation ex-

tended into the energy industry; first materially affecting the petroleum industry and later extending into the electricity and gas industries. For many years, Japan's energy industry was a textbook example of a regulated industry, but the picture has changed. A string of events has taken place, altering the operational landscape, including partial liberalization of the electricity and gas retail markets and the lapse of the Refined Petroleum Import Law.

2.1 Electricity

The Japanese electricity market is not only one of the largest in the world, but also one of the most lucrative. In 2001, total market revenues (*ex self-generation*) were worth more than ¥15 trillion (= \$115.3 billion U.S. dollars; \$1 = ¥130). Tokyo Electric Power Corporation (TEPCO), the world's largest privately owned utility, boasted total assets of more than ¥141.3 trillion (\$1087 billion), volumes of 275.5 TWh, and revenues of ¥5.2 trillion (\$40.1 billion). The sheer mass of TEPCO translated into one of the highest net profit streams in Japan despite high interest payments on debt. The key feature to this arrangement has been the existence of vertically integrated regional monopolies that have been

heavily regulated by the state. After a brief period of government wartime control, nine private power companies were reorganized in 1951 to come into control of most of the generation, transmission, and distribution businesses of electric power throughout the country. Their reclaimed monopoly status was the result of successful lobbying by utilities to return to a system of centralized private power, reminiscent of the early 1930s, not of the highly competitive era of the 1920s.

The greatest strength of the new "1951 system" was that it exploited economies of scale, maintained stable electricity prices for over 20 years, and kept dangerous dioxide emissions in check. Its greatest weakness was that it eventually disproved the theory; the 1951 regulatory structure failed to reconcile rising demand with mounting variable (fuel) and fixed (capital) costs. Over time, the weaknesses became more pronounced. As Fig. 3 indicates, from 1975 to 1985, average monopoly tariffs increased by almost 100% from 11 yen/kWh to 21 yen/kWh. Utility companies argued that the phenomenon was the result of exogenous variable shocks stemming from an overdependence on imported fossil fuels. Although true to an extent (99.8% of Japan's consumed oil is imported, 87% of which is imported

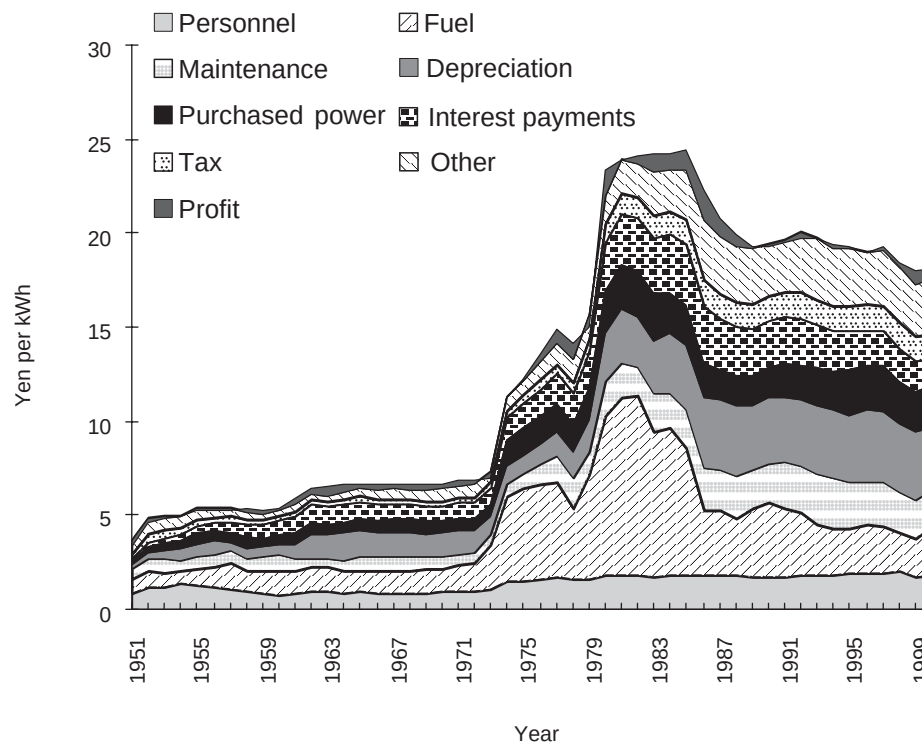


FIGURE 3 Average electricity tariff, fiscal years 1951–1999. Based on data from the 2002 annual report of the Tokyo Electric Power Company; calculations by the author.

TABLE II

Compound Annual Growth in Electricity by Industry Type^a

Industry	Compound growth, fiscal years 1965–2000 (%)		Compound growth, fiscal years 1990–2000 (%)	
	Total volumes	Self-generation	Total volumes	Self-generation
Mining	–3.2	–1.5	–1.5	–5.3
Food	8.2	9.5	3.0	7.5
Textiles	0.3	11.9	–5.6	2.4
Paper	1.7	7.2	–1.2	2.0
Chemicals	0.5	5.9	–0.5	3.1
Oil	3.4	9.9	–4.8	8.1
Rubber	5.4	12.6	0.1	17.2
Ceramics	5.0	7.4	–2.3	4.5
Iron	2.4	5.5	–1.2	2.0
Nonferrous	3.2	1.3	1.4	0.6
Machinery	8.6	30.5	2.0	16.5
Other	7.1	8.3	1.7	9.3
Railroads	3.7	2.8	1.0	0.4
Total (manufacturing)	3.9	6.4	0.4	3.5
Grand total ^b	4.0	6.1	0.7	3.4

^a Source. Energy Data and Modeling Center, *Handbook of Energy & Economics Statistics in Japan*, 2002; calculations by author.^b Grand total equals total manufacturing demand plus railroad demand and other.

from the Middle East), average tariff rates and the unit cost of fuel per kilowatt-hour continued to decouple after 1985. The lack of continuity became more pronounced as market rigidities and rising capital costs prevented a nominal readjustment to pre-1973 tariff levels. Nevertheless, the original presence of “natural monopoly” technologies continued to provide justification for a regulated power industry and accordingly to impose entry barriers, price controls, and obligations to serve. Unlike other industrialized nations, the postwar regulatory regime did not breakdown with the “energy crisis” of the 1970s. Rather, the then-Ministry of International Trade and Industry (MITI) overcame the crises through close cooperation with incumbent utilities, preserving the vertically integrated monopoly structure. The exploitation of economies of scale while controlling market power was not only considered the progressive “public interest” view of the country’s utility regulation until the early-1990s, it was also a means of preserving the political status quo.

Industry restructuring began in earnest with the advent of the Hosokawa Morihiro coalition government in 1993. By that time, electricity prices from 1983 to 1993 were not only perceived to be high domestically, but also internationally. Pretax Japanese electricity tariffs were three times the average

U.S. price in 1993. Broadly speaking, pressure to enact reforms was initiated by several actors, including the incumbent utilities concerned with the threat of self-generators (*jika hatsudensha*) capturing further market share (Table II) the export-led industries burdened with an appreciating currency and mounting operational costs; the potential new entrants seeking to duplicate their success in foreign deregulated markets; and the central government, which was looking not only to maintain its legal powers over the market, but also to placate foreign government pressures (*gaiatsu*).

In response to such growing pressure to lower electricity tariffs, the government decided to undergo regulatory reform. Revisions to the Electric Utility Industry Law were intended to produce a more competitive market, albeit incrementally. Phase one (1995) permitted the nine major power companies (*ex* Okinawa) to act as “single buyers” through a revision of the 31-year-old Electric Utility Industry Law. A bidding structure organized independently of government control was established whereby independent power producers (IPPs) and other self-generators could bid for new contracts for their supplementary power needs. Major players in phase one were predominantly steel makers such as Nippon Steel Ltd. and Kobe Steel Ltd. Phase two (2000)

expanded the scope of liberalization measures to include the following steps:

1. Retail choice for large-lot customers: Competition was opened to utilities servicing customers with contracts for 2 MW or greater and connected to a supply system of 2000 V or higher. Such customers included owners of large-scale office buildings, industrial factories, hotels, hospitals, and department stores. Unlike phase one, customers could now choose their preferred wholesale supplier, whether foreign or domestic.

2. Competitive bidding process for thermal capacity acquisition: The nine incumbent power companies (*ex* Okinawa) were required to place competitive bids when they purchased thermal capacity. Hydroelectric and nuclear power competition were (and are) beyond the scope of the regulatory changes.

3. Profit sharing: Incumbent utilities were no longer required to submit rate case applications for tariff reductions. This served as an incentive to reduce costs voluntarily because the incumbent utility could reserve part of the increased profit margin through reduced service costs.

4. Abolition of Article 21 of the Antimonopoly Law: Without this provision, discriminatory rights to railroad and utility companies (natural monopolies) were no longer upheld. Both the Japan Fair Trade Commission (JFTC) and the Ministry of Economy, Trade, and Industry enacted guidelines that legally prohibited any exclusionary activities inhibiting new entrants.

As of the date of writing of this article, a quantifiable results-oriented assessment of Japan's electricity deregulation suggests only tepid competition, average incumbent tariff declines, and minimal corporate efficiencies. New electricity market entrants account for only 0.85% of total installed capacity nationwide. Structural barriers to market entry, such as extensive backup, wheeling, and maintenance charges, preclude cost-competitive access to many incumbent networks (transmission and distribution). Furthermore, lack of available land for further capacity build and stringent environmental regulations in generation also present strong obstacles to market entry.

Incumbent electric power companies have begun to lower tariff rates in anticipation of future market competition, but, as Fig. 3 indicates, readjustments have been preemptive and voluntary, not forced through direct competition. On an index with 1990 set at 100, average electricity tariffs for Japan's three largest electric power companies (TEPCO, Kansai

Electric Power Co., and Chubu Electric Power Co.) in 2002 fell to 93, suggesting only tepid market pressures to date.

2.2 Gas

In contrast to the electric power industry, the Japanese gas distribution industry is best characterized as vertically truncated, but within a framework of several regional monopolies. Non-Japanese firms have traditionally conducted the "upstream" activities of exploration and supply. The "downstream" activities of refinement and distribution have fallen within the purview of companies such as Tokyo Gas, Osaka Gas, Toho Gas, and Saibu Gas, to name only the four largest companies in terms of volumes, revenues, and assets. Historically, the Japanese gas industry was privately owned but publicly regulated in the postwar era under the Gas Utilities Industry Law. Policymaking and regulatory functions were the responsibility of the MITI (now METI). This situation, however, is beginning to show signs of change, due to the strategic importance of liquefied natural gas (LNG) to Japan's long-term energy needs.

The majority of LNG is used within the electric power sector. In 2000, power plants accounted for nearly 70% of total gas consumption. With the gradual liberalization of the electric power industry in the 1990s, liberalization and restructuring are currently underway in the gas industry as well. The first measures aimed at liberalizing Japan's gas market were adopted in June 1994 and went into effect in March 1995. Revisions were made to the Gas Utilities Industry Law that allowed gas utilities to sell gas on a retail basis to large-lot volume users, thereby circumventing incumbent gas distribution companies. These contestable customers were defined as those consuming more than 2 million cubic meters (MMcm) per year, representing approximately 720 gas users. The changes permitted non-gas companies (power, steel, oil, and other companies with access to gas) to supply large-lot industrial customers for the first time. The revisions also lifted tariff regulations so that parties in the liberalized segment of the market were free to determine price and other contract terms on a negotiated case-by-case basis.

Phase two of the Japan's gas liberalization was adopted in May 1999 and went into effect the following November. These revisions expanded the contestable market to eligible consumers of 1 MMcm per year, or approximately 993 large-lot users. Designated general gas companies were also ordered

to grant third-party access to gas pipelines on a nondiscriminatory basis. The designated companies are Tokyo Gas, Osaka Gas, Toho Gas, and Saibu Gas. The smaller gas distribution companies, such as Shizuoka Gas, Hiroshima Gas, Chubu Gas, and Hokkaido Gas, to name a few, have no obligation to provide access to their infrastructure.

In 2000, a subcommittee established by the METI recommended a pricing formula for third-party access to both the incumbent electric power transmission system and the incumbent gas distribution pipeline system. Known as the “wheeling charge,” this forward-looking cost method allowed for historical comprehensive costs, including depreciation, taxes, and other operational costs, to be added to a stipulated rate in return for third-party access to the transmission network or pipeline. Based on these rules, the four gas companies issued rate schedules for transportation, backup, and storage in March 2000. Further transparency was also imposed regarding the accounting methods and public disclosure requirements to be used by the gas companies in allocating costs among various business segments.

At the time of writing this article, new entrants to the newly liberalized gas industry have been predominantly incumbent electric power companies. TEPCO, for example, as of January 2001, has supplied local city gas companies via its own low-pressure pipeline in Chiba Prefecture; since 2001, Chubu Electric Power Co. has had a joint venture with Iwatani International Corporation and Cosmo Oil to supply LNG via a lorry operation in Mie Prefecture, and Tohoku Electric Power operates a mass natural gas supply business through its subsidiary Tohoku Natural Gas using the Niigata–Sendai pipeline. METI is currently monitoring the outcome of these liberalization measures and subsequent moves by new entrants. The agency is also examining what measures could be taken next and recently impaneled a new committee to develop recommendations for further liberalization.

2.3 Petroleum

Unlike the vertically integrated electric power industry, but similar to the gas industry, Japan’s privately owned petroleum industry has a long history of rigid regulations and failed government direction, resulting in a vertically truncated and horizontally fragmented industry. Incremental deregulation of the Japanese petroleum industry eventually illustrated that regulatory change could lead to greater consumer benefits and corporate consolida-

tion, but not necessarily increased industrial structures or reduced fiscal burdens of the government. The reason for such a regulatory framework stems from the fact that crude oil, natural gas liquids (NGLs), and feedstock consumption meet 42% of Japan’s annual energy needs. Japan accounts for 7% of total world oil consumption. Although the country’s dependence on oil has decreased from 77% in 1973 thanks to a concerted effort to diversify the national energy supply, oil remains a key energy source (Fig. 1). At 99.8%, Japan’s import dependence on oil is extreme and oil is therefore considered a strategic national resource, especially because 80% is sourced from the politically unstable Middle East.

Historically, the petroleum industry was bifurcated into the “upstream” activities, related to the exploration for and production of oil, and the “downstream” activities, related to oil refinement and distribution (wholesale and retail). These two activities coincided in the prewar era, but slowly gave way to strictly “downstream” activities owing to insufficient domestic supply, horizontal market fragmentation, and high fixed costs. Japanese oil firms became heavily dependent on the “majors” (large international oil companies) for upstream operations. It is within this framework that the 1962 Petroleum Industry Law was enacted. Essentially, the law was written to achieve a stable supply of oil by controlling downstream oil refining, effectively authorizing the separation of upstream and downstream activities through permits to establish refineries or to purchase equipment for them. Several laws were subsequently enacted to reinforce the 1962 framework; these include the Petroleum Supply and Demand Balance Law, the Emergency Temporary Law for Stabilizing the Nation’s Livelihood, the Petroleum Stockpiling Law, and the Quality Assurance Law. As a by-product, these laws indirectly managed to protect the fragmented structure of the industry, thereby propping up many of the smaller firms that were dependent on existing regulations to stay in business.

For a brief period during the 1980s, foreign pressure (*gaiatsu*) from Western nations in conjunction with foreign companies caused a media stir in Japan. It was suggested that the majors and independent oil companies could circumvent these laws in order to achieve greater market share in Japan’s abundant downstream activities. Although unsuccessful, what followed was a gradual easing of regulations and laws in order to bend to public pressures to reduce high gasoline, kerosene, and naphtha prices in a domestic market that was

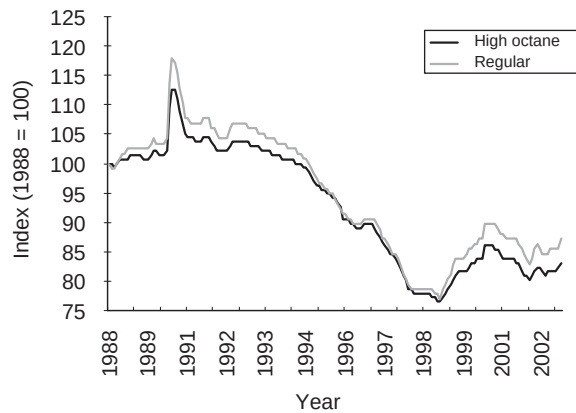


FIGURE 4 Gasoline per liter, fiscal years 1988–2002 (indexed). Based on data from the Oil Information Center; calculations by the author.

experiencing a clear slowdown in economic growth compared to the previous two decades. Technically, such liberalization of the petroleum industry began with the 1986 Refined Petroleum Import Law. The MITI eased its requirement for obtaining a license to import such refined products, answering calls from domestic and international actors to open markets. However, incumbent oil refiners successfully lobbied to limit the impact of the new legislation by requiring only those refineries in Japan with existing domestic production, stockpiling, and quality control to import oil products. The requirement effectively blocked all imports from nonincumbent players, making “petroleum liberalization” a paper tiger.

On March 31, 1996, the Refined Petroleum Import Law expired and the Gasoline Sales Law and Stockpiling Law were also revised. Emphasis shifted away from “energy stability” of imported petroleum to “energy efficiency,” partly aided by advocates of reform in the newly established Hosokawa Morihiro coalition government. As shown in Fig. 4, extreme inefficiencies and ineffective legislation kept gasoline prices high in the 1980s, but real liberalization of the market led to greater market competition and increased consumer welfare gains despite yen depreciation and stable crude oil prices. Indexing average gasoline prices per liter at 100 in 1988, the price of regular and high-octane gasoline, respectively, fell steadily from their Gulf War high of 112 and 118 in 1990, to a historic low of 76.5 in 1999.

The key to such consumer welfare gains emanated from laws enacted to stir competition in the mid-1990s. With the Refined Petroleum Import Law removed, stringent restrictions on petroleum imports were also removed; import licensing was readjusted to require new capacity to decrease from 120 days’

worth of product storage (a clear structural barrier to market entry) to only 60 days’ worth of product. In turn, the abolition of the Specified District System, which prohibited new entrants from targeting fragmented districts in which there were already a number of gasoline stations, allowed for greater competition among retail players. Consequently, incumbent gasoline stations began to lower prices in an effort to build customer loyalty. Corporate profit margins, historically low by world standards, continued to decrease in the newly competitive operating environment. Consequently, consolidation, asset streamlining, and cost cutting were initiated in the industry. In 1999, Nippon Oil and Mitsubishi Oil, for example, merged to form the largest Japanese oil distributor by market capitalization (667 billion yen as of November 22, 2002), sales (estimated at 4 trillion yen in fiscal year 2002), and market share (23%).

3. SOCIAL AND ENVIRONMENTAL IMPACTS

Energy policy does not exist in a vacuum: like all governments, Japan’s national energy policy must coexist with social and environmental issues. How these factors shaped Japan’s changing energy policy landscape is discussed in the following sections. On the one hand, Japan’s fossil fuel dependency did not add to national security. Alternative sources of energy, such as nuclear energy, were encouraged and developed to serve this end. On the other hand, Japan, as a leading nation, had environmental obligations such as those defined in the Kyoto Protocol in 1997. Environmental concerns were not always in conflict with Japan’s wish to free itself from its fossil fuel dependency. However, the goals were not always in tandem.

3.1 Kyoto Protocol

In December 1997, delegates from 159 countries gathered in Kyoto, Japan, for the Third Conference of the Parties (COP3) to the United Nations Framework Convention on Climate Change (UNFCCC). The results of that conference, known as the Kyoto Protocol, addressed important commitments from developed countries to reduce emissions of carbon dioxide and other greenhouse gases between the years 2008 and 2012. Industrial economies were required to cut aggregate man-made emissions of six climate-affecting gases by an average 5.2% below

1990 levels by some point in the 2008–2012 time frame. The six gases, carbon dioxide (CO₂), nitrous oxide (NO_x), methane (CH₄), hydrofluorocarbons (HFCs), perfluorocarbons (PFCs), and sulfur hexafluoride (SF₆), were considered to be ozone-depleting chlorofluorocarbons. As such, developed countries agreed to incorporate into their efforts to implement the Kyoto Protocol measures that would promote, assist, and finance technology transfers in fighting occurrence of such emissions in developing countries.

Japan's emissions reduction target was agreed to be 6% and that for the European Union was 8%; the goal of the United States was to be 7%. As host nation, Japan stressed the importance of the Kyoto Protocol. Conservation and efficiency improvements yielded a double benefit. Conservation policies not only limited the country's overall dependency on imported fossil fuels, they also made export-led companies, such as the Japanese automakers, more competitive through fuel-efficient products.

One strategy to meet the stipulated goals of the Kyoto Protocol was to redouble efforts in the development of alternative sources of energy that reduce or eliminate pollution. If alternatives were palatable enough, significant reductions in the consumption of fossil fuels could be made. Solar, geothermal, fuel cell, and wind power sources were (and are) four such contenders in the implementation of such a strategy. These renewable energy sources currently hold only a small market share of the total primary energy supply. Wind power, for example, had an estimated 0.03% market share in 2000, but is projected to capture a 20% market share by 2020. The METI has been researching and developing practical development of these sources.

3.2 Nuclear Power

In complying with the goals of both the Kyoto Protocol and the METI's long-term energy strategy (energy security, environmental protection, economic efficiency, and economic growth), nuclear power continues to play an important, albeit controversial, role. Despite being the only country to suffer directly the fallout of nuclear weapons, Japan has adopted the peaceful use of nuclear technology to provide a material portion of its electricity generation. Today, nuclear energy accounts for some 30% of the country's total electricity production. However, nuclear accidents, policy contradictions, "not-in-my-backyard" protests, and regional funding issues present recurring obstacles to the government's plans

to perfect a completely indigenous fuel source for the 21st century.

3.2.1 Nuclear Development and Policy:

A Brief History

Following U.S. President Dwight Eisenhower's historic speech, "Atoms for Peace," at the United Nations in 1953, Japan began its nuclear research program. The Atomic Energy Basic Law was enacted in 1955 with the aim of ensuring the peaceful use of nuclear technology in Japan. Democratic methods, independent management, and transparency were (and are) the foundation of nuclear research activities, as well as the promotion of international cooperation. Several nuclear energy-related organizations were established in 1956 under this law to further promote development and utilization, including the Atomic Energy Commission; the Science and Technology Agency; the Japan Atomic Energy Research Institute (JAERI), and the Atomic Fuel Corporation (renamed the Power Reactor and Nuclear Fuel Development Corporation in 1967).

As Fig. 5 indicates, Japan has 52 reactors with a combined installed capacity of 45.7 GW, or approximately 30% of the nation's total installed capacity. The country imported its first commercial nuclear power reactor from the United Kingdom (Tokai-1) in 1966. This gas-cooled (Magnox) reactor built by General Electric Company (GEC) had a relatively small installed capacity of 160 MW; the reactor was finally decommissioned in March 1998. After this unit was completed, only light-water reactors (LWRs) using enriched uranium, either through pressurized water reactors (PWRs) or boiling water reactors (BWRs), have been constructed. Since 1970, 23 PWRs and 28 BWRs (including two advanced BWRs) have been brought online. Initially, Japanese electric power utilities purchased designs from U.S. vendors and built them with the cooperation of Japanese companies, who received licenses to then build similar plants in Japan. Companies such as Hitachi Co Ltd., Toshiba Co Ltd., and Mitsubishi Heavy Industry Co Ltd. developed the capacity to design and construct LWRs. An additional 11 reactors are in the planning stages or are currently under construction.

3.2.2 Reprocessing and Waste Disposal:

National Policy

As already mentioned, one of the obvious goals of Japan's national energy policy is energy security. With the increase in nuclear power as a percentage of its total primary energy source, Japan hopes to

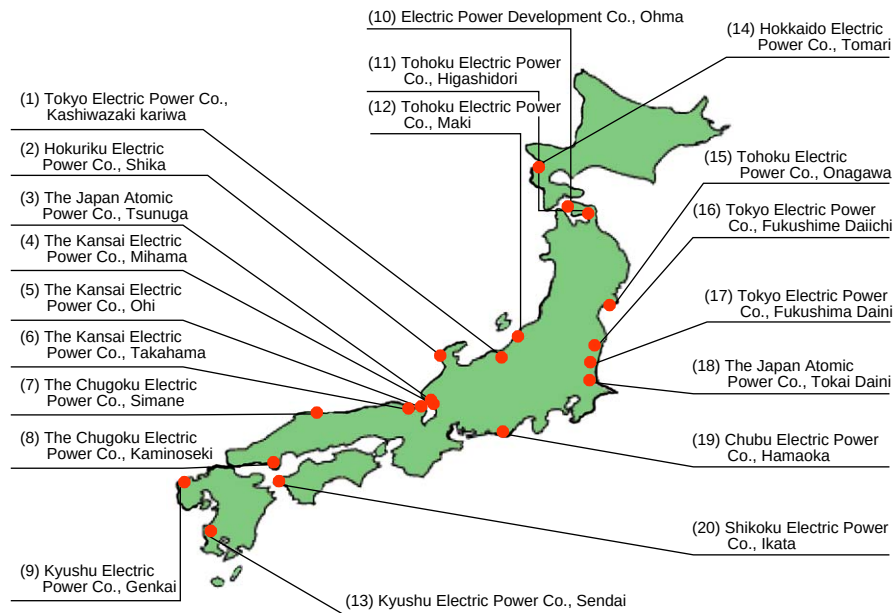


FIGURE 5 Nuclear power plants in Japan. As of 2003, 20 locations are power plant sites (or will be) for 63 units (52 operational, 3 under construction, and 8 in the planning stage), with an expected total output of 59,895 MW (45,742, 3838, and 10,315 MW, respectively). The locations on the map are numbered; the following key details the operational (OP), under-construction (UC), or planning-stage (PS) unit output at each location: (1) seven OP units, each > 1000 MW; (2) one OP unit < 1000 MW and one UC unit > 1000 MW; (3) two OP units, one < 500 and one > 1000 MW, and two PS units, each > 1000 MW; (4) three OP units, one < 500 and two < 1000 MW; (5) four OP units, each > 1000 MW; (6) four OP units, each < 1000 MW; (7) two OP units, one < 500 and one < 1000 MW, and one PS unit > 1000 MW; (8) two PS units, each > 1000 MW; (9) four OP units, two < 1000 and two > 1000 MW; (10) one PS unit > 1000 MW; (11) one UC unit > 1000 MW; (12) one PS unit < 1000 MW; (13) two OP units, each < 1000 MW; (14) two OP units, each < 1000 MW, and one PS unit < 1000 MW; (15) three OP units, each < 1000 MW; (16) six OP units, one < 500, four < 1000, and one > 1000 MW; (17) four OP units, each > 1000 MW; (18) one OP unit > 1000 MW; (19) four OP units, two < 1000 and two > 1000 MW, and one UC unit > 1000 MW; (20) three OP units, each < 1000 MW. The Japan Atomic Power Company Tokai plant closed in March of 1998. Map, plant locations, and output derived from the 2003 data of the Federation of Electric Power Companies.

reduce its import dependency on fossil fuels. However, the challenges facing the Japanese government are much more daunting with the introduction of nuclear power. Plutonium is essential as a major fuel for nuclear power generation. The more plutonium used by a country, the more likely the nation becomes influenced by international politics, especially because plutonium produced within the nuclear reactor can be used for nuclear weapons.

Recovering plutonium through a process known as the “fuel cycle” essentially reprocesses spent fuel. The theory is that once the nuclear fuel cycle is established domestically, nuclear power virtually becomes an indigenous energy. Until now, Japan relied on the reprocessing of spent fuel by European contracts through British Nuclear Fuels (BNFL) and Cogema, with vitrified high-level wastes being returned to Japan for disposal. However, this reprocessing has proved to be expensive and time consuming. In 2005, Japan Nuclear Fuel Ltd. (JNFL) will begin its first commer-

cial enrichment plant at Rokkasho in northern Japan. Its planned capacity is to be 1.5 million separative work units (SWUs)/year (in the nuclear power industry, the separative work unit is a measurement of mass: 1 kg of separative work = 1 SWU). Spent fuel has been accumulating in Rokkasho since 1999 in anticipation of the full operation of the plant (shipments to Europe stopped in 1998).

The plutonium recovered by foreign reprocessing in the United Kingdom (BNFL) and France (Cogema) will be used in LWRs as mixed-oxide (MOX) fuel. MOX fuel was first intended to be used in the Takahama nuclear plant of Kansai Electric Power Company. However, local concerns surrounding the safety of MOX fuel in 2002–2003 created scheduling problems for the implementation of that program.

3.2.3 The Economics of Nuclear Power

The economics of nuclear power generation is largely controversial. Although the fuel cost of generation

is relatively inexpensive ($\geq 20\%$ of total nuclear generation costs), its capital costs are increasingly expensive. In a deregulating market where the marginal cost of new entry is increasingly competitive vis-à-vis incumbent prices, the incumbent competitiveness of Japan's nuclear power remains unclear. Current nuclear projects depend not only on material cost-cutting within tight schedules, but also on operating such generators at reasonably high utilization rates over many years. By way of contrast, fossil fuel-fired power plants are relatively cheap and quick to build, but expensive to operate due to their cost of fuel (currently $\geq 70\%$ of total thermal generation). Thus, the economics of nuclear power further inform the already contentious debate.

3.2.4 "Not-in-My-Backyard" Issues

The "not-in-my-backyard" (NIMBY) attitude is far from being a Japan-specific obstacle to national energy policy, but is one with significant long-term ramifications for the implementation of a successful nuclear power development program in Japan. Simply put, NIMBY refers to a grassroots movement to prevent the further construction and maintenance of nuclear power plants in local communities, where it is believed to represent critical environmental and safety hazards for the general public. A series of nuclear power accidents, mishaps, and scandals have further exacerbated the already volatile and emotional debate on the peaceful use of nuclear energy sources, thus eroding public support in Japan and reinforcing NIMBY sentiments. The accidents involved a sodium leak at the Monju fast breeder reactor (FBR), a fire at the Japan Nuclear Cycle Development Institute (JNC) waste facility connected with the Tokai reprocessing plant, and a 1999 criticality accident at Tokaimura. The latter accident, which claimed two lives and seriously injured three others, was the direct result of workers following an unauthorized procedures manual.

In 2002, an additional scandal—non-accident-related—erupted over an alleged cover-up of safety inspection procedures. Inspection of the shrouds and pumps around the nuclear reactor core, the responsibility of the electric power company, had been contracted out by the power company. In May 2002, questions emerged about data falsification and the significance of reactor shrouds and whether faults in the shrouds were reported to senior management. This incident further sullied the reputation of the incumbent electric power companies and the METI's designs for nuclear power development, in general.

4. CONCLUSION

Japan is a naturally resource-deficient island nation that relies heavily on imported fossil fuels for its energy needs. Stability in the international community continues to be a prerequisite for acquisition and maintenance of Japan's energy supplies. Especially in the case of oil, the fuel on which Japan is most dependent, securing access to stable supplies has been the mainstay of Tokyo's energy policy for over 30 years.

The two oil crises of the 1970s were set in motion by developments on the supply side, but if a third oil crisis should occur, the problem may be a strictly demand-side affair. World oil consumption, especially in Asian countries, continues to rise year after year, calling into question viable energy sources for future consumption. Japan has recognized the need to adopt a strategy that embraces both its imported energy dependency and its ability to offset such problems with indigenous natural forms of energy, such as nuclear, wind, solar, and geothermal power. In advancing the four Es—energy security, environmental protection, economic efficiency, and economic growth—the Japanese government endeavors to convince the Japanese public of the benefits of nuclear power, but with limited success. Nuclear accidents and scandals, both domestically and internationally, have tarnished nuclear power's image as a safe, viable alternative to imported fossil fuels. Moreover, NIMBY protests and the introduction of liberalization have added extra political and economic dimensions to this already contentious subject matter.

Deregulation of the energy industries (electricity, gas, and oil) will likely continue. Already, revisions to the basic laws have shown signs of material change. Prices in all three sectors have gradually fallen. Competition has led to partial consolidation in at least the petroleum industry, with prospects for the other sectors in the years to come. In conclusion, Japan's national energy policy is in a state of constant fluctuation and development. The foremost priority is stability of supply. Especially vital to this goal is maintaining an ample supply of oil and other imported fossil fuels needed to feed the world's second largest economy.

Acknowledgments

The author sincerely thanks the following people for reading an earlier draft of this article and/or for their many helpful suggestions throughout the course of this study: Dr. Chris Rowland and Dresdner Kleinwort Wasserstein (United Kingdom), Mr. Peter C. Evans (Massachusetts Institute of Technology, United States), and Dr. Yuki A. Honjo (JapanReview.net, Japan).

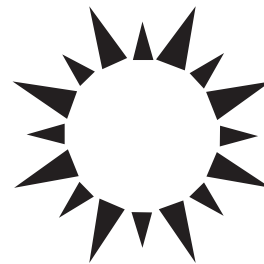
SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • National Energy Policy: United States • National Security and Energy • Oil Price Volatility • World Environment Summits: The Role of Energy

Further Reading

- Evans, P. (1997). "Japan's Deregulated Power Market: Taking Shape." A Cambridge Energy Research Associates (CERA) Global Power Forum Report. Cambridge Energy Research Associates, Cambridge.
- Hein, L. E. (1990). "Fueling Growth: The Energy Revolution and Economic Policy in Postwar Japan." Harvard Univ. Press, Cambridge, Massachusetts.
- International Energy Agency (IEA) (2002). "Energy Balances of OECD Countries." Organization for Economic Cooperation and Development, Paris. (CD-ROM)
- Lesbirel, H. (1998). "NIMBY Politics in Japan: Energy Siting and the Management of Environmental Conflict." Cornell Univ. Press, Ithaca, New York.
- Oyama, K. (1998). The policymaking process behind petroleum industry regulatory reform. In "Is Japan Really Changing Its Ways?: Regulatory Reform and the Japanese Economy" (L. Carlile and M. Tilton, Eds.), pp. 142–162. Brookings Institution Press, Washington, D.C.
- Samuels, R. (1987). "The Business of the Japanese State: Energy Markets in Comparative and Historical Perspective." Cornell Univ. Press, Ithaca New York.
- Scalise, P. (2001). "The Powers That Be: Japanese Electricity Deregulation." Dresdner Kleinwort Wasserstein, Tokyo.

National Energy Policy: United States



MIRANDA A. SCHREURS

University of Maryland
College Park, Maryland, United States

-
1. Energy Politics
 2. George W. Bush's National Energy Policy

Glossary

Btu tax A tax on the heat or energy content of fuels. A British thermal unit (Btu) is defined as the amount of heat necessary to change the temperature of 1 lb of water at sea level by 1°F. The Broad-Based Energy Tax proposed in 1993 by the Clinton Administration would have placed a tax on energy based on its energy or Btu content. A gallon of diesel fuel, has more Btus than a gallon of liquefied natural gas. Thus, under the Clinton Administration proposal, fuels with a high energy content, which tend to be the dirty fossil fuels, would have been taxed at a higher rate than many alternative energies, which have lower average energy contents.

Corporate Average Fuel Economy (CAFE) Standard A miles per gallon standard established by law that manufacturers of cars and light trucks must obtain. In 2003, CAFE standards are set at a minimum of 27.5 miles per gallon averaged across a manufacturer's entire fleet. If a manufacturer does not meet the standard, it must pay a civil penalty of \$5.00 for each 0.1 mile per gallon that the fleet does not obtain multiplied by the number of vehicles the manufacturer produces.

Energy Policy and Conservation Act (1975) This was one of the earliest laws enacted in the United States with the explicit purpose of regulating and reducing energy consumption. The Act also established a Strategic Petroleum Reserve in the Gulf of Mexico. During reauthorization, a national inventory of onshore energy sources was created and a home-heating oil reserve was established in New England.

Public Utilities Regulatory Policies Act (1978) This law was passed in the face of high energy prices in an effort to reduce dependence on foreign oil, to encourage energy efficiency, and to promote alternative energy sources. It required electric utilities to purchase power from independent companies that were able to produce

power at a lower cost than what it would have cost the utility to generate the power. The Act is credited with the development of nonhydro renewable energies and has been especially favorable to the development of natural-gas-fired "cogeneration" plants that produce both electricity and steam.

Efforts to establish a national energy policy in the United States began under Jimmy Carter's presidency during the late 1970s, but then languished for over a decade. Since the early 1990s, there have been renewed attempts to establish a national energy policy. In 1992 a National Energy Strategy was formulated but U.S. dependence on imported energy continued to grow. Under the George W. Bush administration, there have been renewed efforts to establish a national energy policy, but sharp differences of opinion have prevented the passage of new energy legislation, at least for the time being.

1. ENERGY POLITICS

Understanding energy politics in the United States and efforts to develop a national energy policy requires an understanding of the different actors involved and their economic and political interests. There are widely divergent views regarding how to deal with America's large and growing appetite for energy and the pollution this produces.

The fossil fuel industries—the oil, coal, and natural gas producers—tend to support policies that favor the expansion of drilling and mining activities, including in protected lands and offshore. The automobile industry is typically opposed to legislation that mandates fuel efficiency improvements or raises gasoline taxes. The nuclear industry is eager to see a renewal in government support for nuclear

energy. Environmental groups and the renewable energy industries tend to call for greater governmental support for energy conservation and renewable energy and legislation to control pollution from fossil fuel burning. Labor unions tend to be most concerned with the implications of governmental policies on jobs and consumer groups with the prices consumers will have to pay for energy. Each of these groups tends to frame energy issues in a different light, with some groups concerned more with economic security and quality of life concerns, others with national security interests, yet others with environmental protection, and some primarily with jobs.

Energy politics is also closely tied to the nature of the energy source, risk perceptions related to the use of the energy source, and the distribution of energy resources in the country.

Coal mining has been central to U.S. energy politics for well over a century and coal production is on the rise. However, although coal mining was a major employer in the past, the number of coal miners has steadily declined with time. Whereas in 1980 there were an estimated 220,000 coal miners in the country, in 2003 that number is down to approximately 100,000. Coal is a heavily regulated energy source because of the high safety risks for miners and the pollution emitted from coal burning as well as the damage that can be caused to lands from coal extraction.

Coal mining is heavily concentrated in Wyoming, West Virginia, and Kentucky, which together account for more than one-half of U.S. production of coal. Other major coal-producing states include Pennsylvania, Texas, Montana, and Illinois. Efforts to regulate coal production and coal burning in power plants for environmental reasons has resulted in considerable interstate politics, pitting producing states against downwind states suffering from acid rain.

Crude oil production is also heavily concentrated in a few states. Texas, Alaska, and California are the three largest producers, with Louisiana, Oklahoma, New Mexico, and Wyoming also being substantial producers.

Federal and Indian lands are of great interest to energy developers. Approximately 38% of total U.S. coal production was from Federal and Indian lands in 1999. The federal government owns the outer continental shelf, which is the territory that lies approximately 3 nautical miles from the shoreline and extends out 200 nautical miles. In 1998, the outer continental shelf (primarily in the Gulf of Mexico) was the source of close to 25% of domestic natural gas and 20% of crude oil production.

The politics surrounding nuclear energy are distinct from those regarding other energy sources. The future of nuclear energy remains highly uncertain and heavily debated. There are 103 licensed nuclear reactors operating in 65 plants in 31 states. In the 1960s and 1970s, the government provided the nuclear energy industry with substantial subsidies to offset the heavy initial investment required for new plant construction. The government also guaranteed the nuclear industry that it would develop a national nuclear waste depository. There are, however, many obstacles facing the nuclear energy industry. The accident at the Three Mile Island nuclear station in March 1979 that led to the permanent shutdown of reactor number 2 sent chills through the nation and intensified an already strong tendency for local communities and environmental groups to object to the building of new nuclear power plants. According to the report, "Nuclear Energy Policy," high construction costs are another serious problem for the industry; construction costs for reactors completed since the mid-1980s have been between \$2 and \$6 billion, or more than \$3000 per kilowatt of electric-generating capacity (in 1997 dollars). Although no new nuclear power plants have been ordered since the Three Mile Island disaster and over 100 reactors have been canceled, 16 commercial reactors have received 20-year license extensions, with another 14 plants undergoing reviews for such extensions.

Over time, there has been some consensus established among stakeholders on the importance of energy efficiency improvements and energy conservation, but there is no real agreement on how the nation should deal with its large and growing appetite for energy. How much of a role should the government play in ensuring a stable energy supply? To what extent should government policy favor particular energy industries? Is deregulation of the electricity sector a good idea? What mix of energy sources should be achieved? To what extent should environmental considerations play a role in the nation's energy plans?

1.1 Energy Supply and Demand

The United States is the world's largest consumer of energy, accounting for 24% of total world energy consumption; it is a major importer of energy and is the world's largest source of greenhouse gas emissions. Approximately 86% of total fuel consumption is of fossil fuels. The United States is the world's second largest producer of coal after China and the second largest producer of natural gas after Russia. It also is a major producer of oil, representing 9% of

global production and nuclear energy, accounting for 31% of global nuclear electricity production.

Almost three decades after the 1973 oil embargo by the Organization of Petroleum Exporting Countries (OPEC), the United States remains highly dependent on energy imports. In 2001, the United States consumed 97 quadrillion British thermal units (Btu) of energy. Approximately 39% of this was oil, 24% natural gas, 23% coal, 8% nuclear, and 6% renewable energies (primarily hydro and biomass). In contrast, the United States produced approximately 72 quadrillion Btu of energy [33% coal, 28% natural gas, 21% oil, 11% nuclear, and 8% renewables (largely hydro and biomass)]. This means that approximately 30% of all energy consumed in the country is imported.

U.S. dependence on oil imports is especially large, at 53% of all oil consumed. Given that approximately 20% of oil imports are from the Persian Gulf and another 40% from OPEC, U.S. energy markets are very dependent on the politics of the Middle East, a highly volatile region. The United States has also become a net importer of natural gas.

Although there have been substantial improvements in energy efficiency over the past three decades, total energy consumption continues to rise as a result of a growing population and more energy-intensive lifestyles. Thus, although the U.S. economy has become approximately 60% more efficient in the past 30 years, per capita energy consumption levels in the United States are among the highest in the world. On average, an American uses 342 million Btu of energy per year, or almost twice as much as a Japanese (174 million Btu) or a German (172 million Btu). Moreover, as the U.S. population is expected to grow substantially in the next decades due to a combination of immigration and a relatively high birth rate compared with many other advanced industrialized states, energy demand is expected to continue to rise.

1.2 Energy Planning in U.S. History

Unlike many other advanced industrialized democracies, the United States does not produce regular multiyear national energy plans. Historically, the lack of interest in long-term energy planning in the United States stems both from a relative abundance of energy and from a fairly strong tendency toward neoliberal economics. Prior to the 1970s, energy policy change tended to be incremental and largely reactive.

Through the late 1950s, the United States produced at least as much energy as it consumed. In contrast with many countries in Europe, which

nationalized energy industries in the first half of the 20th century, in the U.S. energy production remained in private hands and little consideration was given to long-term energy supply.

In the 1960s, energy consumption began to surpass energy production and by the 1970s, the United States had become a major importer of energy, and especially petroleum. Over the course of the 1970s, in response to the nation's severe environmental problems and sharp increases in energy prices, the government became increasingly involved in regulating the energy industry, promoting energy conservation, and providing incentive schemes for the targeted development of specific energy sources.

1.2.1 *The Impact of the OPEC Oil Embargoes*

The 1973 oil embargo by OPEC sent world oil prices soaring. In reaction to the embargo, in 1975 at Gerald Ford's urging, Congress passed the Energy Policy and Conservation Act, which established Corporate Average Fuel Economy (CAFE) standards for automobiles, extended domestic oil price controls, and created the Strategic Petroleum Reserve, an oil stockpile for emergency situations. Two years later, Congress created the Department of Energy (DOE).

Jimmy Carter was convinced that the nation's energy security demanded a comprehensive energy plan. Carter's 1978 National Energy Plan was the first attempt by a president to establish a national energy policy, one that called for both expanded production of coal and enhanced energy conservation. The plan included numerous measures for the promotion of renewable energies, provisions for energy conservation, and energy taxes. It also led to the establishment of the Public Utilities Regulatory Act, which required utilities to purchase energy from "qualifying facilities," effectively ending the electric utility monopoly on electricity production and helping to foster a market for renewable energy sources.

1.2.2 *The Ronald Reagan Years*

Many of the policy goals laid down in Carter's National Energy Plan, however, came under attack during the years of the Reagan administration. Reagan was eager to down-size government and remove many of the regulations he felt were burdening the energy industry. Although he failed in his effort to close the Department of Energy, he succeeded in winning Congressional support to rescind tax breaks for energy-saving devices and decreased government funding for research and development of renewable energy sources and energy conservation initiatives.

1.2.3 George H. W. Bush's National Energy Strategy and Energy Policy Act

Although George H. W. Bush had a political outlook similar to that of Ronald Reagan, like Carter he did see the need for the creation of a national energy plan. Thus, in 1989 he ordered the DOE to prepare a National Energy Strategy (NES). The NES was a response to growing concerns with global warming and rising energy prices in the late 1980s and early 1990s. The plan was criticized by environmentalists, however, because it included plans for oil drilling in the Arctic National Wildlife Refuge (ANWR) and did little to reduce U.S. carbon dioxide emissions. Indeed, a decade later these same issues remain two of the major points of contention between Republicans and Democrats in the formulation of a national energy plan.

The NES led to the adoption by Congress in 1992 of the Energy Policy Act, which began the opening of the electric utility industry to competition, established energy efficiency standards for federal facilities, mandated that the federal government replace some of its fleet with alternative fuel vehicles, established the (unmet) goal of having 10% of the nation's gasoline use be replaced by alternative fuels (as of the year 2000, only 3.6% of the nation's highway gasoline use had been replaced), called for voluntary reporting of greenhouse gas emissions and the establishment of a national inventory of greenhouse gases, and initiated measures to aid the nuclear energy industry (including a revision of the regulatory framework and environmental standards governing nuclear waste disposal at Yucca Mountain in Nevada).

Although the 1992 National Energy Policy Act was the most important energy legislation to be passed in over a decade, it failed to do much to move the nation toward greater energy independence or to do much to promote a more sustainable energy future.

1.2.4 William J. Clinton's Failed Btu Tax

Little progress in these directions was made under the Clinton administration, either. Clinton and Vice President Al Gore proposed a tax on the heat content of fuels (a Btu tax) in an effort to cut energy consumption (and thereby reduce dependence on imported oil) and to reduce greenhouse gas emissions. The Btu tax, however, was rejected by Congress in favor of a 4.3 cent per gallon increase in the federal gas tax, which, because of the historically low price of gasoline at the time, was politically palatable. The Senate also made known its lack of intention to ratify the Kyoto Protocol that the Clinton administration formally signed in 1998. Had

the Kyoto Protocol been ratified by the Senate, major efforts at energy conservation and the establishment of a carbon emissions trading system would have been necessary.

2. GEORGE W. BUSH'S NATIONAL ENERGY POLICY

Soon after George W. Bush entered office, California was hit by rolling electricity blackouts and sharp increases in electricity prices. In reaction to the California "energy crisis" and fulfilling a campaign pledge, shortly after taking office Bush announced his intention to establish a national energy plan. Vice President Richard Cheney was given the task of leading an energy task force that was charged with drafting a plan. After 3 months of closed-door meetings and with considerable fanfare, in May 2001 the energy task force released its National Energy Policy (NEP) report, "Reliable, Affordable, and Environmentally Sound Energy for America's Future." The report and pursuant energy bills have been the subject of heated Congressional debate and a number of lawsuits.

The NEP report suggests that the nation is facing the most serious energy crisis since the oil embargoes of the 1970s. Moreover, in the coming decades, unless action is taken, "projected energy needs will far outstrip expected levels of production. This imbalance...if allowed to continue, will inevitably undermine our economy, our standard of living, and our national security." To deal with the crisis and future energy needs, the report calls for promoting energy conservation, repairing and modernizing the energy infrastructure, and increasing energy supplies.

Perhaps the strongest theme to come out of the NEP report is the need to increase the nation's energy supply. The NEP calls for oil drilling in ANWR, the promotion of clean-coal technology, nuclear energy development, and natural gas exploration. It also makes mention of the need to encourage renewable energy development, but suggests that renewable energies are unlikely to make a large dent in U.S. foreign energy dependence. A second theme of the report is the potential to make major gains in energy efficiency and new energy sources through technological developments. The report also urges a modernization and expansion of the nation's aging energy infrastructure (oil and natural gas pipelines, refinery capacity, and electricity transmission grids).

In total, the report included 105 recommendations, including many that called for cooperation

with foreign governments to improve the environment for energy investment and to improve the stability and security of supply.

The release of the NEP has had a mixed reception. Conservatives like it. The Heritage Foundation's Senior Policy Analyst for Energy and Environment calls the plan "a step in the right direction" because it calls for meeting the nation's energy needs through developing as yet untapped domestic energy sources, removing interstate transmission restrictions, modernizing energy delivery systems, promoting nuclear energy (which does not produce greenhouse gas emissions) and clean coal technology, and removing regulatory burdens that create market inefficiencies.

In contrast, environmentalists have been highly critical of the NEP. Greenpeace, for example, argues that the NEP is leading the nation "down the wrong road" because the plan fails to take steps to reduce greenhouse gas emissions, calls for 1300 new fossil fuel and nuclear power plants, favors oil extraction in ecologically sensitive areas (including ANWR and the Rocky Mountains), and will use taxpayer subsidies for the nuclear energy and fossil fuel industries.

Several lawsuits have been brought against the Vice President's office as well in relation to the closed-door process by which the NEP was drafted. The National Resources Defense Council (NRDC) sued the Department of Energy, a key member of the task force, under the Freedom of Information Act, for the release of thousands of pages of documents used by the task force in formulating the plan. Although the DOE provided some 13,000 pages that the NRDC has since made available on the Internet for public scrutiny, thousands of additional pages of documents were not released. In February 2002, the NRDC won a District of Columbia court motion requiring the DOE to expedite the release of the remaining documents.

Judicial Watch and the Sierra Club are suing the Vice President's office for the release of additional documents, task force minutes, and computer records that would shed light on who advised the task force and how this may have influenced the document's development. Their lawsuits are motivated by the concern that private industry representatives essentially functioned as members of the advisory group, and under the law, this would mean that the group's deliberations should be made open for public scrutiny. The collapse of Enron in early 2002 and the revelations that Enron had tried to manipulate California's electricity markets for profit have also played into these lawsuits since Cheney met

with Enron's Chief Executive Officer, Kenneth L. Lay, several times while he was heading up the task force. Cheney's office has sought unsuccessfully to have the cases dismissed.

The terrorist attacks on the World Trade Center and the Pentagon on September 11, 2001, and the subsequent U.S. decision to go to war in Iraq, for a time shifted government and public attention to national security concerns.

Events in the Middle East did, however, feed into energy debates and renewed concern about U.S. oil dependence on the Middle East. They strengthened national sentiment on the need to reduce U.S. dependence on imported oil. But clear partisan differences remained on how greater energy independence should be achieved. The Republicans have tended to favor greater domestic production, including production of nuclear energy, whereas moderate Republicans have at times joined Democrats in calling for more energy conservation and support for nonnuclear renewable energies.

Thus, although recognizing the need for some kind of national energy policy plan, the 107th Congress failed to pass comprehensive national energy legislation even though both houses had passed energy bills. Major differences in the Democratic-controlled House and Republican-controlled Senate versions of energy legislation could not be bridged and, thus, Congress ended its session without passing energy legislation. The November 2002 elections returned control of the Senate to the Republicans and strengthened the position of Republicans in the House.

Despite the Republicans' strong showing in the 2002 election and high voter-approval ratings, the Bush administration has a poor environmental image. The administration has done little to win the confidence of environmentalists with its rejection of the Kyoto Protocol, its failure to regulate carbon dioxide emissions or to set more stringent fuel efficiency standards, and its efforts to open protected federal lands to mining, oil and natural gas drilling, logging, and recreational purposes. An August 23, 2003 public opinion poll found that of the 1011 adults surveyed by telephone 53% preferred the Democrats' approach to the environment and 29% preferred Bush's (the respective figures in January 2002 were 43% and 38%). Similarly, 42% of Americans said they preferred the Democrat's approach to energy compared with 33% who said Bush was doing a better job (the respective figures in January 2002 were reversed, with 33% favoring the Democrats and 46% favoring Bush's approach).

In a bid to improve its image on environmental and energy issues, the Bush administration has begun to focus the domestic energy debate on hydrogen fuel development. Hydrogen is being billed as a pollution-free fuel since the energy-generating chemical reaction between hydrogen and oxygen produces only water as a waste product. Hydrogen production, however, requires energy from other energy sources. Natural gas is the most affordable means of producing hydrogen although its cost is still four times as high as the cost of producing gasoline. The administration's hope is that, through technology development, it will be possible to produce hydrogen more affordably, including with the use of coal and nuclear power.

In his January 2003 State of the Union address, President Bush announced his intentions to fund hydrogen fuel technology research and development. The administration targeted \$1.7 billion to be distributed over the subsequent 5 years for the Freedom CAR and Fuel Initiatives, public-private cooperative endeavors for the development of hydrogen fuel cells, hydrogen fuel cell-powered cars, and hydrogen infrastructure. The administration's stated goal is to have hydrogen-powered vehicles and a fuel distribution network in place by 2020.

The initiative has won praise, especially from Republicans and Democrats in states with automobile manufacturers, but has earned more mixed reviews from Democrats and environmentalists who have criticized the administration for placing so much emphasis on a still unproven energy source and providing relatively little investment for available renewable energy technologies. Concern has also been raised that in order to develop hydrogen it will be necessary to build more coal-burning fossil fuel plants and nuclear power facilities.

The 108th Congress took up the energy legislative debate that the 107th Congress failed to complete. The House of Representatives was the first to act. In April 2003, the House of Representatives passed H.R. 6 on a vote of 247 to 175. The bill includes approximately \$19 billion in energy-related industry tax incentives over a 10-year period for alternative fuels, energy efficiency, electricity restructuring, and oil and gas production. The bill provides tax breaks to oil and gas companies to encourage production from marginal wells and offshore drilling, provides millions of dollars for clean coal technology, and permits utilities to more easily write off the cost of new transmission systems. The bill also grants tax credits for solar and wind power and encourages energy-efficiency improvements in homes.

On a vote of 228 to 197, the House killed a proposed amendment that would have removed the provision for oil drilling on 2000 acres of the Arctic National Wildlife Refuge. Moderate Republicans joined Democrats in calling the move ill-conceived, arguing that the Arctic National Wildlife Refuge should be protected as one of the nation's most important pristine areas and that a more effective approach would be to save oil through the introduction of higher automobile fuel efficiency standards. In an earlier vote, however, the House defeated a provision that would have required a 5% reduction in automotive fuel use by 2010 (or an average fuel efficiency standard of approximately 30 miles per gallon). The justification given by opponents of higher auto mileage standards was that it would be bad for the economy as it would make it harder for manufacturers to produce popular sports utility vehicles, would result in layoffs, and would pose safety hazards because the standards would require the production of smaller cars.

Democrats in the House also failed in their efforts to rewrite electricity rules to protect consumers from market manipulation that could lead to huge price swings and ban methyl tertiary butyl ether as a fuel additive despite its role as a potential source of groundwater pollution. The Natural Resources Defense Council has criticized the House bill for providing huge subsidies to the coal, natural gas, and nuclear industries but providing relatively few incentives for less polluting industries.

Attention then shifted to the Senate, which began voting on measures related to its national energy bill. Repeating events in the House of Representatives 3 months earlier, Senate Democrats failed in their effort to win support for a proposal for a 40-mile-per-gallon fuel-economy standard for passenger cars by 2015. Instead, the Republican majority voted to direct the Department of Transportation to work on a new standard, taking into consideration the impact it would have on jobs and consumer safety. The current CAFE standard of 27.5 miles per gallon has not been changed since the 1986 model year.

Deadlock in the Senate on a number of other energy issues was finally broken on July 31, 2003, when the Senate Republicans in a surprise move agreed to a suggestion by Senate Minority Leader Thomas A. Daschle that last year's energy bill (that was negotiated by a Democratic-controlled Senate) should be resurrected. The Republican majority agreed to this proposal, knowing that they would be able to write new provisions into it in the coming months. The Democrats extracted a promise that

separate votes would be taken in 2003 on climate change control and measures to give federal regulators greater authority to oversee utility mergers. In accepting the 2002 Senate Bill, the Republicans gave up on provisions for loan guarantees for new nuclear power plants.

The sense of urgency behind the need for a national energy policy was rekindled by the August 2003 electricity blackout that left New York City and large sections of the Northeast sitting in the dark. The primary cause of the blackout was determined to be an antiquated electricity distribution infrastructure.

In the following months in largely closed-door conferences, Republican leaders of the House and Senate met to try to iron out differences between the House and Senate versions of energy legislation. The controversial House decision to drill in ANWR was dropped from the bill by the Republican leadership after it became clear that the bill stood no chance of being passed in the Senate if it were included, but other controversial elements remained.

The bill proposed by the conferees included \$23 billion in tax incentives primarily for coal mining, oil and exploration, the construction of new transmission lines and power plants, and the building of a natural gas pipeline from Alaska to the Midwest. It also included tax incentives for wind power and biodiesel fuel made from soybeans and a doubling of ethanol fuel mandates. While the House of Representatives passed the bill, it was blocked by a coalition of 32 Democrats, 7 Republicans (mostly from Northeastern states), and one Independent in the Senate. The bill was supported oddly enough by labour unions, energy companies, and renewable energy producers. It was opposed, however, by those concerned with its large price tag (estimated to be over \$30 billion over 10 years). Senators from Northeastern states also objected to the incentives provided for Midwestern coal-fired utility plants because the pollution from these plants affects them. There was also bipartisan opposition to the provision that would have exempted manufacturers of methyl tertiary-butyl ether from product liability lawsuits.

Thus, although there was great demand for new energy legislation, strong differences between Democrats and Republicans, between Midwestern agricultural and mining states and Northeastern states, and between energy producers and environmentalists prevented passage of new energy legislation prior to the adjourning of the 108th Congress for the holiday break.

SEE ALSO THE FOLLOWING ARTICLES

Energy Development on Public Land in the United States • European Union Energy Policy • Fuel Economy Initiatives: International Comparisons • National Energy Policy: Brazil • National Energy Policy: China • National Energy Policy: India • National Energy Policy: Japan • National Security and Energy • Renewable Energy in the United States • Strategic Petroleum Reserves

Further Reading

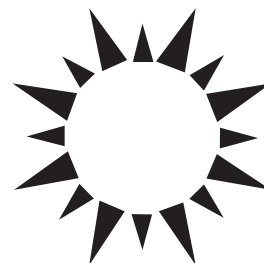
- Alliance to Save Energy. A nonprofit coalition of business, government, environmental, and consumer leaders: <http://www.ase.org>.
- American Council for an Energy Efficient Economy. A nonprofit organization dedicated to energy efficiency: <http://www.aceee.org>.
- Chubb, J. E. (1983). "Interest Groups and the Bureaucracy: The Politics of Energy." Stanford University Press, Palo Alto, CA.
- Davis, D. H. (1993). "Energy Politics." St. Martin's Press, New York.
- Department of Energy (DOE). Created in 1978 and has as one of its missions the protection of national and economic security by providing a diverse and reliable supply of energy. Also responsible for nuclear energy: <http://www.energy.gov>.
- Duffy, R. J. (1997). "Nuclear Politics in America: A History and Theory of Government Regulation." University of Kentucky Press, Lexington, KY.
- Energy Information Administration. A statistical agency of the Department of Energy that provides energy data, forecasts, and analyses: <http://www.eia.doe.gov>.
- House Committee on Energy and Commerce. The House of Representatives' committee that is responsible for the supply and delivery of energy, among many other issues. It also has jurisdiction over the Department of Energy: <http://energycommerce.house.gov>.
- Ikenberry, G. J. (1988). "Reasons of State: Oil Politics and the Capacities of American Government." Cornell University Press, Ithaca, NY.
- Jasper, J. M. (2000). "Nuclear Politics: Energy and the State in the United States, Sweden, and France." Princeton University Press, Princeton, NJ.
- Kraft, M. E. (2003). "Environmental Policy and Politics." 3rd ed. Pearson Longman, New York.
- Landsberg, H. H. (1993). "Making National Energy Policy." Resources for the Future, Washington, DC.
- Natural Resources Defense Council (NRDC). An influential environmental action organization active on energy and environmental issues: <http://www.nrdc.org/>
- Senate Committee on Energy and Natural Resources. The Senate committee that deals, among other policy issues, with energy resources and development, including regulation, conservation, strategic petroleum reserves, and appliance standards; nuclear energy; surface mining; and federal coal, oil, and gas: <http://energy.senate.gov/>.
- Stagliano, V. A. (2001). "A Policy of Discontent: The Making of a National Energy Strategy." Pennwell, Tulsa, OK.
- Tugwell, F. (1988). "The Energy Crisis and the American Political Economy: Politics and Markets in the Management of Natural Resources." Stanford University Press, Palo Alto, CA.

Nationalism and Oil

VÍCTOR RODRÍGUEZ-PADILLA

National Autonomous University of Mexico

Mexico City, Mexico



-
1. Introduction
 2. The Rise and Peak of Nationalism: 1920–1980
 3. Pragmatism Substitutes Nationalism: 1981–2003

Glossary

economic rent What remains of revenue after deducting the production costs of a mineral deposit, including normal return (mean industrial return) on the required capital.

expropriation Obligatory transfer of the property of a particular party to state administration for the public interest, normally in exchange for an indemnity payment.

nationalization Authoritative transfer of private sector means of production to national collective ownership represented by the state for reasons of national security or public interest.

public monopoly Exclusive right of the state to undertake an economic activity such as oil or natural gas extraction.

sovereignty Abstract principle that designates the instance that detains legitimate authority, the only authority with the capacity to enact norms; the international community recognizes that sovereignty over natural resources is inherent to the nation and is exercised by its representative, the state.

state interventionism A growing and direct process of intervention by the state in the economy or a sector of the economy, acquiring property rights and undertaking business activities and productive activity.

Nationalism is a doctrine of a people under foreign domination who try to free themselves and form a sovereign nation. The oil industry evokes such doctrine—a historical phenomenon, a process, a value, a feeling, and a vision of a country. As a doctrine, nationalism claims the nation's ownership of its natural resources, the state's valuation of such resources, and the subordination of the management

of reserves and production by development policies. It is a dogma fed on socialist, communist, anti-imperialist, and anti-capitalist ideas that have been expanded here and there, imitating the achievements made by other countries.

1. INTRODUCTION

The term “nationalism” is commonly used to describe state interventionism that permeated the oil industry since the beginning of the 20th century. It was born in Latin America with the first concessions and was generalized in all continents during the postwar period and culminated by “overturning” the Organization of Petroleum Exporting Countries (OPEC) during the 1970s. The nationalist intervention took on various forms. Taken to its ultimate consequences, it led to the nationalization of oil property and activities and finally to a state-owned monopoly.

In its character of patriotic value, nationalism means defense of sovereignty. Territorial sovereignty, one of the essential bases of nation-state, becomes extendible to the subsoil and, therefore, to the resources contained within. In certain countries, ownership and the state-owned monopolies over oil and natural gas have even been indispensable elements in the consolidation of the nation-states themselves.

Equally, nationalism translates a sentiment against an external action. It evokes the struggle for the liberalization of oil, of which greedy foreign companies have taken control. Just as the imperialist practices of the superpowers awoke the nationalism of oppressed populations, the predatory practices of the multinationals aroused oil nationalism. It was a reply to the plundering, the arrogance, and the arbitrariness of the concession system. And this sentiment has tended to prevail.

However, nationalism also gave birth to independent development following a colonial past. To put oil at the service of development through ownership of the resources, a greater share of income and gradual control of related activities was the fundamental objective proposed by oil-producing countries. That nationalist vision contained a mix of values, interests, and aspirations that converted it into a vital challenge, an image that strongly took root in society.

Contrary to popular belief, nationalism does not exist *per se*. It emerges when the conditions are favorable, when the state cannot exercise sovereignty over its natural resources, and/or when foreign interests threaten that original, inalienable, and imprescriptible right. However, its achievements will be few if the context does not favor the state's intervention in the economy.

Because of these objectives, nationalism has been one of the political factors that have contributed the most to the structural transformations of the international oil industry. It called a halt to the concession system and performed a crucial role in recovering the ownership of natural resources and the administration of complete sectors of the oil industry by oil-producing countries. Following its peak during the 1970s, the nationalist fervor abated but did not disappear altogether. In a worldwide context characterized by the stagnation of oil prices, economic liberalism, and the end of the cold war, the international oil companies have returned to the oil-producing countries, but not on the same terms as before. Nationalism is currently lethargic but will reemerge if the circumstances change and become favorable once again.

2. THE RISE AND PEAK OF NATIONALISM: 1920–1980

History repeats itself from one country to another. A few years after oil exploitation began by multinational companies, a social discontentment arose and grew to become a national movement of rejection. And although the nationalist struggles took on different forms and accomplished diverse goals, they shared the same driving force: the reclaiming of the permanent sovereignty of the nation's natural resources. In 1917, Russia was nationalized without any indemnity whatsoever, but this related to a different logic: the assumption of control by the state of all means of production, not just the oil industry.

2.1 The Nationalist Movements

Nationalism was born in Latin America at the beginning of the 20th century. From there, it radiated all over the world. Historical reasons explain that course. On the one hand, it was the natural expansion zone for the North American oil industry; on the other hand, it was the zone that had achieved greater political consciousness and institutional maturity because it consisted of countries that had ceased to be colonies decades earlier. The savage capitalism of the concession system (Table I) and a heightened political consciousness conjugated to breathe life into the nationalist phenomenon.

Argentina created the first nationalized company in 1923, and Uruguay created the first state-owned refinery in 1931. The Bolivian government nationalized the assets of Standard Oil of New Jersey in 1937 after it was discovered that the company had played a dual role in the Chaco War of 1932–1935, a bloody conflict between Bolivia and Paraguay that was instigated and backed by Standard Oil and its rival, Royal Dutch Shell.

Lázaro Cárdenas expropriated in 1938 following failed attempts to control and put the oil industry at the service of internal development through laws and regulations. The 1917 constitution had established the state-owned ownership of oilfields but maintained the concession system. Standard Oil and Shell took advantage of this situation to divide up the territory, plunder the subsoil, and make Mexico the second-largest oil exporter in 1920 without a single tangible benefit for the country. The companies, although paid off, organized a blockade against Mexican crude and the national oil company (PEMEX).

At the end of the fierce dictatorship that turned the country into a vast petroleum field dominated by Shell, Esso, and Gulf, Venezuela changed the law in 1943 to recover a part of the oilfields and established a less unfair distribution of the oil benefits. But it was not until 1948 that Juan Pablo Pérez Alfonso managed to impose the “50/50” principle on the oil companies. A year later, such a principle was being applied in the Middle East oil-producing countries.

In Iran, the Anglo–Iranian Company rejected that distribution. In response, the Mossadegh government nationalized oil in 1951 and created the National Iranian Oil Company (NIOC). The International Court of Justice at The Hague, Netherlands, ruled in the Mossadegh government's favor, but the process was not completed due to a military coup d'état that ousted the government in 1952. The new authorities ratified state ownership of the oilfields and the

TABLE I

*Inheritance from the Old Colonial Regime***The Concession System**

Using skill, influence, and the support of its country of origin, an oil company obtained from the government of an underdeveloped country the concession for a vast territory to carry out oil exploration operations for 50 years or more. Frequently, it was the men close to power who obtained the concessions, but sooner or later, they were transferred to the oil companies. In the end, the majority of national territory was under concession.

The concession implied the exclusive right to explore, extract, and export the product so that no other investor could do the same in the zone indicated in the mining deed. The concessionaires organized and controlled the operations to suit their own criteria and interests. They decided the volume and nature of the investments, working areas, exploration plans, oilfields to be exploited, production capacity, extraction rates, and exportation volumes. The criteria used were linked to agreements between the oil companies and, in the last instance, to the evolution of a world demand broadly managed by the large multinationals. The role of the state was reduced to that of simple tax collector.

From there arose the dual economic structure. The oil industry was isolated from the rest of the economy. Its evolution depended on the behavior of world demand. From there, some oil areas suffered from overexploitation, and others suffered from stagnation. Also, the oil companies could monopolize very cheap oil to resell it at high prices. The main beneficiary was not the owner but rather the entity that extracted, exported, transformed, and sold the oil as products. The concessionaire ensured the highest benefits without the slightest obligation to satisfy the demands of economic and social development, neither of the country nor of the oil sector itself. In summary, the companies managed a geological heritage that was not theirs to their own convenience. They exercised sovereignty over oil resources of the country that had opened its doors to them. They usurped a right that was not theirs.

Another equally important problem was the manner in which the companies operated the concessions. With the complicity of weak or corrupt governments, they turned the oil areas into regions with their own laws, authorities, and police forces. They created a state within a state. Communities and farmers stripped of their land, workers in misery, accidents of gigantic proportions, destruction of oil fields, the accelerated wastage and exhaustion of reserves, enormous fortunes in the hands of a few, and interference in the internal affairs of the country were some of the concession system. It was not for nothing that this is historically considered as one of the most savage forms of capitalism. The unjust nature of such a system awoke nationalism, which caused its collapse.

Agreements between Large Companies

In 1928, Anglo-Iranian (BP), Royal Dutch Shell, Compagnie Française des Pétroles (Total), Standard Oil of New Jersey (Exxon), and Mobil Oil agreed, through the Red Line Agreement, to jointly penetrate the old frontiers of the Ottoman Empire (Turkey, Iraq, Syria, and the Arab peninsula) with the exception of Kuwait. Each company would make its own investments, independently of the others, in refining, distribution, and marketing. That same year, the three largest companies (Standard Anglo-Iranian, and Shell) signed the Achnacarry Agreement to conserve the market portions, regularize production, and determine world oil price, Mobil Gulf, Texaco, and Chevron joined the pact. These and other subsequent agreements allowed the "Seven Sisters Cartel" and Total to exercise total control over the extraction and exportation of Middle East oil. However, because the collusion among companies limited the aspirations of putting oil at the service of national development, the producing countries rebelled.

existence of the NIOC but maintained the concession system. That nationalization, aborted by the intervention of the International Oil Cartel and the large capitalist powers, had a double effect of dissuading other countries from following its example but confirming the legitimacy of the struggle. It propagated the idea that the state should ensure absolute control of oil operations through nationalization or negotiated agreements. The Iranian experience represented a real basis for the subsequent political changes in the region.

Nationalism gained strength during the 1960s. Certain Latin American countries created upstream monopolies but respected existing rights (e.g., Brazil, Argentina, Chile), whereas others nationalized (e.g., Cuba, Peru, Bolivia). The infection traveled as far as India and Syria. In parallel, numerous national companies were created in Venezuela, Kuwait, Saudi

Arabia, Argelia, Iraq, and Libya, some of which were strengthened by technical assistance from the former Soviet Union and from companies on the margin of the cartel.

A decisive factor in the progress of nationalism was the creation of OPEC in 1960 by Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela at a meeting in Baghdad, Iraq. These were later joined by Qatar in 1969, Indonesia and Libya in 1962, Abu Dhabi in 1967, Argelia in 1969, Nigeria in 1971, Ecuador in 1973, and Gabon in 1974. Born to combat the reduction in posted prices, unilaterally fixed by the oil companies, the cartel was not only dedicated to that. It encouraged the adoption of a single tax treatment in member countries, the recovery of ownership rights, and state control of production. In 1966, OPEC pronounced the principle of the countries' sovereignty over hydrocarbons. In 1968, it

pronounced the principle of national participation in the concessions. The organization's support turned out to be crucial when one of its members took action against the consortia. The example prospered. In 1965, the Latin American State Oil Reciprocal Assistance Organization was created. In 1968, the Organization of Arab Petroleum Exporting Countries (OAPEC) was formed.

Another fundamental factor in the rise of nationalism was the peak in the liberation and decolonization process. A total of 45 countries in Asia, Africa, and Oceania gained independence during the 1960s decade.

The movement of the nonaligned countries contributed to maintaining nationalist feelings on a high—on being constantly present against the political blockades and foreign domination, on the one hand, and for each country's right to liberty, independence, and autonomous development, on the other. However, the persistence of poverty was a reminder that access to independence was not enough to ensure progress. Economic independence must be conquered until a standard of living in accordance with the most elementary rules of human dignity is achieved. Developing countries wanted to make use of their wealth, organize their development, and cooperate on a basis of equality and reciprocal interest. From there arose, in particular, a movement ever more important in the rejection of the control of their natural resources by foreign companies and of reclaiming their national sovereignty. This discussion gained strength at international forums.

Two additional factors encouraged developing countries' aspirations of freeing themselves from the control of the large multinational companies. The first was the appearance and strengthening of new investors—the national companies of the industrialized nations such as France, Italy, Belgium, and Japan and the “independent” U.S. companies such as Occidental, Amoco, Conoco, Getty, and Arco—that were prepared to accept a share of the benefits more in favor of the producing countries with more restricted access to the oilfields as well as offering technical and commercial assistance to the national companies so as to gain certain permits and guarantee cheaper sources of supply. The second factor was the appearance of new production zones with high oil potential such as Libya and Nigeria. This led to fierce competition among companies, and the developing countries used this competition to their advantage to impose their conditions.

Nationalism reached its peak during the 1970s, not only as a consequence of a historical process but

also due to a series of favorable circumstances. On the economic plane were tensions regarding the availability of oil as a result of transportation problems, increased global production costs associated with the development of high-cost oilfields (e.g., the North Sea, Alaska) necessary to respond to the rapid growth in demand, and the need for higher international process to correct imbalances in the U.S. supply system. On the political front were the rise to power or consolidation of nationalist governments and a new Arab–Israeli war. Taking advantage of the political climate, the OPEC countries rebelled against the companies. Advances in one country were reproduced in the others, dealing simultaneously with ownership rights, the distribution of benefits, and administration of oil operations.

In 1970, Libya managed to increase taxes and posted prices. Iran and Kuwait copied the measure. Argelia went further by unilaterally fixing such prices. These events set the criteria to begin negotiations between OPEC and the companies, resulting in the Tehran (1971), Tripoli (1971), and Geneva (1972–1973) agreements. The most important achievement was that the determination of prices ceased to be a decision made only by the oil companies. From that point onward, the OPEC countries intervened in such decisions.

Argelia nationalized the interests of six companies and 51% of the French concessions in 1970. The following year, Libya nationalized BP's assets. In 1972, Iraq nationalized the Iraq Petroleum Company, Ecuador nationalized its oil, and Libya acquired 50% of ENI's assets. In 1973, Saudi Arabia's share in the ARAMCO concessions took effect; from an initial value of 25%, it would reach 51% by 1982. The shah of Iran and the companies agreed to immediate nationalization of all the assets of the Anglo–Iranian Company in exchange for a supply contract guaranteed for a 20-year period. Libya nationalized 51% of the remaining concessions. Nigeria acquired a 35% share of BP–Shell assets.

On October 6, 1973, the Yom Kippur War broke out. Ten days later, OAPEC members reduced their exports and declared an embargo against the United States and the low countries for having assisted Israel against Egypt and Syria. Weeks later, the exporting countries started to fix sovereignty and posted prices without consulting the oil companies. By the end of the year, Iraq had nationalized Exxon's and Mobil's equity in the Basrah Petroleum Company. The following year saw the full nationalization of Iraqi oil.

In 1974, the General Assembly of the United Nations recognized the right to nationalize or

transfer to state the control and execution of mining projects. The Declaration and Action Program for the Establishment of a New International Economic Order stipulated the following:

In order to safeguard its resources, each state is entitled to exercise effective control over them and their exploitation with means suitable to its own situation, including the right to nationalize or transfer ownership of such resources to its nationals, this right being an expression of the full permanent sovereignty of the state. No state may be subjected to economic, political, or any type of coercion to prevent the free full exercise of this inalienable right.

In 1974, Nigeria took a 55% share of all concessions. Kuwait, Qatar, the United Arab Emirates, and Saudi Arabia fixed their shares at 60%. The Saudis also increased the tax rate to 85% and fixed royalties at 20%. In 1975, Venezuela nationalized and Kuwait acquired all concessions, with the oil companies receiving compensation on the value of the assets and a long-term trading agreement with certain advantages in exchange for technological and transportation services. This formula served as an example to other countries.

2.2 From Nationalism to the Application of the Principle of Sovereignty

The “overturn of OPEC” encouraged nationalism in the Third World. The world’s perception of ever more scarce and more expensive oil played in its favor. The rhythm and intensity with which each country imitated the organization depended on political will; the technical, human, and financial capacity of the public sector to develop oil industry activities; the geological potential; the level of production and importance of demand; the weight of the oil bill; external indebtedness; and the shortage of currency. These factors determined the goals and margin for maneuver of the state when faced by the multinationals.

Nationalist efforts were focused in several directions: the recovery of ownership rights, state administration of activities considered as strategic, the creation or strengthening of national companies, the substitution of concessions for the system of contracts, and tightening up on taxes imposed on companies with regard to access, operation, and distribution of benefits.

2.2.1 Recovery of Ownership Rights

Many countries substituted the royal prerogative system for the ownership system. In the first place, the mineral resources, before being discovered, belonged to nobody. The state used its royal

prerogative power, formerly attributable to the king, to concede the right of use, determine the general conditions of the search and exploitation of the oilfields, and designate the right to mine or extract to the company of its choice. In the second place, under the ownership system, the oilfields are the property of the state, which may entrust their exploitation to a state-owned monopoly or grant a contract to a third party, establishing the most convenient terms and conditions on a case-by-case basis. If the resource is already being exploited, the state has the right to nationalize it. The third possibility, the occupation system, postulates that the owner of the soil or the surface is also the owner of the subsoil. This did not echo among the developing countries as being incompatible with history, politics, and the ideology dominant at that time.

2.2.2 Direct Control of Strategic Activities

Certain nations imposed state exclusivity on exploration, development, production, transportation, refining, processing, distribution, marketing, exportation, and importation of crude oil, natural gas, gasoline, and petrochemical products. Others assigned exclusive rights to exploration and production but left the other segments open. Still others limited exclusivity to crude oil imports. The remainder of the countries decided to promote competition in all of the links in the chain.

2.2.3 Incorporation or Strengthening of National Companies

One of the nationalist goals most appreciated by the producing countries was the state’s direct intervention in oil production activities in two ways: (1) as an investor (through a share in the consortia or concessions) and (2) as the direct executioner of the projects. In this manner, the state became not only owner but also entrepreneur. That role led to the creation or strengthening of state-owned companies. These should be capable of taking control of nationalized assets, executing the projects, maintaining the rate of production, and (if necessary) negotiating agreements with the multinational companies to obtain capital, technology, and know-how. During the 1970s, 38 state-owned companies were incorporated and 19 already existing companies were strengthened.

2.2.4 Substitution of the Concession System for the Contract System

Abandoning the concept of concessions had become a question of principle. As a result, production sharing contracts and risk service contracts flourished

in their many forms. However, the title of “concession” did not disappear. It survived, adapting itself to the new circumstances, normally incorporating the state ownership of the oil and an obligatory share of the national company under association agreements. The concession system continued to be used by countries lacking known oil potential that had small, little explored territories and pragmatic governments. Countries that had interesting geological potential or that were already producers were inclined toward production sharing or association contracts. Nationalist countries, with experienced state-owned oil companies and great geological potential, were inclined toward risk service contracts. The companies quickly adapted to the new system because it did not exclude the possibility of limiting risks and making significant profits. Together with the World Bank, they promoted rate of return-based profit-sharing contracts.

2.2.5 Toughening the Conditions of Access, Operation, and Benefit Distribution

Nationalism meant searching for the maximum economic or industrial benefit that a country could obtain from the relationship with the companies. To achieve rapid, efficient, and complete prospecting of the sedimentary basins, as well as optimum extraction of hydrocarbons, the terms of the contracts were reduced, the investment obligations were increased, programmed periods were contracted, delivery of the contracted zones was accelerated, surface rents were increased, and the control and follow-up processes were improved.

To increase the share in the benefits, the tax burden was increased through greater premiums per contract signed and for commercial discovery (bonus), increases in royalties and taxes on benefits, reductions in cost oil (the part of production destined to recovering costs) and profit oil (the part that produces income for the companies), increases in amortization periods, greater state participation in the projects, and the disappearance in the depletion allowance for the reconstitution of oilfields. The tax collection mechanisms multiplied and were perfected with an emphasis on efficient collection of taxes such as the rent-skimming tax, cost stop, price cap, and windfall profit tax.

Rights, taxes, customs tariffs, obligations for national content, limitations for repatriating benefits abroad, commitments to create infrastructure, the use of local goods and services, the hiring of local labor, and the carrying out of social projects were established to prevent the flight of oil profits from national territory. With the intention of dominating

the local oil industry, clauses were imposed relating to the transfer of technology as well as to training the national companies' employees. Finally, measures were demanded to protect the natural environment and the productive and social activities that may be affected by oil projects.

The most notable toughening up was seen in the large and medium oil-exporting countries that remained unnationalized, such as Egypt, Oman, and Qatar, and in the smaller producers that strongly embraced nationalism, such as Ivory Coast, Zaire, and Sri Lanka. Some countries, such as Argentina, India, and Turkey, were more pragmatic, following the trend but without frightening away the oil companies. Following the 1979 oil glut, a second wave of reclamations was observed, driven by the non-OPEC countries of Malaysia, Colombia, and Egypt as well as by other member countries, such as Ecuador, Indonesia, and Gabon, that previously had shown a degree of moderation.

3. PRAGMATISM SUBSTITUTES NATIONALISM: 1981–2003

The toughening of the conditions of access, operation, and distribution of benefits would not have been possible without favorable circumstances. The factors of the oil industry itself were particularly important. The dramatic growth in demand, the substantial increase in oil prices, and the general perception of scarcity caused savage competition among the oil companies to obtain mining rights. The Rome Club's catastrophic projections reflected the climate of that period. The perspective of ever more expensive and scarce oil forced companies to accept less advantageous conditions on the “invest now or go away” principle. Such circumstances did not last long. The prevailing conditions following the second oil glut were very different from those that had existed during the 1970s. The producing countries ceased to enjoy the correlation of forces in their favor due to changes both in the oil industry and in the economic and global political contexts that serve as a backdrop. Under those circumstances, thriving competitions developed among producing countries to attract foreign investment.

3.1 Nationalism Weakened by Lower Capital Availability

The industry structure changed as a result of the nationalist movements in the large exporting

countries. Vertical integration was broken, new markets were created, and new producers that did not join OPEC emerged. When the companies lost control of world reserves, the quantities produced, and the fixing of prices, relationships with the producing countries became highly dependent on the peaks and troughs of the market. From that point onward, the availability of risk capital for exploration and extraction projects depended directly on oil price levels, which were difficult to predict. In a tense market with perspectives of scarcity, the owners of the reserves had more leeway to impose their conditions. In a surplus market, the foreign operators could impose theirs.

Oil price quotations declined during the 1980s. The reduced budgets of the multinationals were not directed at developing countries as a priority but instead were directed at politically safe areas or at stock exchanges to absorb firms with proven reserves. Therefore, the companies began to pressure the producing countries for more favorable conditions. The World Bank and other credit institutions turned the thumbscrews by ceasing to finance public investment in oil. This situation changed little during the 1990s. Price increases during 1995–1996 and 1999–2002 were translated into greater availability of funds, but these were directed toward regions that were politically safe but had potential (e.g., the North Sea), at large producers that opened up their territories (e.g., Russia, Venezuela), at large producers that lowered their pretensions (e.g., Nigeria, Angola, Oman, Egypt), and at the frontier zones in industrialized countries (e.g., the Gulf of Mexico).

3.2 Economic Crisis, Neoliberal Paradigm, and a Unipolar World Close in on Nationalism

Changes in the general frame of reference also contributed to weakening the position of the producing countries. The solution to the delicate economic situation of the 1980s known as the “debt crisis,” characterized by inflation, currency depreciation, and high levels of debt, was sought through structural adjustment programs agreed to with international financial organizations that invariably included measures to enable foreign investment. Credit restrictions of debt service obligations reduced the availability of public funds. This reduction directly affected the national oil companies, especially when such limitations combined with inadequate prices, poor efficiency, and the diversion of

funds for purposes other than productive reinvestment to reduce their response capacity to face an ever growing demand due to industrialization processes, the gradual elimination of social backwardness, and population growth. And although the economic urgency was slowly disappearing, neoliberal politics continued to be applied under the pressure of the so-called “Washington consensus.”

The frame of reference changed completely during the 1990s and the first few years of the new millennium. The North–South confrontation, which until then had formed the backdrop of the relationships between the producers and the companies, changed with the end of the cold war. With the collapse of the Soviet Union and the socialist regimes of Eastern Europe, the political, technical, economic, and human support that had been offered to several developing countries disappeared, leaving them with no other option but to turn to the multinationals.

On confirming U.S. dominance, pressure increased on those countries that had nationalized their oil industries. As a result of the Gulf War in 1990, Iraqi oil exports were administered under the UN “oil-for-food” program. Since 1996, Iran and Libya have been subject to U.S. economic sanctions that particularly affect the oil sector. Hugo Chavez’s government, trying to regain governmental control over *Petroleos de Venezuela* (PDVSA) and restrain the opening up of foreign investment in Venezuela, has been subject to destabilizing pressures since 2001. Also, following the military intervention in Iraq in March 2003, Iraqi oil is under U.S. guardianship.

In parallel, the generalization of the neoliberal paradigm accelerated the opening of economies and the liberalization of activities as well as the flow of investment. The outline of the state decreased notably. The energy sector, one of final defenses of the state-entrepreneur, did not escape the change. The restructure included the opening of reserved areas, the segmentation of activities, the introduction of competition, and the sale of state-owned companies. The effervescence centered in the natural gas and electricity industries but also reached the oil industry. In a kind of historical revenge, the multinational companies and their governments took advantage of the new circumstances to reclaim positions and privileges that they had enjoyed in the past.

All countries ceded, although with varying scope and speed. The need to attract foreign capital was less urgent for those that possessed important oil potential and national companies with balanced trade and economies. For most, the adverse conditions limited their negotiating power. That notable

asymmetry had repercussions in OPEC. Ecuador left the organization in 1992, and Gabon left in 1995. The defense of sovereignty over hydrocarbons had united OPEC, but once ownership rights had been recovered, it was divided by pricing strategies.

3.3 Adjustments to Attract Risk Capital

The politics of openness are best understood by classifying countries based on their oil industry experience, geological potential, and economic health. In that sense, three main groups may be identified:

- *Countries that were completely dependent on external contributions of capital, experience, and technology.* Their negotiating power with the multinationals was very weak. They were highly sensitive in the economic context and to the international oil market situation. They had a modest or unknown oil potential, the production levels of some were low, and their economies generally were poorly developed. Countries in this category included New Guinea, Congo, Guatemala, The Philippines, and Paraguay.

- *Countries that partially or fully controlled exploration and production activities but did not achieve full control of their industry due to, among other factors, budget restrictions that weighed heavily on the national oil companies.* To maintain productions rates, they entered into association agreements and shared production agreements or risk service contracts. This category included medium-sized exporters such as Indonesia, Oman, Nigeria, Qatar, Ecuador, and Egypt as well as importers with solid state-owned companies such as Argentina, Brazil, and India.

- *Nationalized countries that enjoyed a generous geology and that had taken control of all sectors of the oil business.* They were able to undertake vast exploration campaigns, managing to obtain capital, know-how, and technology to carry out such projects. Their experience in commercial circles allowed them to optimize production in advantageous conditions and allowed them to undertake an internationalization process. In the beginning, their relationship with the companies was limited to service contracts, but later they agreed to negotiate agreements regarding access to production. This was the case with Iran, Iraq, Venezuela, Argelia, Kuwait, Saudi Arabia, and Mexico.

With regard to sovereign management of resources, the dynamics in which the countries in the

first two groups evolved were similar; with some exceptions, they had always turned to the companies. Conversely, the countries in the third group, having nationalized and started an internationalization process, evolved in a different dynamic.

3.4 A Step Back for the Small and Medium Producers

The countries in the first group, small or medium producers or those with no production, introduced faster and more vigorous adaptations to attract the oil companies. For these, the principle of sovereignty was limited to ideological debate. They quickly concluded that encouraging exploration to prove the potential existence of hydrocarbons was the main objective in controlling the industry or maximizing the taxation of the eventual production of oil or natural gas.

On the other hand and with some exceptions, the countries in the second group reacted relatively slowly and moderately. The problems arising from the change of context did not seem to make a mark in their nationalist conviction. However, to the extent that the exploration budgets were reduced and the economic reforms deepened, the adjustments in oil contracting became ever more important.

Despite their differences, the small and medium producers were faced with a policy of adjustment whose content and essential components were comparable. They adopted the following measures (among others): assignment of a better cost oil and profit oil for the companies, reduction or elimination of the state's participation in projects, reductions in taxes, suppression of oil price controls and currency exchange, reduction or elimination of royalties, signing of seismic option contracts, elimination of local personnel training, and relaxation of the restrictions related to burning natural gas.

There were some countries that did not want to relax their nationalism and others that did just the opposite. They did not take just one step backward; they took many. For example, they accepted waiving the renegotiation of contracts in case of oil price increases, applying the same advantageous conditions that other companies obtained in negotiations with states, guaranteeing a substantial benefit to foreign companies, and exonerating the operators from the payment of taxes, the elimination of royalties, and so on.

In summary, the small and medium producers consented to sacrificing economic profit and desisted

from the objective of dominating the oil industry. However, they did not waive ownership rights, an eminently political claim.

3.5 Adjustments in the Large Producers

Because the countries in the third group had to wait, the large producers were the last to call the oil companies again. But they did so not to occupy the place they had before nationalization; instead, they did so to participate in the specific activities that the national companies had developed with difficulty due to lack of capital, technology, and experience. The important role that these factors play is seen with greater clarity during times of crisis—when oil prices collapse, debts increase, fiscal resources become stunted, and state company exploration and production budgets collapse. The insufficiency of these key factors explains the advance of pragmatism to the detriment of nationalism.

The first to call the oil companies again, in 1986, was Argelia, a country overwhelmed by huge foreign debt. Its government even proposed the sale of some of the largest oilfields. It was followed by Iran and Iraq. Faced with enormous needs to rebuild their economies after the war, they opened negotiations in 1989 to rebuild and expand production capacity with the help of the multinationals. In the case of Iraq, the granting of sharing production contracts to non-U.S. firms was not far removed from the strategy of obtaining powerful allies such as France and China, a strategy that pressured for the end of sanctions that the UN Security Council had imposed on Iraq following the Gulf War. Under the tutelage of the United States, the Iraqi oil industry would be privatized and the oilfields would be granted in concession.

In 1990, Venezuela began an opening up process through operating service agreements to reactivate inactive or abandoned fields as well as association agreements in liquid natural gas (LGN), heavy-duty oils, and orimulsion. In 1995, profit-sharing agreements were introduced. In 1999, the natural gas industry was opened. And in 2001, a 51% share was established for PDVSA in all exploration and production agreements.

After the Gulf War, Kuwait called on the oil companies temporarily to rebuild its extraction installations. Government and Congress are face to face regarding the possibility of newly granting mining rights to the private sector. Saudi Arabia opened negotiations in 2000 to allow the multinationals to exploit nonassociated natural gas through

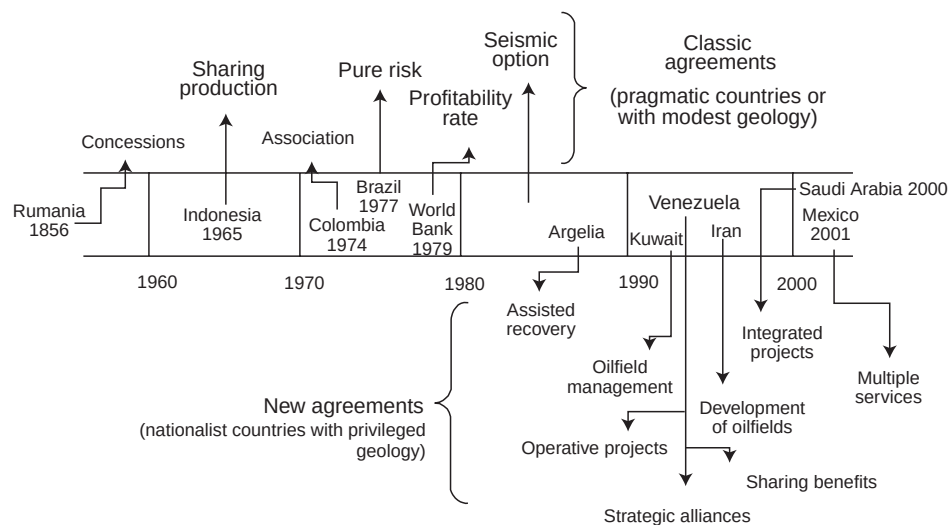
agreements, including its use in the generation of electricity and the desalting of water. Finally, the Mexican government in 2001 announced its intention to enter into service agreements to share the risks and benefits of nonassociated gas and (later) oil, despite Congress's opposition.

What has been the nature of the changes in terms of the principle of sovereignty over natural resources? In the first place, no country has resigned ownership of oil and natural gas. In the second place, and leaving to one side Argelia and Venezuela, countries have not changed their oil legislation to grant mining deeds to the oil companies. What they have done is reinterpret the legal framework in such a way that the state companies may enter into contracts so that the international oil companies contribute capital, technology, and management capability, at the same time assuming part of the risk. As a consequence, new oil agreements have emerged that are unlike the concessions or contracts of the 1960s and 1970s (Fig. 1). Governments have tried to find politically acceptable ways of permitting the local oil industry to participate again, making a few visible institutional adjustments so as not to awake nationalist opposition, although not always successfully (as in the cases of Kuwait and Mexico).

3.6 The Opening of Economically Planned Countries

In the framework of Perestroika, a set of reforms aimed at modernizing a country internally and bringing it closer to the West as foreign policy, the Soviet Union began in 1985 by signing a series of technical assistance and technology transfer agreements, culminating in 1990 with the signing of exploration agreements with the Elf, Chevron, and Total companies. Since the fall of the Soviet Union in 1991, the 11 republics have been conceding various types of mining rights.

In the Eastern European countries, the return to a market economy and the changes in the Soviet oil industry, above all concerning the oil supply and the technical cooperation they received, arose mainly from the opening up of their territories to foreign capital. The first country to invite bids for the granting of exploration and production permits was Bulgaria in 1989. In mid-1991, all of the other countries (Lithuania, Estonia, Poland, Czechoslovakia, Hungary, Rumania, and Albania) had already signed agreements with the oil companies or were in negotiations to do so.



Other agreements: Technical assistance, transfer of technology, service contracts, joint ventures, etc.

FIGURE 1 Diversification of exploration and production agreements.

In the developing countries with planned economies, the opening up began before the fall of the Berlin Wall, such as in the cases of China (1979) and Vietnam (1986). Although the call to the Western companies is recorded in the framework of a broader economic opening, in both cases the need for capital and technology to develop offshore oilfields was observed. In the case of Cuba, the opening up was a logical consequence of the cooling off of relations with Moscow. In 1990, the Cuban authorities decided to turn to the European companies in the search for marine oilfields. Moreover, as a test of autonomy, in 1989, Mongolia created a state-owned company to negotiate association agreements with Western companies. Finally, Laos and Cambodia, which were practically unexplored, signed their first contracts in 1991.

3.7 Nationalism Weak but Not Extinct

The nationalist movements allowed the correction of a very unfavorable situation for underdeveloped countries. However, the international context changed rapidly, and the majority of countries did not have time to provide themselves with the means necessary to take economic control of their economic resources.

The following important conclusions may be drawn from the exercise of sovereignty over natural resources over the past three decades or so by developing countries:

- The management of the local oil industry, in terms of sovereignty, is a valid objective. Past experience has not demonstrated that such a thesis is erroneous.

- Although the state is the owner of the subsoil's resources, this does not guarantee economic control over hydrocarbons. The state must be able to directly execute such projects.

- The state directly manages the oil industry, but this does not automatically guarantee said economic control. Certainly, it is indispensable to have an operational sovereign instrument over hydrocarbons (i.e., to have a state-owned oil company), but that is not enough.

- The design of contracts and efficient fiscal mechanisms is a necessary but not sufficient condition to guarantee the correct exploitation of oilfields, the recovery of the greater parts of the profits, and the transfer of technology and know-how.

To achieve the economic control of natural resources required by developing countries, besides the aforementioned ownership rights, a state-owned company and an efficient taxation system are required to reach a high level of economic, social, political, and institutional development.

Also, such economic control is reached only when the product is delivered to the end user. Therefore, some countries, such as Kuwait and Iran, have allowed the state companies to invest and operate in the consuming countries in both the oil and energy sectors. This is an impossible task if the state

companies do not have sufficient room for maneuvering on the operative, administrative, financial, and strategic fronts. Paradoxically, state control over the oil business is refined with autonomy and internationalization such as in the case of Venezuela.

Nowadays, the balance of power is different from what it was during the 1970s. Total sovereignty over hydrocarbons, as it was conceived in the past, seems difficult to reach. Approximately two decades ago, steps were made in the opposite direction: transfer of assets to the private sector and downsizing or disappearance of state companies. This does not mean that nationalism is dead.

The persistence of a market dominated worldwide by purchasers, neoliberal economic policies, the conditions imposed by international financial organizations, and government pressure that supports and promotes the large oil companies are some of the factors that limit sovereignty over hydrocarbons. Military aggression only annihilates it, such as in the case of Iraq. In those cases, the existence of nationalism is ensured. It will exist insofar as the state's right to exercise permanent sovereignty over the nation's natural resources is not guaranteed or is threatened.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Economic Geography of Energy • Environmental Injustices of Energy Facilities • Geopolitics of Energy • National Security and Energy • Oil Industry, History of • OPEC, History of • War and Energy

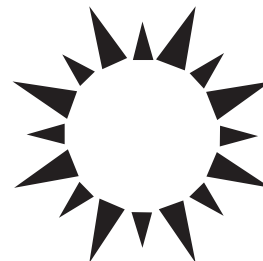
Further Reading

- Aïssaoui, A. (2001). "Algeria: The Political Economy of Oil and Gas." Oxford University Press, Oxford, UK.
- Alnasrawi, A. (1991). "Arab Nationalism, Oil, and the Political Economy of Dependency." Greenwood, New York.
- Angelier, J. P. (1976). "La rente pétrolière." *Energie et Société*, Paris.
- Elm, M. (1992). "Oil, Power, and Principle: Iran's Oil Nationalization and Its Aftermath." Syracuse University Press, Syracuse, NY.
- Ferrier, R. (1994). "The History of the British Petroleum Company I and II." Cambridge University Press, Cambridge, UK.
- Ghadar, F. (1984). "The Petroleum Industry in Oil-Importing Countries." Lexington Books, Lexington, MA.
- Linde, C. (2000). "The State and the International Oil Market: Competition and the Changing Ownership of Crude Oil Assets." Kluwer Academic, Boston.
- Mommer, B. (2002). "Global Oil and the Nation-State." Oxford University Press, Oxford, UK.
- Penrose, E. (1968). "The Large International Firms in Developing Countries: The International Petroleum Industry." Allen & Unwin, London.
- Philip, G. (1982). "Oil and Politics in Latin America: Nationalist Movements and State Companies." Cambridge University Press, Cambridge, UK.
- Sampson, A. (1976). "Les Sept sœurs: Les grandes compagnies pétrolières et le monde qu'elles ont créé." A. Moreau, Paris.
- Tanzer, M. (1969). "The Political Economy of the International Oil and the Underdeveloped Countries." Beacon, Boston.
- Taverne, B. (1999). "Petroleum, Industry, and Governments: An Introduction to Petroleum Regulation, Economics, and Government Policies." Kluwer Law International, The Hague, Netherlands.
- Terzian, P. (1983). "L'étonnante histoire de l'OPEP." Editions Jeune-Africa, Paris.
- Vandewalle, D. (1998). "Libya since Independence: Oil and State-Building." Cornell University Press, Ithaca, NY.
- Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money, and Power." Simon & Schuster, New York.

National Security and Energy

WILFRID L. KOHL

Johns Hopkins University
Washington, DC, United States



-
1. Introduction
 2. Oil: Background
 3. Dimensions of Oil Security
 4. Reliable Electricity Supply and Price Volatility
 5. Natural Gas and Price Volatility
 6. Terrorism and Energy Security

Glossary

Achnacarry agreement Second oligoplistic accord reached by major international oil companies at Achnacarry castle in Scotland to control downstream marketing of oil and divide it according to market shares existing in 1928.

futures market An organized market, such as the New York Mercantile Exchange, in which contracts are bought and sold to deliver a specified quantity of oil (or other commodities) at a specified future date at a price to be paid at the time of delivery. Futures contracts allow the holder to hedge risk.

Hubbert curve U.S. petroleum geologist King Hubbert predicted (correctly) in 1956 that U.S. oil production followed a bell-shaped curve and would peak in approximately 1970.

International Energy Agency Established in 1974 within the Organization for Economic Cooperation and Development to promote energy cooperation among consumer countries including oil crisis management; currently has 26 member countries.

Kyoto Protocol International agreement signed in 1997 that sets binding emissions reductions of greenhouse gases with an average 5.2% reduction below 1990 levels for industrial countries. As of 2002, the protocol had not yet entered into force. The European Union and Japan have ratified it, but the United States has declined to participate.

Organization of Petroleum Exporting Countries (OPEC) - Founded in 1960 in Baghdad; members are Algeria, Indonesia, Iran, Iraq, Kuwait, Libya, Nigeria, Qatar, Saudi Arabia, United Arab Emirates, and Venezuela.

Red Line Agreement Agreement in 1928 on oil production in most of the former Ottoman Empire by American

and European major international oil companies; established structure of the Iraq Petroleum Company and was an important foundation of the international oil company cartel.

Seven Sisters A cartel of seven international oil companies that dominated the world oil market after 1928 until the rise of OPEC; five of the companies were American—Standard Oil of New Jersey (Exxon), Socony Vacuum (Mobil), Standard of California (Chevron), Texaco, and Gulf—two were European—Anglo-Iranian Oil, which became British Petroleum, and Royal Dutch/Shell.

Strategic Petroleum Reserve U.S. government oil reserve established under the Energy Policy and Conservation Act of 1975 to be used to mitigate effects of oil market disruptions; contained approximately 650 million barrels as of 2003 located in four underground salt caverns along the Texas and Louisiana Gulf Coast.

Texas Railroad Commission A state agency that in 1932 acquired authority to regulate oil production via a system of market demand prorationing within the state.

Traditionally, energy policy has sought security of supply, affordability, and limited impact on the environment. Until recently, energy security has been dominated by oil security, since oil has been the leading fuel and is subject to the influence of the OPEC cartel and the geopolitics of the world oil market. This article reviews the history of oil, how it became a strategic commodity, and the importance of oil today. It analyzes the multiple dimensions of oil security: the long-term outlook for oil supply, OPEC market power, oil and the economy and the costs of oil market disruptions, the role of oil imports and the balance of payments, oil crisis management, oil and environmental security, and the link between oil, foreign policy, and geopolitics. The article concludes by considering other energy security concerns: shortages and price volatility in electricity and natural gas, and potential threats to energy facilities by terrorist attacks.

1. INTRODUCTION

Energy policy is a subset of economic policy, foreign policy, and national and international security policy. Traditionally, energy policy has sought security of supply, affordability, and limited impact on the environment. According to the Bush administration's "National Energy Policy," energy should be "reliable, affordable, and environmentally sound." Until recently, energy policy was dominated by oil policy since oil has been the leading fuel in the United States and most industrial economies, and it has been subject to price volatility and the vagaries of the world oil market that is greatly influenced by the Organization of Petroleum Exporting Countries (OPEC) cartel. Most of this article therefore reviews the elements of oil security. It begins with a brief historical review of the oil market and how oil developed as a strategic commodity with an eye to its military uses. After World War II the focus shifted to its rapidly expanding role in the civilian economy and how to deal with oil supply interruptions and their macroeconomic effects as well as possible resource constraints.

Other types of energy security concerns have surfaced, including shortages and the effects on consumers of high electricity and natural gas prices as a result of problems with deregulation. In the future, there could be potential threats to energy facilities by terrorists in the wake of the September 11, 2001, attacks. These aspects are discussed toward the end of the article.

2. OIL: BACKGROUND

2.1 History

Early concerns about oil and security date back to World War I. In 1913, Winston Churchill, then First Lord of the Admiralty, decided to convert the British Navy from using coal to bunker oil, which was a cleaner and more efficient fuel. To ensure adequate oil supply, the British government took a majority share in the Anglo-Persian Oil Company (later renamed British Petroleum). This was one of the first major government interventions in the oil industry. Other navies soon followed the British example. The appearance of tanks and other motorized military vehicles highlighted oil's important new role as a strategic commodity. During World War I, allied nations established government controls over oil

supply. Just after the war, the U.S. government established three Naval Petroleum Reserves.

Following World War I, as oil became more important in the civilian economy in transportation and industry, the U.S. Geological Survey predicted that the United States was about to run out of oil because of insufficient domestic reserves. The U.S. government in turn supported the efforts of American oil companies to gain access to concessions in the Middle East. In 1928, the Red Line and Achnacarry agreements provided mechanisms for coordination of supply and markets by a secret cartel of seven American and European international oil companies (the seven sisters), which dominated the international oil market until well after World War II. Following the discovery in 1930 of large oil fields in east Texas, which produced an oil surplus, state regulation of production began in 1933 under the mantra of conservation by the Texas Railroad Commission to help stabilize oil prices and the health of the domestic industry. (The Texas Commission later provided a model for OPEC.)

World War II demonstrated the growing importance of oil in modern warfare to fuel fighting ships, freighters, submarines, tanks, troop transports, and airplanes. Access to oil was a major reason behind Germany's invasion of the Soviet Union and Eastern Europe and Japan's advances on the Dutch East Indies and Southeast Asia, including the Japanese attack at Pearl Harbor. (The United States had placed a *de facto* oil embargo on Japan. U.S. naval forces could have interfered with Japanese efforts to secure oil supplies in Southeast Asia.) Meanwhile, the future importance of the Middle East, where major new oil fields had been discovered in the 1930s, was clearly recognized. In 1943, President Roosevelt announced that the defense of Saudi Arabia was vital to the United States and extended Lend-Lease aid. After the war, the U.S. government supported the efforts of four American oil companies (Exxon, Mobil, Texaco, and Chevron) to form a consortium to develop Saudi oil, the Arabian-American Oil Company (Aramco).

The first postwar oil crisis occurred in 1951 in Iran when a nationalist leader, Mohammed Mossadegh, seized power and decided to nationalize the Anglo-Iranian Oil Company (BP). Because of the strategic importance of Iran, the CIA helped to stage a counter coup that ousted Mossadegh 2 years later and restored the Shah. At the same time, the U.S. government took the lead in organizing a new Iranian oil consortium, which for the first time included participation by American oil companies.

During the Suez crisis in 1956 (the second postwar oil crisis), when Egypt's President Nasser nationalized the Suez Canal and Israel invaded the Sinai and began moving toward the Canal, Britain and France intervened and sent troops. Concerned about escalation of the crisis and Soviet involvement, the United States offered to provide oil to its European allies if they would agree to a cease-fire and withdraw, which they did. Excess U.S. oil production capacity enabled this diplomatic action. By the early 1970s, this excess capacity had disappeared.

After the postwar recovery, a boom period of U.S. and global economic growth stimulated increasing demand for crude oil and petroleum products. In the United States, this was driven by a rapid expansion of highway transportation and growing environmental concerns over coal-fired power plants. However, new oil supplies were also becoming available, especially from the Middle East. OPEC was founded in 1960 to put pressure on the international major oil companies not to reduce prices further in a period of increased competition in the oil market from independents and more than adequate supply. However, by the early 1970s, demand increases and rising U.S., European, and Japanese imports led to a much tighter market. This set the stage for stronger OPEC actions in the oil crises of the 1970s.

In October 1973, when the United States supported Israel in the Yom Kippur war, Arab nations responded with an oil embargo and production cutbacks. OPEC quadrupled oil prices, raising concerns about the security of oil supply to the Organization for Economic Cooperation and Development (OECD) nations and the long-term adequacy of oil resources. In response, the United States created a Strategic Petroleum Reserve (SPR) and took the lead in establishing the International Energy Agency (IEA) associated with the OECD. The principal purpose of the IEA was to establish an emergency oil-sharing system among member countries to assist in managing future supply interruptions.

The Iranian revolution caused an oil supply shortfall in 1979 and produced another oil crisis as oil prices doubled, led by the spot market and adopted by OPEC. The outbreak of the Iran–Iraq war in 1980 exerted further upward pressure on oil prices, which reached a high of \$34/barrel for the Saudi benchmark crude in 1981. This action turned out to have negative consequences for OPEC because the high oil prices in the first half of the 1980s encouraged fuel switching, the development of more efficient technologies, and a strong increase in non-OPEC oil production, which, along with decreasing demand due to recession,

reduced OPEC's market share from more than 30 million barrels/day (mbd) in the 1970s to as low as 16 mbd by the mid-1980s. Following Saudi Arabia's decision in 1985 to abandon its role as "swing producer," oil prices collapsed in 1986 to less than \$10/barrel, although prices returned to the \$15–20 range the next year after OPEC regrouped and reinstated production quotas.

By the time of the outbreak of the Persian Gulf War in 1990, the world oil market had become more competitive. Players included national oil companies, international majors, and larger numbers of independent oil companies. Oil was now traded like other commodities in forward and futures markets, providing a mechanism to hedge price risk. Market transparency was also stronger because of the expanding application of information technologies, which enabled faster adjustment to market trends. Also, several governments, led by the United States, had built up strategic oil stocks for use in emergencies.

The Persian Gulf crisis (1990–1991) demonstrated the traditional aspects of oil diplomacy and military force along with market-based mechanisms. After Iraq invaded Kuwait in October 1990, the United Nations (UN; urged by the United States) placed an embargo on oil shipments from Iraq, removing more than 4 mbd of crude from the world market. A U.S.-led UN expeditionary force was deployed to Saudi Arabia during the fall to protect Saudi Arabia and to prepare for military action to push Iraq out of Kuwait. Although some of the oil shortfall was made up by increased production from Saudi Arabia and other producers, oil prices escalated to nearly \$40/barrel in part because of uncertainty regarding what would happen next. Although other motives may have played a role in the U.S. action, clearly the threat posed by Saddam Hussein to the security of the world oil market, and to the world economy, was a paramount concern. When air strikes against Iraq commenced in January 1991, the United States, Germany, and Japan initiated a drawdown of strategic oil stocks that was coordinated by the IEA. As a consequence of both actions, oil prices declined to normal levels closer to \$20/barrel. During this oil crisis many buyers and sellers hedged oil transactions on the futures market.

At the end of the 1990s, another oil price collapse and then a price shock took place, which were much more the result of market forces. After a misjudgment of the oil market by OPEC at the end of 1997 when it expanded production as Asia was going into recession, oil prices plunged in 1998 and early 1999 to approximately \$10 barrel, causing serious damage

to the economies of oil-producing countries. OPEC rallied, and acting with greater cohesion beginning in March 1999, it lowered production, causing prices to rebound and then increase to more than \$30/barrel in 2000. In 2001, OPEC struggled to manage falling prices set off by a global recession made worse by the attacks of September 11, 2001, and the war on terrorism.

2.2 Importance of Oil Today

Although the importance of oil in the world economy has diminished somewhat since the 1970s, oil remains the largest source of primary energy in industrial countries at 40% (compared to 55% in 1980). The oil intensity [i.e., the amount of oil required to produce a unit of gross domestic product (GDP)] of industrial economies has fallen considerably due to fuel switching, structural change, and advances in energy efficiency. In 2000, oil imports represented only 4% of the total value of OECD imports, compared to 13% in 1981. However, oil is still very important, accounting for two-thirds of international energy trade. Oil reserves are less abundant than those of coal and natural gas and less evenly distributed. Most oil reserves are located in developing countries, which makes the oil market more amenable to cartel control than other commodity markets, hence the influence of OPEC.

For the United States, oil accounts for 39% of primary energy consumption, which in 2000 amounted to 19.5mbd of petroleum products (Fig. 1). Of this amount, approximately 10 mbd was derived from crude oil and product net imports. Most oil (approximately 13mbd) is consumed in the transportation sector. The United States accounts for approximately 25% of the world's oil consumption. The U.S. oil demand is expected to grow approximately 1.5% annually until 2020, led by

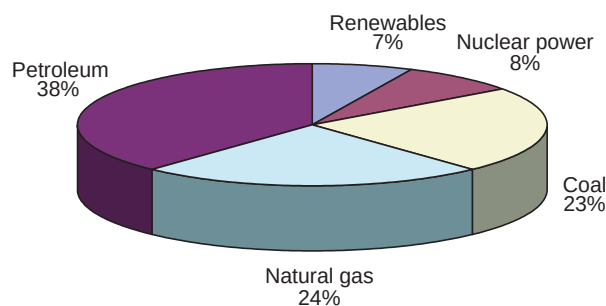


FIGURE 1 U.S. primary energy consumption, 2000. From the Energy Information Administration (2002).

growth in the transport sector. Net imports, which amounted to 53% of crude and product needs in 2000, are projected to increase to 62% in 2020. Imports come from a variety of countries, both OPEC and non-OPEC (Fig. 2). The U.S. domestic oil production was approximately 9.5mbd in 2000. The U.S. onshore production is scheduled to decline, but production in the Gulf of Mexico is expected to increase, which will probably lead to stable or slightly expanded production in the next few years.

In peacetime, the U.S. military uses less than 5% of national consumption of crude oil and petroleum products. In case of military conflict, this requirement would increase. However, the military requirement represents a relatively small percentage of domestic oil consumption, and it is dwarfed by the needs of the civilian economy. Most of the incremental military demand in wartime would likely come from overseas in one or more war zones, and most of the needed supplies would be purchased abroad. In case of a large-scale conflict involving the domestic homeland, the military has mechanisms in place to procure oil and other fuels for defense needs. This would involve invoking the Defense Production Act and the Energy and Security Act. The Department of Defense maintains petroleum stocks in the United States and abroad.

3. DIMENSIONS OF OIL SECURITY

3.1 Long-Term Outlook for Oil Supply

The oil shocks of the 1970s and their associated uncertainties led to major questioning about future oil resource scarcity, as highlighted by the publication

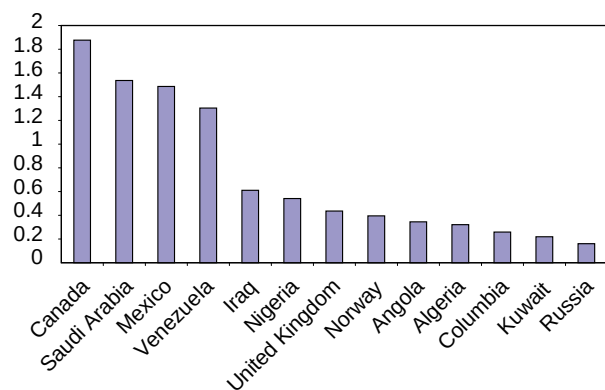


FIGURE 2 U.S. oil imports, source countries (million barrels/day, first 6 months of 2002). From *Monthly Energy Review*, Department of Energy, Washington, DC.

of the Club of Rome study, "The Limits to Growth." After the oil price collapse of 1986, this view was less prevalent. However, at the end of the 1990s debate had begun on the future of oil supplies, spurred by articles in prominent journals by geologists Colin Campbell and Jean Laherrere, who argued on the basis of applying the Hubbert curve to the world that conventional oil production would peak in the first decade after 2000 and set the stage for a coming oil crisis. (King Hubbert was a famous U.S. geologist who correctly predicted that U.S. oil production would peak in approximately 1970.) Their argument was based on the proposition that all large oil fields have already been found and that world reserve data are inaccurate, especially in the Middle East, where several OPEC member countries suddenly and mysteriously increased their reserve figures in the late 1980s. According to these pessimists, world recoverable oil reserves at the end of 1998 were estimated at 1800 billion barrels, and world production could peak by 2010 if not sooner. A similar perspective is presented by geologist Kenneth S. Deffeyes in his book, "Hubbert's Peak: The Impending World Oil Shortage."

Among the leading opponents of this view are economists M. A. Adelman and his associate Michael Lynch, who pointed out serious limitations of the Hubbert curve and emphasized the role of investment and new technology in expanding reserves. Oil reserves are better viewed as inventory, which is replenished by investment. Depletion is constantly delayed by new knowledge and advances in production technology. Oil supply forecasts have tended to be dominated by a pessimistic bias.

In 2000, the authoritative U.S. Geological Survey (USGS) published its latest estimates of world oil and gas reserves outside the United States that have the potential to be added during the period 1995–2025. The estimated volumes of undiscovered conventional oil are 20% greater than the 1994 estimate. The potential addition to reserves from reserve growth (e.g., due to applications of new technology) is also very large. When the new mean global estimates are combined with previous estimates for the United States, the USGS contends that worldwide ultimately recoverable reserves (URRs) of conventional oil total 3.021 trillion barrels, and natural gas liquids (frequently added to oil estimates) total an additional 324 billion barrels. URRs include cumulative production to date, identified remaining reserves, undiscovered recoverable resources, and estimates of "reserve growth" in existing fields. The 1994

USGS estimate listed URRs of conventional oil at approximately 2.3 trillion barrels.

The new USGS estimates have been adopted by two organizations that regularly publish widely respected energy market forecasts. Both the IEA in its 2001 "World Energy Outlook" and the U.S. Energy Information Administration (EIA) in its "International Energy Outlook 2002" take an optimistic view of future world oil supply and conclude that proven oil reserves are adequate to meet demand until 2020, with a world production peak to occur sometime thereafter. In addition, the world has very large reserves of unconventional oil, including Venezuelan heavy oil and Canadian tar sands, which will become economic to produce at higher oil prices. In short, the world appears to have sufficient oil reserves for the foreseeable future.

An important factor in the increased estimates for oil reserves, especially for oil reserve growth, is the recent advances in oil production technology, which have improved success rates and lowered costs of finding oil and production. These advances include three- and four-dimensional seismic for locating and evaluating underground deposits; directional drilling; floating production systems; deep-water platforms; and the general widespread application of computers and information systems by oil companies and service contractors to improve analysis, management, and communications. Direct production costs for the international oil companies are estimated to average \$3–6/barrel worldwide.

3.2 OPEC Market Power

The world's conventional oil reserves are concentrated in a relatively few countries. Most of these countries are members of OPEC, established in 1960, which controls approximately two-thirds of global oil reserves and in 2002 slightly less than 40% of world oil production. OPEC has a formal organization, meets approximately four times per year, and attempts to act as a cartel to manage world oil supply and prices. Since 1982, it has set production quotas for its members. As an organization of 11 countries, not firms, OPEC has not always been successful at maintaining cohesion. Members have political as well as economic objectives and are known to cheat on their quotas. However, since the 1970s OPEC has been successful most of the time in using its market power generally to maintain the price of oil well above the costs of production (which are less than \$2/barrel in the Persian Gulf or approximately \$4/barrel including finding and development costs.)

It has also been successful at either causing or capitalizing on oil market disruptions to obtain very high prices. Basic to its success in this regard is the fact that oil demand and, to some extent, supply are inelastic and do not respond quickly to changes in price. There are few readily available substitutes for oil, especially in the transportation sector.

OPEC, however, has made mistakes. As previously noted, after locking in very high prices in 1981, which caused a recession and a precipitous decline in oil demand, along with an increase in non-OPEC production, OPEC lost market share and prices collapsed in 1986. However, OPEC regrouped and benefited from an oil price spike at the time of the Persian Gulf War in 1990, and Saudi Arabia increased its production following the UN embargo against Iraq. During the early and mid-1990s in a more competitive oil market, prices were more volatile and ranged from approximately \$18 to \$20/barrel. Following a bad decision by OPEC to increase production at the end of 1997, oil prices collapsed in 1998 and early 1999. However, OPEC—led by Saudi Arabia, the largest producer—reasserted cohesion in its March 1999 decision and succeeded in obtaining support from non-OPEC states Norway and Mexico, which resulted in lower production and a price recovery in 1999 and 2000—initially to more than \$30/barrel before settling back into the mid-\$20 range. Thus, OPEC has shown that it still has market power and is able to use it, even if it is not a perfect cartel. It has never been able to agree on an enforcement mechanism for its decisions. In economic parlance, OPEC is an imperfect monopolistic cartel of the von Stackelberg type. OPEC has a large enough market share to influence prices, but its influence is limited by the existence of other competitive suppliers.

Looking to the future, OPEC's market share is expected to grow, which will give it even more leverage on the world oil market. Both the EIA and the IEA predict that OPEC production will increase from 30 mbd in 2000 to approximately 60 mbd in 2020 to meet growing oil demand, which would represent a market share of more than 50%. Non-OPEC production is also projected to grow but less rapidly.

3.3 Oil and the Economy and the Cost of Oil Market Disruptions

Research has shown that sudden increases in oil prices are linked to inflation, increased unemployment, and higher interest rates. Eight of 10 post-World War II recessions in the United States

(including several recent ones) were preceded by oil price shocks (Fig. 3). A study by the Stanford Modeling Forum concluded that the first oil shock of 1973–1974 reduced U.S. GNP in 1975 by 3–5.5%, and the second shock in 1979–1980 reduced U.S. GNP in 1981 by 2–4%. An OECD study reported the real income losses to all OECD countries due to the second oil shock at approximately 5% in 1980 and 8% in 1981. However, there is asymmetry in the oil–economy relationship. Whereas rising oil prices tend to retard economic growth, declining oil prices or an oil price collapse (such as occurred in 1986 and again in 1998) do not necessarily stimulate growth, apparently because there are more economic adjustment costs and coordination problems associated with rising prices.

David Greene, Donald Jones, and Paul Leiby of the Oak Ridge National Laboratory analyzed the macroeconomic impact of oil price increases on the U.S. economy. They contend that there are three types of economic losses:

Loss of potential GDP: When the price of oil is increased by monopoly power, oil becomes more scarce. The economy is able to produce less output with the same resources of capital, labor, materials, and land.

Macroeconomic adjustment costs: When prices increase rapidly, there are additional transitional costs because wages and prices are not able to adjust rapidly enough to the higher oil prices to permit the economy to operate at full employment. There is also the possibility of adjustment required by changes in monetary policy.

Wealth transfers from U.S. oil consumers to foreign oil exporters: These transfers are equal to the quantity of U.S. oil imports multiplied by the difference between the monopoly price and the competitive price of oil. These transfers go to OPEC and non-OPEC producers and are a cost to the U.S. economy.

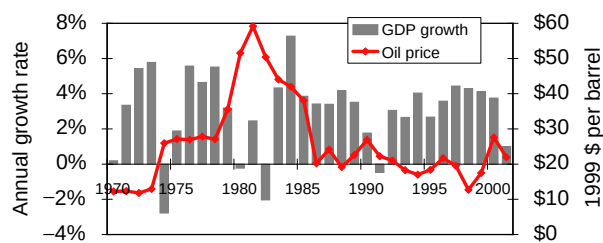


FIGURE 3 Oil price and economic growth, 1970–2001. From Greene and Tishchishyna (2000) and data updates, 2001.

In the most recent update of this analysis, David Greene and Nataliya Tishchishyna estimate the costs to the U.S. economy of oil dependence and oil monopoly power over 30 years at approximately \$7 trillion 1998 dollars, which as they point out is approximately equal to the payments on the national debt during the same period. (This estimate is controversial since it assumes a competitive oil price of approximately \$10/barrel. There are a range of views regarding what oil prices would be absent the OPEC cartel.) Whether or not this figure is correct, the costs of oil price shocks to the economy have clearly been substantial.

Greene and Tishchishyna define oil dependence as the product of a noncompetitive world oil market dominated by OPEC; high levels of U.S. oil imports; the importance of oil to the U.S. economy, where it is critical in the transportation sector; and the absence of readily available or affordable substitutes. The transportation sector is key because on-road vehicles account for two-thirds of U.S. oil consumption and there are no substitutes immediately available. In their view, the cost of oil dependence to the United States is a market failure because of the cartelization of the world oil market.

The United States can take some actions to reduce the costs of this dependence, including developing advanced technologies to increase efficiency in the transport sector, pursuing alternative energy technologies, and improving the technology of oil exploration and recovery to increase oil supply. Furthermore, the United States can diversify its sources of oil to more non-OPEC countries, a trend that is already under way.

It should be noted that the economic significance of oil consumption in the U.S. economy is much less today than it was in the past. In 1999–2000, the oil cost share of GDP was 1.4–2.0% of GDP. This compares with a 4–6% share in the early 1980s. Changing economic structure and improvements in efficiency of oil use have both contributed to this development. As discussed by Brown and Yücel, the relationship between oil prices and economic activity also weakened by the end of the 1990s.

3.4 Role of Oil Imports and the Balance of Payments

In the media and political debates, oil security is frequently equated with the level of U.S. oil imports, which has been rising. (Net U.S. oil imports were 55% of oil consumption in 2001.) The transfer of wealth to pay for oil imports is a factor, as noted

previously. However, equally if not more important is the level of overall oil consumption, which is related to the oil intensity of the U.S. economy.

In recent oil market disruptions, the increase in oil prices has been the most important factor. Any supply shortfalls have not lasted more than a few months. Since crude oil is traded on a world market, the amount of oil a country imports has little effect on the price of oil in the short term since that price is set on a world market. The vulnerability of the United States and other countries to oil shocks depends more on the level of oil consumption and the price of oil.

Arguments are often made that rising U.S. oil imports are harming the U.S. economy, and therefore national security, because of their contribution to the trade deficit. Although U.S. oil imports are increasing, they have amounted in recent years to 8–10% of the value of total U.S. merchandise imports. This is certainly acceptable for a complex economy the size of that of the United States. By comparison, several European countries and Japan import more oil (more than 90% of their oil needs) than does the United States, and this has not harmed their economies. (The situation can be different for a developing country with a simpler economy and reduced ability to pay for a large amount of oil imports.) The changing role of oil imports in the trade deficit is more a function of changes in the imports and exports of other commodities than it is in the role of oil imports. In short, the United States will need to get used to increasing levels of oil imports. Reducing import dependence would be very costly in the short to medium term. However, as noted previously and later, there are reasons to consider ways to reduce oil consumption overall in the future.

3.5 Oil Crisis Management

An international framework for oil crisis management is provided by the IEA, established in 1974 and located in Paris, where it is associated with the OECD. The 26 member countries are obligated to hold oil stocks equivalent to 90 days of imports. The heart of the agency is the formal oil-sharing mechanism, which is intended to mitigate future supply disruptions. The system is triggered by either a 7% or a 12% or greater supply shortfall in the daily rate of oil supplies to one or more member countries, whereupon a request can be directed to the secretariat that the oil-sharing mechanism be triggered. The 7% case would normally be dealt with by demand restraint/conservation measures, but the 12% plus case is more serious and might require

sharing of oil supplies. In any case, the governing board would need to be convened to decide on the appropriate response by a weighted majority vote. The system has never been triggered, although a request was considered in 1979.

With growing awareness that small market disruptions can yield sharp oil price spikes, IEA countries agreed in 1984 to a less formal system of international consultations aimed at release of government-controlled emergency stocks known as Coordinated Emergency Response Measures by member countries that possess such stocks (mainly the United States, Germany, and Japan, although other countries are considering the idea). This framework was first used in January 1991 for a coordinated stock release at the beginning of military action in the Persian Gulf War. In 1995, the governing board reached agreement that coordinated stock draw should be given priority in response to market disruptions regardless of size and before activation of the formal allocation mechanism.

The other purposes of the IEA are to promote the development of alternative energy sources and energy efficiency, to gather data and conduct studies on the world oil market and other energy markets, to encourage cooperation with nonmembers, and to assist in the integration of energy and environmental policies.

The European Union (EU) also has a requirement that member countries hold stocks of petroleum products covering at least 90 days of average daily consumption for the previous calendar year. Unlike the IEA, the EU lacks the necessary power to manage the stocks in a crisis. Member states own and control their own stocks. They are obliged to consult each other before releasing stocks from the strategic reserve. The EU is considering adding to its strategic oil reserves and managing them on a more centralized basis.

Stocks in IEA net importing countries have been declining during the past 15 years or so at the same time that OECD oil import dependence has been increasing. Meanwhile, the world oil market is growing, and much of the new oil demand will come from developing countries in Asia, which do not possess strategic oil stocks. Global oil security would be increased if Asian countries established minimum emergency stock requirements, as has been recommended by the Asia Pacific Energy Research Center.

The U.S. SPR was established in 1975 to help protect the nation against future oil market disruptions. Managed by the Department of Energy (DOE), the SPR oil was mainly acquired during the 1980s

and is stored in underground salt domes along the coast of the Gulf of Mexico. At the end of 2003, the SPR contained 650 million barrels of oil. President George W. Bush has instructed the secretary of energy to proceed with further filling of the reserve up to its 700 million barrel capacity using principally royalty oil from federal offshore leases.

The president determines whether there is a severe supply interruption that warrants use of the SPR. If so, the secretary of energy announces an auction sale of a certain amount of oil at a price determined by the market. Bidders (oil companies) send their offers to the DOE and, after selection, submit a letter of credit or a cash deposit. The oil is then delivered through a pipeline or a marine terminal. The whole process can take 3 or 4 weeks to complete. The amount of import coverage afforded by the SPR depends on the level and rate of increase in imports. With U.S. net imports of approximately 11 mbd at the end of 2003, the current SPR offers approximately 53 days of import coverage if all imports were cut off, which is a very unlikely case given that U.S. oil imports come from a diversified group of nations.

There has been controversy about how and when to use the SPR. In the first Bush administration, the president refused to order an SPR drawdown despite a sharp escalation in oil prices in the fall of 1990 after Iraq invaded Kuwait (although the DOE did proceed with a SPR test run.) However, in January 1991, a drawdown of 33.75 million barrels of oil was authorized at the beginning of Desert Storm in coordination with stock draw by other allies under the IEA (later reduced to 17.3 million barrels of oil actually sold). That release, coupled with the success of the air war, helped to reduce oil prices and restore stability to the market. In September 2000, President Clinton authorized release of oil to bolster supplies at a time of high oil prices and low inventories of heating oil in New England before the start of winter. This was actually an oil swap, not a sale, because companies accepted obligations to take oil and return slightly more 1 year later. During the 2000 election campaign, candidate G. W. Bush criticized this action and argued for retaining the SPR for truly emergency uses.

In Germany, the government works closely with industry to administer a government-owned reserve, the Erdölbevorratungsverband, which stocks both crude oil and products to meet the 90-day IEA import requirement. Most other European countries rely on mandates to industry to hold private emergency stocks. In Japan, the Ministry of International Trade and Industry supervises government and private stocks.

3.6 Oil, Transportation, and Environmental Security

In the United States, transportation accounts for 27% of overall energy consumption and 68% of petroleum use. The petroleum share of transport energy in 1999 was 96%. Most transportation energy use (76%) comes from highway vehicles—light vehicles and trucks—with the rest coming from aircraft, boats and ships, pipelines, rail, and off-highway construction and farm vehicles. For European countries, transportation is also the most important and fastest growing sector for oil use, whereas the situation in Asia is more diversified, with more oil being used in industry and power generation. Oil use in transportation contributes importantly to air pollution and to greenhouse gases that are responsible for global climate change.

Transportation in the United States is a source of every major pollutant except sulfur dioxide. Of the criteria pollutants listed in the U.S. Environmental Protection Agency's (EPA) National Ambient Air Quality Standards, the oil/transport contribution (1999) is as follows: carbon monoxide, 77%; nitrogen oxides (NO_x), 55.4%; volatile organic compounds (VOCs), 47%; and very small amounts of particulate matter (PM-10) and sulfur dioxide. (Carbon dioxide is not yet regulated.) Although much progress has been made in reducing emissions from vehicles, and lead has been essentially phased out of gasoline, challenges remain. The most important air pollution challenge is low-level ozone, which is formed in the atmosphere by complex photochemical reactions involving VOCs and nitrogen oxides in the presence of sunlight. Exposure to high levels of ozone can cause coughing, eye irritation, sore throat, headache, and chest pain, especially for children and old people who have asthma conditions. Efforts to limit criteria pollutants by the EPA have been made under the Clean Air Act, originally passed in 1970 and amended several times.

There are still many U.S. cities that do not meet minimal standards for ozone, of which the precursors come from vehicle tailpipe emissions and evaporative emissions. The major problem seems to be NO_x emissions, which are increasing. (NO_x emissions also come from electric power plants.) Another problem is that although engines are much cleaner today, vehicle miles traveled continually increases as more cars are purchased. More efficient engines also tend to produce more NO_x. A third problem is that sport utility vehicles are being purchased in increasing numbers, and they have not been regulated as

stringently as light-duty vehicles. They are less efficient and produce more emissions.

The Clean Air Act Amendments of 1990 provided for tier 1 standards on emissions from mobile sources covering nitrogen oxides in addition to carbon monoxide, hydrocarbons, and particulate matter for light-duty vehicles beginning with model year 1994. More stringent tier 2 standards will apply to all passenger vehicles starting in 2004, including a reduction in the sulfur content of gasoline in order to ensure the effectiveness of emission control technologies. Standards will also be tightened for heavy-duty vehicles (trucks) beginning in 2004, including a tighter combined NO_x and VOC standard. This will be followed in 2007 by new "ultra-low" sulfur content requirements for diesel trucks and specific NO_x emissions control technologies. Europe is not far behind the United States in strengthening its emissions requirements for mobile sources.

Oil use is also a major contributor to greenhouse gas emissions. According to the EIA, CO₂ emissions (the leading greenhouse gas) from industrialized countries accounted for 51% of the global total in 1999, followed by developing countries at 35% and transition economies in Eastern Europe/FSU at 13%. In the industrialized world, oil use contributed almost half (49%) of carbon dioxide emissions. The transportation sector is an important contributor. In the United States, transport contributed 33% of carbon dioxide emissions in 1999.

The Bush administration rejected participation in the Kyoto Protocol in March of 2001 and put forward an alternative voluntary plan to reduce the carbon intensity of the U.S. economy in spring 2002. However, it has not offered to regulate carbon dioxide emissions, which were absent from the administration's "Clear Skies Proposal" to strengthen limits on three other pollutants. In what may be a harbinger of the future, the state of California legislature passed a bill in July 2002, subsequently signed by the governor, calling for a plan by 2005 to reduce greenhouse gas emissions from cars and light trucks. Methods of achieving reductions are not specified but presumably might include fuel efficiency, hybrid engines, fuel cells, or other technologies. Cars and light trucks emit approximately 40% of greenhouse gas emissions in California. As has happened in the past, once again the state of California may be blazing a new trail in environmental regulations.

Currently, the future of the Kyoto Protocol is uncertain; it has been ratified by the EU and Japan but needs to be approved by the Russian parliament

before it can come into force. Several European countries have already been experimenting with carbon taxes.

3.7 Oil and Geopolitics, Foreign Policy, and International Security

Oil has implications for foreign policy, international security, and geopolitics. However, the implications are more international than national, except for the United States, for which the two are intertwined. The fact that two-thirds of the world's oil reserves are located in the Persian Gulf states makes them targets for aggression. The end of the Cold War and the disintegration of the Soviet Union removed the Soviet threat to the region. However, the Gulf remains an unstable place, subject to revolutions and regional wars, as the Gulf War of 1990–1991 demonstrated when Iraq attacked and annexed Kuwait and threatened Saudi Arabia. In that war, a UN military force was assembled, led by the United States, to protect Saudi Arabia and force Iraq to relinquish Kuwait. Although there were probably several reasons for the U.S. action, surely the protection of access to Gulf oil supplies was a major reason. It appears that the United States, as the world's remaining superpower, is the only country that can play this role.

A U.S. security role will continue to be needed in the Gulf, where there is no regional balance of power. That role is strengthened but has also become more complicated following the U.S. invasion and occupation of Iraq in 2003. Any threat to Gulf oil supplies would pose a threat to the world economy if supplies were cut off and oil prices escalated. The United States will oppose any serious threat to the world economy because it is increasingly interconnected with that economy, whatever the level of Persian Gulf imports into the United States (which is much lower today). Europe and Japan are much more dependent on imports of Gulf oil than is the United States.

As noted in "Energy Security," the DOE's 1987 report to the president, "Increased dependence on insecure oil supplies reduces flexibility in the conduct of U.S. foreign policy." A key example of this is U.S. relations with Saudi Arabia, for a long time considered a key ally in the Middle East presumably because of its leading role as an oil producer. However, Saudi Arabia is an autocratic regime, not a democracy, with weaknesses in the area of human rights and treatment of women, and it is the country that produced many of the terrorists who participated in the September 2001 attacks on the World Trade Center and the Pentagon. However, the "oil

factor" reduces U.S. leverage and freedom of action vis-à-vis that country.

The globalization of the world oil market is widening the context of international oil security in the future. For example, the rapidly growing economies of Asia will import increasing amounts of oil from the Middle East in the years ahead. Oil tankers en route to Asia must pass through the narrow Strait of Malacca near Singapore. If the strait were closed by terrorists or military action, tankers would be required to add considerable distance and expense to deliver their oil, which would immediately raise freight rates worldwide. Will the United States sometime be asked to police this vital Strait?

Central Asia, as another example, is a region of increased interest for future oil production, assuming pipeline routes can be constructed to ship the oil to Western markets. However, the new republics around the Caspian Sea are not democracies and could become unstable during future transfers of power or revolutions or ethnic strife. A security framework for the region has yet to be devised.

Oil interacts with foreign policy in other ways. The United States maintains trade sanctions on many countries throughout the world, including Iraq (subject to UN sanctions after the Gulf War), Iran, and Libya. These three countries are major oil producers but they are off limits to U.S. companies for major oil development projects. Leaving Iraq aside as a special case, there is some question whether these sanctions are working effectively if they are not multilateral and accepted by other countries.

3.8 Strategies to Reduce Oil Dependence

The best long-term strategy to reduce the costs and risks of oil dependence lies in research and development of affordable alternatives to petroleum, especially in the transportation sector, in which there is a need for new technologies and fuels. Early efforts focused on developing vehicles that use compressed natural gas or liquefied petroleum gas or on alcohol fuel additives, such as ethanol or methanol, that can be blended with conventional gasoline. Electric vehicles have been tried, but they have been limited by the distances they can operate before batteries need to be recharged—and a breakthrough in battery technology has so far been illusive. Hybrid vehicles, which combine a small conventional internal combustion engine with an electric motor and offer increased efficiency and lower emissions, appear to be the best candidates for a transition period. Ultimately, fuel cell vehicles, which produce no

emissions except water and can run at high efficiency, are the best hope, although they would require a new infrastructure to provide hydrogen or a way to create it on board the vehicle. Fuel cells are still very expensive, but the second Bush administration has invested in a new government–industry partnership to promote their development.

It is unclear how much time this will take, but by 2010–2015 fuel cell vehicles could be available on the market, and by 2020 they could be expanding their market share. Such a development could have a major impact on reducing future world oil demand, although the transition to a new transportation technology will take a considerable amount of time.

4. RELIABLE ELECTRICITY SUPPLY AND PRICE VOLATILITY

Because electricity cannot be stored, security of supply is critical. Previously, this was ensured by monopoly utilities under government regulation. With the beginning of liberalized markets, there are new questions. Liberalization shortens contracts and threatens adequate investment. The California electricity crisis of 2000–2001 brought the problem sharply into focus.

The California crisis occurred 2 years after the state reformed its power market. In June 2000, in a period of exceptionally hot weather and in which there were some problems with grid operation, there were rolling blackouts in the San Francisco Bay area. Later in the summer, electricity prices tripled in the southern part of the state, and San Diego Gas and Electric asked the Federal Energy Regulatory Commission for price controls in wholesale markets. High wholesale prices led to questions about market power. The governor signed legislation placing rate caps on residential and small commercial users in southern California. In December, the price of electricity spiked to 30 ¢/Kwh. Pacific Gas and Electric (PG&E) and Southern California Edison, facing high pool prices and unable to raise retail prices, later defaulted on \$12 billion of debt. PG&E eventually declared bankruptcy.

In early 2001, California experienced a series of short-duration rolling blackouts. Meanwhile, Governor Davis signed an emergency order authorizing California's Department of Water Resources to become a temporary buyer of power, allowing it to float revenue bonds to finance power purchases under

long-term contracts. The state's experiment with deregulation and a power pool had apparently failed.

The details of what happened in California are still being investigated. There was a combination of bad planning and bad luck. A 3-year drought in the Northwest had reduced hydroelectric capacity normally available for import into California. Natural gas prices spiked during the same period because of insufficient pipeline capacity at the border and within the state. There was physical congestion in north–south transmission. When pool prices escalated, market power among wholesale producers (many of them out of state) became a problem. However, one of the most significant facts was long-term underinvestment in electric power plants in a state in which population, and therefore electric demand, was increasing. This was not an optimal condition on which to proceed with full-scale liberalization of the electric market.

As highlighted by a recent IEA study, "Security of Supply in Electricity Markets," during the past 20 or 30 years the electric systems of most OECD countries have generally maintained adequate reserve margins, even to the point of overinvestment, which ensured security of supply but at additional costs to consumers. In a liberalized market in which market players bear the costs and risks, the incentives to overinvest are removed. At the same time, there may be market imperfections that hinder the ability to achieve reliable supply, including limited demand-side response to market conditions, price distortions, policy barriers for certain technologies or the use of certain fuels, or cumbersome licensing procedures for new power plants.

The United States is midway in the complex process of restructuring the electric industry, previously regulated at both federal and state levels. Some states are further along in the process and have done better than California at making competition work (e.g., Pennsylvania and New York). Other states are holding back. The Federal Energy Regulatory Commission, which has played an important role in the process, is pressing utilities and states to form regional transmission organizations and has issued an important new proposed rule on standard market design. Meanwhile, Congress is considering new energy legislation, which may provide more guidance at the federal level.

In Europe, the United Kingdom and the Nordic countries have led the movement toward liberalized markets. Also, the EU has begun a process with its 1997 electricity directive to move the EU countries collectively toward more competition in the electric industry.

The issue of security of the U.S. electric supply was highlighted more broadly by a major power blackout on August 14, 2003, which lasted up to 2 days and affected some 50 million people in the Midwest and Northeast as well as in Ontario, Canada. The blackout reportedly began with downed power lines at a utility in Ohio. A U.S.–Canada Task Force is investigating the causes. The incident may lead to a stronger government role in enforcing reliability standards in the electric industry and in modernizing an aging power grid.

5. NATURAL GAS AND PRICE VOLATILITY

Natural gas is the fastest growing energy source worldwide, and its share of total energy consumption, according to the EIA, is projected to increase from 23% in 1999 to 28% in 2020. For the United States, the gas share is predicted to increase from 23.6% in 2000 to 26.5% in 2020. Gas is the cleanest of the fossil fuels, and it is the preferred choice for new electric power plants driven by gas turbines. Coal is likely to remain the lead fuel for power generation, but natural gas will expand its share considerably.

Gas is more difficult to transport than oil and requires pipelines or special liquefied natural gas (LNG) tankers. For this reason, natural gas tends to be traded in regional markets (e.g., North America, Europe, and Asia). However, although the amount of gas traded internationally (20%) is much less than the amount of oil (50%), international gas trade is also growing via international pipelines and expanded LNG trade.

There is no shortage of natural gas in the world. The USGS reports in its 2000 “World Petroleum Assessment” that only 10% of worldwide gas resources have been produced (compared to 25% for oil). Considerable volumes of gas remain to be discovered and developed. A major problem has been that much of the world’s gas is located far from demand centers. More than half of the world’s remaining reserves are in the former Soviet Union, Middle East, and North Africa. Two countries, Russia and Iran, hold approximately 45% of the world’s reserves. Thus, there is some concentration of reserves.

Western Europe, which holds only approximately 2% of world gas reserves (mostly in the North Sea), imports approximately 40% of its requirements,

mainly from Russia and Algeria. European gas demand is likely to expand rapidly during the next 20 years, which will require increased imports. The European gas industry is also in the process of deregulation following an EU directive in 1998 that set forth a staged plan of achieving more competitive gas markets.

In 1995, the IEA study on gas security concluded that most European countries could withstand gas supply interruptions, but it also noted that gas infrastructure is less flexible than that of oil. Could gas supply security become an issue in the future? If so, it might be desirable to establish a framework for gas security with obligations to hold strategic stocks and develop demand restraint programs. Since the power industry will be using increasingly more gas, reliable gas supply is crucial to the functioning of the electricity industry.

The United States has considerable gas supplies, but it imports approximately 15% of its gas needs from Canada. It also receives other gas imports via LNG. Although LNG shipments are currently small, they are projected to increase.

The U.S. gas industry has been deregulated and functions mostly on short-term contracts and spot sales. On the whole, the liberalized U.S. gas market has worked well. However, there have been some problems. In 2000, natural gas prices increased dramatically and were pushed even higher by very cold weather in November and December. Prices remained high through the first half of 2001. The mean price range during the period was \$2.53–7.85 per million Btu, up from \$1.98 per million Btu during 1995–1999. The spot price reached more than \$10 per million Btu at the Henry Hub at the end of December 2000. What was troubling was the length of time that prices stayed at high levels. High-demand growth and cold weather, plus inadequate gas reserves in storage, explain part of the problem. So does the inelastic short-term supply response to price. Later in 2001, gas prices declined because of the slowdown in the economy and milder temperatures.

Although gas pipelines were used to capacity during the crisis period, there were apparently few infrastructure constraints except in and near California, where transmission capacity was not adequate to transport all the gas needed and natural gas prices spiked higher than elsewhere in the U.S. market (and contributed to the California electricity crisis).

In 2003, at the request of the Secretary of Energy, the National Petroleum Council published an

industry study, "Balancing Natural Gas Policy: Fueling the Demands of a Growing Economy." The study notes that traditional North American gas producing areas can only meet 75% of projected U.S. gas demand in 2025. This includes Canadian production which is reaching a plateau. The gap could be filled by encouraging increased production in the Rocky Mountains, by Arctic gas (which requires a new pipeline), and by expanded LNG imports. At the same time there will be a need to expand gas infrastructure and to deploy new technologies to increase the efficiency of gas use. The study notes that gas price volatility will likely continue, reflecting the variable nature of supply and demand in a free market.

6. TERRORISM AND ENERGY SECURITY

The September 11, 2001, attacks against the United States by the Al-Qaeda group have raised the specter that future terrorist assaults could be made against U.S. energy infrastructure. The subject is not new. Oil companies operating in Columbia have suffered attacks against pipelines and other facilities by guerillas operating in that strife-torn country. However, it is new for the United States. Although electrical/nuclear power plants might be considered especially vulnerable targets, certainly oil refineries and, to a lesser extent, oil and gas pipelines are potential targets as well. New efforts are being made to increase physical security around these facilities, but it is a daunting task. Cyber attacks against energy systems are also a possibility.

The analysis of energy infrastructure vulnerabilities has begun and will require extensive government-industry cooperation. A new Office of Energy Assurance was created soon after September 11, 2001, in the DOE, but it may be folded into the new Department of Homeland Security if endorsed by Congress.

Although physical security can be strengthened at large power plants and petroleum refineries, it will undoubtedly be impossible to protect power lines and pipeline systems from attack. Instead, strategies need to be developed to promote redundancies and reserves in our energy systems to be able to respond to interruptions and reinstate normal service as quickly as possible.

A study of this subject by the U.S. Energy Association, "National Energy Security Post 9/11," recommends some initial actions, which require the application of advanced energy technologies.

First, detailed risk assessments need to be made of North American energy system vulnerabilities to terrorism, which might be called probabilistic vulnerability assessments, and corresponding countermeasures. Second, because many existing control systems are linked to the Internet and are therefore vulnerable to cyber attacks, steps should be taken to develop secure, private communications networks for energy systems, with adequately protected back-up systems. Third, with regard to the widely dispersed electricity grid, which is particularly vulnerable, a more flexible strategy of intelligent "islanding" is needed to minimize the impacts of an attack, along with self-healing mechanisms. For major energy facilities, there is a need to develop and install sensors to detect dangerous biological and chemical agents, which might be released in cooling towers, and to identify and deploy countermeasures. Contingency planning is needed for all kinds of attack scenarios against centralized and local facilities.

SEE ALSO THE FOLLOWING ARTICLES

Geopolitics of Energy • Gulf War, Environmental Impact of • Inflation and Energy Prices • National Energy Policy: United States • Nationalism and Oil • Oil Crises, Historical Perspective • Oil Industry, History of • Oil Price Volatility • OPEC, History of • OPEC Market Behavior, 1973–2003 • War and Energy

Further Reading

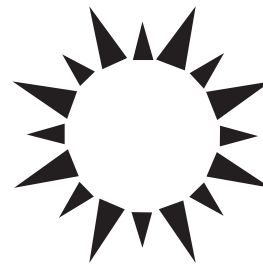
- Adelman, M. A. (1995). "The Genie out of the Bottle: World Oil since 1970." MIT Press, Cambridge, MA.
- Asia Pacific Energy Research Centre. (2003). Emergency oil stocks, and energy security in the APEC region, APEC No. 00-RE-01.2. Asia Pacific Energy Research Centre, Tokyo.
- Campbell, C., and Laherrere, J. (1998, March). The end of cheap oil. *Sci. Am.* 278(9), 78–83.
- Deffeyes, K. S. (2001). "Hubbert's Peak: The Impending World Oil Shortage." Princeton Univ. Press, Princeton, NJ.
- Energy Information Administration (2001). "U.S. Natural Gas Markets: Mid-term Prospects for Natural Gas Supply." Energy Information Administration, Washington, DC.
- Energy Information Administration (2002). "International Energy Outlook, 2002." Energy Information Administration, Washington, DC.
- Greene, D. L., and Tishchishyna, N. I. (2000). Costs of oil dependence: A 2000 update, ORNL/TM-2000/152. Oak Ridge National Laboratory, Oak Ridge, TN.
- Greene, D. L., Jones, D. W., and Leiby, P. N. (1997.). The outlook for U.S. oil dependence. *Energy Policy* 26(1), 55–69.

- International Energy Agency (2001a). "World Energy Outlook, 2001 Insights." Organization for Economic Cooperation and Development, Paris.
- International Energy Agency (2001b). "Oil Supply: The Emergency Response Potential of IEA Countries in 2000." Organization for Economic Cooperation and Development, Paris.
- International Energy Agency (2002). "Security of Supply in Electricity Markets." Organization for Economic Cooperation and Development, Paris.
- National Energy Policy Development Group (2001). "National Energy Policy." National Energy Policy Development Group, Washington, DC.
- National Petroleum Council (2003). "Balancing Natural Gas Policy: Fueling the Demands of a Growing Economy," vol. 1, Summary. Washington, DC.
- Quarterly Review of Economics and Finance* 42 (2002). [Special issue on the oil market]
- U.S. Energy Association (2002, June). "National Energy Security Post 9/11." U.S. Energy Association, Washington, DC.
- U.S. Geological Survey (2000). "World Petroleum Assessment." U.S. Geological Survey, Washington, DC.
- Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money and Power." Simon & Schuster, New York.

Natural Gas, History of

CHRISTOPHER J. CASTANEDA

California State University
Sacramento, California, United States



-
1. Origins
 2. Natural Gas in Fredonia, New York
 3. Early Commercial Utilization
 4. Natural Gas in Pittsburgh
 5. Natural Gas in the Southwest
 6. Long-Distance Pipelines
 7. Natural Gas in the Great Depression
 8. Appalachian Gas and Federal War Planning
 9. Gas in the Postwar Era
 10. Deregulation

Glossary

Appalachia The mountainous region stretching from northern Mississippi to southern New York and commonly characterized by rural communities and poverty.

British thermal unit (Btu) The amount of heat required to change the temperature of 1 lb of water 1°F at sea level.

Federal Energy Regulatory Commission (FERC) The successor to the Federal Power Commission; created in 1977.

Federal Power Commission (FPC) The federal regulatory agency responsible for regulating the interstate natural gas industry; created in 1920 and abolished in 1977.

hydrocarbons Organic compounds that are composed entirely of carbon and hydrogen. Petroleum products are composed of hydrocarbons, and methane, or natural gas (CH₄).

Insull, Samuel (1859–1938) A leader in the American public utility industry during the early 20th century. He was born in London and served as Thomas Edison's secretary as a youth.

manufactured coal gas A fuel gas for illuminating and heating purposes produced by heating coal in a retort and capturing the resulting vapors for distribution.

waste gas Natural gas, considered to be a nuisance in the production of oil.

satisfied approximately 25% of U.S. energy demand. It has been used for electric power generation, industrial heat processes, domestic heating and cooking, and transportation fuel. Natural gas is composed primarily of methane, a hydrocarbon consisting of one carbon atom and four hydrogen atoms (CH₄). As a “fossil fuel,” natural gas is rarely pure. It is commonly associated with petroleum and often contains other hydrocarbons, including butane, ethane, and propane. In the United States, substantial natural gas utilization did not begin until after the discovery of large quantities of both crude oil and natural gas in western Pennsylvania during 1859.

1. ORIGINS

Natural gas was observed and utilized in limited quantities during ancient times. References in literature to burning springs, burning bushes, or perpetual lights suggest that natural gas was used, albeit rarely, for heating. In ancient China, burning gas springs heated brine water in order to extract salt, and there were flaming gas springs in Greece and Rome. Recorded observations of burning springs in France, Italy, and Russia also exist. The philosopher Plutarch and theologian St. Augustine described lights that may have been produced by burning natural gas.

In colonial America, both George Washington and Thomas Jefferson observed natural gas springs. During the autumn of 1770, Washington participated in an expedition along the Ohio and Kanawha rivers in West Virginia and Ohio. Near the present-day town of Pomeroy, Ohio, Washington described a location “wch. the Indians say is always a fire.” About perhaps the same site, Thomas Jefferson recorded his observations of “a hole in the earth...from which issues constantly a gaseous stream.” Other visitors to these springs reported that hunters used them to cook food.

Natural gas is a vital fuel for modern society. During the last 50 years of the 20th century, natural gas

Through the early 19th century, these “burning springs” had little practical or widespread use. Most importantly, there was no practical method of either capturing the gas emanating from the springs or storing and redirecting its flow through piping.

2. NATURAL GAS IN FREDONIA, NEW YORK

Residents of Fredonia, New York, were perhaps the first Americans to use natural gas for lighting on a regular basis. Gas springs in the vicinity of Fredonia had been observed in the early 1800s, but it was not until the mid-1820s that a local gunsmith named William Aaron Hart organized an apparently successful effort to utilize gas from the local gas spring to provide light for local homes and establishments.

Some accounts written much later state that citizens of Fredonia used natural gas to illuminate their town when the French military leader Marquis de Lafayette visited. Lafayette toured America during the years 1824–1825, and he traveled to New York in the summer of 1825. Of his visit to Fredonia, Lafayette’s private secretary, A. Levasseur, recorded that they had observed a great many lights in the town. The local newspaper featured a story on the same events and noted lamps and chandeliers that provided illumination in the town during Lafayette’s visit. Contemporary reports of Lafayette’s visit do not mention gaslights at Fredonia; only the accounts of this event written much later mention gas lighting. Lafayette’s secretary did note gas lighting at other locations. While in a Boston theater, Levasseur recorded observations of “gas blazing abundantly from numerous pipes, and throwing floods of dazzling light over the hall.” These lights were fueled by manufactured coal gas, however, and not natural gas.

After Lafayette’s visit, William Hart continued to develop his interest in natural gas in Fredonia. During 1827, he began work on a plan to supply natural gas to a lighthouse at nearby Barcelona Harbor. After the U.S. government granted him a contract for this service, he installed a primitive gas works. It consisted of a fish barrel placed over the gas spring located at Westfield along Lake Erie. The barrel served as a “gasometer,” or gasholder. Hart sealed the gasometer and transported the gas for one-half mile through hollowed out pine logs to the lighthouse. Gas from the spring provided enough fuel to illuminate 144 burners and create a bright light.

3. EARLY COMMERCIAL UTILIZATION

It was not until Colonel Edwin Drake discovered oil in Titusville, Pennsylvania in 1859 that natural gas became a significant source of energy in the United States. Although Drake had been searching for oil, he found natural gas as well; oil and natural gas are often found in the same geologic structures. Natural gas discovered in eastern Pennsylvania was marketed to regional customers. Therefore, the Drake discovery heralded the beginning of both the modern U.S. oil and natural gas industries.

Prior to Drake’s discoveries, there were few successful long-term attempts to utilize natural gas for either industrial or commercial purposes. By the mid-19th century, only those factories or towns located very near a natural gas well could utilize the fuel. The difficulty of containing a natural gas spring, storing the gas, and transporting it over long distances limited its utility. For example, significant natural gas discoveries such as the high-volume well discovered by William Tomkins in 1841, near Washington’s burning spring on the Canadaway Creek, attracted some attention but little commercial interest. Alternatively, manufactured coal gas plants could be built and operated anywhere as long as coal, or the feedstock, was readily available. Thus, the manufactured coal gas industry developed much

TABLE I

Introduction of Manufactured Gas to Major Cities^a

Year	City
1816	Baltimore
1825	New York City
1829	Boston
1832	Louisville
1835	New Orleans
1836	Philadelphia
1843	Cincinnati
1846	St. Louis
1849	Chicago
1854	San Francisco
1867	Kansas City
1867	Los Angeles
1871	Minneapolis
1873	Seattle

^aThe dates primarily reflect the year manufactured gas was first produced in the city for commercial use. In some cases, however, the date reflects when a city charter was granted, and charters were sometimes granted before and even just after gas service began.

more quickly than that of natural gas in the 19th century. By the mid-19th century, many towns and cities had a manufactured gas plant and a local distribution system that provided some coal gas for residential and business lighting. (see Table I).

The earliest recorded use of gas for industrial purposes in the United States occurred in 1840, near Centerville, Pennsylvania. The gas was used to distill salt from brine water. Gradually, in the 1860s and 1870s, local deposits of natural gas were utilized for a variety of industrial heating applications. Even in Fredonia, New York, where some residents and shop owners utilized natural gas for lighting beginning in the mid-1820s, a formal natural gas company was not organized for many years. In 1858, businessmen established the Fredonia Gas Light and Water Works Company to operate the local gas wells and discover new ones.

Natural gas was not used on a large scale until the 1880s, and gas wells were most likely to be abandoned when oil was not concurrently discovered. An example of early abandonment occurred in 1865 when a 480-foot drilling effort struck a natural gas reservoir near West Bloomfield, New York. The operators estimated the gas flow to be about 2000 cubic feet (mcf) per day; they directed the gas into a large balloon and attempted to measure the flow by calculating the time required to fill it. Because the investors were disappointed that oil was not discovered, they abandoned the project. Not everyone was disappointed that this well contained only natural gas. Several businessmen formed the Rochester Natural Gas Light Company and purchased the same gas well in 1870. The nearest town desiring natural gas was Rochester, about 25 miles away. The company constructed a pipeline system to connect the well with the town. They built a pipeline out of Canadian white pine. The 2- to 8-foot log segments were planed to a uniform 12.5-inch exterior diameter, and they were bored for an 8-inch interior diameter. Construction and maintenance of the wood pipeline system was particularly problematic, but the company began transporting natural gas during the winter of 1872. Consumers in Rochester discovered quickly that hotter burning natural gas was not easily interchangeable in their burners with manufactured coal gas. This situation resulted in lower than expected natural gas demand. Continuing problems with gas transportation facilities caused significant problems for the company; rotting and leaking wood pipelines simply prevented the adequate transportation of natural gas from well to consumer. Soon, the company stopped operations.

A more successful attempt to transport natural gas took place in 1872. New natural gas discoveries created a demand for specialized gas pipelines. In this case, a 2-inch wrought-iron line was constructed and used to connect a gas well 51 miles away. The line transported "waste gas" from nearby oil fields to Titusville. This pipe transported 4 million cubic feet (mmcf) per day to 250 customers, both residential and industrial.

The primary obstacle to the expansion of the natural gas industry in the mid-19th century was inadequate pipeline facilities and technology, not lack of supply. Hollow log pipelines leaked and disintegrated, but cast and wrought iron lines also suffered from significant intrinsic defects. Wrought iron lines in the period 1872–1890 were typically less than 8-inches in diameter, and the pipe segments were attached with couplings tightened with screws. Gas leaks were common problems. Most of the gas transported in pipelines during this period flowed under the natural pressure of the well without the aid of additional compression.

4. NATURAL GAS IN PITTSBURGH

Pittsburgh became the first major U.S. city in which industry utilized large volumes of natural gas for industrial heat processes. Abundant Pittsburgh area coal deposits and the importation of iron ore from the Juanita region in central Pennsylvania (and later from the Mesabi range) facilitated development of a substantial iron industry originally fueled by coal. Extensive coal burning for industrial heat created significant air pollution, and Pittsburgh became known as the "Smoky City." Contemporary newspaper articles noted the black smoke produced by burning coal. In 1884, *The New York Times* reported that natural gas would be used in Pittsburgh's industries to reduce coal smoke pollution.

The earliest recorded use of natural gas in a Pittsburgh iron works occurred in 1870–1871, but widespread natural gas utilization did not commence until the early 1880s, after the development of nearby gas wells. Entrepreneurs then organized new regionally based gas firms. One group of Pittsburgh area manufacturers established the Chartiers Valley Gas Company in 1883 to transport natural gas from local gas fields to their glass and steel plants. This company's first line extended from the Hickory gas field to Pittsburgh. The wrought-iron line was the first "telescoping" pipeline, meaning that a smaller diameter pipe installed at the well's origin led to a larger diameter pipe in the city. The telescoping line system

was useful for lowering the gas line pressure as gas flowed into the city. For pipe less than 12 inches in diameter, the typical connection was a screw coupling. Pipe segments were threaded on the outer edge of each length end that turned into a screw coupling.

As of 1886, the Chartiers firm also laid claim to operating the largest continuous pipe in the world. The company installed a 16-inch pipe extending from the Murrysburg gas field to Pittsburgh. After 8 miles, the 16-inch line was fed into a 6-mile-long 20-inch pipe, and it in turn fed into a 5-mile section of 24-inch cast-iron tubing, tested at 300 pounds per square inch (psi). The National Tube Works constructed this line; J. P. Morgan controlled National Tube.

By the late 1880s, Pittsburgh had become the locus of the American steel and coal industry, and it was also the center of the natural gas industry. In 1886, there were 10 iron and steel mills using natural gas in their puddling furnaces, with many more planning to convert to gas. Six glass-making factories and reportedly every brewery in Pittsburgh used natural gas instead of coal. The Sampson Natural Gas Crematory also used the invisible fuel. Pittsburgh received its natural gas from the lines of six companies tied into 107 regional gas wells. Five hundred miles of pipeline transported natural gas from wells to the city, including 232 miles of line within the Pittsburgh city limits.

As natural gas utilization increased, local engineers addressed technological problems associated with its transportation. Solomon R. Dresser focused attention on drilling and oil-field-related technology. In 1880, he formed S. R. Dresser & Company and conducted pioneering work in pipe coupling. In 1887, Dresser received a patent for using a rubber ring in pipe joints to create a leakproof coupling. Although this method proved not entirely satisfactory, less than a year later Dresser designed a second coupling that was more effective. He developed a two-part mechanical device that pulled the pipe segments together. Between the tightened sections, an internal rubber ring created a seal. Dresser proved the leakproof qualities of this coupling method when he developed his own gas field near Malta, Ohio and used the couplings in a gas line that extended into town. The Malta Natural Gas Line established Dresser as a leader in the natural gas business, and his couplings attracted widespread favor; gas companies located throughout the country ordered them. As much as 90% of the gas pipeline industry used these couplings into the 1920s.

Improved couplings not only reduced leakage, they also lessened the possibility of explosions. In

Pittsburgh, city ordinances prohibited gas lines from operating at pressures higher than 13 psi. This pressure limitation was intended to reduce the leaks prevalent in more highly pressurized lines. Leaks often resulted in accumulations of gas in cellars, leading to explosions and fires. Within the city, regulating valves further reduced the gas pressure. To prevent leaking gas from ending up in residential cellars, the Chartiers Valley Company used its patented "broken stone escape system." This system involved laying a pipe in a trench filled with dirt to the center of the pipeline. Workers then placed about 9 inches of broken stone on top of the line. A layer of tarpaper was then placed over the stone; dirt covered the tarpaper. The stone barrier was placed adjacent to every city lamppost so that escaping gas could vent through the stone. In addition, gas firms used "escape pipes," very small diameter lines leading from each pipe joint to a lamppost. Inspectors checked each escape pipe for possible leaks and identified the joint to which each escape line was connected. A system of 4-inch pipes distributed gas to individual residences. In these pipes, gas pressure was limited to about 5 psi. As the gas entered homes, an additional regulator/shutoff valve lowered gas pressure again to about 5 ounces per square inch, so that gas could be burned satisfactorily in gaslight fixtures.

George Westinghouse, inventor of the railroad air brake and a resident of Pittsburgh, also became involved in the expanding natural gas industry. He explored for natural gas close to home. He drilled for gas in his own backyard located in a fashionable Pittsburgh neighborhood. In late February 1884, a small volume of gas began flowing from the well. The workers continued drilling to a depth of about 1560 feet. At 3 a.m. one morning, Westinghouse awoke to the sound of a tremendous explosion and the loud sound of hissing gas from the well.

Westinghouse needed a company organization to proceed with his new plan of selling his natural gas to Pittsburgh-area customers. He purchased a moribund company, the Philadelphia Company, to produce the fuel. As President and Director of the company, Westinghouse watched the firm become one of the largest gas businesses in the Pittsburgh area. For additional supply, the company leased substantial gas production acreage in western Pennsylvania. By 1887, the Philadelphia Company supplied approximately 5000 residential and 470 industrial customers with gas from about 100 natural gas wells located on 54,000 acres.

Westinghouse's financial participation in the natural gas business brought his inventive mind in touch with

some of the major problems of this new industry. Between the years 1884 and 1885, he applied for 28 gas-related patents, and during his lifetime he applied for a total of 38 gas equipment patents. Some of Westinghouse's most important inventions for natural gas included a system for enclosing a main gas line in residential areas with a conducting pipe to contain gas leaks. Westinghouse also developed a method for "stepping down" the highly pressurized gas in main trunk lines to lower pressure in residential areas. To prevent accumulations of gas in homes and shops after gas service was shut down and then restarted, Westinghouse patented a pressure regulator and cutoff valve that automatically restricted gas flow when the pressure dropped below a particular point. Tragedies nonetheless occurred. On the morning of January 31, 1885, two major explosions at Thirty-fifth and Butler streets in Pittsburgh nearly leveled an entire city block, killing two and injuring 25 others, some severely. The first explosion occurred at George Hermansdorfer's butcher shop after an accumulation of gas in his cellar; two or three people were badly burned. People rushed to investigate the explosion when a second, larger explosion occurred nearby. Subsequent explosions caused substantial injury to life and property, damaging as many as 15 buildings. Local residents threatened a riot against the gas company, and a representative of the Fuel Gas Company made a stunning admission: the pipes had not been tested before the gas was turned on. Efforts to develop gas regulators and emergency shutoff valves were absolutely required to ensure that this fuel could be utilized safely.

Andrew Carnegie, Pittsburgh's foremost entrepreneur, understood that natural gas had superior heating characteristics. He wrote: "In the manufacture of iron, and especially in that of steel, the quality is also improved by the pure new fuel. In our steel rail mills we have not used a pound of coal for more than a year, nor in our iron mills for nearly the same period." The iron and steel maker also noted that natural gas burned much more cleanly compared to coal.

By 1885, 150 companies had charters to sell gas in Pennsylvania, but the future of natural gas was not certain. Gas fields tended to exhaust themselves within several years after discovery. Selwynn Taylor, a Pennsylvania mining engineer, believed that most regional natural gas fields would soon be exhausted, and the price of coal would rise to the levels existing prior to the discovery of regional gas fields. His beliefs were typical of the time: existing natural gas fields and current production, transportation, and distribution systems simply could not supply enough gas to satisfy the demand and natural gas was

ultimately unreliable. Fears of short-lived wells aside, gas discoveries in other Appalachian states, first in Ohio and then West Virginia, made this fuel economically significant to the entire region. Industries located in cities such as Buffalo, Cleveland, Toledo, and Cincinnati all began using natural gas from nearby wells. Waste and poor planning, however, led to many failed ventures. In one episode, the Indiana Natural Gas & Oil Company had built the longest pipeline to date in 1891. The transmission system consisted of two parallel 8-inch lines extending from northern Indiana gas fields to Chicago, a distance of approximately 120 miles. These lines transported natural gas at 525 psi. The supply quickly declined and the lines were soon removed from service. Episodes such as this but on a smaller scale were repeated throughout the region.

Similar supply problems in Indiana continued. During the late 19th century, an area covering 7000 square miles included a large number of producing natural gas fields. Despite attempts to regulate the production and flow of natural gas, unrestrained gas demand soared in the state. By 1907, many of Indiana's once productive natural gas fields had expended their valuable fuel, and many natural gas customers had to return to manufactured gas utilization. Gas discoveries in Oklahoma and in the eastern and southern Kansas gas fields suffered similar stories of rapid development followed by depletion. Episodes such as these characterized the natural gas industry, as opposed to manufactured gas, as fairly undependable.

By the turn of the century, the natural gas industry was most developed in the Appalachian region. Productive gas wells in West Virginia, Pennsylvania, New York, Kentucky, Tennessee, and Ohio led to the establishment of regional gas firms that built pipelines to serve local markets in the entire region. Natural gas was used primarily for industrial purposes, but, where available, its higher heating content meant that it was a superior cooking and heating fuel, although appliances for these purposes were still not widely available until later in the 19th and early 20th centuries. Natural gas was a promising fuel, but its limited availability and dependability forced entrepreneurs to proceed cautiously with plans to develop fields and build pipelines.

5. NATURAL GAS IN THE SOUTHWEST

The discovery of massive southwestern natural gas fields and technological advancements in

long-distance pipeline construction dramatically altered the early 20th-century gas industry market structure. In 1918, drillers discovered a huge natural gas field that became known as the Panhandle Field, situated primarily in North Texas. In 1922, drillers discovered the Hugoton Field, located in the common Kansas, Oklahoma, and Texas border area (generally referred to as the midcontinent area). The combined Panhandle/Hugoton fields became the nation's largest gas-producing area, comprising more than 1.6 million acres. It contained as much as 117 trillion cubic feet (tcf) of natural gas and accounted for approximately 16% of total U.S. reserves in the 20th century.

As oil drillers had done earlier in Appalachia, they initially exploited the Panhandle Field for petroleum only while allowing an estimated 1 billion cubic feet/day (bcf/d) of natural gas to escape into the atmosphere. As new gas markets appeared, the commercial value of southwestern natural gas attracted entrepreneurial interest and bolstered the fortunes of existing firms. These discoveries led to the establishment of many new southern firms, including the Lone Star Gas Company, Arkansas Louisiana Gas Company, Kansas Natural Gas Company, United Gas Company, and others, some of which evolved into large natural gas companies.

The sheer volume of these southwestern gas fields and their distance from distant urban markets emphasized the need for advancements in pipeline transport technology. In particular, new welding technologies allowed pipeline builders in the 1920s to construct longer lines. In the early years of the decade, oxyacetylene torches were used for welding, and in 1923, electric arc welding was successfully used on thin-walled, high-tensile-strength, large-diameter pipelines necessary for long-distance compressed gas transmission. Improved welding techniques made pipe joints stronger than the pipe, and seamless pipe became available for gas pipelines beginning in 1925. Along with enhancements in pipeline construction materials and techniques, gas compressor and ditching machine technology improved as well. Long-distance pipelines became a significant segment of the gas industry beginning in the 1920s.

These new technologies made possible for the first time the transportation of southwestern natural gas to midwestern markets. Soon, the southwest supplanted Appalachia's position as the primary region for marketable gas production. Until the late 1920s, virtually all interstate natural gas transportation took place in the northeast, and it was based on Appalachian natural gas production. In 1921, natural gas produced in West Virginia accounted for approxi-

mately 65% of interstate gas transportation whereas only 2% of interstate gas originated in Texas. Most interstate gas flowed into western Pennsylvania and Ohio. Appalachian fields experienced serious depletion in the 1920s, however, and various state legislators attempted to prohibit out-of-state gas exportation. These attempts to corral natural gas for intrastate utilization were largely unsuccessful.

Between the mid-1920s and the mid-1930s, the combination of abundant and inexpensive southwestern natural gas production, improved pipeline technology, and increasing nationwide natural gas demand led to the creation of the new interstate gas pipeline industry. Metropolitan manufactured gas distribution companies, typically part of large holding companies, financed most of the pipelines built during this era. Despite the high cost of the long-distance lines, access to natural gas even for mixing with existing manufactured gas could be a profitable venture. Natural gas was so abundant it was often substantially less costly than coal gas.

In 1927, Cities Service built the first long-distance line originating in the Panhandle field. This 250-mile, 20-inch pipeline connected the Panhandle field with a Cities Service gas distributor in Wichita, Kansas. Standard Oil (New Jersey) also participated in several significant pipeline ventures during these years. The first of these was Colorado Interstate Gas Company. Organized in 1927 by Standard, Cities Service, and Prairie Oil & Gas, this firm built a 350-mile, 22-inch line originating at the Texas–New Mexico border and extending to Denver. In California, natural gas from the Buena Vista field in the San Joaquin Valley fueled industry and commercial establishments in Los Angeles, and in 1929, Pacific Gas & Electric (PG&E) constructed a 300-mile pipeline from the Kettleman field north of Los Angeles to bring natural gas to San Francisco. San Francisco was one of the first major urban areas to switch from manufactured gas to natural gas. Because the same volume of natural gas had nearly twice the heating content as coal gas, burners and airflow valves in stoves and water heaters had to be adjusted to accept the natural fuel. With near military precision, PG&E divided San Francisco into 11 districts that were successively converted to natural gas. Six hundred trained men divided into 35-member crews converted PG&E's service area within 5 months. The conversion of 1.75 million appliances cost \$2 million, but natural gas was less costly for the utility to market compared to coal gas.

New long-distance gas lines and expensive conversion programs were necessary if gas utilities were

TABLE II

Estimated Waste of Natural Gas in the United States in Billions of Cubic Feet^a

Year	Natural gas waste ^b		Total U.S. natural gas consumption
	Total U.S.	Texas panhandle	
1919	213	n/a	256
1920	238	n/a	286
1921	193	n/a	248
1922	233	n/a	254
1923	416	n/a	277
1924	343	n/a	285
1925	324	n/a	272
1926	417	220	289
1927	444	405	296
1928	412	351	321
1929	589	294	360
1930	553	252	376

^aSource. Federal Trade Commission, "Report to the Senate on Public Utility Corporations," Senate Document No. 92, 70th Congress, 1st Session, Part 84-A, 1935, pp. 93 and 95.

^bWaste means gas production that was flared or vented and otherwise not utilized. n/a, data not available.

going to meet consumer demand. The new holding companies marshaled tremendous amounts of capital to build pipelines, extend service, and promote gas utilization. They also became adept at advertising and marketing. Trained salesmen, company servicemen, and even co-opted plumbers touted gas. During the 1920s, utility companies offered for sale a wide variety of gas-powered appliances, including space heating units, water heaters, stoves, and even gas-powered refrigerators. By 1926, about 50,000 automatic water heaters had been installed in homes, but gas appliances were not inexpensive.

Another use for natural gas beginning in the late 19th century was carbon black production. Produced by burning natural gas, carbon black was used for coloring in paint and inks. It was also used as a reinforcing agent in rubber and automobile tires. Natural gas produced in fields not connected by pipelines to urban markets was a prime candidate for carbon black production. Even by the late 1930s, about two-thirds of the amount of marketable gas produced was either flared, vented, or used to make carbon black. (see Table II). But greater profits awaited entrepreneurs willing to finance pipelines connecting gas fields to urban and industrial gas markets. Urban natural gas utilization also brought forth efforts to develop a standardized odorant. Unlike coal gas,

which typically has a distinct smell, natural gas is odorless. Thus, a leak or inadvertently opened valve might allow odorless gas to accumulate in an enclosed space and asphyxiate people, or explode. Experiments with odorants date to at least 1885, and in 1930, the Bureau of Mines conducted experiments with mercaptan, which later became the standardized gas odorizer.

6. LONG-DISTANCE PIPELINES

By the late 1920s, four public utility holding companies dominated the U.S. gas industry and sought to control interstate gas transportation as well. Two of the largest holding companies, Columbia Gas and Standard Oil (New Jersey), distributed more than half of the gas sold in the entire Appalachian region. Henry Doherty's Cities Service dominated the lower midwest. The largest public utility conglomerates included Middle West Utilities, Inc. and Insull Utility Investments, Inc., both controlled by Samuel Insull and headquartered in Chicago. By the late 1920s, Insull's empire included 248 gas, coal, and electric power firms serving 4741 communities in 30 states.

Planning for the first 1000-mile pipeline began in 1926 when Samuel Insull and associates discussed the possibility of building a natural gas pipeline connecting southern gas fields with Chicago area gas utilities. They sponsored engineering studies, considered a pipeline route, and examined potential gas acreage. In April, 1930 they first incorporated as the Continental Construction Corporation; a year later the company changed its name to the Natural Gas Pipeline Company of America (NGPL). NGPL's proposed 24-inch line would extend 980 miles from north Texas to Chicago. Commonly referred to as the "Chicago pipeline," this line would allow Insull to convert Peoples Gas Light & Coke Company's service area from dependence on manufactured coal gas to cleaner, hotter burning, and less expensive natural gas.

The NGPL venture was jointly planned, financed, and controlled by three utility holding companies and three other oil firms. The three holding companies were Samuel Insull's Insull & Sons, Henry Doherty's Cities Service, and Standard Oil of New Jersey. NGPL purchased its gas supply from gas fields controlled by the pipeline's owners. Standard Oil (NJ) agreed to furnish 25% of NGPL's requirements indirectly through the Canadian River Gas Company. Canadian River was a partnership of Cities Service and Prairie Oil & Gas, Standard's partners in

the Colorado Interstate line. The Texoma Natural Gas Company supplied the remaining 75% of NGPL's gas requirements.

Henry L. Doherty & Company contracted to build the NGPL line. Construction began in August, 1930, and the main line was completed 12 months later. A total of 418 million pounds of steel pipe buried 6 feet transported gas at 600 psi. Construction costs for the main line, nine compressor stations, and telephone lines totaled \$35 million. Although NGPL's major market was Insull's Chicago area utilities, some gas was also sold to gas distributors in Kansas and other states. The first gas deliveries in Chicago commenced on October 16, 1931, and by January 1, 1932, the line was delivering 55 mmcf/d with an originally designed total capacity of 175 mmcf/d. With access to abundant volumes of natural gas, Chicago became the largest U.S. city to convert its utility distribution system to "mixed gas," and later to straight natural gas. Peoples Gas Light and Coke Company first began producing a mixed gas with a 800-Btu content. Mixed gas, a mixture of lower Btu coal gas and higher Btu natural gas provided a hotter burning flame than did coal gas alone, for both cooking and heating. Peoples Gas Light and Coke Company began charging for gas based on a price per "therm" (1 therm=100,000 Btu) basis, rather than by volume; natural gas had nearly twice the Btu rating compared to an equal volume of manufactured gas.

Peoples Gas Light and Coke Company organized a massive campaign to merchandise gas house-heating equipment. The company placed full-page and three-quarter-page advertisements in newspapers serving Chicago and in 50 outlying communities; advertisements appeared on billboards, streetcars, and shop windows. In addition, the utility hired 270 company-trained salesmen, 60 heating engineers, and 14 sales directors to promote gas consumption. Within the first 10 weeks of the promotion, Peoples Gas Light and Coke Company installed about 10,000 conversion burners, and the company made 30,000 gas installations during the gas sales promotion. Servicemen adjusted existing residential furnaces to accept the higher Btu mixed gas. In order to convert appliances, gas mains required cleaning to remove oil residue and other impurities from the manufactured gas.

Also during this time, a consortium led by North American Light & Power Company, which owned gas and electric properties throughout the midwest, purchased from Odie R. Seagraves and William L. Moody III (Moody-Seagraves Interests) the begin-

nings of the pipeline these two men had planned to build from Seagraves's Hugoton gas field properties to Omaha, Nebraska. The North American Light & Power Company joined the Lone Star Gas Company and United Light & Power Company in a partnership to purchase the Moody-Seagraves project and rename it the Northern Natural Gas Company. North American financed the construction of Northern Natural, which was completed 1931. The 1110-mile, 24- and 26-inch line transported gas to various cities along its path to Minneapolis via Omaha.

During the 1930s, a third group of entrepreneurs formed a third pipeline to connect southwestern gas fields with midwestern customers. They incorporated the Panhandle Eastern Pipe Line Company. By 1936, it was transporting gas from the Texas Panhandle through an affiliated firm to Detroit, Michigan.

7. NATURAL GAS IN THE GREAT DEPRESSION

In the late 1920s and early 1930s, the most well-known public utility figure was Samuel Insull, a former personal secretary of Thomas Edison. Insull's public utility empire headquartered in Chicago did not fair well in the economic climate that followed the 1929 Wall Street stock market crash. His gas and electric power empire crumbled, and he fled the country. The collapse of the Insull empire symbolized the end of a long period of unrestrained and rapid growth in the U.S. public utility industry. In the meantime, the Federal Trade Commission (FTC) had launched a massive investigation of the nation's public utilities, and its work culminated in New Deal legislation that imposed federal regulation on the gas and electric industries. The Public Utility Holding Company Act (1935) broke apart the multitiered gas and electric power companies and the Federal Power Act (1935) and the Natural Gas Act (1938), respectively, authorized the Federal Power Commission (FPC) to regulate the interstate transmission and sale of electric power and natural gas.

During the Depression, the gas industry also suffered its worst tragedy in the 20th century. In 1937, at New London, Texas, an undetected natural gas leak at the Consolidated High School resulted in a tremendous explosion that virtually destroyed the Consolidated High School, 15 minutes before the end of the school day. Initial estimates of 500 dead were later revised to 294. Texas Governor Allred appointed a military court of inquiry that determined an

accumulation of odorless gas in the school's basement, possibly ignited by the spark of an electric light switch, created the explosion. This terrible tragedy was marked in irony. On top of the wreckage, a broken blackboard contained these words, apparently written before the explosion: "Oil and natural gas are East Texas' greatest mineral blessings. Without them this school would not be here, and none of us would be here learning our lessons." Although many gas firms already used odorants, the New London explosion resulted in the implementation of new natural gas odorization regulations in Texas.

8. APPALACHIAN GAS AND FEDERAL WAR PLANNING

During World War II, the Pittsburgh, Youngstown, and Wheeling areas contained hundreds of steel mills and metallurgical factories, as well as rubber and chemical plants that required large volumes of natural gas. Natural gas was vital to these factories because it burned at a constant specific temperature, providing high-quality product manufacture. Approximately 660 Appalachian area factories used a substantial amount of natural gas, and wartime energy demands put further pressure on Appalachian gas reserves. Appalachian natural gas production had peaked in 1917 at 552 bcf of natural gas, or about 63% of total U.S. gas production; this percentage declined to approximately 15% by the late 1930s. The decline resulted from diminishing Appalachian gas reserves as well as a proportionate increase in southwestern-produced gas. By 1943, Appalachian production alone was insufficient for meeting regional industrial, commercial, and residential demand. (see Table III).

The intense drain on Appalachian reserves stimulated private entrepreneurial efforts to increase production and build new pipelines. At the same time, some industry executives were already looking forward to a burgeoning gas industry after the war. During one meeting held during 1942, J. French Robinson, a prominent gas utility executive, stated that "in the postwar sunshine of abundant materials for our use, we will be able to realize the potential values of natural gas to all this nation as never before." Patriotic fervor aside, the business of war stimulated both industrial production and entrepreneurial ambition.

To direct the federal government's wartime energy policy, Roosevelt chose Harold I. Ickes, who was then Secretary of the Interior. On May 28, 1941, Ickes

TABLE III

Natural Gas Production by Region, 1912–1970^a

Year	Region (%) ^b			Total marketed production (tcf) ^c
	Appalachia	Southwest	Other	
1912	74	22	2	0.56
1920	55	34	11	0.80
1922	46	37	17	0.76
1924	31	45	24	1.14
1926	26	50	24	1.31
1928	21	57	22	1.57
1930	17	61	22	1.94
1935	16	65	19	1.92
1940	15	68	17	2.66
1945	10	73	17	3.91
1950	6	80	14	6.28
1960	3	87	10	12.80
1970	2	90	8	21.90

^aSource. U.S. Bureau of Mines, "Natural Gas Annuals and Minerals Yearbook (Government Printing Office, Washington, D.C.), various years; and Energy Information Administration, "Natural Gas Production and Consumption," Energy Data Reports, DOE/EIA-0131 (Government Printing Office, Washington, D.C., 1978). Also see, David Gilmer, "The History of Natural Gas Pipelines in the Southwest," *Texas Business Review* (May–June, 1981), p. 133.

^bAppalachia includes Pennsylvania, Ohio, West Virginia, and Kentucky (and New York for 1920 only). Southwest includes Texas, Louisiana, Oklahoma, and Kansas.

^ctcf, trillion cubic feet.

assumed his new position as the first Petroleum Coordinator for National Defense; this agency was later renamed the Petroleum Administration for War (PAW). In this role, the new "oil czar" exercised special emergency powers over much of both the oil and gas industries. Despite initial industry fears, Ickes implemented a cooperative relationship with the energy industry during wartime. The PAW created a Natural Gas and Natural Gasoline Division to be responsible for the gas industry. E. Holley Poe, a former executive of the American Gas Association, headed the division. His charge was maintaining natural gas production and deliverability, particularly in the Appalachian region. Poe also attempted to marshal support for joint-industry cooperation while administering the wartime industry. The PAW's authority over natural gas was relatively modest compared to that of the Supply Priorities and Allocation Board (SPAB). The SPAB, which later merged into the War Production Board (WPB), had much broader powers over industry. Regarding

natural gas, the agency dictated specific gas sales allocation orders to gas pipelines.

During late 1941, representatives of the natural gas industry, military, PAW, WPB, and the American Gas Association met several times in different cities to discuss recommendations for restricting some classes of natural gas consumption and maintaining production levels during the war. J. A. Krug, Chief of the WPB Power Branch, was particularly concerned about potential shortages in Appalachia, southern California, and the midcontinent areas. He proposed a special "Limitation Order" for conserving natural gas. The order had two major goals: (1) to increase production and (2) to curtail nonessential consumption. Major General H. K. Rutherford wrote a letter of support and noted the critical situation faced by war industries dependent on natural gas.

In early February, 1942, the WPB issued Order L-31. This action called for voluntary compliance with pooling arrangements "to achieve practicable maximum output in the area or areas in which a shortage exists or is imminent." The order authorized the WPB to integrate natural gas systems, curtail gas sales when necessary, and reallocate existing gas sales. The WPB actively encouraged pipelines to transport gas at 100% load factor, to use gas storage fields whenever possible to free up pipeline capacity for gas transmission, and to develop curtailment schedules. Six months later, the WPB issued Order L-174, which imposed the same restrictions on the manufactured coal gas industry.

The PAW and WPB also addressed the Appalachian gas production problem. First, the PAW set guidelines for a new drilling program, M-68, for developing a nationwide oil and gas drilling program "consistent with the availability of material and equipment." This program limited drilling of gas wells to not more than 1 every 640 acres. Industry leaders objected to M-68, believing that it would stymie efforts to maintain current production levels. In response, the PAW issued new spacing provisions that permitted drilling one well on each 160 acres for specified deep horizons and one to each 40 acres for shallow wells.

The importance of Appalachian natural gas supply to the war effort was reflected in the disproportionate number of wells drilled there. Between 1942 and 1945, approximately 70% of all gas wells drilled in the country were drilled in Appalachia, even though overall production levels did not rise significantly. Wartime demand simply sped up the depletion of Appalachian gas fields. Government drilling and consumption regulations could not reverse this situation.

TABLE IV

Natural Gas Prices and Demand, 1945–1970^a

Year	Marketed production	
	Trillions of cubic feet	Average wellhead price (cents/million cubic feet)
1945	4	4.9
1950	6	6.5
1955	9	10.4
1960	13	14.0
1965	16	15.6
1970	22	17.1

^a Source. American Gas Association, Gas Facts. (Various years.)

9. GAS IN THE POSTWAR ERA

In the period following World War II, the natural gas industry expanded rapidly. A new round of long-distance pipeline construction made natural gas available throughout the nation. Natural gas fueled factories, electric power-generating facilities, and provided heat for homes and cooking. Demand for gas fuel rose dramatically as it became available. (see Table IV). In this postwar era, entrepreneurs organized several long-distance gas pipeline firms to connect southwestern gas supply with northeastern markets. New pipeline firms organized to sell natural gas to northeastern markets. One group of entrepreneurs purchased the so-called Big Inch and Little Big Inch pipelines from the United States government and converted them for natural gas transportation. The government had financed these lines during World War II to transport oil from the Texas Gulf Coast to the New York refinery area. Under new private ownership, the newly named Texas Eastern Transmission Corporation and affiliated lines delivered natural gas for the first time to Philadelphia, New York, and Boston. Two other new pipelines built either during or immediately after the war, the Tennessee Gas and Transmission Company and Transcontinental Gas Pipe Line Company, also began delivering southwestern-produced natural gas to northeastern customers in major urban areas.

Other pipelines extended from southwestern gas fields to growing urban markets on the West Coast and in the Southeast. California is a large producer of natural gas, but rapid population and infrastructure growth fueled the demand for more of it. El Paso Natural Gas became the first interstate pipeline to deliver natural gas to California, followed by

Transwestern Pipeline Company in the early 1960s. The Northwest Pipeline Company began transporting natural gas produced in the San Juan Basin in Colorado and New Mexico to customers in Seattle after 1956. In 1959, Florida Gas Transmission Company delivered the fuel to Floridians. By the mid-1950s, therefore, the beginnings of a national market for natural gas emerged. During the last half of the 20th century, natural gas consumption in the U.S. ranged from about 20 to 30% of total national energy utilization. However, the era of unrestricted natural gas abundance ended in the late 1960s.

The first overt sign of serious industry trouble emerged when natural gas shortages appeared in 1968–1969. Economists almost uniformly blamed the shortages on gas pricing regulations instituted by the so-called Phillips Decision of 1954. This law had extended the FPC's price-setting authority under the Natural Gas Act to the natural gas producers that sold gas to interstate pipelines for resale. The FPC's consumerist orientation meant, according to many economists, that it kept gas prices artificially low through federal regulation. Gas producers consequently lost their financial incentive to develop new gas supply for the interstate market, and shortage conditions developed.

10. DEREGULATION

The 1973 OPEC oil embargo exacerbated the growing shortage problem as factories switched boiler fuels from petroleum to natural gas. Cold winters further strained the nation's gas industry. The resulting energy crisis compelled consumer groups and politicians to call for changes in the regulatory system that had constricted gas production. In 1978, a new comprehensive federal gas policy dictated by the Natural Gas Policy Act (NGPA) created a new federal agency, the Federal Energy Regulatory Commission, to assume regulatory authority for the interstate gas industry.

The NGPA also included a complex system of natural gas price decontrols that sought to stimulate domestic natural gas production. These measures appeared to work almost too well and contributed to the creation of a nationwide gas supply "bubble" and lower prices. The lower prices wreaked additional havoc on the gas pipeline industry because most interstate lines were then purchasing gas from producers at high prices under long-term contracts. Some pipeline companies had also invested tremendous amounts of money in expensive supplemental gas

projects such as coal gasification and liquid natural gas (LNG) importation. The long-term gas purchase contracts and heavy investments in supplemental projects contributed to the poor financial condition of many gas pipeline firms. Large gas purchasers, particularly utilities, also sought to circumvent their high-priced gas contracts with pipelines and purchase natural gas on the emerging spot market.

Once again, government was forced to act in order to bring market balance to the gas industry. Beginning in the mid-1980s, a number of FERC orders, culminating in Order 636 (and amendments), transformed interstate pipelines into virtual common carriers. This industry structural change allowed gas utilities and end-users to contract directly with producers for gas purchases. FERC continued to regulate the gas pipelines' transportation function, but pipelines ceased operating as gas merchants as they had for the previous 100 years. Restructuring of the natural gas industry continued into the early 21st century as once-independent gas pipeline firms merged into larger energy corporations.

Natural gas is a limited resource. Although it is the most clean burning of all fossil fuels, it exists in limited supply. Estimates of natural gas availability vary widely, from hundreds to thousands of years.

TABLE V

Natural Gas Production and Consumption in the United States^a

Year	Total dry production	Total consumption
1970	21,014,292	21,139,386
1972	21,623,705	22,101,452
1974	20,713,032	21,223,133
1976	19,098,352	19,946,496
1978	19,121,903	19,627,478
1980	19,557,709	19,877,293
1982	17,964,874	18,001,055
1984	17,576,449	17,950,524
1986	16,172,219	16,221,296
1988	17,203,755	18,029,588
1990	17,932,480	18,715,090
1992	17,957,822	19,544,364
1994	18,931,851	20,707,717
1996	18,963,518	21,966,616
1998	19,125,739	21,277,205
2000	19,072,518	22,546,944

^aSource. Energy Information Agency, "Supply and Disposition of Natural Gas in the United States, 1930–2000," Historical Natural Gas Annual, Government Printing Office, Washington, D.C.

In millions of cubic feet.

Such estimates are dependent on technology that must be developed in order to drill for gas in more difficult geographical conditions, and actually finding the gas where it is expected to be located. Methane can also be extracted from coal, peat, and oil shale, and if these sources can be successfully utilized for methane production, the world's methane supply will be extended another 500 or more years.

Since 1970, natural gas production and consumption levels in the United States have remained reasonably stable. During the 1980s, both consumption and production levels dropped about 10% from the 1970 levels, but by the later 1990s, production and consumption were both on the rise. (see Table V). In the absence of aggressive conservation programs, unexpected shortages, or superior alternative energy sources, natural gas consumption will continue to increase.

For the foreseeable future, natural gas will continue to be used primarily for residential and commercial heating, electric power generation, and industrial heat processes. The market for methane as a transportation fuel will undoubtedly grow, but improvements in electric vehicles may well dampen any dramatic increase in demand for engines powered by natural gas. The environmental characteristics of natural gas, however, should retain this fuel's position at the forefront of desirability of all fossil fuels, while supplies last.

SEE ALSO THE FOLLOWING ARTICLES

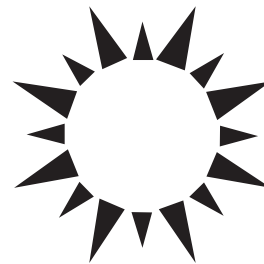
Coal Industry, History of • Electricity Use, History of • Manufactured Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Processing and Products • Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Nuclear Power, History of • Occupational Health

Risks in Crude Oil and Natural Gas Extraction • Oil Industry, History of

Further Reading

- Bragdon, E. D. (1962). "The Federal Power Commission and the Regulation of Natural Gas: A Study in Administrative and Judicial History." Ph.D. Dissertation. Indiana University.
- Castaneda, C. J. (1999). "Invisible Fuel: Manufactured and Natural Gas in America, 1800–2000." Twayne Publishers, New York.
- Castaneda, C. J., and Smith, C. M. (1996). "Gas Pipelines and the Emergence of America's Regulatory State: A History of Panhandle Eastern Corporation, 1928–1993." Cambridge University Press, New York.
- DeVane, D. A. (1945). Highlights of the legislative history of the Federal Power Act of 1935 and the Natural Gas Act of 1938. *George Washington Law Rev.* XIV (Dec. 1945).
- De Vany, A. S., and Walls, W. D. (1995). "The Emerging New Order in Natural Gas: Markets vs. Regulation." Quorum Books, Westport, Connecticut.
- Frey, J. W., and Ide, H. C. (1946). "A History of the Petroleum Administration for War, 1941–1945." Government Printing Office, Washington, D.C.
- Herbert, J. H. (1992). "Clean Cheap Heat: The Development of Residential Markets for Natural Gas in the United States." Praeger, New York.
- MacAvoy, P. W. (2001). "The Natural Gas Market: Sixty Years of Regulation and Deregulation." Yale University Press, New Haven.
- Peebles, M. W. H. (1980). "Evolution of the Gas Industry." New York University Press, New York.
- Rose, M. H. (1995). "Cities of Light and Heat: Domesticating Gas and Electricity in Urban America." University of Pennsylvania Press, University Park.
- Sanders, E. (1981). "The Regulation of Natural Gas: policy and politics, 1938–1978." Temple University Press, Philadelphia.
- Stotz, L., and Jamison, A. (1938). "History of the Gas Industry." Stettiner Brothers, New York.
- Tarr, J. A. (1998). Transforming an energy system: The evolution of the manufactured gas industry and the transition to natural gas in United States (1807–1954). In "The Governance of Large Technical Systems" (O. Coutard, Ed.), pp. 19–37. Routledge, London.
- Tussing, A. R., and Barlow, C. C. (1984). "The Natural Gas Industry: Evolution, Structure, and Economics." Ballinger Publ., Cambridge.

Natural Gas Industry, Energy Policy in



MICHELLE MICHOT FOSS

Institute for Energy, Law and Enterprise, University of Houston
Houston, Texas, United States

1. Introduction

2. Examples of Natural Gas Policy

Glossary

associated/dissolved natural gas A type of natural gas that occurs in crude oil reservoirs either as free gas (associated) or as gas in solution with crude oil (dissolved gas).

dry natural gas The natural gas that remains (1) after the liquefiable hydrocarbon portion has been removed from the gas stream (i.e., gas after lease, field, and/or plant separation) or (2) after any volumes of nonhydrocarbon gases have been removed, when they occur in sufficient quantity to render the gas unmarketable. Dry natural gas is also known as consumer-grade natural gas. The parameters for measurement are cubic feet at 60°F and 14.73 pounds per square inch absolute.

methane A colorless, flammable, odorless hydrocarbon gas (CH₄); the major component of natural gas. It is also an important source of hydrogen in various industrial processes. Methane is a greenhouse gas.

natural gas A gaseous mixture of hydrocarbon compounds, the primary one being methane. The U.S. Energy Information Administration measures wet natural gas and its two sources of production (associated/dissolved natural gas and nonassociated natural gas) and dry natural gas (which is produced from wet natural gas).

nonassociated natural gas A form of natural gas that is not in contact with significant quantities of crude oil in the reservoir.

wet natural gas A mixture of hydrocarbon compounds and small quantities of various nonhydrocarbons existing in the gaseous phase or in solution with crude oil in porous rock formations at reservoir conditions. The principal hydrocarbons normally contained in the mixture are methane, ethane, propane, butane, and pentane. Typical nonhydrocarbon gases that may be present in reservoir natural gas are water vapor, carbon dioxide, hydrogen sulfide, nitrogen, and trace amounts of helium. Under

reservoir conditions, natural gas and its associated liquefiable portions occur either in a single gaseous phase in the reservoir or in solution with crude oil and are not distinguishable at the time as separate substances. The Securities and Exchange Commission and the Financial Accounting Standards Board refer to this product as natural gas.

Natural gas has come a long way as a good with intrinsic value. This resource was once considered a mere by-product of oil and thus not worth the significant capital investment required to find, gather, treat, transport, and distribute it. The relative abundance, wide dispersion, and cleanliness of natural gas have propelled it to the forefront of the fossil fuels, so that natural gas today is poised to become a global commodity, a bridge fuel to the next energy future, and a source of molecular building blocks to new materials. The key policy challenges lie in differentiating the various markets associated with natural gas molecules and fashioning competitive supply, demand, and pricing mechanisms; designing appropriate policy approaches for components of the natural gas value chain that bear public service attributes and thus affect the public interest; and mobilizing capital investment while balancing efficiency and equity concerns.

1. INTRODUCTION

The term “natural” is used to distinguish gaseous and associated liquid hydrocarbons that are produced in the earth from those that are “manufactured,” typically from coals (and often referred to as “town gas”). Methane (one carbon and four hydrogen atoms) is the most abundant molecular component of a natural gas stream, typically comprising as much

as 85%, depending on whether the natural gas occurs with crude oil or separately. Other molecules in the natural gas stream, in varying proportions, include ethane, propane, butane, carbon dioxide, oxygen, and nitrogen, and impurities such as hydrogen sulfide and rare gases that must be removed before the natural gas can be used. Liquefied petroleum gas (LPG), a propane/butane mixture, is stripped out of the natural gas stream during field production; LPG can be shipped by pipelines or separately by truck or tanker for uses such as petrochemical manufacturing, charcoal starter fuel in backyard grills, or vehicle transport fuel. In many countries, particularly those with short heating seasons, LPG comprises the principal domestic fuel source for water and space heating and cooking. Natural gas liquids (NGLs)—ethane and larger molecules—also are stripped out as feedstocks for petrochemicals (methane can also be used instead of the heavier molecules for petrochemical applications, but much larger volumes of methane are required). The NGL molecules can also be transported by pipeline, truck, or tanker. Other than pipeline delivery, methane can be liquefied and thus transported economically over large distances in liquefied natural gas (LNG) ships or by truck; LNG cargoes may contain molecules other than methane if there is no processing at the point of liquefaction. Methane can also be compressed and transported shorter distances by truck as compressed natural gas (CNG); eventually, depending on emergence of viable technologies and cost structures, it may be possible to transport CNG on tankers. Compression and liquefaction are required for methane to be used as a vehicle fuel, whereas LPG can be used directly.

Worldwide, the natural gas industry has grown rapidly in recent years. Nations with rich natural gas resources have aggressively added new petrochemical capacity. These huge investments tend to be quite lumpy and the products are subject to intense global competition (with commensurate impacts on the feedstock molecules), leading to the well-known cyclicity in these businesses. Likewise, LNG investments are also sizable, lumpy, and subject to global forces, and are also fast growing. For nations that do not have large enough domestic demand relative to the size of their resource base, or that have not developed petrochemicals capacity for conversion of natural gas to other products, LNG is an important means of deriving value for their natural resource endowments through international trade. LPG as a domestic energy source has also grown, proving to be a relatively cheap (in terms of local infrastructure)

and clean replacement for biomass. In particular, LPG has replaced wood (and is therefore preferable in places where deforestation has been rampant) and animal dung. LPG also represents an improvement in both cleanliness and safety over kerosene, also used as domestic fuel. Where LPG is in wide use, typical policy challenges include pricing and transparency in market organization. Because LPG is most often found as a domestic fuel in poorer countries, the tendency has been toward government control of final prices charged to customers and heavy subsidization of these prices. These policy approaches are usually neither fair nor effective; they are also costly to governments. In many countries, lack of transparent market organization for LPG contributes to widespread theft, increasing distortions and injuring the consumers least able to make alternative choices.

Above all, it is the growth in demand for piped methane as a clean convenient option for consumer energy markets that has garnered the greatest interest. Thus, from hereon “natural gas” refers to pipeline delivery of methane and the myriad operational and commercial aspects involved. Generally, worldwide, natural gas/methane infrastructure systems consist of upstream, midstream, and downstream operational elements. Upstream operations involve two components:

1. Exploration. This consists of activities that lead to discovery of new natural gas resources. Exploration risk is one of the strongest forms of risk.
2. Production. Extraction of discovered supplies from hydrocarbon fields either with crude oil (as associated natural gas) or separately (nonassociated natural gas). If natural gas is associated with crude oil, it must be separated.

Midstream operations are more extensive:

1. Gathering. Collection of natural gas production from multiple wells connected by small-diameter, low-pressure pipeline systems and delivery to a processing plant or long-distance pipeline.
2. Processing and treatment (if necessary). Processing is the separation of heavier molecules and unwanted substances such as water from a methane gas stream. If the gas stream contains impurities such as hydrogen sulfide, then treatment is required.
3. Storage. Containment of supplies, usually in depleted underground reservoirs or caverns such as those associated with salt domes. Storage can be located either near production or near demand.
4. Transportation. Delivery of gas from producing basins to local distribution networks and

high-volume users via large-diameter, high-volume pipelines. In countries that are federal republics, pipeline systems may be distinguished by whether they cross state or provincial boundaries (for example, in the United States, interstate pipelines as opposed to intrastate systems that operate within the state jurisdiction). Countries vary greatly with respect to allowable pipeline specifications for heat content, as measured by the British thermal unit (Btu) in the United States, Canada, and elsewhere, and as related to the presence of molecules other than methane in the piped stream. For example, pipelines transport and distribute methane or “dry” gas in Canada and the United States, which means that heavier molecules are removed before the natural gas stream enters the pipeline system. Exceptions do exist, such as the Alliance Pipeline, which transports “wet” gas from British Columbia to Chicago, where molecules other than methane are stripped out in processing for use in other markets. Pipeline standards generally are set for safety reasons.

5. Liquefaction, shipping, and regasification. Known collectively as the LNG value chain, this entails conversion of gas to a liquid form via refrigeration, resulting in a cryogenic fluid (temperature -256°F) for transportation from a producing country or region to a consuming country or region via ship. LNG is stored until it is returned to the gas phase (regasification, using vaporization) for pipeline transportation within the consuming region. In the United States, LNG is also used as a storage form of natural gas produced from domestic fields, until the gas is needed, primarily for peak use. Both storage and transport of LNG are done at nearly atmospheric pressure.

Finally, the infrastructure system consists of downstream operations:

1. Distribution. Retail sales and final delivery of gas via small-diameter, low-pressure local gas networks operated by local distribution companies (often termed gas utilities).

2. End use and conversion. Direct use or conversion for use in other forms (petrochemicals, electric power, or vehicle fuels).

The following commercial elements serve to bind the operating segments of the natural gas infrastructure system, linking suppliers, transporters, and distributors with their customers:

- Aggregation. Consolidation of supply obligations, purchase obligations, or both as a means of

contractually (as opposed to physically) balancing supply and demand.

- Marketing. Purchase of gas supplies from multiple fields and resale to wholesale and retail markets. Retail marketing constitutes sales to final end users (typically residential, commercial, industrial, electric power, and public sectors).

- Capacity brokering. Trading of unused space on pipelines and in storage facilities.

- Information services. Creation, collection, processing, management, and distribution of data related to all the other industry functions listed here.

- Financing. Provision of capital funding for facility construction, market development, and operation start-up.

- Risk management, Balancing of supply, demand, and price risks.

Altogether, the upstream, midstream, downstream, and commercial elements constitute the natural gas value chain. The various segments are highly interdependent but, in an open, competitive market, they also can be highly competitive. The policy challenges associated with increased worldwide use are numerous. Frameworks for efficient discovery and optimal production are the first hurdle. Efficient and equitable mechanisms, often at odds, for pipeline transportation and local distribution are the second major hurdle. Methane is of little use in consumer energy markets without pipeline infrastructure. These large systems tend to be characterized by strong technical economies of scale and high barriers to entry. Particular problems also emerge with respect to the public interest/public service component of these facilities, mainly with respect to reliability and pricing of systems that are usually operated in monopoly, duopoly, or limited competitor regimes. In all cases, a specific policy quandary is how best to achieve prices for natural gas transportation and distribution through tariff designs that yield something close to what competitive markets might be able to achieve, with contestability (potential competition) usually providing a basis for market-based pricing in larger markets.

A third challenge is development of transparent markets for natural gas supply and consumption. If pipelines are an essential feature, a central question is whether molecules have intrinsic value or whether the combination of pipeline and molecule together must constitute the service. Increasingly, the trend has been to separate infrastructure and product (often termed “unbundling”) and to search for ways

of providing competitive access to pipeline systems for multiple suppliers and users of natural gas (often termed “third-party access” or “open access”). In these cases, pipelines become like toll roads, priced through tariff design, whereas molecules become priced in discrete, competitive markets. When pipelines become subject to third-party access regimes, the commercial activities described earlier and associated with pipeline operations, i.e., the linking of suppliers to buyers, can be separated into competitive business activities. Once subjected to competitive markets, methane molecules become commoditized. With all market participants as price takers, new sources of risk and new policy challenges arise. In recent years, the rapid commoditization of methane in the United States and Canada triggered growth in marketing and trading (both of the physical product as well as financial derivatives) as separate businesses. With methane a commodity, and pipeline capacity close to being commoditized, balancing supply and demand across a market becomes both more efficient and more fragmented, as new combinations of activities and relationships across the value chain become established. The fuel’s growing importance in the international economy, as natural gas becomes globalized via LNG shipments and disparate national and regional markets become linked, has meant new incentives for technical improvements in supply and demand management and balancing. The principal segments of a natural gas infrastructure system—exploration, production, transportation and distribution—share substantial capital requirements and comparable, albeit different, risks. The long lead times required for development of each sector’s assets present both industry leaders and policymakers with the problem of adequately anticipating changes in supply and demand. These projections must be accurate and timely to attract the necessary long-term investment consistently and to minimize market disruptions and distortions. Inadequate projections create conflicts to the extent that they result in supply–demand imbalances, which neither industry nor government has the flexibility to correct in a timely manner.

Both the international trade linkages and the evolution of market-based policies for natural gas mean timely and accurate data and information on supply, demand, and prices, a fourth requirement. A fifth and increasingly complicated challenge is dealing with integration, with respect to industry organization and international trade. Industry organization can encompass both vertical (meaning

up and down the value chain) and horizontal (meaning over some geographic or market extent) integration. Paradoxically, the forces for integration within a natural gas industry often occur in spite of policy objectives that seek to instill deintegration and competition as part of transitions to competitive markets. Integration of physical infrastructure across international boundaries has grown rapidly with increased demand for piped methane. As transportation and information technologies have improved, so have the opportunities for system linkages—first within a country, then among geographically contiguous nation states, and increasingly across the globe. With improved physical and commercial linkages comes an ever greater need for more complex, sophisticated, and coordinated policy solutions, posing new dilemmas in international trade.

Natural gas policy across nations is most easily differentiated by how the value chain is organized and operates with respect to the balance between markets and government—that is, sovereign ownership of or control over the critical segments of the value chain. The high degree of interdependency across the value chain segments, the propensity toward integration, the fact that, in most cases, large deposits of natural gas are associated with oil (a strategic commodity for many producing nations), the energy security aspects of natural gas supply and delivery, and the public interest/public service concepts embedded in pipeline infrastructure have all resulted in a strong pattern of government ownership, control, or, in the least, intervention in natural gas enterprises and markets. In most countries, the natural gas value chain has been developed through integrated, sovereign owned, or heavily controlled enterprises. The rare exceptions are the United States, Canada, and Australia, all of which have allowed the natural gas system infrastructure to emerge through the activities of private, investor-owned companies. For the United States and Canada, this experience extends roughly 100 years; for Australia, this has been the practice since the 1960s. The United States is even more unique in that private ownership of much of the resource base is allowed; indeed, it is a powerful tradition. Federal and/or state ownership of the natural resource is limited to certain onshore lands and offshore, and even in these cases, development of sovereign-owned and controlled resources has always been through competitive acquisition of leases in organized auctions and private investment in exploration and production (a tradition also maintained in Australia,

where states and territories play a comparable role, and in Canada, which has limited private ownership of natural gas resources in southern Alberta province; the vast majority of the resource base, both onshore and offshore Canada, is controlled by the provincial crown governments). Of great interest is the transition in many countries away from sovereign ownership and/or control of integrated natural gas enterprises, as well as the push for ever more competitive markets in many locations, including Canada and the United States.

To a large extent, these transitions have been driven by specific needs for increased efficiency and to introduce innovations, to solve fundamental problems in pricing and service, and to attract investment into the natural gas value chain. As with any industry, natural gas market development requires sufficient supply availability and enough demand to justify the infrastructure systems to connect buyers and sellers. To attract this investment, governments have experimented with policies designed to stabilize the investment environment by optimizing participant choice at predictable prices that reflect, or at least attempt to mimic, actual supply and demand conditions. During the past 15 years or so, the progression toward competitive markets has meant movement toward market determination of investment, and operation of assets subject to real-time supply-and-demand interactions. Under these conditions, actionable information must be timely, accurate, and transparent. For such information to be truly actionable, the decision maker also must have timely access to whatever system capacity the information prompts that decision maker (supplier, customer, or intermediary) to demand. Finally, competitive markets must comprise systems wherein this information and capacity cannot be dominated or manipulated by a few anticompetitive participants. These conditions are difficult to create, implement, and maintain, and imply new and changing roles for market participants and government overseers.

2. EXAMPLES OF NATURAL GAS POLICY

2.1 United States

2.1.1 Overview

Natural gas exploration and production on private lands, including environmental and safety controls, are regulated at the individual state level by

conservation commissions. Exploration and production on state lands are controlled by separated state agencies charged with management of those lands. Exploration and production on federal lands, onshore or offshore, are managed by federal agencies. Natural gas exploration and production, gathering, and processing are all viewed to be “workably competitive” industries and are not regulated for prices. Tariffs for transportation within state boundaries on intrastate pipelines are regulated by state public utility commissions (PUCs). The PUCs also license new intrastate pipelines. Tariffs for transportation across state boundaries in interstate pipelines are regulated by the Federal Energy Regulatory Commission (FERC). The FERC also issues licenses (certificates of public need) for new interstate pipelines. Tariffs for natural gas distribution to final customers are regulated by PUCs. The most competitive gas service in the United States is for industrial customers. The least competitive service is to residential customers.

The FERC is governed by five appointed commissioners and operates as an independent authority. Enabling legislation for the FERC dates back to the 1930s (it was created as the Federal Power Commission and charged principally with development of water and hydroelectric facilities). The FERC’s authority to regulate interstate natural gas commerce is embodied in the 1938 Natural Gas Act. The individual state PUCs were established at various times, generally between the late 1800s through the 1970s. Each state has a separate enabling legislation for the formation of its PUC. Notable exceptions are the state of Nebraska, which does not regulate natural gas, and Texas, where natural gas is regulated by the Texas Railroad Commission rather than the PUC. The PUCs also vary with regard to numbers of commissioners, whether commissioners are elected or appointed, and sizes of staffs and budgets. The FERC and state PUCs are funded through fees charged to regulated industries.

The style of regulation in the United States traditionally has been “cost of service” or “rate of return,” which involves a determination of revenue requirements and rate structures based on costs provided by the regulated firms. A regulated company may be a local distribution company (gas utility), an intrastate pipeline, or an interstate pipeline. The regulated company’s revenue requirements are the total funds that the company may collect from ratepayers (customers). Revenue requirements are calculated by multiplying the company’s rate base by an allowed rate of return (ROR) and adding this

product to the company's operating costs (OCs), as shown in the following formula:

$$\text{Revenue requirement} = (\text{rate base} \times \text{ROR}) + \text{OC} \\ - (\text{taxes} + \text{depreciation})$$

The rate base is the total value of the company's capital investments, which may include construction work in progress. The allowed rate of return constitutes a profit sufficient to pay interest on accumulated debt and to provide a "fair" return to investors. A fair return is determined through a comparable earnings test (whereby a company's earnings are measured against those of a firm facing comparable risks), a discounted cash flow approach (whereby a company's capital costs are estimated by analyzing conditions in the financial market), or some other method. Operating costs include expenses for purchased gas, labor, management, maintenance, and advertising. The costs of taxes and depreciation are also part of a company's revenue requirements. The regulatory process can be generally described as follows:

- A regulatory commission (a PUC or the FERC) first seeks to determine how much of an applicant's capital stock should be included in the rate base, then attempts to determine which elements of test-year costs and revenues should be allowed for regulatory purposes and whether to allow specific changes since the test year. The final step is to determine what the fair rate of return is for the company.

- States and the FERC have legal rules for deciding what should be included in the rate base, although the same is not necessarily true for the method of calculating allowed rate of return.

- States may vary from each other and from the FERC according to the particular sets of rules that are used (for example, to calculate rate base) and the impact of these rules on rate case decisions. However, over the course of the long history of natural gas regulation in the United States, the states and the FERC have generally shared practices fairly quickly.

- All regulators are constrained in their abilities to calculate cost of capital. This is due in part to general disagreement within the industry of how market cost of capital should be computed, and in part because commissions are not well equipped to deal with the complexities surrounding these issues. As a result, a critical component of a rate case proceeding is a commission's reliance on historical information, or precedent, as well as the testimony of interveners, parties with specific interests in the outcome of rate cases (principally large customers and consumer

advocates representing small business and residential users; competing regulated firms may also intervene).

- All U.S. regulatory commissions hear rate cases, issue blanket rulings that set broad policy parameters, and act as judges and adjudicators on disputes.

With the implementation of unbundling (separation of pipeline transportation from natural gas sales, with nondiscriminatory open access or third-party access) in 1992, the FERC and many of the states now encourage market-based rates for transportation service. With respect to distribution, many states are experimenting with "incentive-based" regulation designed to encourage more efficient operation and capital cost decisions than has historically been achieved with cost-of-service regulation.

2.1.2 History

When comparing the United States to other countries, an important difference is that the U.S. natural gas system has always been characterized by the participation of private companies, with regulation as a substitute for competition to moderate private monopoly market power. For the most part, regulation has been directed toward pipeline transportation and local distribution. However, during periods of U.S. history, cost-of-service-style regulation was also applied to natural gas production at the wellhead with disastrous results. Table I illustrates the complex progression of regulatory policy eras in the United States.

In the early days of the U.S. natural gas industry, the construction and operation of natural gas distribution systems tended to be concentrated around local deposits of natural gas. Cities and towns that were near the early discoveries of natural gas in the late 1800s were often the centers of intense competitive activity as companies struggled to build competing systems. Because all of the early natural gas companies were private and because the intense competition reduced the returns to shareholders, state-level regulation of local distribution companies (LDCs) through public utility commissions evolved. The strategy was to stabilize investment returns to shareholders while attempting to mimic most of the benefits of competition to customers through regulation (competition by substitution, as it is often called). The form of regulation typically used was "cost of service," in which regulators granted a rate of return that was deemed to be reasonable to the LDCs. In exchange for a limited return, the LDCs enjoyed a monopoly franchise for service in a city or

TABLE I

Natural Gas Industry Regulation/Deregulation in the United States

Policy history	Impact
Competitive local distribution company industry	Emergence of natural gas industry as local distribution companies were established to provide town gas and later natural gas service
State public utility commissions, 1885–1927	Formation of Massachusetts Gas Commission in 1885 through regulation of intrastate pipelines (in all 48 states by 1927); followed the Supreme Court ruling (<i>Munn v. Illinois</i> , 1877) that established the basis for regulating monopolies (grain elevators and warehouses) as public utilities
Development of interstate transportation, 1930s	Technological advances (mechanized trenching and arc welding) allowed construction of long-distance pipelines to transport natural gas from large producing fields in the southwestern United States to key market in the Northeast and upper Midwest
Federal regulation of interstate transportation (Public Utility Holding Company Act and Federal Power Act of 1935; Natural Gas Act of 1938)	Interstate Commerce Act of 1887 provided the basis for federal intervention. A U.S. Supreme Court decision in 1934 (<i>Nebbia v. New York</i> , dealing with milk prices) expanded the basis for public utility regulation. Disputes centered on pricing natural gas in cross-state sales activities and market power of interstate public utility holding companies. The Federal Power Act established and authorized the Federal Power Commission; the natural gas industry was thus comprehensively regulated, from the burner tip, to intrastate transmission, to interstate transmission, by state and federal jurisdictions
Federal regulation of wellhead prices (U.S. Supreme Court <i>Phillips</i> Decision, 1954)	Dispute regarding pricing of natural gas produced in Oklahoma for delivery in Michigan led to cost-of-service regulation at the wellhead, with the FPC as agency with authority
Beginning of wellhead price decontrol (Natural Gas Policy Act, 1978; Public Utility Regulatory Policy Act of 1978; Powerplant and Industrial Fuel Use Act of 1978)	The FPC's inability to deal with the scope of wellhead regulation and provide sufficient adjustment to increase price ceilings and encourage production, as well as disparity in pricing natural gas sold in interstate markets, with gas sold in unregulated intrastate markets, and resulting curtailments, all led to decontrol. The Natural Gas Policy Act of 1978 extended wellhead price ceilings to the intrastate market, introducing the process of deregulation by loosening certification requirements to facilitate gas flows
First stage of open access for interstate pipelines (FERC Orders 436 and 500, 1985)	"Phased decontrol," with surplus conditions created by the Natural Gas Policy Act; the need for flexible pricing and transportation led to "special marketing programs" that released gas from long-term contracts into price-discounted supply pools. FERC Order 436 in 1985 created the open-access era and Order 500 provided some resolution for take-or-pay liabilities
Final restructuring rule for interstate pipelines (FERC Order 636, 1992)	FERC Order 636 continues separation of merchant and transportation functions of interstate pipelines
Regulation of interstate transportation and related services (FERC Order 627, 1999)	In Order 637, the FERC moved to improve competitive markets, mainly targeting capacity and operations for captive customers and to deal with pipeline affiliate issues

town but also had the obligation to serve all customers within that franchise.

The discovery of huge natural gas deposits in Texas and Oklahoma fundamentally changed the U.S. natural gas industries. Companies began to build long-distance pipelines to carry natural gas from the southwestern United States to the Northeast

and Midwest, where gas was needed for winter heating. Almost immediately, disputes arose among individual states with regard regulatory jurisdiction over interstate sales of natural gas. By 1938, the U.S. government was prepared to step into the conflict. Passage of the Natural Gas Act (NGA) that year gave the Federal Power Commission (FPC) regulatory

authority over interstate natural gas commerce. This action was consistent with the philosophy of the times. Following the Great Depression, there was considerable mistrust of large businesses and greater faith in the ability of government to intervene and solve problems.

The NGA treated the interstate pipelines as natural monopolies. (It should be noted, however, that in debating the NGA, the U.S. Congress deliberated on contract carriage as an alternative approach.) The economics of early pipeline construction and operation and conditions in the early natural gas markets were considered to be such that it was unlikely for many companies to build competing facilities. As a result, the pipelines acted as “merchants,” contracting with natural gas producers for supply and also with local distribution companies for deliveries. Disputes related to the price of natural gas in the interstate market did not end, however, and the federal government intervened again, this time through the Supreme Court. In the landmark *Phillips* decision in 1954, the Court concluded that the FPC should also have regulatory authority over the price of natural gas at the wellhead. By this time, thousands of natural gas wells had been drilled in Texas, Oklahoma, Louisiana, and other states, including the beginnings of the U.S. offshore industry. The task faced by the FPC was daunting, and the ability of federal regulators to perform efficiently was limited. Distortions began to show up immediately, most importantly in the difference between prices for natural gas in the regulated interstate market and prices in the unregulated intrastate market (meaning gas produced and sold within the boundaries of individual states). Demand for natural gas had grown and prices were rising in the intrastate market. As a consequence, producers shifted their strategies so as to sell more gas in that market. By the time of the oil embargoes and supply shocks in the early 1970s, insufficient amounts of natural gas were committed to the interstate market. During the severe winter in 1976, shortages and curtailments of natural gas supplies occurred all over the eastern United States. To make matters worse, because the interstate pipelines controlled all transactions, there was no way for natural gas producers to engage in sales directly with customers.

Broad dissatisfaction with how the natural gas sector was managed led to an unwinding of federal regulatory control. By the 1970s, public opinion regarding government management of economic activity, including energy, had begun to erode. The

preference for market-based solutions was increasing. Already the United States was engaged in major transformations to reduce government intervention in other sectors, such as airline transportation, telecommunications, and banking. The first step was to remove regulatory control of natural gas at the wellhead, with the Natural Gas Policy Act (NGPA) of 1978, which also transformed the FPC into the Federal Energy Regulatory Commission. The strategy chosen by the U.S. government was flawed, with a tremendously complicated schedule for decontrol of natural gas from different formations, by year of discovery (vintage), and so on. The U.S. Congress created more than 200 different categories of natural gas. Overall, natural gas prices rose rapidly in response to demand. After a period of time, higher natural gas prices caused demand to fall as customers, especially large industrial users, shifted to cheaper fuels. Demand adjustments were further complicated by two additional laws that had been passed to deal with 1970s energy crises, the Public Utility Regulatory Policy Act (PURPA) and the Powerplant and Industrial Fuel Use Act (PIFUA), which, together with the NGPA, encompassed the National Energy Act. PURPA encouraged experimentation with “cogeneration” of electricity using natural gas at industrial facilities, which sold their electric power to electric utilities at a price that (presumably) reflected the costs utilities would avoid by not building new electricity generation capacity themselves. PIFUA, however, prohibited natural gas use in most industrial applications and by the electric utilities. The resulting confusion and fluctuations in prices created havoc on both sides of the interstate pipeline merchant contracts. Both pipelines and producers were left holding contracts with take-or-pay obligations that led to severe financial strain for many companies. Because interstate pipelines still acted as merchants, bottlenecks existed all over the natural gas system, preventing efficient transactions from taking place.

In 1983, the FERC began to put into place the policies that have led to the restructured natural gas market that we see in the United States today. Through a series of actions, the FERC began to dismantle the interstate pipeline merchant function. Pipelines came to be treated as common carriers, conduits through which any seller or buyer could ship gas. Natural gas came to be treated as a commodity, whereas previously oil companies had treated natural gas as a by-product with no intrinsic value. Pipeline construction technology had changed dramatically over the years, and many parts of the

United States were served by more than one interstate pipeline, reducing the natural monopoly advantage. The advent of sophisticated computer information systems allowed real-time trading of natural gas, and financial instruments (futures contracts and other mechanisms) enabled suppliers and customers to manage commodity price risk. Although conditions today are vastly different—natural gas in recent years has enjoyed a growing share of the U.S. energy mix, albeit with great price elasticity for certain kinds of demand—there is no doubt that policies in the past constrained market growth of the industry. Indeed, the position that Canada enjoys as a major exporter of natural gas to the United States (approximately 15% of U.S. consumption) is a direct outcome of the 1976 shortages.

2.1.3 Issues

Several issues remain following restructuring to restore and enhance competition in the U.S. gas system.

2.1.3.1 Market Disruptions The FERC's actions to implement open access on U.S. interstate pipelines created a "wholesale market" for natural gas, with competitive pricing, trading, and marketing activities; price risk management (the New York Mercantile Exchange established a natural gas futures contract in 1993); market mechanisms to facilitate trading of unused pipeline capacity; and a national standards board (the Gas Industry Standards Board) to facilitate commercial activity. Beginning in 2000, surging prices for natural gas and electric generation constraints as a result of extended drought in the Pacific Northwest resulted in collapse of the electric power market in California. Disparities in natural gas prices between the California internal market and other U.S. locations and related improprieties in natural gas trading, along with the bankruptcy of Enron Corporation, led to a general collapse in the "energy merchant" segment, including credit downgrades, additional bankruptcies, and severe losses in market capitalization. Energy merchant businesses deal in unregulated wholesale market activities, including construction and operation of competitive, unregulated infrastructure assets. Many energy merchants were affiliated with regulated natural gas interstate pipelines and utilities. Continued conflict around issues that emerged during these events heavily impacted natural gas markets in the United States and Canada. A subsequent natural gas spike in 2003, with ancillary concerns regarding natural gas field production trends and disputes regarding how natural gas price information is compiled and

communicated in the marketplace, continue to retard further policy, regulatory, and commercial activity.

2.1.3.2 Lack of Competitive Service Lack of competitive service to small residential and commercial customers occurs at the end of distribution systems. Unbundling and open access to facilitate deliveries of competitive supply for smaller customers have not materialized as expected following FERC's restructuring rule. These initiatives generally are in the domain of state jurisdictions, but federal/state coordination, always an issue, is required in some instances.

2.1.3.3 Barriers to Entry of New Pipelines The certification process, already deemed to be too onerous given competitive market conditions, has been further complicated by the market disruptions of the 2000s and lack of financial capital available from distressed energy merchant businesses and parent pipeline companies.

2.1.3.4 Uncertainty with Regard to Electric Power Restructuring Following early success with natural gas, some states (notably California in 1994) and the FERC (through Orders 888 and 889), and with encouragement through the Energy Policy Act (EPA) of 1992, proceeded to experiment with similar unbundling and open-access approaches for electricity grids. The 1992 EPA supported creation of a bulk, wholesale market for electric power. The electric power market collapse in California and natural gas market disruptions have stymied further initiatives by the FERC for a segment of the energy sector that is considered to be crucial to growth of the natural gas industry.

2.2 Canada

The Canadian natural gas system parallels that of the United States, with an important exception. All natural gas resources in Canada are controlled by the provincial crown governments. Exploration and production activities are carried out by private, competing firms under the rules and regulations established by provincial energy ministries. This contrasts with the United States, where roughly two-thirds of natural gas resources and production are in the private domain (held either by companies or individuals).

Like the United States, Canada's transportation and distribution systems are owned and operated by private (investor-owned) companies regulated to

control for any monopoly power. Regulation of long-distance, interprovincial pipelines is carried out by the National Energy Board (NEB), which receives its authorization from the federal National Energy Board Act. The NEB, like the FERC, licenses new pipelines, sets tariffs for transportation, adjudicates disputes, and sets broad policy parameters with blanket rulings. Intraprovincial pipelines and local distribution systems are regulated by provincial energy utilities boards (EUBs). Each province has enabling legislation for its EUB. As with the state PUCs and the FERC in the United States, the EUBs and the NEB use similar methods for regulating their client industries, and are funded by these industries. The NEB and EUBs have also, traditionally, used cost-of-service ratemaking like the U.S. commissions do. Canada is a significant exporter of gas to the United States, supplying about 17% of U.S. (the lower 48 states) current demand. The NEB licenses and regulates all natural gas export activity from Canada.

Canada began to restructure its natural gas system ahead of the United States in the early 1970s with the Western Accord, which eliminated control of the natural gas supply by Canada's monopoly interprovincial pipeline, TransCanada. The Agreement on Natural Gas Markets and Prices in the 1980s was a statement in principle of support for a market-based natural gas system. The Open Access Order in 1986 unbundled the Canadian system and allowed contract carriage on Canada's pipelines. Since these steps were taken, the NEB has consistently encouraged market-based rates for transportation.

At the provincial level, EUBs followed the NEB with open access and market-based tariffs. Nearly every local distribution system in Canada offers some form of competitive supply to its core customers (residential and small commercial), with LDCs in Ontario moving toward full open-access systems for core customer service.

2.3 Britain

Like the United States and Canada, Britain's national gas grid evolved to transport "town gas" manufactured from coal by the state-owned British Gas (BG). However, unlike its North American counterparts, Britain's grid was not converted to natural gas until the 1960s. As the first gas came ashore from West Sole field in the North Sea in 1967, British Gas began the process of transforming its business. By comparison, discovery of natural gas in the United States at the turn of the 20th century, first in the Appalachians

and later in the Southwest (Texas and Oklahoma), had already launched the U.S. natural gas industry. Britain has moved much more quickly than the United States to allow competition to residential and small commercial customers.

Restructuring of the British natural gas sector involved a series of steps:

1. Passage of the Oil and Gas (Enterprises) Act of 1982, which laid out the process for privatization.
2. Privatization of British Gas with the Gas Act of 1986 and creation of The Office of Gas (Ofgas) as regulator.
3. Initiation of competition in the contract market (large industrial and electric utility users who use more than 2500 therms/year) in 1989.
4. Accelerated competition in the contract market, with government targets and issuance of licenses to qualified customers taking more than 2500 therms/year at a single site in 1992.
5. Creation of the gas cost index in 1992 to limit cost passthroughs by British Gas to domestic customers who use 2500 therms/year or less.
6. Passage of the Competition and Service (Utilities) Act in 1992, which included natural gas service standards.
7. Passage of the Gas Act of 1995, which laid out targets for full competition in the domestic sector by 1998.
8. Creation of the Network Code in 1996 to ensure smooth operation of the liberalized United Kingdom gas industry.

The Oil and Gas (Enterprises) Act was passed in the year 1982 to set out the aims and objectives of the privatization of British Gas. The main issue was to break down the monopoly British Gas had in the gas supply market, so enabling the introduction of competition in gas supply. The year 1986 saw the implementation of the Gas Act, which brought about fundamental changes necessary for the privatization program to go ahead. First, Ofgas, the Office of Gas Supply, was established as the regulatory body over the gas industry. Its role is to monitor the privatization of the gas industry and to ensure customers' rights are recognized in every aspect. Second, British Gas was reorganized into a public limited company (Plc) as a monopolist with secure markets only for the medium term. BG Plc shares were released onto the stock market.

The mechanism for competition within the gas market had been implemented, with major new private sector companies created. Some were established by oil companies active in the North Sea;

others are ventures by the electricity companies and some are truly independent. The Gas Act emphasized the need for vigorous competition among the new independent suppliers with the objective of undercutting British Gas prices. The goal was to introduce competition in stages, with the industrial sector first, followed by commercial and then residential users.

2.3.1 Development of the Nonfranchise Market

In 1989, the first competition for large contract customers in industry and power generation appeared. The gas supply market did not really open up until 1992, when gas consumers buying over 2500 therms/year (million Btu, or mmBtu, per year) at a single site qualified to convert to independent suppliers of gas. The government's target was to have 60% of the contract market open to competition. By 1993, independents claimed about a 50% share of the contract market. In 1995, Ofgas reported that British Gas held a roughly 27% share. The complete breakdown of British Gas' supply monopoly occurred 1998, when every consumer of gas, including approximately 19 million domestic users, was given the option to choose their own supplier. Legislation approved in 1995 (the Gas Act of 1995) allowed liberalization of the residential market to begin in 1996.

Competition in the domestic gas supply market was introduced in stages. Under an amended Gas Act 1995, Ofgas became responsible for securing competition in gas supply to domestic customers, and issued licenses to competing companies for the transportation, shipping, and supply of gas. The Ofgas Technical Directorate enforced regulations on the quality and calorific value of gas, approved and stamped gas meters, and tested disputed meters. It also provided advice on a wide range of other technical issues. The principal job was to ensure British Gas did not take unfair advantage of its monopoly powers. This was done by limiting the prices the company could charge and setting standards for customer services. Two price controls were used (modifications have been made since restructuring was initiated). The first was a transportation and storage price control, which capped the prices TransCo, which is part of BG Plc and is the public gas transporter and operator of the pipeline system, can charge British Gas Trading and other users for transporting gas in its pipelines and storing gas. The pipeline system is operated on a contract carriage basis (suppliers arrange for transportation from beach to final customer), but with the recognition that pipeline transportation is monopolistic. The

second control involved a supply price control that limited the prices British Gas Trading, now part of Centrica, could charge domestic consumers.

With the initiation of privatization in 1992, a gas cost index was used as part of the regulatory formula that limited BG's allowable revenue for the 18 million customers who use 2500 therms or less of gas per year. Ofgas introduced the index after reviewing the price control set by the government when BG was privatized in 1986. This had allowed BG to pass through to domestic customers the average actual cost of buying gas for all customers, which is about 40% of the total allowed price of gas. With introduction of the index, this part of the allowed price became based on the level of the index and was no longer linked to actual gas costs. In addition to price and service regulation, Ofgas oversaw conformance with the Network Code, established in 1996, which sets out the rights and responsibilities for TransCo and all gas shippers using the existing grid for transportation and storage. Under the code, shippers are liable for any imbalances in their daily shipments. With the network code fully operational by 1998, tariff formulas for gas suppliers will no longer apply. Gas prices will be left to market forces, including those charged to Britain's 18 million small commercial and residential customers.

2.3.2 Implementation of Domestic Competition

Key to final implementation in 1998 of domestic competition was the 1997 price control review for BG's domestic supply business. Proposals developed by Ofgas in the summer of 1996 allowed the average domestic customer to realize a reduction in gas bills, with continuing real price reductions for the following 2 years. In addition, TransCo put forth proposals for transportation/distribution tariffs. The main proposals were (1) a 3-year price control period, (2) continuation of the retail price index, (RPI) $-X$, price control on supply costs, with X set at 5% (currently 4%), (3) direct passthrough to customers of British Gas' gas purchase costs and transportation costs, (4) service standards to remain as now, but with increased compensation payments when specified standards are not met, and (5) scope for British Gas to make a profit margin of 1.5% on turnover built into prices to allow a reasonable return to shareholders, as British Gas faces the introduction of competition into the domestic market. The price of gas was made up of three components: the cost of purchasing gas from producers, the costs of transportation, and supply costs. The proposals focused on six areas in particular: the treatment of gas costs,

transportation charges, BG's operating costs in its domestic supply business, BG's profit margin, the structure of the price control, and the duration of the price control.

2.3.3 Issues

A number of primary issues were associated with Britain's gas restructuring. Take-or-pay (TOP) contract resolution was essential. As with U.S. and Canadian companies before restructuring, BG held contracts with North Sea gas producers for supply. With falling gas prices, as a consequence of competition, BG and its suppliers faced a situation in which the value of gas in the contract was higher than it was in the marketplace, rendering the contract unsustainable; new competitors in the United Kingdom market were able to contract for the cheapest gas supplies, thus renegotiation of the TOP contracts was necessary for implementation of restructuring, but the issues over reapportioning costs became politically difficult.

Setting the value of "X" in the price cap formula was difficult. Determination of this value relative to performance of gas companies is not simple, although it is less conflictual compared to establishment of the allowed rate base for rate of return (or cost of service) regulation traditionally used in the United States and Canada. Two other issues involved the low margins to gas suppliers and the slow implementation of domestic competition. With the onset of competition and falling gas prices, suppliers in the United Kingdom faced low and diminishing profit margins in their businesses. As in the U.S. and Canada, fear and uncertainty about reliability of gas service and the ultimate cost of service, particularly to residential customers, has delayed full implementation of domestic competition.

There was also the issue of coordination with other European initiatives. Britain leads Europe in creating a competitive natural gas market. In 1997, the European Union Council established a directive allowing large customers (above 25,000 cubic meters) to select from competitive suppliers with a target of the year 2000 for implementation by member states. Initially, each member state is to grant third-party access (TPA) to 20% of the market (reducing the customer threshold if necessary), with TPA reaching 28% after 5 years and 33% in 10 years. States have moved very slowly with this directive, with few countries establishing independent regulators or forcing incumbent monopolies to respond in spite of court actions.

TABLE II

Natural Gas Policy in Latin America

Country	Private participation in exploration and production	Private participation in pipelines	Independent regulator
Mexico	No	Yes	Yes
Colombia	Yes	Yes	Yes
Venezuela	Yes	No (pending)	No (pending)
Brazil	Yes	Yes	Yes
Bolivia	Yes	Yes	Yes
Peru	Yes	Yes	Yes

2.4 Latin American Experience

The Latin American experience with natural gas policy is summarized in Table II. During the past decade or so, experimentation in the region was active in tandem with other economic and political reforms. The disintegration of Argentina's economy and uncertainty surrounding other country regimes, as well as dissatisfaction with results of market reforms, have, as of this writing, posed a number of stumbling blocks to continued development. The region is rich in natural gas resources, but much of this supply is "stranded" as a result of inadequate domestic markets and the expense of exporting natural gas in the form of LNG (Trinidad and Tobago being a notable exception). With respect to exploration and production, Latin American countries have been characterized by the presence of sovereign, state-owned oil companies, such that privatization was an initial requirement in many cases (Mexico is unique in not pursuing this step). Argentina and Mexico represent distinct cases that reflect the range of issues in the region.

2.4.1 Argentina

Until its recent financial and economic travails, Argentina had made great strides to build a competitive market for natural gas. Most of that country's initiatives have survived the depression of the past few years, but market participants have suffered throughout the period of energy economic malaise.

In the late 1980s, the government of Argentina began to privatize state-owned energy companies as part of an economic reform drive to combat hyperinflation and a chronically underperforming domestic economy. These efforts included the

privatizations of Gas del Estado, which controlled the Argentine natural gas transportation and distribution grids, and Yacimientos Petroliferos Fiscales-Argentina (YPF-A), the national oil and gas company. Argentina now has two main transportation companies, Transportadora del Gas Norte (TGN) and Transportadora del Gas Sur (TGS), both owned and operated by consortia of Argentine, U.S., Canadian, and European companies. Eight investor-owned distribution systems are now in operation, also with combinations of Argentine and foreign direct investment. The transportation system and portions of the distribution systems operate under open-access conditions following the Canadian and U.S. systems. As in Canada and the U.S., physical, economic bypass is allowed.

The regulatory framework in Argentina is provided by Ente Nacional Regulador del Gas (ENARGAS), an independent national commission. ENARGAS maintains several objectives:

1. To protect consumers' rights.
2. To promote competition.
3. To guarantee long-term investment.
4. To regulate distribution and transport services.
5. To guarantee fair and nondiscriminatory tariffs.

For tariff regulation, ENARGAS uses a variation of the British RPI-X methodology and incentive mechanisms to reward performance by the private operators. Prices to final customers are unbundled and the passthrough method is used to guarantee price transparency. The formula is specified as follows:

$$\begin{aligned} \text{Total price} &= \text{gas price} + \text{transportation price} \\ &\quad + \text{distribution margin} \end{aligned}$$

Subsidized tariffs are allowed for certain customers but must be justified to the Presupuesto Nacional, the federal budget authority. No dumping of gas is permitted. Tariffs vary with distance and type of contract (firm or interruptible) as in the United States, Canada, and Britain. It is possible to adjust distribution tariffs as often as every 6 months in correlation with the U.S. price index and productivity factors. (Because the Argentine dollar was pegged to the U.S. dollar as an antiinflation strategy, U.S. price movements are used in a variety of Argentine economic sectors.) It is also possible to adjust distribution tariffs seasonally. The overall tariff structure is revised every 15 years.

ENARGAS has several roles (regulator, enforcer, judge) and functions principally to advise users; publish information on the industry's evolution and

maintain a comprehensive library; approve transportation and distribution tariffs; authorize the building and expansion of the transportation and distribution networks; foster open access and non-discriminatory treatment; normalize security and gas quality, measurement, and odorization processes; establish the rules for new permits; implement sanctions; perform regular inspections of sites; require periodical informs to the companies that participate in the industry; guarantee competitive business practices; and resolve controversies between companies and third interested parties through public consultations. The commission resolves disputes using public consultations and general resolutions. This counters the style of U.S. and Canadian commissions, which use open hearings and technical conferences to allow input into decision making and for dispute settlement. ENARGAS, an independent regulatory commission with no influence from Argentina's energy ministry, has three appointed commissioners. The commission uses both internal and external auditors to monitor its activities and reports to the national executive branch of the Argentine government.

Argentina's model is closest to the U.S. upstream regime. The potential exists in Argentina for a healthy gas (and power) commodity market with trading and marketing activities used by producers, marketers, and large customers (including local distributors) to hedge against natural gas price volatility. Yet, today YPF still controls around 65% of marketed natural gas production and about 75% of gas produced in fields. Thus, even the privatized YPF remains extremely powerful.

2.4.2 Mexico

Mexico has pursued a strategy of reserving upstream petroleum and gas exploration to Petroleos Mexicanos (Pemex), the national oil company. During the early days of Mexico's industry, oil and gas exploration was carried out by private foreign and Mexican companies. Disputes between the Mexican government and foreign operators, and political imperatives following Mexico's revolution, resulted in the 1938 nationalization of Mexico's oil industry. Article 27 of the regulatory law to Mexico's constitution stipulates that Pemex has sole control of the production of oil and gas and the products derived from the raw resources.

During the 1970s, hydrocarbon production did not keep pace with economic modernization, so that by 1973 Mexico found itself to be a net importer of crude oil. Critical discoveries restored Mexico's

stature as an oil producer and exporter. Investment in upstream activities continued until the early 1980s, when Mexico's external debt crisis, exacerbated by falling world crude prices, triggered a contraction in government spending. The collapse of crude prices in 1986 impacted upstream activity even more. Spending by Pemex on exploration and production dropped from its peak of approximately 86% of Pemex's total budget in 1982 to less than 60% in 1988. Historically, crude oil has been given priority because of its export potential and value. However, two factors contributed to an effort to increase natural gas production. One, in the late 1970s, was the desire to increase gas sales to the United States, which led to construction of the Cactus-Reynosa pipeline as a result of negotiations between Pemex and Border Gas, a consortium of U.S. companies. The 2-billion cubic feet (bcf)/day project was never realized because of disputes about pricing. The second factor was concern about inefficient utilization of energy. Energy use was, and is, highest in the energy sector. Expenditures were made to gather and transmit gas, especially from the huge Bay of Campeche fields, and to reduce wasteful gas flaring, which declined from 26% of production in 1970 to less than 3% by 1989. Domestic consumption of natural gas continued to grow in the 1980s, but with relatively little new investment in natural gas production and transmission made by Pemex. The result is Mexico's current situation of inadequate domestic production capacity to satisfy natural gas demand.

In the 1980s, in response to the critical economic situation Mexico faced after the oil market crash and currency devaluations, the Mexican government began to implement market reforms. Public opinion and political will for privatizing Pemex have historically been weak, but a series of major accidents, chronic shortages, and unreliable service forced the managers at Pemex to take action. The government gradually removed the obligations on Pemex to provide everything from roads to hospitals and schools as part of its social obligations to the state. Pemex began to reduce its huge employment from more than 250,000 to just over 133,000 today. In 1992, Pemex was reorganized into four functional subsidiaries for exploration and production, refining, natural gas and basic petrochemicals, and secondary petrochemicals. The government also changed Pemex's tax status by creating a corporate tax rather than controlling all of Pemex's revenues and returning some portion to the company for reinvestment. The corporate tax rate for Pemex

remains high (more than 60%) and Pemex still does not have independence with respect to its budget planning.

In 1995, further, more dramatic steps were taken to reform Mexico's energy sector. The regulatory law to the constitution was changed to allow private investment in natural gas transportation, distribution, and storage, in recognition of the importance of this fuel to Mexico's economic development. A regulatory commission has been created (the Comisión Reguladora de Energía, or CRE), charged with the privatization of assets formerly controlled by Pemex. Rules have been established for pipeline tariffs and first-hand sales of imported gas from the United States (although Pemex is expected, at some point, to resume its bid to be a net exporter of gas). The CRE handles all auctions for state-owned assets (portions of the Pemex pipeline grid and local distribution networks) that are to be turned over to private operators, and uses a price cap formula similar to the Ogas RPI-X to regulate tariffs. The CRE has five appointed commissioners. A law passed by the Mexican Congress in 1995 established the CRE as an independent entity, but Mexico's energy ministry retains a great deal of influence. Like ENARGAS, the CRE handles conflicts and disputes with operators through private consultations rather than through the public meetings typical in the United States and Canada.

In spite of all of these reforms, deep problems exist in Mexico's oil and gas sector. Pemex has watched its market share of exported oil erode as other countries moved aggressively to lure private investment into their upstream businesses. The investment demands on Pemex for improvement and expansion projects are huge. Although the company has had some success with foreign placements of debt, many questions remain about Pemex's ability to finance capital improvements. Finally, although the effort to attract private investment into Mexico's natural gas pipeline and distribution segments continues, Pemex remains the sole supplier of natural gas, which will restrict growth of Mexico's natural gas market.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, Energy Policy in • Greenhouse Gas Emissions from Energy Systems, Comparison and Overview • Markets for Natural Gas • National

Energy Policy: United States • Natural Gas, History of • Natural Gas Processing and Products • Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil-Led Development: Social, Political, and Economic Consequences

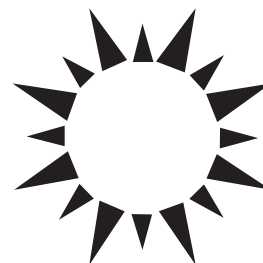
Further Reading

- Foss, M. M. (1995). "Natural Gas in the Twenty-First Century: Adjusting to the New Reality." Doctoral dissertation, University of Houston, Houston, Texas.
- Foss, M. M. (1998). Worldwide transitions: Energy sector reform and market development. *Natur. Resourc. Environ.* (Spring 1998).
- Foss, M. M. (2000). Perspectives on the international exploration business. In "International Oil and Gas Ventures: A Business Perspective" (G. Kronman and T. O'Connor, Eds.), pp. 11–39. American Association of Petroleum Geologists, Tulsa, Oklahoma.
- Foss, M. M. (2002). Latin American gas and gas-fired power shows progress, risk. In "Natural Gas and Electric Industries Analysis" (R. Willett, Ed.), pp. 230–247. Financial Communications Company, Houston.
- Tussing, A., and Barlow, C. (1984). "The Natural Gas Industry: Evolution, Structure, and Economics." Ballinger Publ. Co., Cambridge, Massachusetts.
- U.S. Energy Information Administration (EIA). (1992–1998). "Natural Gas Issues and Trends." Available on the Internet at www.eia.doe.gov.

Natural Gas Processing and Products

RICHARD G. MALLINSON

University of Oklahoma
Norman, Oklahoma, United States



-
1. Requirements for Gas Processing
 2. Gas Dehydration
 3. Gas Sweetening
 4. Hydrocarbon Recovery and Fractionation
 5. Conversion to Transportable Chemicals and Fuels

Glossary

acid gas The acidic constituents of natural gas, H_2S , CO_2 , mercaptans (RSH), CS_2 , and COS .

dew point The temperature and pressure at which a gas mixture begins condensing to form a liquid phase; for a fixed composition, the point is reached by lowering the temperature or increasing the pressure except when the mixture exhibits retrograde condensation.

ethane, propane, butanes, liquefied petroleum gas (LPG), and natural gasoline Natural gas liquid products for which there are various purity grades; they are defined by a combination of maximum and minimum compositions, vapor pressures, and boiling points.

gas conditioning The process of removal of contaminants, usually acid gases, nitrogen, and water, from natural gas.

gas dehydration The process of removal of water from natural gas to achieve a low dew point that will avoid condensation of water during further processing, transportation, and use.

gas sweetening/treating The process of removal of acid gas components from natural gas.

gas to liquids (GTL) Processes for the chemical conversion of methane to other chemicals and fuels that are predominantly liquids; most frequently refers to Fisher-Tropsch-type processes that preferably produce hydrocarbons from C_5 to C_{30} .

liquefied petroleum gas (LPG) A natural gas liquid product composed primarily of a mixture of propane and butanes.

natural gas liquids (NGL) The higher hydrocarbons separated from natural gas that are usually produced

and stored as liquids under pressure at ambient temperature.

natural gasoline A natural gas liquid product that consists primarily of pentanes and higher alkane hydrocarbons (C_5^+); it is a liquid at ambient temperature and atmospheric pressure and has a low octane number.

retrograde condensation The phenomenon where a hydrocarbon gas mixture of a particular composition will pass through the dew point and begin condensation of a liquid phase when the pressure is lowered.

sour gas Natural gas that has at least a small percentage of acid gases, specifically sulfur-containing components.

Gas processing is the preparation of raw natural gas as it is produced from the reservoir for transportation to markets for utilization. Traditionally, this has been primarily the removal of chemical constituents that are not desired due to safety reasons (e.g., H_2S , Hg), operability (e.g., water due to condensation and corrosion), and/or economics (e.g., CO_2 that lowers the heating value or those components that may be profitably separated and marketed as a separate product, natural gas liquids, C_2 to C_5^+ hydrocarbons). The final product for this has been compressed gas that is fed into pipeline networks. Increasingly, the processing required to transport the gas to markets for end use has included the production of liquefied natural gas by complex refrigeration processes as well as the chemical transformation of the methane into solid or liquid chemicals and fuels that are commonly produced from natural gas, are much more easily transported than the gas itself, and satisfy significant markets for the products. The most commonly produced are ammonia and its derivative fertilizers and methanol, but there is strong promise for production of liquid hydrocarbon fuels on much larger scales to satisfy increasing demand, especially for cleaner fuels.

1. REQUIREMENTS FOR GAS PROCESSING

Feed gas to a processing plant can have extremely large variations in composition from one location to another. Produced gas, subject to removal of liquids and solids with separators only at the field locations, is frequently saturated with water due to intimate contact between gas and water in the reservoir formation. Gas associated with oil production may be saturated with hydrocarbons from ethane to C_{10}^+ . In both cases, the amounts will vary with the pressure and temperature of the gas. In the case of water, this is commonly a proportion up to several percentage points. For hydrocarbons, this can be well above 10%. For the acid gases, primarily CO_2 and H_2S but including other sulfur-containing species such as mercaptans (RSH) and COS, reservoirs in which the majority of the gas volume is comprised of these species is possible. At some point, the high percentage of these contaminants will make the gas uneconomical for recovery, but dealing with several percentage points H_2S and CO_2 above 10% is common. The other contaminant commonly found in gas at these percentage levels and higher is nitrogen (and rarely helium). Lower levels of contaminants that must be removed are possible but less common, with mercury being one example (generally removed using solid adsorbents).

The requirements of gas processing are to remove the contaminants mentioned previously to appropriate levels. These levels are specified by contract and determined by the market for which the gas is destined as well as for safety and operability considerations. Typical "pipeline quality" gas specifications are discussed here.

Water must be removed to levels that will ensure that no condensation of the water will occur in any of the subsequent operations or distribution of the gas. This is both to minimize corrosion and to prevent damage to equipment due to two-phase flow where liquid water droplets can damage machines that compress or expand the gas and cause erosion by impingement on surfaces at high velocities. Dehydration also prevents the formation of methane hydrates, solid crystalline complexes of methane with water molecules, that can plug flow lines as well as damage equipment. To avoid condensation, a dew point for the gas is selected based on a specified temperature and pressure. This represents the combination of the minimum temperature and the maximum pressure to which the gas may be expected to be subjected, and at those conditions where no

water will condense (e.g., $0^\circ C$ and 6.7 MPa). Typical values of allowable water content at such conditions are in the range 50 to 110 mg/m³ and vary by location due to climate and other factors.

For sulfur content, specifications for H_2S , mercaptans, and total sulfur are common, with H_2S and mercaptans each having limits in the range of 6 to 24 mg/m³ and total sulfur having limits from 115 to 460 mg/m³. These limits, although needed in part due to corrosiveness of the sulfur compounds, are due in significant part to their noxious and toxic properties.

Specifications for CO_2 and nitrogen are primarily set due to the fact that they are inert and have no heating value and, therefore, reduce the heating value of the product gas. There are typically limits of a few percentage points on each, but their amounts are also indirectly constrained by a heating value specification of the product gas.

The specifications for hydrocarbons are somewhat more complex. For methane, there is some minimum specification (e.g., 75%). For higher hydrocarbons, there may be individual maximum limits (e.g., 10% ethane, 5% propane, 2% butanes). These limits are based on ensuring against condensation, in the same fashion as for water but also for the combustion processes for which most pipeline gas is used. The presence of higher hydrocarbons changes the required air/fuel ratio for combustion and, when outside of design limits, can cause incomplete combustion with additional pollution and higher or lower temperatures, among other operational problems. From the standpoint of the gas processor, there are frequently economic incentives to recover the higher hydrocarbons from the gas. When present in sufficient quantities, the value of the hydrocarbons can exceed the cost of their recovery. Typical products of value include ethane, propane, butanes, and natural gasoline. These are collectively called natural gas liquids and, as mentioned previously, are frequently present in near saturation conditions (at their dew points) in gas associated with oil production. Each of these products has its own specifications that put minimum and maximum limits on the hydrocarbon constituents and may include heating value and boiling point ranges.

For liquefied natural gas (LNG), the product gas will have specifications with considerably lower contaminant levels to avoid contaminants that could condense or freeze at the very low temperatures involved in production of LNG ($-161^\circ C$). For gas to chemicals and fuels processes, the feed gas may have similar specifications to pipeline gas except that sulfur (and possibly other trace contaminants) will be

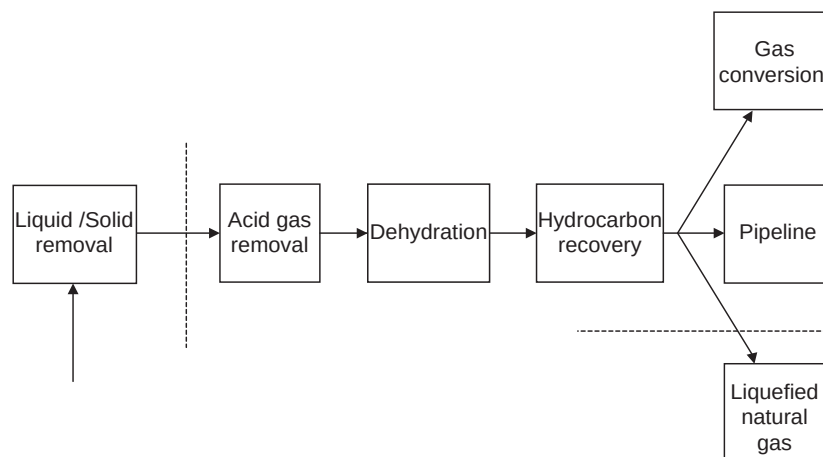


FIGURE 1 Block diagram of major gas-producing units.

further restricted due to its ability to deactivate catalysts used for the conversion reactions.

Figure 1 illustrates a block diagram of the processing units that may be found in a gas processing plant. The presence, order, and relative size of the process units will be dependent on the composition of the feed, the gas product specifications, and the specific processes selected for each unit. Gas conversion plants are considered separately given that these processes are a follow-on to the type of plant shown in Fig. 1.

2. GAS DEHYDRATION

Gas dehydration is accomplished by means of one of two primary processes (or occasionally both). Absorption of water vapor from the gas into a hygroscopic liquid is the predominant method, but adsorption of water onto solid desiccants is also used, particularly for achieving very low dew points.

The hygroscopic liquid is most frequently triethylene glycol (Fig. 2), but other glycols are sometimes used. Glycols are quite stable, but the presence of oxygen can make the solutions highly corrosive. Water is highly soluble in the glycol, and when a water-containing natural gas stream is placed in contact with a glycol solution, the water is absorbed from the gas phase into the liquid phase. The other important property is that the boiling point of the glycols, especially triethylene glycol, is much higher than that of water (288°C vs 100°C at atmospheric pressure). Therefore, the glycol–water mixture may be heated to revaporize the water at temperatures far below the boiling point of the glycol, resulting in

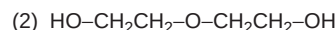


FIGURE 2 (1) Ethylene, (2) diethylene, and (3) triethylene glycol.

little covaporization of the glycol. In addition, aliphatic hydrocarbons are absorbed only in very small amounts, although aromatics are very soluble. Acid gases that may be present are also absorbed in small amounts, although their presence can affect the solubility of water in glycol during the absorption and can also result in absorption of aliphatic hydrocarbons.

Figure 3 presents a process flow diagram for a glycol dehydration system. The glycol contactor is a vertical cylindrical column with dimensions of from approximately 0.5 m up to several meters in diameter, with larger sizes required for larger gas flow rates being processed. Inside the column are usually 8 to 10 contacting trays placed above each other approximately 0.6 m apart. Although trays are common, packing materials that allow the liquid to trickle down over their surface to create a high gas–liquid contacting area are also used and can be preferred in a number of applications, particularly for increasing the capacity of existing trayed columns. The total height of the column is approximately 5 to 8 m and is not greatly affected by the flow rates. The lean (low water content) glycol solution is introduced at the top of the column and flows across each tray and down to the next. The water-containing natural gas is introduced into the

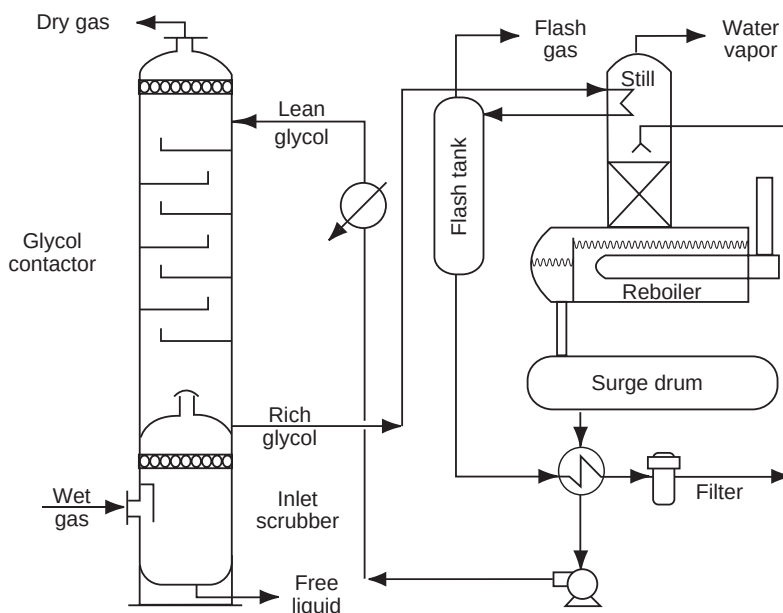


FIGURE 3 Typical glycol gas dehydration process diagram. Reprinted from Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," Vols. 1–2, 11th ed. GPSA, Tulsa, OK.

bottom of the column and flows upward through perforations or caps in the trays where the gas and liquid are highly mixed. This mixing provides good contact between the gas and the liquid to allow the transfer of the water from the gas to the liquid. This countercurrent flow pattern provides the best approach to equilibrium for the removal of water from the gas phase. The gas has less water remaining at each successive tray until it leaves the column at the top. The liquid increases its water content as it flows down through the column until it is discharged at the bottom. Typical applications require approximately 0.25 L of TEG/kg water removed and a lean (inlet) glycol concentration of approximately 99%. The contacting generally takes place at near ambient temperatures ($\sim 38^{\circ}\text{C}$) and at convenient pressures, with as little lowering of the gas pressure as possible (2–6 MPa).

The glycol must then be regenerated to return it to its lean concentration for reuse in this continuous process. The water-rich glycol flows from the bottom of the column to the still/reboiler, where it is heated to approximately 200°C and the water is boiled off at atmospheric pressure and vented. Any absorbed aromatics are also vaporized, and emissions handling equipment may be needed in this case. The lean glycol is then cooled and pumped back to the glycol contactor.

A number of variants of this process that improve performance, particularly under challenging circum-

stances such as where especially low dew points are required, are in commercial operation. In that case, a vacuum still can achieve higher lean glycol purity, say above 99%.

Solid desiccant dehydration makes use of the attraction between water in the natural gas and the surface of certain solid materials. In these adsorption processes, the surface properties are such that a weak bond will form between the solid surface and water when exposed to water during the gas phase. The bond strength is a function of temperature with high water-adsorbing capacity at ambient operating temperatures and lower capacity as the temperature increases. These materials can be fabricated into particles with diameters in the range of 1 to 5 mm and with high active surface areas and water adsorption capacities from 20% to more than 40% by weight. The most common materials are silica gels, activated aluminas, and crystalline molecular sieve zeolites.

Although it is a more expensive process than glycol dehydration, the solid desiccant process can achieve significantly lower dew points (as low as -60 to -100°C). Adsorption is also useful for simultaneous removal of acid gases.

The solid particles are placed as a bed in a vertical cylindrical column. The wet feed gas flows down through the column at a velocity designed to provide an appropriate residence time for the adsorption to take place at high rates, but it is limited by the loss of

pressure due to the friction of the gas flow through the void spaces between the particles that increases as the velocity increases. The adsorption takes place at a convenient gas pressure, say 2 to 6 MPa, at near ambient temperature.

The process operates continuously in a cyclic manner with two columns in parallel, as shown in Fig. 4. One column accepts the flow of natural gas for a period of time that allows the water-adsorbing capacity of the volume of particles in the column to be reached. A larger column of particles has a higher capacity and operates for a longer time but commonly for less than 12 h. During this time period, the other column is in the process of being regenerated, with the water removed by flowing a stream of hot ($\sim 315^{\circ}\text{C}$) regeneration gas over the particle bed. The regeneration gas is a portion of the dried gas that, after passing over the regenerating bed and taking up the “desorbed” water, is cooled back to ambient temperature. This stream passes below its dew point; thus, water condenses and is removed in the knockout drum. The gas, now saturated with water, is put back into the feed gas. It should be pointed out that at the high regeneration temperature, the regeneration gas can accept significantly more water than it can at ambient temperature; thus, only a relatively small portion must be reprocessed. When the adsorbing column is saturated, valves are switched to redirect the wet feed gas to the

regenerated column and to begin regeneration of the “saturated” column.

Other processes for dehydration are available for commercial application; however, these are generally limited to specialty applications such as very small gas flow rates. Membrane permeation is a technology that continues to advance, although limitations on selectivity cause unacceptable losses of feed gas for dehydration. These alternative process choices are generally more costly for the most common gas dehydration applications.

3. GAS SWEETENING

A wide array of processes are available for the removal of the acid gases (primarily H_2S and CO_2 but including all sulfur compounds such as mercaptans, CS_2 , and COS). The selection of the most appropriate one (both technically and economically) depends primarily on the feed gas composition, specifically which acid gases are present (or absent) and at what concentrations as well as the final product specifications. The major categories of sweetening processes are amine reaction processes, physical absorption processes, combination processes, and cyclic processes. Figure 5 shows a matrix that considers some of the criteria for process selection.

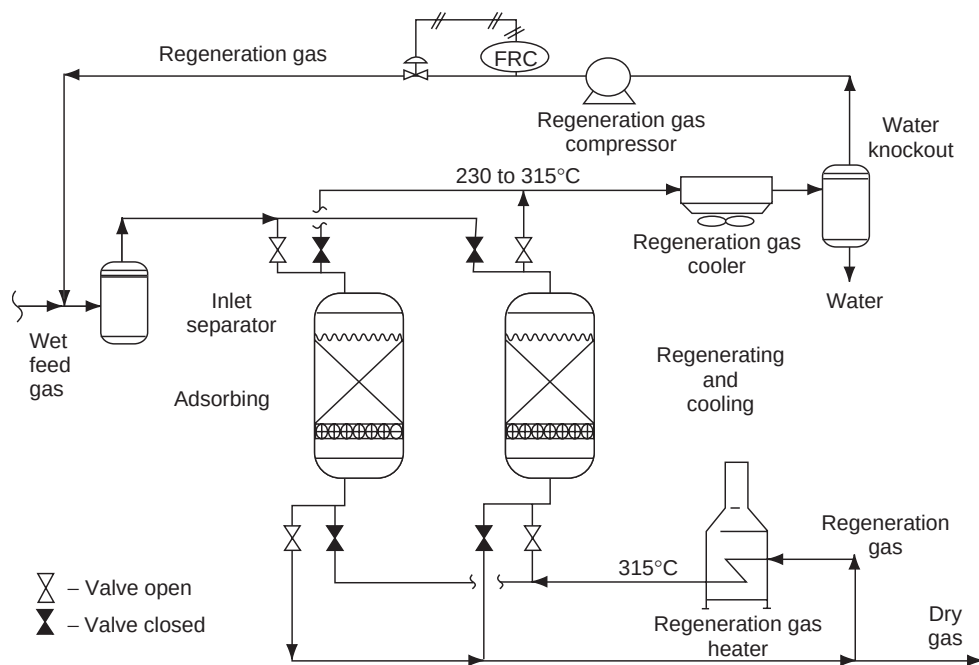


FIGURE 4 Solid desiccant dehydration process diagram. Reprinted from Gas Processors and Suppliers Association. (1998). “GPSA Engineering Data Book,” Vols. 1–2, 11th ed. GPSA, Tulsa, OK.

	Normally capable of meeting 0.25 grain ^a H ₂ S	Removes mercaptans and COS	Selective H ₂ S removal	Solution degraded (by)
Primary amine	Yes	Partial	No	Yes (COS, CO ₂ , CS ₂)
Secondary amine	Yes	Partial	No	Some (COS, CO ₂ , CS ₂)
Tertiary amine	Yes	Partial	Yes ^b	No
Combination	Yes	Yes	Yes ^b	Some (CO ₂ , CS ₂)
Physical solvent	Maybe ^c	Slight	Yes ^b	No
Solid bed	Yes	Yes	Yes ^b	No
Liquid redox	Yes	No	Yes	CO ₂ at high concentrations
Sacrificial	Yes	Partial	Yes	No
^a 0.25 grains H ₂ S/100 scf ≈ 5.7 mg/m ³ (4 ppmv). ^b Some selectivity exhibited. ^c Can make 0.25 grain under some conditions.				

FIGURE 5 Matrix of sweetening process selection factors. Data from Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," Vols. 1–2, 11th ed. GPSA, Tulsa, OK.

The absorption of the acid gases into aqueous solutions is low due to low solubilities, but the placement of a basic amine into the solution allows the absorbed acid gas to react to form an acid–base complex. This creates a high capacity for absorption that depends on the concentration of the amine. Figure 6 shows the dominant amines used in these solutions. For a primary amine (RNH₂), one molecule of acid gas is complexed with each amine molecule. For a secondary amine (R₂NH), two molecules of acid gas are complexed, and for a tertiary amine (R₃N), three are complexed, except that CO₂ is not complexed with the tertiary amine, giving it a preferential selectivity for sulfur (Fig. 7). In practice, only approximately one-third of the theoretical complexation capacity (moles of acid gas complexed per mole of amine) is realized.

Figure 8 shows a process flow diagram for an amine reaction-type gas sweetening process. The sour gas enters the column from the bottom and flows up and out the top while the lean (low acid gas content) amine solution flows down from the top to the bottom of the column. As with dehydration columns, the countercurrent flows of the two streams maximize the transfer of acid gas constituents to the liquid phase. The column internals are trays or packing in similar fashion to the dehydration column. The gas leaving the column may be passed through a water scrubber to recover more volatile amines (e.g., MEA, DEA). The gas at this point is essentially saturated with water and would then pass to a dehydration system. The column operates at near ambient temperature and a convenient gas pressure, say 38°C and 2 to 6 MPa. The acid gas-rich amine solution leaves the bottom of the column, is heated, and flows to the upper part of the stripping column. This also is a gas–liquid contactor that

Monoethanolamine	HOCH ₂ CH ₂ NH ₂
Diethanolamine	(HOCH ₂ CH ₂) ₂ NH
Triethanolamine	(HOCH ₂ CH ₂) ₃ N
Methyldiethanolamine	(HOCH ₂ CH ₂) ₂ NCH ₃
Diisopropanolamine	(CH ₃ HOCHCH ₂) ₂ NH
Diglycolamine	HOCH ₂ CH ₂ OCH ₂ CH ₂ NH ₂

FIGURE 6 Major amines used in gas sweetening.



FIGURE 7 Simplified amine acid gas chemistry.

creates surface area between the gas and liquid phases that allow, at a higher temperature of approximately 110 to 130°C, the sour gases to be stripped into the gas phase. The bottom of the stripping column includes a reboiler where some of the heating of the amine solution occurs. The stripped or lean amine solution is then cooled and pumped back to the sweetening column. Both columns' sizes depend on the volumetric flow rates of the gas and liquid streams, with larger flow rates requiring larger diameters (in the range of 0.4 m to several meters), and on the amount of sour gas that is transferred. The amine concentration in the aqueous solution is typically in the range of 20 to 30%, and this defines the capacity of the solution and the amount that is needed to achieve the designed sweetening specification (e.g., liters of amine solution per cubed meter of gas treated). The acid gas leaving the stripper may be vented or burned if sulfur is absent or sufficiently low, but if significant amounts

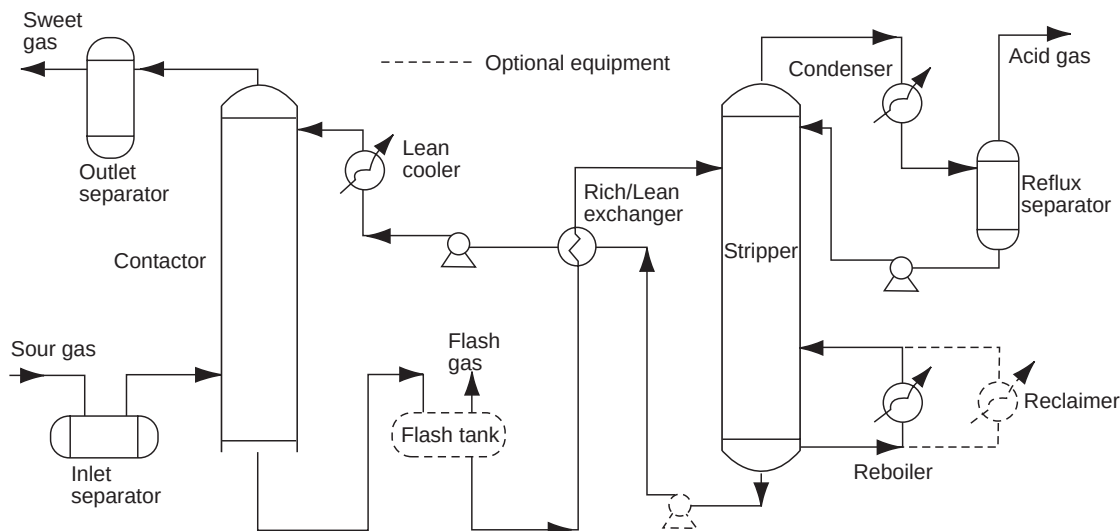


FIGURE 8 Example of an amine reaction-type sweetening process flow diagram. Reprinted from Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," Vols. 1–2, 11th ed. GPSA, Tulsa, OK.

are present, the sulfur-containing gas will have to be recovered.

Physical absorption processes operate in the same fashion as does glycol dehydration. A solvent that has a high capacity for acid gases is contacted with the sour natural gas in an absorption column (Fig. 3). In some cases, the solvent may be selective for sulfur compounds, may also absorb water so that the need for a separate dehydration step is eliminated, and/or may also absorb hydrocarbons advantageously. The regeneration of the solvent is generally different from the case with glycols in that it does not depend on boiling off the acid gases. Instead, pressure reductions allow the absorbed gases to vaporize, and a stripping column may be used to provide gas–liquid contacting to enhance the vaporization. Major processes using this method include the Selexol process using a polyethylene glycol derivative as the solvent, the Fluor solvent process using anhydrous propylene carbonate, and the Rectisol process using methanol.

Combined processes use both reacting and absorbing materials. Shell's Sulfinol process uses an aqueous amine along with Sulfolane, a proprietary physical solvent.

In general, cyclic processes operate by contacting with a solid adsorbent (e.g., iron sponge, zinc oxide, molecular sieve zeolites) that must be regenerated in similar fashion to the solid desiccant dehydration process or by contacting with a liquid solution (e.g., Chemsweet, Sulfa-check, Sulfatreat) that must be changed and regenerated at the end of the cycle.

Other technologies are available, including iron chelation redox-type processes and membrane permeation processes that may have applications in specific circumstances.

The major process for sulfur recovery is the Claus process with a number of possible variations. The sulfur-containing gas is first burned with air in a thermal reaction furnace to convert approximately one-third of the sulfurous gas to SO_2 . At the high temperature ($\sim 1000^\circ\text{C}$), some elemental sulfur is formed. The gas is then cooled below the sulfur dew point to allow much of it to condense. The gas is then reheated to approximately 400°C and passed over an alumina-type catalyst that allows the reaction of H_2S and SO_2 to form sulfur and water. This reaction is equilibrium limited in that complete conversion is not thermodynamically feasible (lower temperature favors higher conversion) because the catalyst requires a high temperature ($\sim 400^\circ\text{C}$) to actively carry out the reaction. To achieve higher conversion to sulfur, the converted gas stream is then cooled below the dew point of sulfur that then condenses, and the remaining gas is once again heated and passed over a second catalytic reactor. This may be repeated for further conversion in a third catalytic reactor to achieve a high enough conversion to meet requirements for release of the remaining gas to the atmosphere, although further cleanup of this gas by combustion and (possibly) scrubbers may be necessary.

The primary use for the recovered sulfur is in the manufacture of sulfuric acid.

4. HYDROCARBON RECOVERY AND FRACTIONATION

The first objective of hydrocarbon recovery in natural gas processing is the reduction of the hydrocarbon dew point so that condensation cannot occur during subsequent processing, transportation, and use of the gas. Of course, the removal of these hydrocarbons causes a loss of volume of the product gas, called shrinkage, as well as a loss of heating value given that the higher hydrocarbons have a greater heat content per volume than does methane. If the higher hydrocarbons cannot be sold or used for fuel, this is a substantial loss of value that must be minimized. Historical practice frequently has been to burn those products to simply dispose of them. (Indeed, much the same could be said about natural gas found with oil, although this has changed substantially during the past decade or so.) The second objective is to maximize recovery of the higher hydrocarbons as various liquid products collectively referred to as natural gas liquids. This is now quite common, with the enhanced economic value more than offsetting the loss of value to the natural gas product.

Earlier in the article, it was mentioned that the required product dew point for water occurred at the combination of the highest pressure and the lowest temperature. For hydrocarbons, the situation is more complex in that gaseous hydrocarbon mixtures undergo the phenomenon of retrograde condensation. This is when, at a fixed temperature and composition, a lowering of the pressure may cause condensation. Thus, lowering the pressure moves the system toward the dew point rather than away from it (as is the case for most other systems, e.g., water vapor in natural gas). Thus, the operative criteria for higher hydrocarbon removal is their removal to levels that ensure that condensation cannot occur at the lowest expected temperature at any pressure. The point at which this occurs is referred to as the cricondentherm.

The removal of the higher hydrocarbons proceeds through cooling of the gas at a pressure of convenience to maintain as high a pressure for pipeline flow as possible. Figure 9 shows the normal boiling points of the major hydrocarbons found in natural gas. As can be seen, condensation and recovery of greater amounts of lighter hydrocarbons requires lower temperatures, and this is true at any pressure. Thus, the lowest temperature required will be dictated by the composition of the gas, the fractions of specific components to be recovered (or

n-Pentane	36.07
i-Pentane	27.85
n-Butane	-0.5
i-Butane	-11.72
Propane	-42.04
Ethane	-88.60
Methane	-161.49

FIGURE 9 Boiling points of natural gas hydrocarbons at atmospheric pressure (in °C).

dew point to be achieved), and the operating pressure.

If the pressure is relatively low, a refrigeration process is used to chill the gas sufficiently to cause condensation of the higher hydrocarbons. These are typically propane and higher with small amounts of ethane. The condensate is then removed in a gas-liquid separator and flows to a stabilizer column that makes sure that any lighter hydrocarbons that are dissolved in the condensate, such as methane, ethane, and small amounts of higher hydrocarbons, can revaporize to an equilibrium composition and will not vaporize during later liquid processing, storage, and transportation. The stabilizer column is a vertical cylindrical column with internal trays or packing that promotes gas-liquid contact and mixing, as discussed earlier. The gas from the cold separator and the stabilizer column are then combined to form the product gas stream.

An alternative process is the use of a solvent to absorb the higher hydrocarbons, analogous to solvent dehydration. Lean oil absorption can be used as the process. The lean oil is simply another hydrocarbon stream of a sufficiently low vapor pressure that it remains in the liquid phase through the absorption and regeneration. Regeneration is carried out by heating the rich oil to allow the higher hydrocarbons absorbed from the natural gas stream to revaporize. Such plants are relatively uncommon today.

If the gas is available at sufficiently high pressures, and especially if greater recovery of natural gas liquids is desired for economic reasons, the cooling of the gas is achieved by allowing it to expand, losing some of its pressure. This is the mainstay of hydrocarbon recovery processing. There are two alternatives for this, with the simplest being the JT expansion making use of the Joule-Thomson effect where the expansion occurs as the gas flows through a restrictive valve. The design determines the degree of cooling required to achieve the amount of natural gas liquids condensation required, and additional

refrigeration can be provided if needed. This type of plant, although common, is relatively inefficient in that there can be no work recovered from the expansion.

Most newer plants, particularly those that wish to maximize recovery of gas liquids, allow the gas to expand through a turbine where the gas cooling is achieved and some work is recovered from the turbine. Usually, the turbine shaft drives a compressor also used in the process. The basic process flow diagram is shown in Fig. 10. The feed gas is first chilled by cooler gas streams in the plant, and any condensed liquids are removed in the low-temperature cooler before reaching the turbo expander. The pressure is reduced in the expander, from approximately 7 to 4 MPa in the example shown, while the temperature drops from 51 to -77°C . After the additional liquid condensate is removed in the very low-temperature separator (called the demethanizer), the cold gas is warmed by chilling the incoming feed gas. The product gas can now be partially repressurized by the compressor that is driven by the expander turbine. Additional recompression may or may not be required. There are many variations of this process to accommodate specific composition and recovery requirements, for example, the use of external refrigerants to achieve lower temperatures for greater ethane recovery. These processes can recover as much as 80 to 90% of the ethane (and virtually all of propane and higher hydrocarbons) in the natural gas.

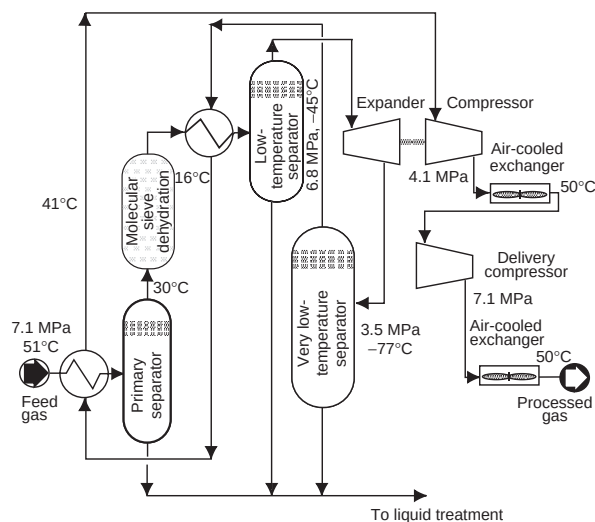


FIGURE 10 Turboexpander plant diagram. Reprinted from Rojey, A., Jaffret, C., Cornot-Gandolphe, S., Durand, B., Julian, S., and Valais, M. (1997). "Natural Gas Production Processing Transport." Editions Technip, Paris.

The processing of the gas to pipeline quality has now been completed. The liquid streams will require further separations at low temperature to produce an optimum product set. There are many possibilities for configurations that depend on both the composition of the liquids and the value of the various products. Products are defined by their composition, boiling range, and vapor pressure and may have several grades. Some major products include natural gasoline, commercial butane, commercial propane, and high-purity ethane. These products are produced by separation in one or more fractionation columns that contain trays or packing. The columns can appear in a wide range of sizes depending on the quantities of the components. Typical sizes are in the range of 0.5 to 3.0 m in diameter and 8 to 15 m in height. Higher purities require taller columns that give greater opportunity for the gas and liquid to mix. Temperatures are typically below ambient temperature, and pressures are typically in the range of 1 to 5 MPa. The conditions are selected to maximize the difference in volatilities of the components to be separated in each column. The volatility differences control whether a component appears primarily in the gas or in the liquid leaving a column. In a four-column fractionation, the raw natural gas liquids are fed to the first column, the deethanizer, with primarily ethane vaporizing and leaving from the top of the column and the remaining liquid flowing from the bottom of the column to the depropanizer, where primarily propane vaporizes and leaves from the top. The remaining liquid flows from the bottom of the column to the debutanizer, where the butanes leave from the top and flow to the fourth column that separates isobutane from normal butane. The liquid leaving the bottom of the debutanizer is the C_5^+ natural gasoline.

The product streams may flow by pipeline to subsequent processes or plants for use or to storage as liquids at ambient temperature and sufficient pressure to maintain them in the liquid state.

5. CONVERSION TO TRANSPORTABLE CHEMICALS AND FUELS

Natural gas has been a major feedstock for production of chemicals since early in the 20th century with the development of gas-solid catalytic chemistry. This is despite the intrinsic stability of the methane molecule that requires the conversion of methane to

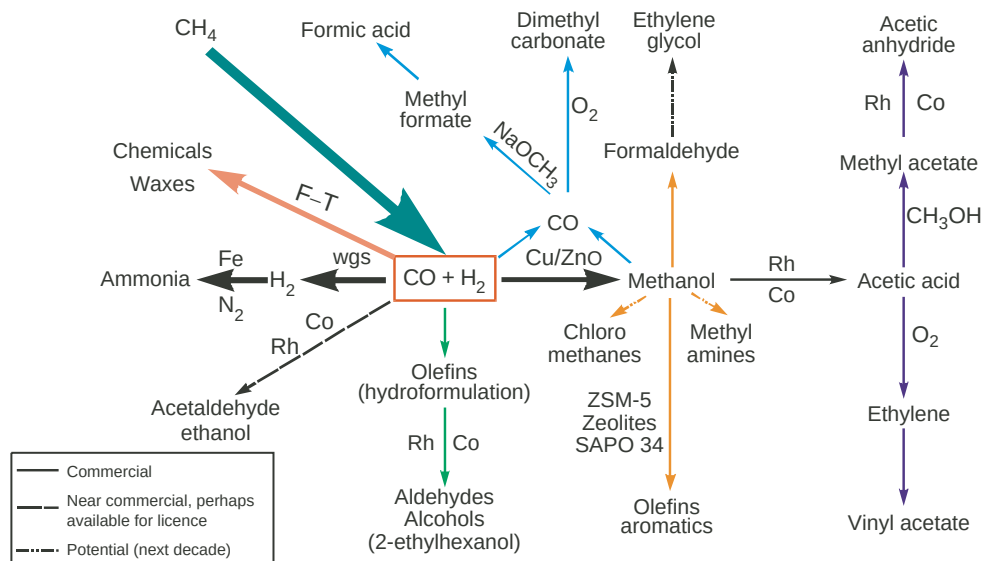


FIGURE 11 Chemical pathways for gas conversion.

carbon monoxide and hydrogen, known as synthesis gas, before further conversions to useful chemicals. Figure 11 illustrates the chemical pathways to form primary and secondary industrial chemicals from natural gas via the synthesis gas intermediate. Historically and through the current time, methanol and ammonia are among the most produced industrial chemicals. Ammonia represents the fixation of nitrogen and is the precursor for most nitrogen fertilizers. Methanol is used as a solvent as well as a feedstock for many other chemicals. Lastly, the production of liquid hydrocarbons via Fischer-Tropsch has a recent history of production from natural gas but a long history of production starting with coal. It appears that processes for production of liquid fuels from Fischer-Tropsch synthesis (FTS) are poised to undergo rapid expansion as markets for clean fuels and chemical feedstocks increase. All of these products satisfy one of the requirements for monetizing many more remote natural gas resources in that they are significantly more transportable than natural gas that is often located far from demand that can use pipelines and the relatively limited, but also increasing, markets for LNG.

The first step in the synthesis of ammonia, methanol, and liquid hydrocarbons is the production of synthesis gas. The primary means of carrying out this conversion has been methane steam reforming. This is shown in Fig. 12 as reaction 1. Steam reforming has the attribute of producing large amounts of hydrogen, one-third of which comes from the water. However, it is a very energy-intensive

- (1) $\text{CH}_4 + \text{H}_2\text{O} \leftrightarrow \text{CO} + 3\text{H}_2$
- (2) $\text{CO} + \text{H}_2\text{O} \leftrightarrow \text{CO}_2 + \text{H}_2$
- (3) $\text{CH}_4 + 1/2 \text{O}_2 \leftrightarrow \text{CO} + 2\text{H}_2$
- (4) $3\text{H}_2 + \text{N}_2 \leftrightarrow 2\text{NH}_3$
- (5) $\text{CO}_2 + 3\text{H}_2 \leftrightarrow \text{CH}_3\text{OH} + \text{H}_2\text{O}$
- (6) $\text{CO} + 2\text{H}_2 \leftrightarrow -\text{CH}_2- + \text{H}_2\text{O}$

FIGURE 12 Natural gas conversion chemical pathways.

reaction. The reaction takes place on supported nickel metal catalyst pellets placed in tubes through which the natural gas and water feed flows. The thermodynamics require very high temperatures to achieve good conversions, and the reaction is conducted at a temperature of approximately 800°C. In addition, the reaction is highly endothermic, requiring large heat inputs. The combination of these two requirements leads to placing the reactor tubes inside a direct-fired combustion furnace where radiative heat transfer can provide the needed heat input. Conversion remains incomplete, and a secondary reformer with a different nickel catalyst is used at temperatures of 900 to 1100°C. Less heat input is required because the remaining conversion is smaller, and this may be supplied by feeding oxygen to cause the heat-releasing exothermic partial oxidation reaction, shown as reaction 3 in Fig. 12.

For ammonia production, the secondary reformer may use air for its oxygen supply because the accompanying nitrogen will be converted to ammonia. Also for ammonia production, the CO will be

unused, so it is converted to CO_2 by the water-gas shift reaction (reaction 2 in Fig. 12) that produces another mole of H_2 for each methane that was converted (for a total of four). Newer plants may make use of molecular sieve zeolite adsorbents to adsorb only H_2 for producing a pure hydrogen stream. In many current plants, the CO must be reduced to very low levels because it is a poison for the ammonia synthesis catalyst. This is done by converting the residual CO back to methane in a methanation reactor. The CO_2 is also removed using sweetening methods discussed earlier, leaving a nitrogen and hydrogen gas stream ready for the ammonia synthesis process.

Because of the energy intensity of steam reforming and its high H_2/CO ratio that is not needed for FTS (reaction 6 in Fig. 12), large-scale FTS plants are designed to use either partial oxidation or a combined reforming to produce synthesis gas with lower energy intensity. However, partial oxidation generally uses pure oxygen that adds to process cost and complexity. Combined reforming uses less oxygen by combining partial oxidation with steam reforming in several configurations. This balances the heat required by the endothermic steam reforming reaction against the heat released by the exothermic partial oxidation reaction, although high temperatures are still required.

Research is under way on membranes that separate the oxygen, particularly on certain ceramic oxygen transport membranes. This would eliminate the need for expensive cryogenic oxygen plants and make the reforming unit very compact.

Ammonia synthesis takes place on a supported iron catalyst. The overall reaction is shown as reaction 4 in Fig. 12. This is also an equilibrium limited reaction that is favored by low temperatures and very high pressures, taking place at approximately 400°C and 40 MPa. A process diagram for ammonia synthesis is presented in Fig. 13. The feed gas is compressed and preheated before flowing to the top of the converter. The reaction is exothermic and the gas temperature increases as the reaction proceeds, so the catalyst beds are separated into two or three distinct zones. Between each zone, cool hydrogen gas is injected to mix with and cool the reacting gas so as to allow a higher conversion. The conversion remains relatively low, and the product gas is cooled to below the ammonia dew point. The product gas then passes through the separator where liquid ammonia is recovered, and the remaining gas is recycled to the reactor feed.

Methanol synthesis takes place over a Cu/ZnO catalyst at a pressure of approximately 5 MPa and a temperature of 200 to 250°C . The reaction, shown as reaction 5 in Fig. 12, requires an H_2/CO ratio of 2, but this may be higher in practice. A process diagram is shown in Fig. 14. Feed gas is compressed and flows to a multiple-bed catalytic reactor in similar fashion to that in ammonia synthesis. This reaction is also equilibrium limited, and cool feed gas is added between catalyst zones to allow higher conversion. As with ammonia synthesis, methanol synthesis achieves a relatively low conversion of the synthesis gas feed, the product gas is cooled to condense the methanol and water reaction products along with

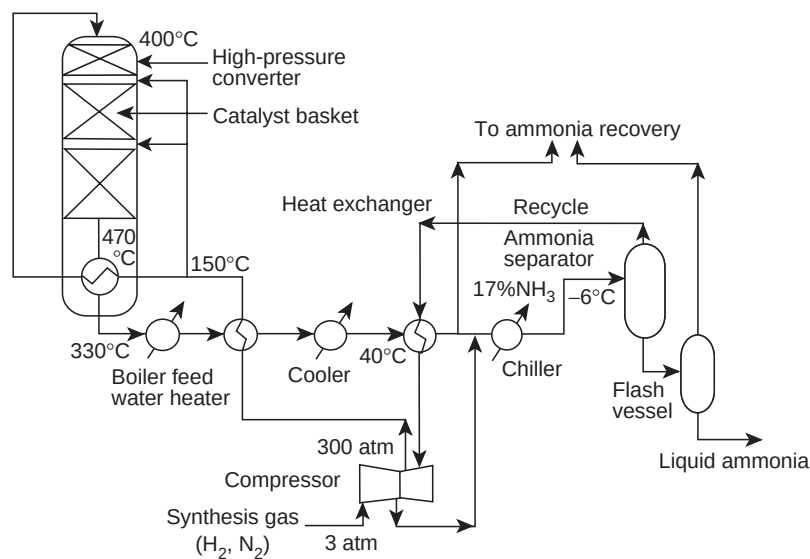


FIGURE 13 Ammonia synthesis process diagram.

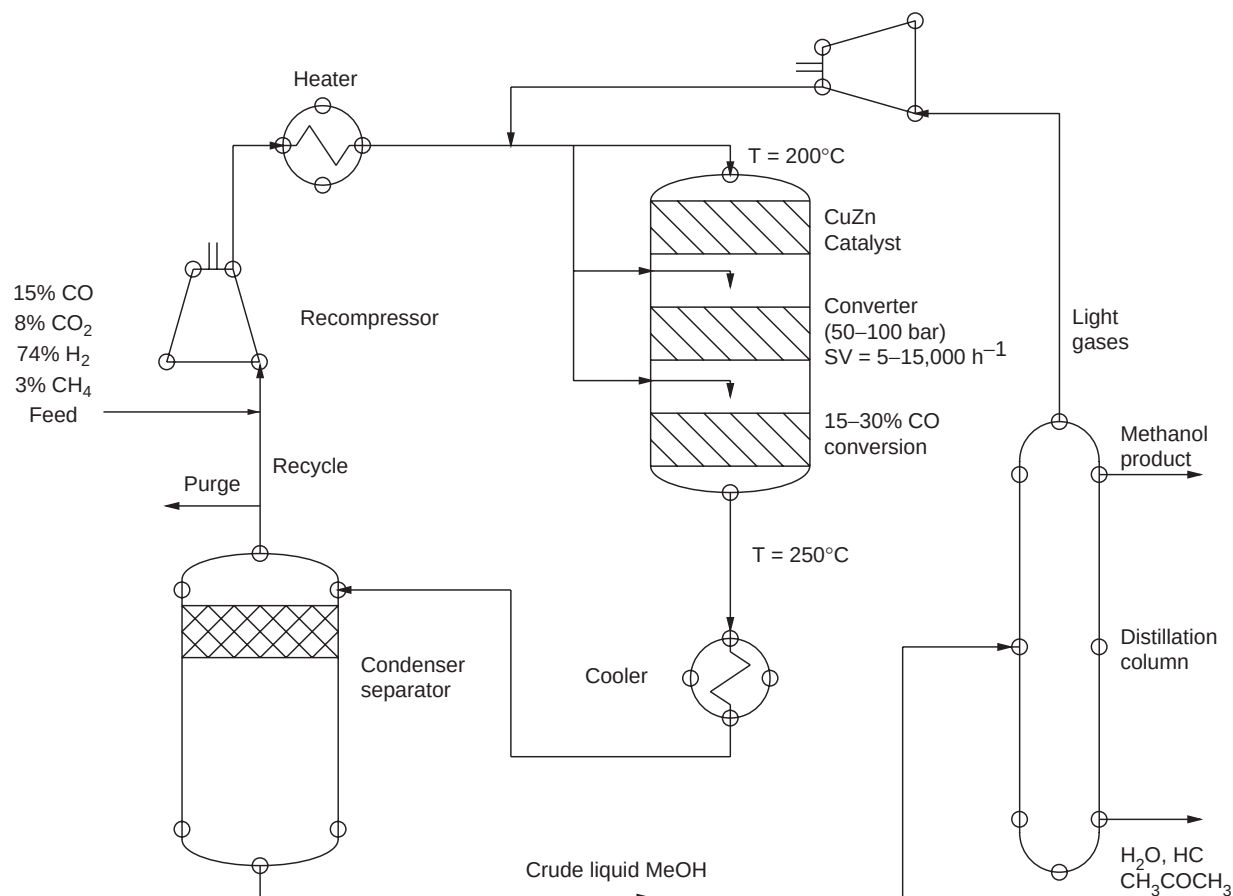


FIGURE 14 Methanol synthesis process diagram.

small amounts of by-products, and the gas is recycled to the feed. The crude product is distilled to produce various purity grades of methanol.

FTS has a complex chemistry that is simplified in reaction 6 in Fig. 12. The primary objective of large-scale Fischer–Tropsch gas-to-liquids (GTL) processes is to produce liquid hydrocarbons suitable for fuel use, and this is represented in reaction 6 by the production of the hydrocarbon-repeating unit $-CH_2-$. Under some conditions, such as high temperatures, oxygenates and olefins may be produced for specialty chemicals use. The reaction produces a wide range of molecular weight products that are predominantly straight chain alkanes. The product distribution produces some light gases (e.g., methane, ethane, propane) that are undesirable, some light liquid alkanes generally referred to as naphtha (a desirable feedstock for production of olefins), longer chain alkanes that are suitable for diesel fuel, and even longer chain molecules that are solids at room temperature called waxes that may be converted into diesel fuel and other useful, but smaller market,

products. The relative amounts of these may be adjusted by altering the catalyst and reaction conditions, but only within certain ranges. These are produced under so-called low-temperature synthesis conditions using supported cobalt- or iron-based catalysts. There are trade-offs between the two catalysts, with cobalt appearing to be more desirable despite its higher cost due to generally higher reaction rates and the fact that iron also catalyzes the water–gas shift reaction that undesirably consumes CO. A process diagram is shown in Fig. 15, which includes a complete process from natural gas to liquid hydrocarbons. The feed natural gas and oxygen both flow to the bottom of the fluidized bed synthesis gas (FBSG) reactor that uses combined reforming. The synthesis gas stream then flows to the hydrocarbon synthesis (HCS) reactor. The reactor operates with a catalyst suspended as a slurry in the liquid product with the gas feed bubbling up through the reactor. The reaction is exothermic, so heat is removed from the reactor with internal piping through which water is pumped and converted to

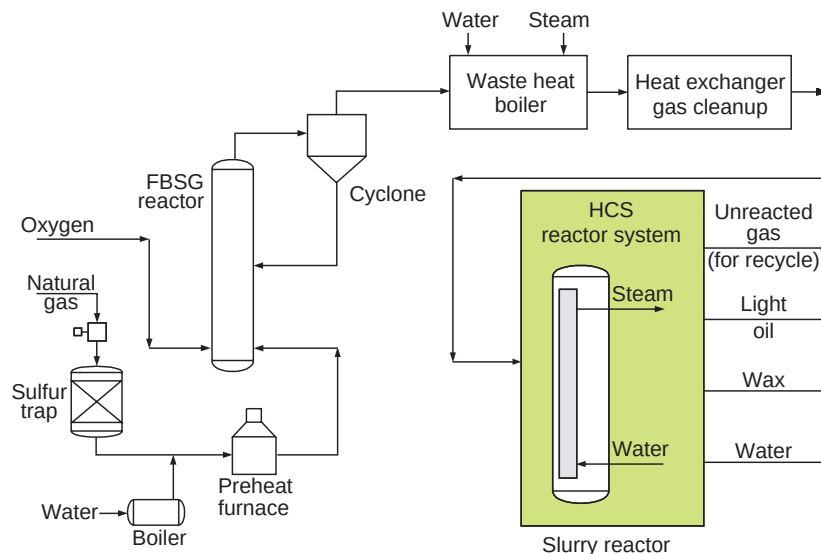


FIGURE 15 Example of a Fischer-Tropsch hydrocarbon synthesis process diagram. FBSG, fluidized bed synthesis gas reactor; HCS, hydrocarbon synthesis reactor.

steam. This maintains the desired reaction temperature of 200 to 250°C at the operating pressure of approximately 2.5 to 5.0 MPa. Relatively high conversions are achieved so that unreacted gas (and produced light hydrocarbons) is not recycled; rather, it is burned and its heat is recovered. The product liquid with catalyst flows out of the reactor, where the catalyst is separated and returned to the reactor suspended in a portion of the product hydrocarbon liquid. The water produced by the reaction is removed as a separate liquid phase, and the hydrocarbon liquid product is further refined to make final products ready for transportation by liquid pipelines or tankers.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Natural Gas • Natural Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Resources, Global Distribution of • Natural Gas

Resources, Unconventional • Natural Gas Transportation and Storage • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Drilling

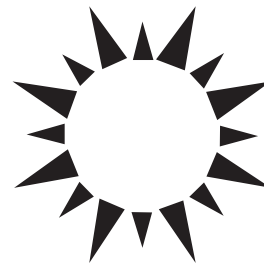
Further Reading

- Campbell, J. M. (1984). "Gas Conditioning and Processing," vols. 1–2, 6th ed. Campbell Petroleum Series, Norman, OK.
- Farrauto, R. J., and Bartholomew, C. H. (1997). "Fundamentals of Industrial Catalytic Processes." Blackie Academic and Professional, London.
- Gas Processors and Suppliers Association. (1998). "GPSA Engineering Data Book," vols. 1–2, 11th ed. GPSA, Tulsa, OK.
- Kohl, A. L., and Riesenfeld, F. C. (1985). "Gas Purification..," 4th ed. Gulf Publishing, Houston, TX.
- Rojey, A., Jaffret, C., Cornot-Gandolphe, S., Durand, B., Julian, S., and Valais, M. (1997). "Natural Gas Production Processing Transport." Editions Technip, Paris.
- Satterfield, C. N. (1991). "Heterogeneous Catalysis in Industrial Practice." 2nd ed. McGraw-Hill, New York.
- Twigg, M. V. (1989). "Catalyst Handbook." 2nd ed. Wolfe Publishing, London.
- Wender, I. (1996). Reactions of synthesis gas. *Fuel Process. Technol.* 48, 189.

Natural Gas Resources, Global Distribution of

RONALD R. CHARPENTIER

United States Geological Survey
Denver, Colorado, United States



-
1. Introduction
 2. Worldwide Distribution
 3. Impact of Distribution
 4. Data Sources

Glossary

associated/dissolved gas The natural gas that is in or from oil fields, i.e., fields with a gas/oil volume ratio of less than 20,000 cubic feet per barrel. The natural gas may be in separate pools within the oil field, floating on the oil as a gas cap, or dissolved in the oil.

cumulative production The total volumes of natural gas that have been produced.

endowment The sum of cumulative production, reserves, reserve growth, and undiscovered resources.

future resources The sum of reserves, reserve growth, and undiscovered resources.

liquefied natural gas (LNG) The liquid form of gas; can be stored under low temperature and high pressure for transportation by ship.

nonassociated gas The natural gas that is in or from gas fields, i.e., fields with a gas/oil volume ratio of at least 20,000 cubic feet per barrel.

reserve (or field) growth The increases of estimated gas volume that commonly occur as oil and gas fields are developed and produced.

reserves The estimated quantities of natural gas, expected to be commercially recovered from known accumulations relative to a specified date, under prevailing economic conditions, operating practices, and government regulations. Reserves are part of the identified (discovered) resources and include only recoverable materials.

resource A concentration of naturally occurring natural gas in or on the earth's crust, some of which is currently or potentially economically extractable.

stranded gas Discovered natural gas that is partially or completely isolated from markets because of distances

or insufficient transportation infrastructure, and thus is undeveloped or underdeveloped.

undiscovered resources Those resources postulated from geologic information and theory to exist outside of known oil and gas fields.

Natural gas is a mixture of hydrocarbon compounds that are in a gaseous state at standard conditions of temperature and pressure. Natural gas is very abundant worldwide. The greatest concentrations of conventional natural gas are in the Middle East and in the countries of the former Soviet Union. Because natural gas burns more cleanly and produces less greenhouse emissions compared to coal or oil, it is likely to provide a larger portion of the energy supply in the future.

1. INTRODUCTION

Some natural gas is found in fields that produce little or no volume of liquids. Such gas is called nonassociated gas; the fields that produce nonassociated gas have a gas/oil volume ratio of at least 20,000 cubic feet per barrel. Natural gas is also found in oil fields (fields with a gas/oil volume ratio of less than 20,000 cubic feet per barrel). In an oil field, natural gas can occur in separate pools or as a gas cap floating on the denser oil, in which cases it is termed "associated" gas. Oil field gas can also be dissolved in the oil, in which case it is termed "dissolved" gas.

1.1 Resource Classification

In the following discussion, conventional natural gas resources are classified in four categories: cumulative

production, reserves, reserve growth, and undiscovered resources. Cumulative production is that volume of natural gas that has already been produced. Reserves are those volumes remaining in already discovered oil and gas fields that are known to reasonable certainty. Reserve (or field) growth includes those volumes in already discovered oil and gas fields that are not known with certainty, but can be expected to be added to reserves in the future. Undiscovered resources are those volumes of natural gas in conventional oil and gas fields that have not yet been discovered. These volumes of gas are not static; as gas is discovered or produced, volumes shift from one category to another (e.g., from undiscovered resources to reserves, or from reserves to cumulative production). The sum of these four categories is called the endowment of natural gas. The future resources are the sum of the reserves, the reserve growth, and the undiscovered resources—that is, the endowment minus what has been already produced.

Many terms are used to identify categories and subcategories of resources. These terms are not always used consistently and must be approached with caution. The discussion here deals with conventional resources of natural gas, that is, those resources that are in discrete fields (usually with distinct gas/water or oil/water contacts). The bulk of historical natural gas production has been from conventional reservoirs. Unconventional natural gas resources, such as coal bed gas, tight gas sands, fractured shale gas, and gas hydrates, are commonly classified separately. They are treated here as unconventional gases.

1.2 Uncertainty of Estimates

Even more so than for oil, estimates of natural gas resources are subject to considerable uncertainty. This is true for all categories, including estimates of past production. Recorded gas volumes sometimes include total gas volume and at other times do not include that portion of natural gas that is noncombustible (nitrogen, CO₂, etc.). Generally, noncombustible components are less than 15% of gas volumes, but locally can be much greater. Throughout this article, total gas volumes are given (i.e., the noncombustible volume is included). Past production, as reported by the appropriate government agencies, may or may not include amounts of gas that were flared or reinjected into the reservoir to sustain pressure for additional oil recovery.

The amount of reserve growth in discovered fields is not well known, but is probably very significant. Especially in the case of stranded gas (discovered natural gas that is partially or completely isolated from markets because of distances or insufficient transportation infrastructure), there is little incentive to invest in the costs to evaluate fields fully. Although most studies of reserve growth have been based on more easily accessible data from the United States, studies of reserve growth outside the United States show that such growth occurs worldwide. Gas reserve growth outside the United States reported from 1977 to 1991 was at a rate greater than was growth of gas fields in the United States during that period.

2. WORLDWIDE DISTRIBUTION

A majority of the data used in this article on the worldwide distribution of conventional natural gas resources are from the U.S. Geological Survey (USGS) and the U.S. Energy Information Administration. The USGS provides probabilistic estimates of reserve growth and undiscovered resources, but only the mean estimates are cited here. The original USGS estimation of reserve growth was at the world level; for the analysis here, estimates of reserve growth by region were generated by allocating the USGS world reserve growth estimates proportional to the volumes of reserves in each region.

Table I and Fig. 1 show how conventional natural gas resources of the world are distributed into the eight regions used by the USGS. Note that 62% of the resource endowment is concentrated in just two regions—(1) the former Soviet Union and (2) the Middle East and North Africa. Also note that only in the North American region has a large proportion of the resource already been produced. The map in Fig. 2 shows how the resources that remain available are concentrated primarily in a part of the world that stretches from the Middle East, through the Caspian Sea, to West Siberia. Table II presents the natural gas production and consumption by region for the year 2000. The major trade movements of natural gas are shown in Fig. 3. These eight regions are discussed herein in decreasing order of their endowment; Tables I and II and Figs 1–3 should be referred to throughout the discussions.

The world distribution of unconventional gas resources is not well understood. Most of the exploration and production of unconventional resources has been in the United States. Geologic studies show that similar unconventional accumulations

TABLE 1

World Conventional Natural Gas Resources by Region^a

Region ^b	Trillion cubic feet						World cumulative production (%)	World reserves (%)	World reserve growth (%)	World undiscovered resources (%)	World future resources (%)	World endowment (%)
	Cumulative production	Reserves	Allocated reserve growth	Undiscovered resources	Future resources	Endowment						
1. Former Soviet Union	389.20	1682.95	1203.79	1611.26	4498.01	4887.20	22.21	35.12	32.89	31.01	32.95	31.73
2. Middle East and North Africa	73.49	1835.57	1312.96	1369.93	4518.47	4591.96	4.19	38.30	35.87	26.36	33.10	29.82
3. Asia Pacific	56.50	344.52	246.43	379.34	970.29	1026.79	3.22	7.19	6.73	7.30	7.11	6.67
4. Europe	204.30	279.71	200.08	312.37	792.15	996.46	11.66	5.84	5.47	6.01	5.80	6.47
5. North America	938.60	256.51	415.45	681.50	1353.46	2292.06	53.56	5.35	11.35	13.11	9.92	14.88
6. Central and South America	55.26	223.61	159.95	487.19	870.75	926.01	3.15	4.67	4.37	9.38	6.38	6.01
7. Sub-Saharan Africa and Antarctica	7.43	106.57	76.23	235.29	418.09	425.52	0.42	2.22	2.08	4.53	3.06	2.76
8. South Asia	27.59	63.07	45.11	119.61	227.79	255.38	1.57	1.32	1.23	2.30	1.67	1.66
Total	1752.37	4792.52	3660.00	5196.49	13649.01	15401.38	100.00	100.00	100.00	100.00	100.00	100.00

^a All estimates reported here are mean estimates. Gas resource totals from U.S. Geological Survey. Allocated reserve growth from U.S. Geological Survey estimates, allocated (for non-U.S.) proportional to volumes of remaining reserves.

^b Numbers correspond to maps in Figs. 2 and 3.

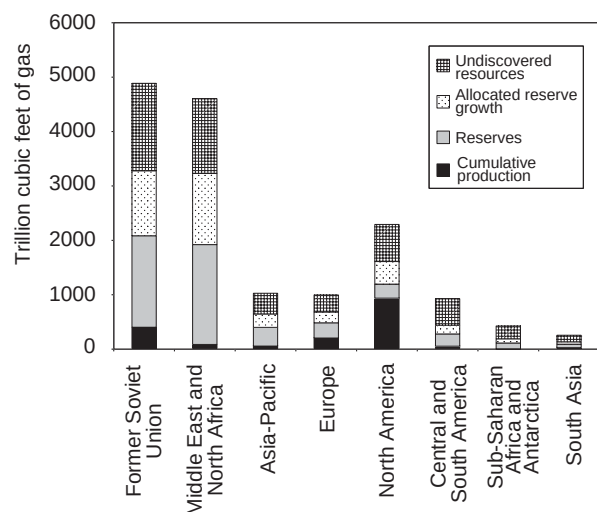


FIGURE 1 Conventional natural gas endowment (cumulative production, reserves, allocated reserve growth, and undiscovered resources) by region (see data in Table I).

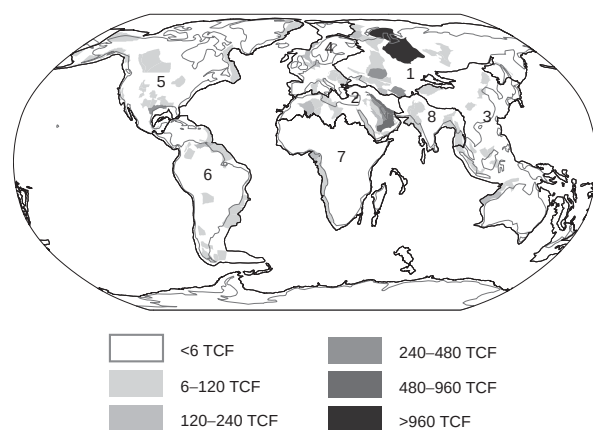


FIGURE 2 Natural gas of the world by geologic province as assessed in the U.S. Geological Survey World Petroleum Assessment 2000. The volumes are a sum of cumulative production, reserves, and undiscovered resources. Reserve growth was not included in the volumes. The shaded regions indicate conventional gas in trillions of cubic feet (TCF). Region key: (1) The former Soviet Union, (2) Middle East and North Africa, (3) Asia-Pacific, (4) Europe, (5) North America, (6) Central and South America, (7) sub-Saharan Africa and Antarctica, and (8) South Asia.

probably occur in many places worldwide, but quantitative estimates are not available.

2.1 Former Soviet Union

The former Soviet Union (FSU) has 33% of the world's future natural gas resources. In 2000, this region accounted for 29% of the world's natural gas production but had only 24% of the natural gas

demand; thus, even though a major user of natural gas, the FSU is also a major exporter. Almost all of the export volume is to Europe through pipeline systems.

Of the 25.43 trillion cubic feet (10^{12} cubic feet, TCF) of production in 2000, 20.63 TCF (81%) was from Russia, mostly from the West Siberian Basin. The West Siberian Basin also has 63% of the FSU's natural gas reserves and 40% of the estimated undiscovered gas. The area of the Caspian Sea also contains very large estimated volumes of natural gas resources, accounting for 12% of the reserves and 21% of the estimated undiscovered gas. The Amu-Darya Basin of central Asia (Turkmenistan, Uzbekistan, Afghanistan, and Iran) has large reserves (9% of the FSU total) and estimated undiscovered gas (10% of the FSU total), but the gas is far from demand centers, so much of the gas is stranded. The Barents Sea has had few previous discoveries, but has significant estimated undiscovered gas (15% of the FSU total).

2.2 Middle East and North Africa

The Middle East and North Africa region also has 33% of the world's estimated future natural gas resources. This includes 38% of the world's reserves and 26% of the world's estimated undiscovered gas. In 2000, this region accounted for 13% of world production, but only 10% of the demand. The region is a major exporter to Europe through trans-Mediterranean pipelines and by transportation of liquefied natural gas (LNG). Much of the gas in the Arabian-Persian Gulf area is stranded, but some is exported to Japan and South Korea as LNG. The Arabian-Persian Gulf area accounts for 87% of the region's reserves and 90% of the estimated undiscovered gas. It includes the world's largest gas field, North Field (over 350 trillion cubic feet of natural gas), in Qatar. Algeria exports a significant volume of natural gas to Europe (69% of the region's exports) by both pipeline and as LNG.

2.3 North America

North America has 10% of the world's estimated future natural gas resources. Of all eight regions listed in Table II, North America is both the largest producer and the largest consumer of natural gas. Over half of the natural gas ever produced in the world has been from this region. It is the only region for which unconventional natural gas has been a significant portion of the production.

The United States is the largest producer (71% of North America's 2000 production) and the largest

TABLE II

World Natural Gas Production and Consumption in the Year 2000^a

Region ^b	Trillion cubic feet		World production (%)	World consumption (%)	Surplus or deficit (%)
	Gas production	Gas consumption			
1. Former Soviet Union	25.43	20.54	28.89	23.51	+ 5.38
2. Middle East and North Africa	11.51	8.87	13.07	10.15	+ 2.92
3. Asia Pacific	7.29	8.20	8.28	9.38	-1.10
4. Europe	10.91	16.85	12.39	19.29	-6.90
5. North America	26.79	27.21	30.43	31.14	-0.72
6. Central and South America	3.43	3.30	3.89	3.78	+ 0.11
7. Sub-Saharan Africa and Antarctica	0.57	0.34	0.64	0.39	+ 0.25
8. South Asia	2.11	2.06	2.40	2.36	+ 0.04
Total	88.03	87.38	100.00	100.00	

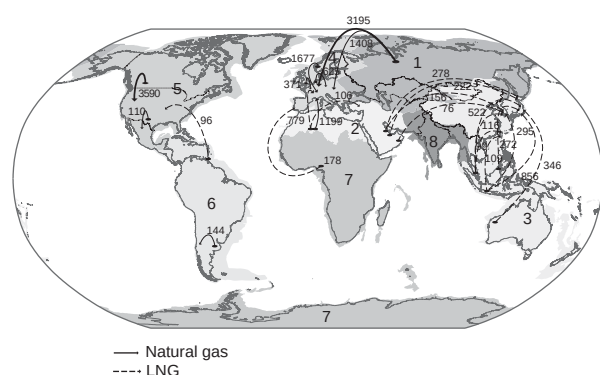
^aData from Energy Information Administration (2001).^bNumbers correspond to maps in Figs. 2 and 3.

FIGURE 3 Major natural gas trade movements in billions of cubic feet (small numbers) for the year 2000. Region key (large numbers): (1) The former Soviet Union, (2) Middle East and North Africa, (3) Asia-Pacific, (4) Europe, (5) North America, (6) Central and South America, (7) sub-Saharan Africa and Antarctica, and (8) South Asia. Modified from British Petroleum Company (2002).

consumer (83% of North America's 2000 consumption). About 60% of U.S. production is from the onshore and offshore areas of the Gulf of Mexico. Almost all of the gas imported into the United States (94% of U.S. imports) comes from Canada by pipeline. Mexico imports only 3% of its demand, which is supplied by the United States.

2.4 Europe

Europe has 6% of the world's future natural gas resources. It is a net importer of natural gas,

having consumed almost 6 trillion cubic feet of gas more than it produced in 2000 (35% of its consumption). Of the gas imported into Europe in 2000, 66% was from the former Soviet Union and 30% was from Algeria. Significant resources of European natural gas occur in the North Sea and adjoining areas. The largest producers in the North Sea area are the United Kingdom (35% of Europe's 2000 production), The Netherlands (24%), Norway (17%), and Germany (7%). Of these, Norway, The Netherlands, and the United Kingdom are net exporters. Almost all other European countries are net importers. Gas production in other parts of Europe is primarily from the Po Valley of northern Italy and Romania.

2.5 Asia-Pacific

The Asia-Pacific region has 7% of the world's future natural gas resources. The region accounts for 7% of world natural gas reserves, 7% of the estimated undiscovered gas, and 9% of the demand. Japan and South Korea, with little production, import LNG from Indonesia, Malaysia, Australia, and Brunei, and also from the Arabian-Persian Gulf. Areas of significant production are Malaysia and Indonesia, with 53% of the region's 2000 production. They have 42% of the reserves and 42% of the estimated undiscovered gas in the region. Northwest Australia also has significant resources—25% of the region's reserves and 29% of the region's estimated undiscovered gas—but some of the reserves are stranded. China has

significant resources also—13% of the region's reserves and 23% of the estimated undiscovered gas.

2.6 Central and South America

Central and South America have about 6% of the world's future natural gas resources. There is little interregional import or export, except for some LNG exported from Trinidad and Tobago to the United States. Argentina and Venezuela are the major natural gas producers of the region, accounting for 39 and 28%, respectively, of the region's production in 2000. Venezuela has 60% of the region's gas reserves and 21% of the estimated undiscovered gas. Argentina has 11% of the region's reserves and 8% of the estimated undiscovered gas. Other significant producers include Trinidad and Tobago, Brazil, Colombia, and Bolivia. Although Brazil currently has only 3% of the region's reserves, it has 40% of the estimated undiscovered gas, most of it in basins along the South Atlantic margin. Most of the region's natural gas is consumed in the country of production, but Argentina, Trinidad and Tobago, and Bolivia are net exporters. Chile and Brazil are net importers of natural gas.

2.7 Sub-Saharan Africa and Antarctica

Sub-Saharan Africa has about 3% of the world's estimated future natural gas resources. The Niger Delta of Nigeria dominates natural gas statistics for sub-Saharan Africa. Most (78%) of the natural gas already produced in sub-Saharan Africa was from the Niger Delta. The Niger Delta also has 82% of the reserves. Nigeria is the only major exporter of natural gas to other regions, that by transportation of natural gas liquids (NGLs). Of its exported gas, most (90%) is to Europe and 10% is to the United States.

In the past decade, deepwater oil exploration off the Atlantic coast of Africa has been very successful; several giant oil fields have been discovered. Improved geologic understanding about the South Atlantic coastal basins suggests that the undiscovered gas may be more widespread than was obtained from past production—52% of the estimated undiscovered gas is in Nigeria and 18% is in (mostly offshore) Angola. With an expansion of deepwater infrastructure and technology, gas production from these areas may become more economically viable, both from oil fields (associated/dissolved gas) and from gas fields (nonassociated gas).

No natural gas resources have yet been discovered in Antarctica. Even if exploration were allowed, any resources found would be very unlikely to be economic in the next few decades.

2.8 South Asia

South Asia has only about 2% of the world's future natural gas resources. South Asia is not well connected to other parts of the world by gas transportation routes and the gas production and consumption are roughly in balance. The region includes two of the world's largest deltas, the Indus and the Ganges-Brahmaputra, both with significant undiscovered resources of natural gas. The largest producers are Pakistan (about 40% of the region's production), India (38%), and Bangladesh (16%). Each of these countries consumes what it produces, with minimal imports or exports of natural gas. In contrast, Burma exports almost half of the gas that it produces.

3. IMPACT OF DISTRIBUTION

Although, natural gas resources are abundant and widespread, they have been underutilized compared to oil resources. Only about 11% of the recoverable conventional gas resource has been produced. Unconventional natural gas is even more abundant, but much of it is not likely to be developed in the next few decades. Tables III and IV list the major producing and consuming countries of the world. Russia and the United States dominate both lists. Russia is a net exporter of natural gas and the United States is a net importer. Some of the major consumers (Germany, Japan, Italy, and France) are only minor producers. Some major producers are only minor consumers (Algeria, Norway, Turkmenistan, and Malaysia).

The geographic distribution of supply is different from that of demand, making transportation critical. Areas such as northern Alaska, northwest Australia, and central Asia have large volumes of discovered natural gas that are poorly connected to areas of gas demand. Transportation to areas of high demand requires the construction of either natural gas pipelines or of facilities to use high pressure and low temperature to liquefy natural gas so that it can be transported by ship as LNG. The gas stranded in these areas could also become more relevant to world supply with improvements in gas-to-liquid technology.

TABLE III

Top 20 Producers of Natural Gas in Year 2000^a

Country	Production (trillion cubic feet)
Russia	20.63
United States	18.99
Canada	6.47
United Kingdom	3.83
Algeria	2.94
Netherlands	2.57
Indonesia	2.36
Iran	2.13
Uzbekistan	1.99
Norway	1.81
Saudi Arabia	1.76
Turkmenistan	1.64
Malaysia	1.50
United Arab Emirates	1.41
Mexico	1.33
Argentina	1.32
Australia	1.12
Qatar	1.03
Venezuela	0.96
China	0.96

^aData from U.S. Energy Information Administration (2001).

TABLE IV

Top 20 Consumers of Natural Gas in Year 2000^a

Country	Production (trillion cubic feet)
United States	22.55
Russia	14.13
United Kingdom	3.38
Canada	3.29
Germany	3.10
Ukraine	2.78
Japan	2.75
Italy	2.49
Iran	2.22
Saudi Arabia	1.76
Netherlands	1.72
Uzbekistan	1.51
France	1.43
Mexico	1.38
Argentina	1.17
Indonesia	1.08
United Arab Emirates	0.97
Venezuela	0.96
China	0.96
Pakistan	0.86

^aData from U.S. Energy Information Administration (2001).

4. DATA SOURCES

Information about natural gas resources is not static. Fortunately, several sources of information are easily accessible. The data about discovered resources used in this article came primarily from the U.S. Energy Information Administration. Their Web site (www.eia.doe.gov) provides extensive data on resources and energy usage for all energy sources (not just oil and natural gas) for the United States and the rest of the world. The data for reserve growth and undiscovered resources came primarily from the U.S. Geological Survey's *World Petroleum Assessment 2000*. The USGS has a Web site (energy.usgs.gov) where that assessment and additional extensive data for the United States and the rest of the world may be accessed.

Other sources of information can be found in several petroleum industry magazines (such as *Oil and Gas Journal* and *World Oil*) that publish annual statistical summaries. The British Petroleum Company (BP) also produces an annual statistical summary and has a Web site (www.bp.com) where extensive data is available.

SEE ALSO THE FOLLOWING ARTICLES

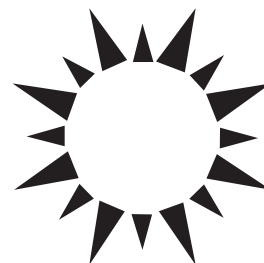
Depletion and Valuation of Energy Resources • Markets for Natural Gas • Natural Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Processing and Products • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil and Natural Gas Resource Assessment: Geological Methods • Oil and Natural Gas Resource Assessment: Production Growth Cycle Models

Further Reading

- British Petroleum Company (BP). (2002). "BP Statistical Review of World Energy June 2002." BP, London. Also available on the Internet at <http://www.bp.com>.
- DeGolyer and MacNaughton Consultants (2000). "Twentieth Century Petroleum Statistics," 56th ed. DeGolyer and MacNaughton, Dallas, Texas.
- Energy Information Administration (EIA) (2001). "International Energy Outlook 2001." U.S. Department of Energy,

- Washington, D.C. Also available on the Internet at <http://www.eia.doe.gov>.
- International Energy Agency (IEA). (2000). "World Energy Outlook 2000." Organisation for Economic Co-Operation and Development, Paris.
- PennWell Corp. (2000). "International Petroleum Encyclopedia." PennWell, Tulsa, Oklahoma.
- U.S. Geological Survey (USGS) World Energy Assessment Team (2000). "U.S. Geological Survey World Petroleum Assessment 2000—Description and Results." U.S. Geological Survey Digital Data Series DDS-60 (four CD-ROMs). USGS, Reston, Virginia. Also available on the Internet at <http://energy.usgs.gov/>.

Natural Gas Resources, Unconventional



VELLO A. KUUSKRAA

Advanced Resources International, Inc.
Arlington, Virginia, United States

-
1. Background
 2. Definitions for Unconventional Gas
 3. Resource Estimates for Unconventional Gas
 4. Unconventional Gas Development in the United States
 5. Conclusion

Glossary

coalbed methane Natural gas stored within a coal seam (coal reservoir) requiring dewatering and/or lowering of pressure for its release and production.

gas hydrates Natural gas stored as a clathrate within the molecular structure of water requiring the injection of heat or the lowering of pressure for dissociation and gas flow.

gas shales Natural gas stored within an organically rich shale interval (reservoir) requiring natural fractures and artificial well stimulation for acceptable rates of gas flow.

geopressured methane Natural gas stored as dissolved gas in highly pressured (geopressured) saline brines requiring the lowering of pressure for gas release and flow.

tight gas sands Natural gas stored within a low-permeability (tight) reservoir requiring natural and/or induced well stimulation for acceptable rates of gas flow.

unconventional gas Natural gas stored in a variety of geologically complex (unconventional) reservoirs requiring advanced exploration and production technology for their commercial development.

Unconventional natural gas has become a major source of energy supply in the United States. A large, favorable resource base, advances in knowledge, and steady progress in extraction technology have enabled significant portions of this resource to be converted to producible reserves. However, worldwide development and assessments of unconventional gas are still limited, concentrated primarily on coalbed methane and selected tight gas plays. This

article provides a status report on our understanding of the size and nature of worldwide unconventional gas resources. The intent is that the information and insights in this article will be of value in accelerating the development of these large, complex natural gas resources.

1. BACKGROUND

For some time, geologists and resource analysts have known that large volumes of natural gas are locked in geologically complex, unconventional reservoirs, such as tight gas sands, gas shales, and coalbed methane (CBM). These unconventional gas reservoirs exist in the United States, Canada, and many other areas of the world. In the past, the natural gas resources stored in these geologically complex reservoirs have been judged as being too costly and technically difficult to recover to be a viable source for future natural gas supplies.

With advances in technology and understanding, the status and outlook for unconventional gas resources has become much more positive. Today, in the United States, unconventional gas provides 5.4 trillion ft³ (Tcf) annually and accounts for 27% of domestic natural gas production, with each of the three unconventional gas sources—tight gas sands, coalbed methane, and gas shales—making a significant contribution, as shown in Fig. 1. The importance of unconventional gas is underscored by the recognition that 8 of the largest 12 U.S. natural gas fields exclusively produce unconventional gas (see Table I). Most significant, as set forth in the U.S. Department of Energy (DOE)/Energy Information Agency's Annual Energy Outlook 2003, unconventional gas is expected to become, by 2015, the largest single source of U.S. natural gas supply, surpassing

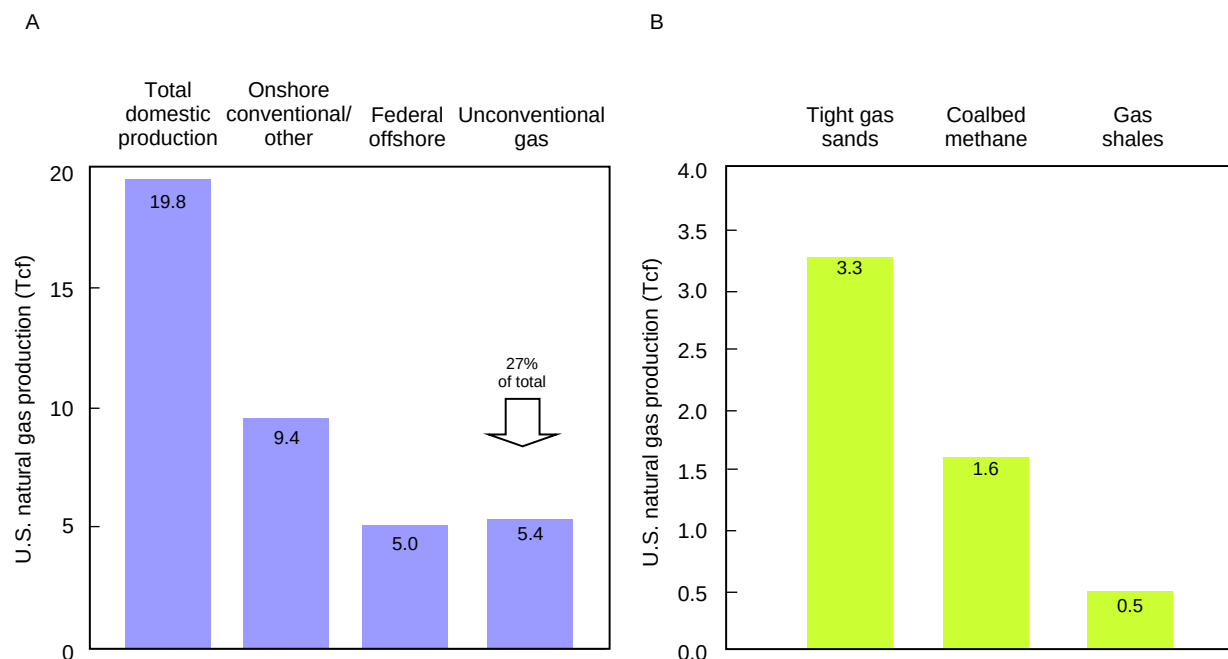


FIGURE 1 Sources of U.S. natural gas production in 2001 (Tcf). (A) From: Conventional/Offshore—Energy Information Agency Annual Reserve Report 2001. Unconventional—Advanced Resources International database. (B) From: Advanced Resources International Database.

TABLE 1

Eight of the 12 Largest U.S. Natural Gas Fields Are Unconventional Gas Fields

Rank	Field name	Basin/State	Type of resource	Year 2001 production (billion cft ³)
1	Blanco/Igancio-Blanco	San Juan, NM/CO	CBM/Tight gas sands	812
2	Basin	San Juan, NW Powder River,	CBM/Tight gas sands	617
4	Wyodak/Big George	WY	CBM	224
5	Carthage	East Texas, TX	Tight gas sands	203
6	Jonah	GGRB, WY	Tight gas sands	176
8	Giddings	East Texas, TX	Tight gas sands/chalk	152
9	Wattenberg	Denver, CO	Tight gas sands	143
12	Newark East	Ft. Worth, TX	Gas shale	114

Source. Advanced Resources International, Energy Information Agency 2001 Annual Reserve Report.

production from onshore conventional gas fields and offshore sources (see Fig. 2).

Worldwide, the pursuit of unconventional gas is also growing in importance. Significant drilling for coalbed methane is under way in Canada's Western Canadian Sedimentary Basin; Australia is actively producing tight gas sands from the Cooper Basin and coalbed methane from the Bowen Basin, and China, India, and South Africa (among other countries) are actively exploring for new coalbed methane plays.

Still, little is known about the ultimate size, nature, and producibility of these vast natural gas resources. The purpose of this article is to expand this base of knowledge.

2. DEFINITIONS FOR UNCONVENTIONAL GAS

Unconventional natural gas is defined as conventional natural gas (primarily methane) trapped in

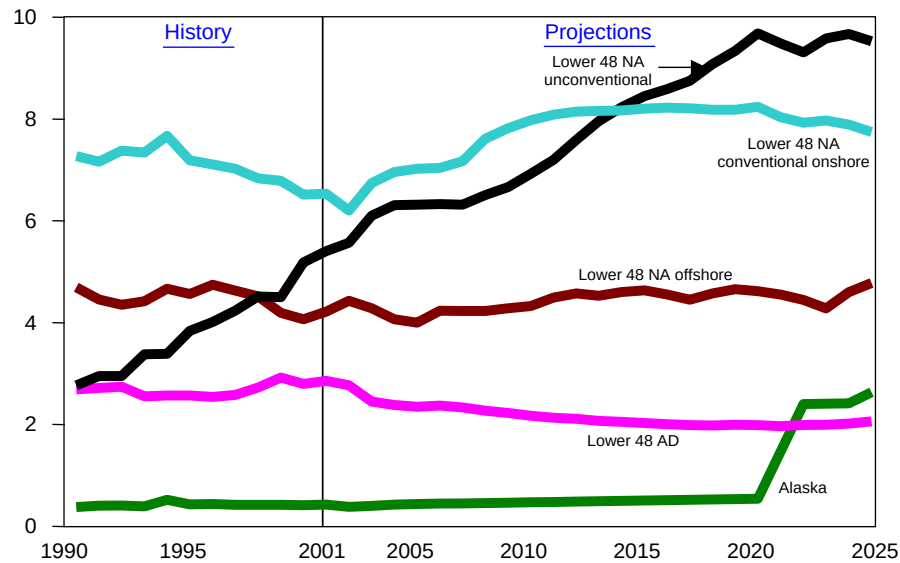


FIGURE 2 U.S. natural gas production, 1990–2025 (Tcf). Source: Advanced Resources International, Energy Information Agency (2003).

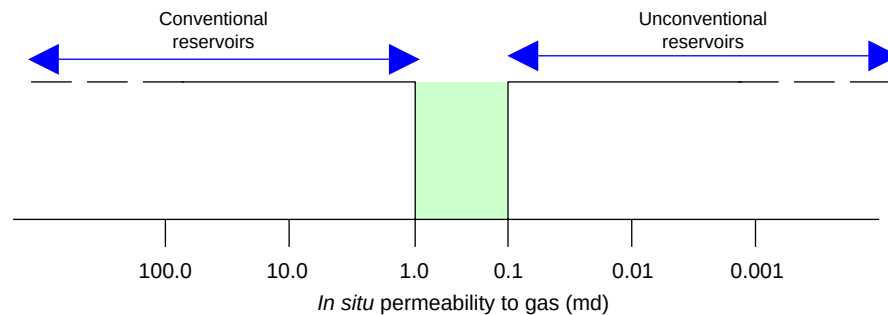


FIGURE 3 Permeability continuum for conventional through unconventional gas reservoirs.

geologically complex, nonconventional reservoirs, such as tight (low-permeability) sands, gas-bearing shales, and coalbeds. In the longer term, more speculative resources, such as gas hydrates, ultra-deep gas, and geopressed methane, may be included in this resource category.

2.1 Tight Gas Sands

Tight gas sands are low-permeability (tight) gas-bearing reservoirs that have an *in situ* matrix permeability to gas of less than 0.1 md, exclusive of natural fracture permeability. Figure 3 illustrates the permeability continuum and criteria for conventional and unconventional gas reservoirs. In contrast to conventional gas reservoirs where the gas is held in structural or stratigraphic traps, tight gas reservoirs are areally extensive, often called continuous-type

accumulations. These reservoirs occur in nearly all petroleum provinces, are usually abnormally pressured, and are often found in basin-center settings. Although they are commonly called tight gas sands, low-permeability reservoirs are found in a variety of rock types, including sandstone, siltstone, carbonate, limestone, dolomite, and chalk.

The porosity of lenticular and marginal-marine tight gas sand reservoirs is low, generally less than 12%, and the reservoir pores have high capillary pressure as a result of postdepositional cementation of conventional reservoir rocks. A second type of tight reservoir, such as a chalk, has high porosity of 15 to 35% but low permeability to gas because the rock is very fine-grained.

Artificial well stimulation, typically hydraulic fracturing, is required to attain commercial rates of gas flow from tight gas sands, unless the well

encounters an extensive cluster of natural fractures that enhances permeability. Identifying these naturally enhanced areas of permeability, the so-called “sweet spots” in an otherwise tight accumulation, is a major research and exploration objective for the industry.

2.2 Gas Shales

Gas shales are self-sourcing reservoirs where a major portion of the resource is adsorbed on and within the organic matter in the shale. To be productive and have adequate flow paths, the gas shale reservoir must be naturally fractured. Because of these conditions, gas shales are often called organic-rich gas shales or fractured gas shales.

Gas shale accumulations tend to be regionally pervasive, continuous-type accumulations and are abnormally pressured (generally underpressured). The quality of gas shales is judged by the total amount of organic matter (organic-rich shales are those containing more than 2% organic matter) and its degree of thermal maturity. Vitrinite reflectance values of over 1.5% are preferred for identifying areas with potential for thermogenic gas. Also important is the type of organic matter within the shales, the thickness and depth of the reservoir, and particularly the regional and local natural fracture intensity of the gas shale accumulation. These factors, in combination, determine the gas-producing potential of a gas shale reservoir.

Because the per well reserves of gas shales are often modest, low well costs, large-scale develop-

ment (for economies of scale), induced fracture stimulation, and natural fracture (sweet spot) exploration are important for commercial, sustained production of natural gas from gas shales.

2.3 Coalbed Methane

Large volumes of methane, along with water and other gases, are generated in the process of converting organic matter to coal. This in-place generated gas, as well as deeper gas that may have migrated into the coal reservoir, is adsorbed (stored) upon and within the molecular structure of the coal. As such, coalbeds are both a source and a reservoir for natural gas. The amount of gas stored in the coal is a function of the coal's rank, pressure, temperature, and degree of gas saturation. High coal ranks and pressures, in general, lead to higher gas storage capacity. The storage capacity is commonly determined by using a sorption (or gas content) isotherm, as shown in Fig. 4.

The permeability of the coal matrix is nonexistent for all practical purposes. Gas moves through the angstrom-size pores in the coal by diffusion, an extremely slow flow process in most cases. As such, the primary flow paths for methane in coal seams are provided by the coal's cleat system, as well as its natural fracture overprint.

Dewatering and subsequent depressurization of the coal reservoirs are used to release the gas sorbed in the coal. Artificial well stimulation, such as hydraulic fracturing or cavitation, is used to create

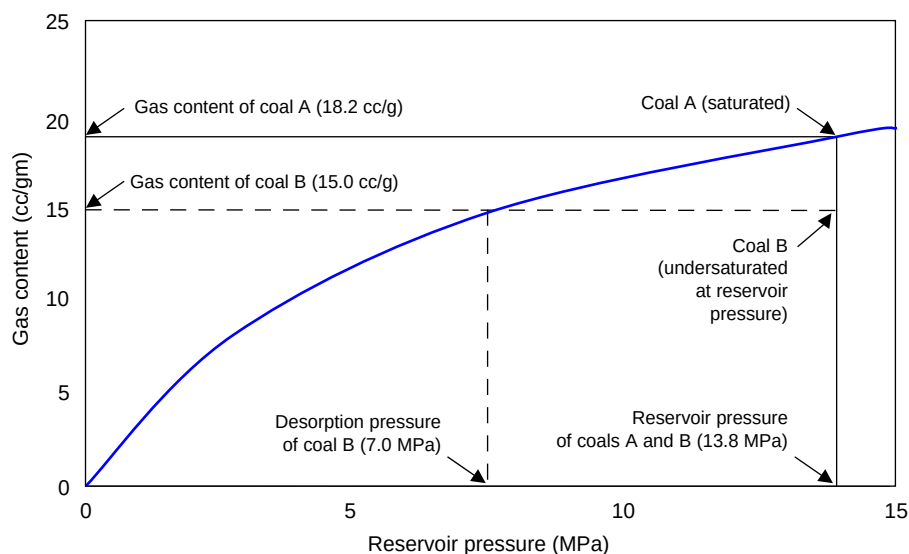


FIGURE 4 Idealized coalbed gas sorption isotherm showing relation between reservoir pressure and gas content for a saturated (coal A) and undersaturated (coal B) coal.

additional flowpaths to accelerate gas flow from the coal face and cleat systems to the wellbore.

As may be surmised from the above definitions for tight gas sands, gas shales, and coalbed methane, several features are common to unconventional gas resources:

1. First, the unconventional deposits may contain large, even massive volumes of gas in-place, often at low concentrations;
2. Second, despite the large volume of gas in-place, only limited areas will have sufficient richness and permeability for commercially developable reserves; and
3. Third, continued improvements in exploration and production technology, such as natural fracture detection, advanced reservoir diagnostics, and lower-cost well drilling and completion practices, are essential if these resources are to be effectively developed and converted to reserves.

Despite their widespread, global presence, the development of unconventional gas resources, particularly tight gas sands and gas shales, has been largely limited to the United States and much of the information in this article is based on this experience. The intent is that the information and insights in this article will be of value in developing the large, complex unconventional gas resources of the world.

3. RESOURCE ESTIMATES FOR UNCONVENTIONAL GAS

As more has been learned about the nature and location of unconventional gas, its resource estimates have increased and gained in confidence.

3.1 Initial Resource Estimates

The initial resource estimates for global unconventional gas were prepared by Kuuskraa and Meyers in the early 1980s, at a time when only limited information was available on the in-place and recoverable resource potential from tight gas formations, coalbed methane, fractured gas shales, or geopressured methane accumulations. Even less was known about methane hydrates or ultradeep gas.

As shown in Table II, the authors estimated that 570 Tcf of unconventional gas may be technically recoverable from the United States and Canada out of a gas in-place estimate of 2000 Tcf. The estimates for other areas of the world of 850 Tcf of technically

TABLE II

Technically Recoverable Resource Estimates for Unconventional Gas (Tcf)

	United States/Canada	Other areas	Total
Gas shales	50	250	300
Tight gas sands	460	400	860
CBM	60	200	260
Total	570	850	1420

Source. Kuuskraa and Meyers (1983).

recoverable resources were much more speculative and conservative and were based on analogues from the U.S. experience.

3.2 Global Gas Resources Workshop

The second comprehensive set of resource estimates for unconventional gas was provided by the "Workshop on Global Gas Resources," convened in Vail, Colorado, in September 1994. The purpose of the workshop was to evaluate the worldwide gas resource base and the volume of resources that could be added with significant progress in technology. The workshop participants included presentations by 32 resource experts from 15 countries, representing international and domestic producers, state-owned energy companies, government organizations, universities, and research institutions.

The workshop participants concluded that advanced technology and unconventional resources could increase the 1993 U.S. Geological Survey (USGS) benchmark resource estimate for global natural gas by nearly 3000 Tcf, as shown in Table III and Fig. 5. Because unconventional resources in tight formations, gas shales, and coalbeds were not included in the 1993 USGS benchmark estimate (except for North America), the bulk of the 3000 Tcf of additional resource (producible with advanced technology to 2015) was estimated to be from unconventional gas.

Detailed estimates for unconventional gas were provided for the United States, Canada, and China by Fisher, Young and Drummond, and Miao, respectively. An overall review of global unconventional natural gas was provided by Kuuskraa. Of particular note were the closing remarks by Fisher on U.S. natural gas resources, setting forth long-term expectations for adding another 10,000 Tcf to the global gas resource base from unconventional gas and technology progress:

...the explicit incorporation of technology in resource estimates has doubled the U.S. estimate of ultimate recovery of natural gas...If the experience in the United States were extrapolated to the global resource base, the world reserves would exceed 20,000 Tcf, and they would be accessible at lower costs.

—Fisher, 1994

3.3 Resource Estimates

As the demand for natural gas has grown, considerably more attention has been placed on understanding

TABLE III

Summary of Workshop Estimates for Natural Gas Resources (Tcf)

	USGS benchmark remaining resources, ^a 1993	Technology- enhanced resources, by 2015
North America	1394	2259
South America	715	987
Europe	617	783
Africa	813	1000
Middle East	2697	2799
Asia and Australasia	963	1733
Former Soviet Union	4361	4928
Total	11,560	14,489

^aTotal remaining resources are the sum of identified reserves and undiscovered gas (mean).

the size, location, and producibility of unconventional gas.

3.3.1 Rogner

Rogner, in 1997, published his latest assessment of global unconventional gas resources (see Table IV). Even after a considerable number of assessments, he noted that the data on unconventional gas is still limited, that estimates “are fraught with geological uncertainty,” that “the technology implications for the eventual production of unconventional gas are poorly understood,” and that “the data are speculative and should be read as such.” Still, the information published by Rogner represented a significant step forward in quantifying both the gas in-place and the range of recoverable resources for unconventional gas. Rogner estimated the potential volume of unconventional gas in-place at over 35,000 Tcf globally.

Rogner estimated 10,000 Tcf of coalbed methane gas in-place, with over 3000 Tcf of gas in-place estimated for North and Central America. His regional distribution of coalbed methane was based on the geographical distribution of anthracite and bituminous coal deposits around the world, assuming average gas content values for these two coal types.

Rogner used the U.S. estimates for in-place natural gas from shale formations as an analogue to estimate 17,000 Tcf of global resources in gas shales, with

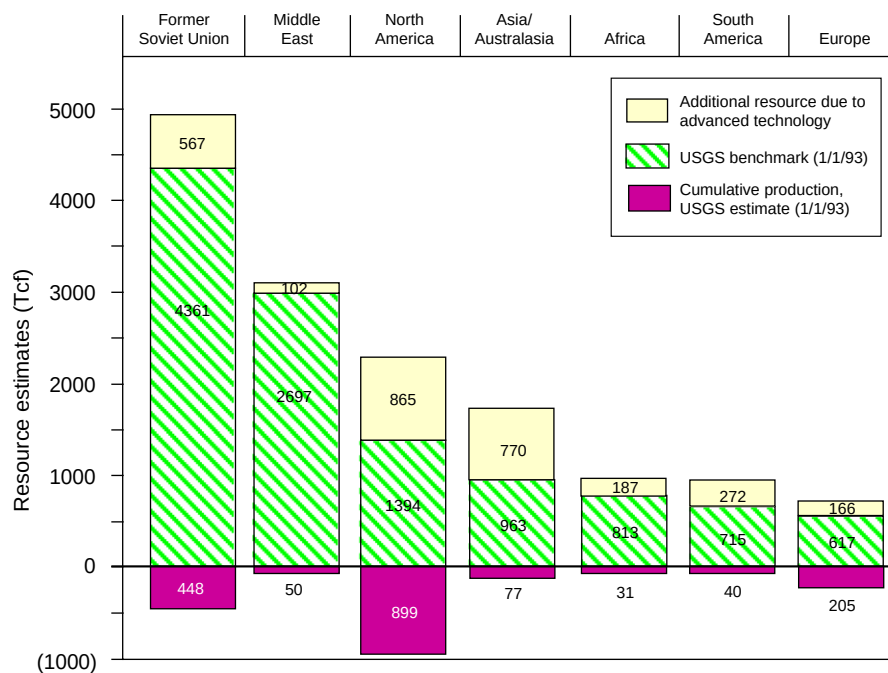


FIGURE 5 Global Gas Resources Workshop estimates for natural gas (Tcf).

TABLE IV

Estimates of Unconventional Natural Gas In-Place by Type (Tcf)

Region	Coalbed methane	Gas from fractured shales	Tight gas sands	Methane hydrates
North America	3270	4160	1480	260,000
Latin America and the Caribbean	40	2290	1400	190,000
Western Europe	170	550	380	30,000
Central and Eastern Europe	130	40	90	—
Former Soviet Union	4290	680	980	180,000
Middle East and North Africa	—	2750	890	10,000
Sub-Saharan Africa	40	300	850	20,000
Centrally planned Asia and China	1320	3820	380	20,000
Pacific OECD	510	2500	760	60,000
Other Pacific Asia	—	340	600	10,000
South Asia	40	—	210	20,000
World	9810	17,440	8020	800,000

over 4000 Tcf of gas in-place in North and Central America.

The 8000 Tcf global resource estimate and the regional distributions for tight gas sands were based on published data on conventional natural gas, with nearly 1500 Tcf estimated for North and Central America.

The global estimate for methane hydrate of 800,000 Tcf in-place was from an earlier estimate by MacDonald. A world map of locations of known and inferred gas hydrates in marine sediments and in continental permafrost was used to establish the regional distributions for this resource.

3.3.2 Nakicenovic

Nakicenovic, drawing on work by Rogner and others, published information on natural gas reserves and technically recoverable resources, including historical consumption of conventional and unconventional gas (see Table V). Reserves are assumed to be recoverable with available technologies at prevailing market conditions. Resources are technically recoverable volumes in addition to reserves, with less certain geological assurances, lacking economic feasibility, and requiring advances in technology for transforming these resources into reserves. The estimate of 33,200 Tcf of reserves and resources for unconventional gas was chosen to correspond to the highest plausible values from the literature. Nakicenovic postulated that potentially the resource base for unconventional gas could be twice as large as that for conventional gas and that the global natural gas resource base will be larger than that for oil.

3.3.3 Kuuskraa

In 1998, Kuuskraa published a country by country estimate for global coalbed methane, both for in-place and for technically recoverable resources (see Table VI). Kuuskraa estimated a CBM in-place value of 3000 to 8000 Tcf and a technically recoverable value of 510 Tcf. Kuuskraa's 1998 estimate for recoverable CBM resources is double his value published in the early 1980s. His most recent estimates were based on a combination of detailed CBM basin studies and information on the rank and depth of coal basins around the world. Future basins studies, particularly of the large CBM resources in Russia and the speculative CBM resources in Canada, the United States, and other countries, and advances in recovery technology would likely add considerably to these estimates of recoverable resources.

3.3.3.1 United States The United States has over 500 Tcf of high-quality coalbed methane resource (see Fig. 6). An additional 230 Tcf of CBM is in the deep coals of the Green River Basin and another 1000 Tcf of hypothetical CBM resource may exist in Alaska. Of this, 150 Tcf is estimated as being technically recoverable, including 29 Tcf of proved reserves plus past production. The estimates for technically recoverable CBM in the United States have steadily increased in the past 10 years and will likely continue to do so in future years. (Additional details on U.S. CBM development are provided in Section 4.)

TABLE V
Global Hydrocarbon Reserves and Resources

	Consumption		Remaining resource base		
	1860–1998	1998	Reserves	Resources	Total
Oil (billion barrels)					
Conventional	782	21	770	970	1940
Unconventional	47	2	970	2580	3550
Gas (Tcf)					
Conventional	2228	76	5700	10,400	16,100
Unconventional	28	N/A	8600	24,600	33,200

Sources. Nakicenovic *et al.* (1996, 1998); World Energy Council (1998); Masters *et al.* (1994); Rogner *et al.* (2000).

TABLE VI
International Coalbed Methane Resources (Tcf)

Country	CBM resource in-place (Tcf)	CBM recoverable resource (Tcf)
Russia	550–1550	N/A
China	350–1150	70
United States	500–1730	150
Australia	310–410	60
Canada	570–2280	140
Indonesia	210	30
Western Europe	120	N/A
Southern Africa ^a	100	20
India	90 +	20
Poland/Czech Republic	70	N/A
Turkey ^a	50	10
Ukraine	50	N/A
Kazakhstan ^a	40	10
Total	3010–7840	510

^a Assumes 20% recovery of gas in-place.

3.3.3.2 Australia Commercial-scale CBM projects exist in the Bowen and Gunnedah Basins, providing over 100 million ft³ per day of production, with exploration under way in the Clarence-Moreton and Sydney Basins. The CBM resource in these four basins is estimated at 260 Tcf of gas in-place, plus 50 Tcf for the Galilee Basin, bringing the total to 310 Tcf. An additional 100 Tcf of gas-in-place may exist in the deep coals of the Cooper Basin. Approximately 60 Tcf of the CBM resource is estimated as recoverable, although considerable uncertainties exist on well productivities and economics. Given the substantial size of the resource, CBM production in Australia will likely be demand-

limited. New gas pipelines to industrial gas users, such as to Mt. Isa, and the completion of a southeastern Australia gas grid are opening new markets, helping boost to both conventional gas and coalbed methane consumption.

3.3.3.3 Canada Canada has 570 Tcf of coalbed methane resources in-place, in formations that meet certain minimum reservoir parameters. This high-graded portion represents approximately one-quarter of the total Canadian CBM resource of 2280 Tcf. Using U.S. coal basins as the analogue, approximately 140 Tcf, in the higher quality reservoir settings, is judged to be potentially recoverable. The largest portion, 100 Tcf, is from the Cretaceous-age coals in the plains and foothills of Alberta. Approximately 35 Tcf of CBM is from British Columbia and 5 Tcf is potentially recoverable from Nova Scotia and Prince Edward Island.

3.3.3.4 China China has a large CBM resource base, estimated at 1150 Tcf, of which 350 Tcf is judged to be in areas geologically favorable for exploration. Of this, 150 Tcf is in eight main CBM exploration areas and perhaps 70 Tcf is technically recoverable. Much of the CBM resource, though located close to urban industrial markets in eastern China, occurs in structurally complex basins. The premier CBM area in China is the eastern margin of the Ordos Basin, where the structural setting is simpler and favorable permeability has been measured. Two new gas pipelines from the Ordos Basin to Beijing and Xian could accommodate 200 million cubic feet (MMcfd) of future CBM production. In addition, China is the world's largest coal producer and emits the highest volumes of mine-sourced coalbed methane into the atmosphere, an estimated

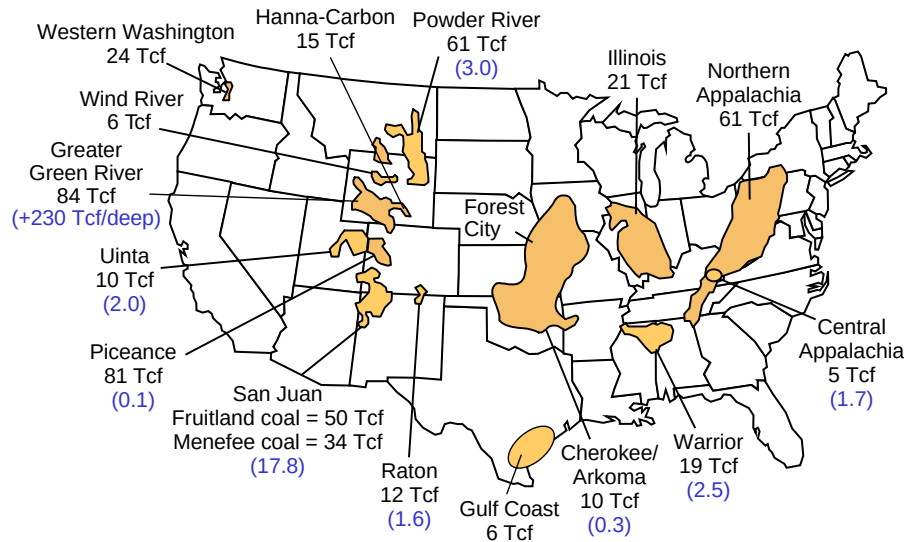


FIGURE 6 Coalbed methane resources of the United States. Values indicated (in Tcf) represent gas in-place; numbers in parentheses indicate resources that have been produced or are reserves (in Tcf) as of 2002; shaded areas indicate coal basins. Source: Advanced Resources Database.

2 billion cubic feet (Bcfd). Capture and use of this coal mine methane would provide a valuable source of clean energy and significantly reduce the emissions of a powerful greenhouse gas.

3.3.3.5 India India has 90 Tcf of coalbed methane resource in-place. Of this 44 Tcf is in six Damodar Valley coal fields, 40 Tcf is in the Cambay Basin of Gujarat, and 6 Tcf is in the Sohagpur and Satpura Basins. Additional CBM resources exist in other Gondwana and Tertiary coal basins of the Godavari and Mahanadi Valley that have yet to be appraised. An estimated 20 Tcf of CBM may be technically recoverable from the Damodar Valley and the Cambay Basin. Production of India's coalbed methane resource will be assisted by a strong natural gas market and plans for expanding its natural gas pipelines.

3.3.3.6 Indonesia A new entry on the world coalbed methane scene, Indonesia is estimated to have 210 Tcf of CBM resources, primarily in the Sumatra and East Kalimantan coal basins. Of this, approximately 30 Tcf is estimated as being in sufficiently high-quality settings to be potentially recoverable. Given its large conventional gas resources, uncertainties about markets, and lack of current activity, development of Indonesia's CBM may progress slowly.

3.3.3.7 Poland/Czech Republic The coalbed methane resources of Poland and the Czech Republic

are estimated at 70 Tcf. This includes 47 Tcf in the Upper and Lower Silesian Basins of Poland, 20 Tcf in the Czech Republic, and 3 Tcf in the Lublin Coal Basin. Additional CBM resources may exist in the deep coals (below 1500 m) of Poland. Because of uncertainties with respect to permeability and gas saturation, no estimates of recoverable resources exist.

3.3.3.8 Russia Ultimately Russia may hold the largest volume of coalbed methane, though much of it is still poorly defined. The Kuznetsk Coal Basin (Kuzbass), on the southern slopes of the west-central Siberian plain, covers an area of 10,000 mi². The numerous coal seams in this basin extend to below 5000 ft in the basin center and hold an estimated 450 Tcf of CBM resource. The Pechora Basin, located on the Arctic Circle adjacent to the Barents Sea, holds an estimated 175 billion tons of high- to low-volatility bituminous coals and an estimated in-place CBM resource of 100 Tcf. Given the lack of data, no estimates are available on technically recoverable resources. The massive Tungusk Basin in the center of Siberian Russia that extends from the Arctic Circle to the Mongolian border may hold over 1000 Tcf and ultimately may become one of the largest CBM basins in the world.

3.3.3.9 Ukraine The Ukraine has 50 Tcf of coalbed methane in-place, primarily in the Donetsk Coal Basin. Given the high geologic complexity of the basin, the limited data, and the extensive past

mining in the basin, estimates of recoverable resources are still speculative.

3.3.3.10 Western Europe Western Europe is estimated to hold 120 Tcf of coalbed methane in place. The Carboniferous coal basins in Western Europe share many similarities: moderately gassy coals, concerns over natural degassing, numerous thin coal seams distributed over a large stratigraphic interval, and structurally complex settings. Coalbed methane has been tested in France, Germany, Great Britain, Belgium, the Netherlands, and Spain. Thus far, the results have been inconclusive, providing little basis for estimates of recoverable resources.

3.3.3.11 Other Countries Coalbed methane exists in numerous other countries, with still limited information on quality and producibility. Prominent are Turkey with 50 Tcf, Kazakhstan with 40 Tcf, and Southern Africa with 90 Tcf of gas in-place. Approximately 20% of this in-place resource may be technically recoverable. South America (Brazil, Colombia, and Venezuela) has no publicly available assessments for CBM resources. Future exploration may add these and other countries to the world's productive CBM ledger.

3.4 Dynamic Resource Base

The review of unconventional gas resource estimates provides a valuable insight. As more is learned about

the nature of these resources and the technologies by which to produce them, the resource estimates have changed dramatically.

- For geologically sound unconventional gas resources, such as tight gas sands, gas shales, and coalbed methane, the historical review shows that the recoverable resource and reserve estimates have steadily increased.

- For geologically uncertain resources, such as geopressured methane, little additional information has been obtained and the expectations for recoverable resources are low.

- For the very speculative unconventional gas resources, such as gas hydrates, much too little is known about their in-place gas concentrations and reservoir settings to provide scientifically reliable estimates on recoverable volumes.

An example of the expected growth in the U.S. unconventional gas resource base is shown in Fig. 7 and Table VII. Figure 7 shows the expected growth in tight gas sands, gas shales, and coalbed methane resource between the years 2000 and 2025, stemming from an assumed continually improving knowledge base and technology. Table VII provides an update to Fig. 7, using the latest information for the U.S. unconventional gas database maintained by Advanced Resources International.

Looking ahead, it is useful to recognize that the information and expectations for unconventional gas represent a "snapshot in time." As new plays are

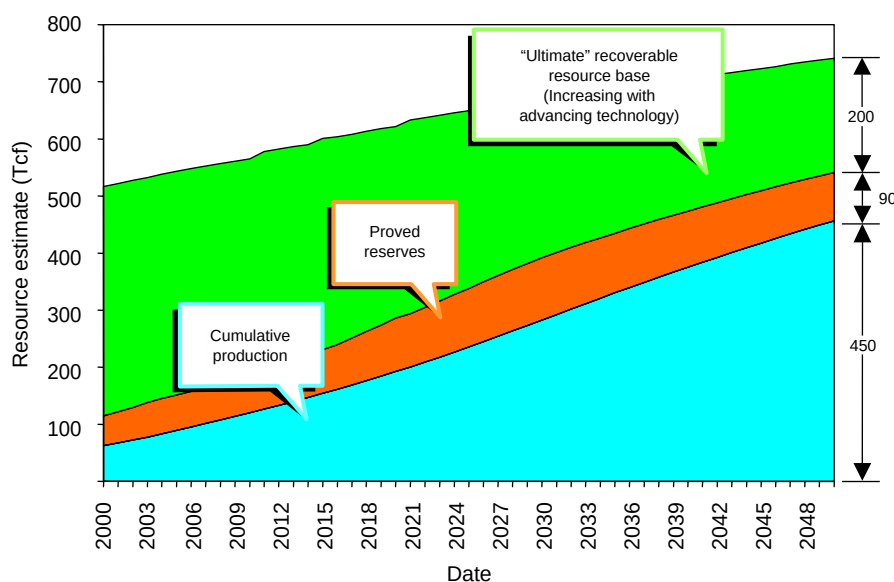


FIGURE 7 Projected growth of unconventional gas resources, years 2000 to 2050. *Source:* Modified by Advanced Resources from National Petroleum Council (1992).

TABLE VII

Estimates of Technically Recoverable U.S. Unconventional Gas Resources (Tcf)

	Year 2002 technology			Year 2025 technology:
	Proved reserves/production	Undeveloped resource	Ultimate resource	Ultimate resource
Tight gas sands	149	292	441	520
Gas shales	13	62	75	90
Coalbed methane	29	95	124	150
Total	191	449	640	760

tested, additional wells are drilled, and more efficient technologies are developed, the outlook for this complex natural gas resource will most likely improve, for two reasons:

1. First, the ultimate size or productivity of the unconventional gas resource base is not yet known. Improved geologic knowledge continues to expand its size by adding new gas plays. Technology progress helps increase recovery from the already defined plays. An example of this is the Barnett gas shale play in the Fort Worth Basin that is now estimated to have 20 Tcf of recoverable resource, up from 1.4 Tcf in the initial estimate prepared over a decade ago (as further discussed below).

2. Second, the economic outlook for unconventional gas resources, basins and plays will likely continue to change significantly with time. With the aid of technology, an unconventional gas basin or play can move from a high cost, marginal resource to a major, highly productive gas reserve. An example of this is the San Juan Basin coalbed methane play that originally was assessed by resource analysts as unaffordable, with costs of over \$5 per million ft³. Today, this important gas play accounts for 9 Tcf of remaining proved gas reserves plus 9 Tcf of past production (as of 2001).

At the other extreme is geopressured methane that was touted, 20 years ago, to hold 1000 years of natural gas supply. No production or reserves have been developed nor are any expected from this resource.

The sequence of resource assessments for the Barnett Shale gas play, illustrates the impact that improved knowledge and technology progress can have on the outlook for an unconventional gas play.

3.4.1 Initial Assessment

The first assessment for the Barnett Shale (Table VIII, column 1) was prepared using data available as of

TABLE VIII

Successive Assessments of Recoverable Barnett Shale Gas Resources

	Initial assessment, 1990	USGS special assessment, 1996	Latest assessment, 1998
Development intensity (acres/well)	320	320	80–320
Completed wells			
Productive	74	180	300
Unproductive	12	30	50
Play (mi ²)	2439	2439	2439
Future wells	4792	4668	10,148
Success rate	0.86	0.86	0.86
EUR/well (Bcf)	0.35	0.84	0.35–1.50
Technically recoverable resources (Tcf)	1.4	3.4	10.0

1990, after the first group of 86 wells was drilled in this gas shale play. The initial wells had relatively low estimated ultimate reserves (EURs) of 0.35 Bcf per well, and the field was being developed on wide well spacings of 320 acres per well. Using this initial data, the Barnett Shale was estimated to have 1.4 Tcf of technically recoverable natural gas.

3.4.2 Second Assessment

The second assessment by the USGS in 1996 (Table VIII, column 2) used a higher EUR per well, due to improved well-completion practices, of 0.84 Bcf per well but still assumed that wide well spacing of 320 acres would be the norm. The improved EUR per well raised the recoverable resource estimate for the Barnett Shale to 3.4 Tcf.

3.4.3 Third Assessment

The third assessment (Table VIII, column 3) incorporated a more optimum, “intensive development” model for the Barnett Shale. It recognized the very limited drainage being achieved by past Barnett Shale wells and that technology progress was leading to higher EURs in the naturally fractured portions of the play. This assessment established that 10 Tcf of technically recoverable gas exists for the Barnett Shale, with much of the economic potential of this resource residing in the higher productivity, sweet spots in the basin that would be developed at spacings of 80 acres per well. Recent updates place the recovery potential at 20 Tcf.

4. UNCONVENTIONAL GAS DEVELOPMENT IN THE UNITED STATES

4.1 Historical Review

Serious pursuit of unconventional gas in the United States began more than 20 years ago, following concerns that the nation was running out of conventional gas supplies. A Federal Power Commissions task force identified large volumes of tight sands gas in-place in three large Western basins. The Bureau of Mines noted that considerable volumes of methane were being vented for safety reasons from coal mines, wasting a valuable resource. In addition, public interest was stirred by major articles in

Fortune and *The Wall Street Journal* that a new natural gas resource—geopressed aquifers—could provide gas for 1000 years.

These initial findings helped launch a series of more thorough assessments of unconventional gas and studies of how advances in technology could make these large in-place resources productive, as further discussed below.

4.1.1 Tight Gas Sands

Large quantities of natural gas resources were known to exist in geologically complex, tight (low permeability) formations. However, using available well completion technology, the production of gas from these tight formations, for the most part, was too low to support economic recovery. A handful of independents were exploring certain of the tight gas sand basins, with generally a poor record of success. The broad consensus was that significant advances in exploration and well completion technology would be required to develop this large, complex gas resource.

In response, the DOE, the Gas Research Institute (GRI), and industry established research and development programs to unlock the gas resource held in these tight rocks. Spurred by advances in technology, the development of tight gas sands expanded rapidly from the initial limited efforts in the Appalachian and San Juan Basins to the major Rocky Mountain gas basins and into Texas and the midcontinent. By 2001, tight gas sands were producing 3.3 Tcf per year. Proved tight gas reserves were 48 Tcf from a cumulative 80,000 drilled wells (not including the

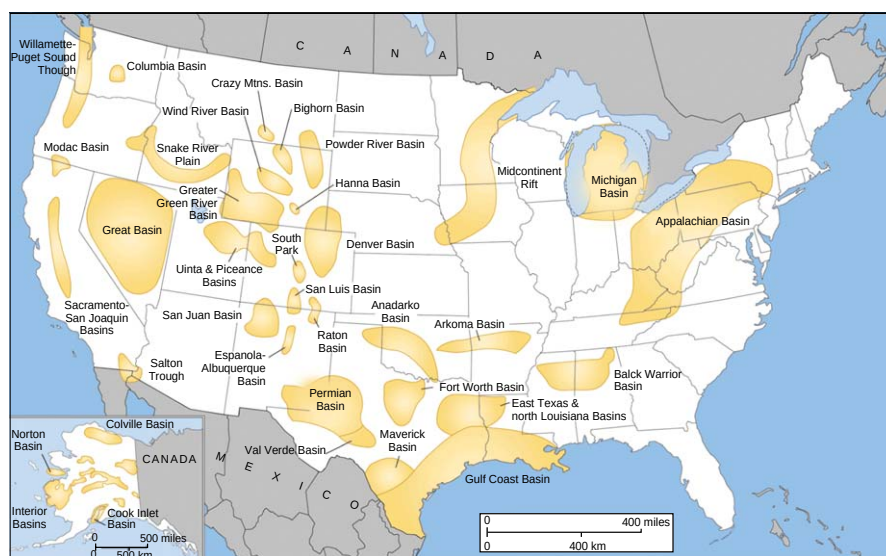


FIGURE 8 U.S. tight gas sand basins. Modified by Advanced Resources from Law (2003).

TABLE IX

In-Place and Technically Recoverable U.S. Tight Gas Sand Resources (Tcf)

Basins	Resource in-place			Remaining recoverable resources	
	FPC (1973)	Kuuskraa <i>et al.</i> (1978)	NPC (1980)	USGS ^a (1995/2002)	Advanced Resources International (2002 technology)
Rockies					
Greater Green River	240	91	136	70	108
Piceance	150	36	49	19	21
Uinta	210	50	20	<i>a</i>	11
Wind River	—	3	34	N/A	33
Denver	—	19	13	<i>1</i>	1
San Juan	—	15	3	16	13
Other					
Anadarko	—	—	—	N/A	19
Appalachia	—	—	—	45	16
East Texas	—	81	22	6	15
Gulf Coast	—	—	—	N/A	19
Northern Great Plains/ Williston	—	74	148	43	21
Permian	—	24	19	N/A	8
Other	—	30	480	<i>12</i>	7
Total	600	423	924	212	292

Sources. Federal Power Commission. Supply-Technical Advisory Task Force (1973); National Petroleum Council (1980). Source for values in *italics*: U.S. Geological Survey (1995).

^aIncluded with data for Piceance Basin.

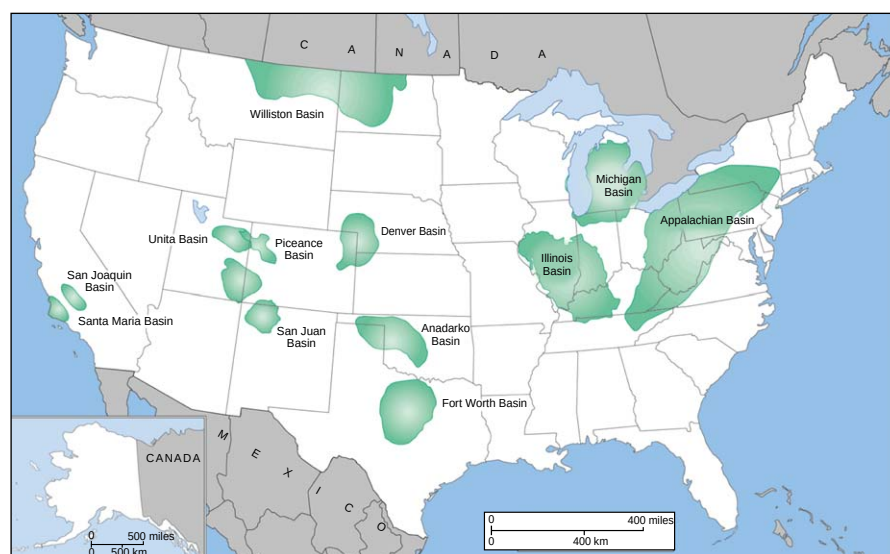


FIGURE 9 U.S. gas shale gas basins. Modified by Advanced Resources from Law (2003).

numerous, older low-producing tight gas wells in the Appalachian Basin), with approximately 101 Tcf of tight gas having been produced. Figure 8 shows the

location of U.S. tight gas sand basins. Table IX provides a historical review of tight gas sand resource estimates.

TABLE X

In-Place and Technically Recoverable U.S. Gas Shale Resources (Tcf)

		Remaining recoverable resources	
Basins	Resource in-place	USGS (1995/2002)	Advanced Resources International (2002 technology)
Appraised			
Appalachian Basin (Devonian shale)	248	12	19
Michigan Basin (Antrim shale)	35	19	12
Fort Worth (Barnett shale)	5	N/A	17
San Juan Basin (Lewis shale)		10	10
Speculative			
Illinois Basin (Albany shale)	160	2	2
Other shale basins	135	1	2
Total	583	44	62

Source. For resource in-place data: National Petroleum Council (1992). Source for data in italics: U.S. Geological Survey (1995).

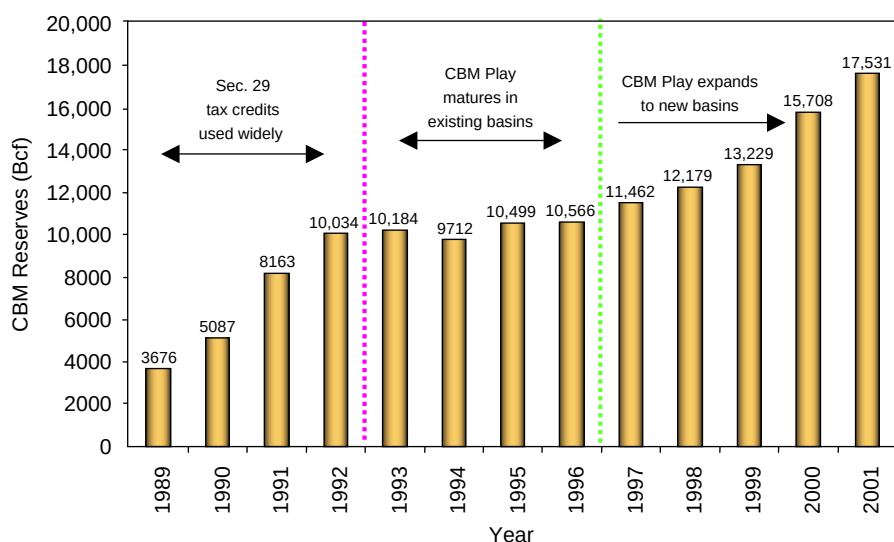


FIGURE 10 U.S. coalbed methane: 10% of U.S. natural gas proved reserves. Source: Advanced Research International Database.

4.1.2 Gas Shales

Twenty years ago, only the Appalachian Basin gas shales were being produced. Wells were being completed open-hole, with little definition of productive pay zones, and were being stimulated with nitroglycerine (a remnant of early 1900s technology). Much of the activity was centered in the Big Sandy area of eastern Kentucky. Little understanding existed about the key gas storage and production mechanisms or about geologically similar gas shale plays in other parts of the country. Gas shale drilling averaged only 200 wells per year and the play was in decline.

By 2001, annual gas shale production had reached 500 Bcf. Proved reserves are 7 Tcf, with another

6 Tcf having been produced. Stimulated by Section 29 tax credits and the expansion into new gas shale basins in Michigan and north Texas, 20,000 gas shale wells were drilled from 1978 to 2001 with a peak of 1700 gas shale wells completed in 1992. Figure 9 shows the location of U.S. gas shale basins. Table X provides estimates of the resource in-place, proved reserves, and remaining technically recoverable gas shale resources.

4.1.3 Coalbed Methane

The combination of building a scientific base of knowledge, developing appropriate technology, and providing economic incentives launched the birth of

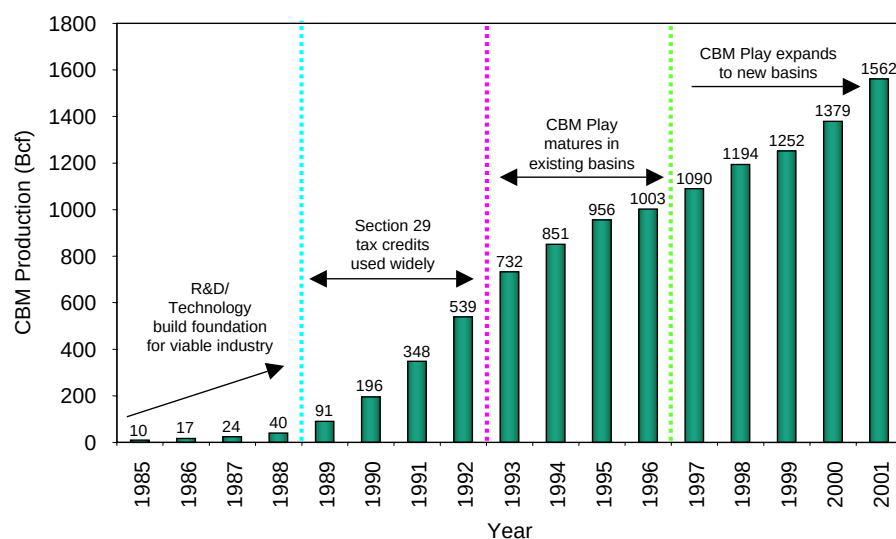


FIGURE 11 U.S. coalbed methane: 8% of U.S. natural gas production. *Source:* Advanced Resources International Database.

TABLE XI

In-Place and Technically Recoverable U.S. Coalbed Methane Resources (Tcf)

Basin	Resource in-place	Production through 2001	Proved reserves 2002	Remaining recoverable resources (2002 technology)	Ultimate resources (2025 technology)
Rockies					
Greater Green River	314	^a	^a	6.2	8
Uinta	10	0.3	1.7	6.5	10
Piceance	81	0.1	^b	7.3	9
Powder River	61	0.6	2.4	36.0	50
Raton	12	0.2	1.4	6.2	9
San Juan	84	8.7	9.1	17.2	42
Other (Hanna, Wind River)	21	^a	^a	^a	^b
Other					
Northern Appalachia	61	^a	^a	5.0	6
Central Appalachia 1.3	5	0.3	1.4	3.4	6
Warrior	19	1.3	1.2	3.2	7
Cherokee/FC/Arkoma	10	^a	0.3	3.0	4
Illinois	21	^a	^a	0.6	1
Other (Gulf Coast, Western Washington)	30	^a	^a	^a	^b
Total	729	11.5	17.5	94.6	152

^a Less than 0.05 Tcf or N/A.

^b Assuming year 2025 technology.

a new natural gas industry: coalbed methane. Much of the early development was by independent producers, such as Devon Energy and Meridian Oil, who subsequently saw their gas production and reserve holdings rise sharply.

Coalbed methane production climbed from essentially zero in the mid-1980s to 1.6 Tcf in 2001. Proved reserves are 17.5 Tcf, with another 11.5 Tcf having already been produced (see Figs. 10 and 11). The development and continuing adaptation of

technology enabled the industry to remain profitable and vigorous, even after the expiration of Section 29 tax credits in 1992. Today, several new coalbed methane basins and plays are being actively developed, including the Powder River (Wyoming), Raton (Colorado), and Uinta (Utah), providing a base for continued growth. Table XI provides estimates of the in-place and technically recoverable coalbed methane resource for the United States.

4.1.4 Geopressured Methane

An aggressive research and development program assembled considerable geologic and reservoir knowledge for this resource. And, although no commercial natural gas production was established, the research and development program helped bring a strong dose of reality and understanding regarding the viability of this gas resource. It also helped dispel the speculation that "1000 years of natural gas" was at hand.

4.2 Assessment of Results

Unconventional gas offers one of the great success stories of progress in technology. A poorly understood, high-cost energy resource, one that the U.S. Geological Survey had not even included in its national appraisals of future gas resources (until their 1995 assessment), is now providing major volumes of annual gas supplies and helping meet growing domestic natural gas demand.

5. CONCLUSION

The discussion and resource estimates for unconventional gas presented in this article are a snapshot in time. The ultimate size or productivity of this geologically complex natural gas resource is not yet known. However, it is useful to document what is known as the foundation for continually improving the base of knowledge and the expectations for this large potential source of future natural gas supplies.

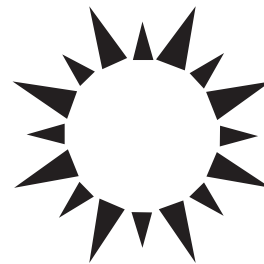
SEE ALSO THE FOLLOWING ARTICLES

Depletion and Valuation of Energy Resources • Natural Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Processing and Products • Natural Gas Resources, Global Distribution of • Natural Gas Transportation and Storage • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil and Natural Gas Resource Assessment: Geological Methods • Oil and Natural Gas Resource Assessment: Production Growth Cycle Models

Further Reading

- Brandenburg, C. E., Kuuskraa, V. A., and Schraufnagel, R. A. (1998). "Coalbed Methane and Technology: Relating the U.S. Experience into International Opportunities," Presented at the 17th Congress of the World Energy Council, Houston, Texas, September 13–18, 1998.
- Fisher, W. L., Pedersen, A. H., and Rosenberg, R. B. (1994). "Workshop Consensus," Proceedings of the Global Gas Resources Workshop, September 19–21, 1994. International Centre for Gas Technology, p. ix–x.
- Kuuskraa, V. A. (1998). Barnett shale rising star in Fort Worth basin. *Oil Gas J.* 96, 67–76.
- Kuuskraa, V. A., and Meyers, R. F. (1983). Review of world resources of unconventional gas. In "Conventional and Unconventional World Natural Gas Resources" (C. Delahaye and M. Grenon, Eds.), pp. 409–458. IIASA Collab. Proc. Ser., CP-83-S4, International Institute of Applied Systems Analysis (IIASA), Laxenburg, Austria.
- Law, B. E., and Curtis, J. B. (2002). "Unconventional Petroleum Systems," AAPG Bulletin, Vol. 86, No. 11, November, ISSN 0149-1423.
- Nakicenovic, N., and Riahi, K. (2001). "An Assessment of Technological Change across Selected Energy Scenarios." Reprinted from a publication by the World Energy Council on behalf of the Study Group on Energy Technologies for the 21st Century. RR-02-005.
- Rogner, H. H. (1997). An assessment of world hydrocarbon resources. *Annu. Rev. Energy Environ.* 22, 217–262.
- Schmoker, J. W., Quinn C. J., Crovelli, R. A., Nuccio, V. F., and Hester, T. C. (1996). "Production Characteristics and Resource Assessment of the Barnett Shale Continuous (Unconventional) Gas Accumulation, Fort Worth Basin, Texas." USGS Open-File Report 96-254. U.S. Geological Survey.
- U.S. Geological Survey National Oil and Gas Resource Assessment Team (1995). "National Assessment of U.S. Oil and Gas Resources," USGS Circular 111S.1995. U.S. Geological Survey.

Natural Gas Transportation and Storage



FARUK CIVAN

University of Oklahoma

Norman, Oklahoma, United States

-
1. Introduction
 2. Natural Gas
 3. Natural Gas Transportation
 4. Natural Gas Storage
 5. Gas Transportation and Storage Services, Operation, and Economic Value
 6. Hazards from Natural Gas Transportation and Storage
 7. Natural Gas Security and Risk Issues

Glossary

aquifer A porous permeable sedimentary formation located belowground and containing groundwater.

base or cushion gas The minimum quantity of natural gas that must be retained in gas storage to ensure the production of natural gas from storage at a sufficient rate and pressure; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions).

deliverability The rate at which natural gas can be withdrawn from gas storage; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions) per unit day.

hydrate A cagelike crystalline solid material similar to ice, formed by combining gas with water.

injectivity The rate at which natural gas can be injected into gas storage; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions) per unit day.

inventory A survey of the total amount of gas contained in storage; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions).

liquefied natural gas (LNG) The liquid state of natural gas, essentially containing methane and little ethane.

liquefied petroleum gas (LPG) The liquid form of propane and butane.

million British thermal units (MMBtu) A measure of heat generated by gas combustion.

permeability A physical parameter expressing the fluid (gas or water) transmitting capacity of the interconnected porous space available in depleted oil and gas reservoirs and water aquifers.

pipeline A long pipe used to transmit fluids.

porosity The volume fraction of the porous space occupied by fluids (gas and water) in depleted oil and gas reservoirs and water aquifers.

reservoir An underground geologic permeable rock formation; contains fluids such as oil, gas, and water in pore spaces.

salt cavern A cavity created in a deep salt bed or dome following the dissolution of salt by circulating hot water.

standard or base conditions Thermodynamic conditions considered as the reference conditions for volume measurement, usually temperatures of 0°, 15.6°, or 25°C and 1.0 atmosphere of pressure.

storage Facilities of various types and capacities used to store natural gas.

storage measures The various means of expressing the properties and conditions of a gas storage area and its gas content.

thousand standard cubic feet (Mscf) A measure of gas volume.

total capacity The largest quantity of natural gas that can be loaded into a particular gas storage area; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions).

working gas capacity The difference between the total and base gas capacities of gas storage.

working (top) gas The maximum quantity of natural gas that a particular gas storage can hold and still allow for circulation; usually measured in volume expressed at 15.6°C (60°F) and 1.0 atmosphere of pressure (standard conditions).

Natural gas is an important commodity traded frequently in the energy market. It can be utilized as a source of energy by direct consumption as fuel or

by other means, including after conversion to hydrogen for use in fuel cells or after conversion to various chemicals. Because it is a gas and a low-density substance at ambient pressure and temperature conditions, safe and economic handling of natural gas during transportation, storage, and utilization requires special measures and sophisticated technologies. Therefore, the worldwide trade and utilization of natural gas is inherently limited by the difficulty of handling gas between the source and market.

1. INTRODUCTION

Natural gas is gaining increasingly great potential as an abundant and environmentally friendly energy source; it is available worldwide from onshore and offshore reserves in various forms, including as gases or as condensates in underground geological reservoirs, and as solid hydrates in hydrate-bearing sedimentary formations present in permafrost regions and ocean floors. Thus, technologies are being developed rapidly for its economical production and utilization. Although transportation is necessary as a means of delivering gas from production to market, gas storage ("parking" gas temporarily) at appro-

priate locations is instrumental in compensating between the surge and peak effects in the gas supply-and-demand value chain. Accurate measurement of natural gas using reliable meters is another important issue involved in custody transfer and inventory accounting in natural gas trading. Effective engineering and management of the gas transportation and storage facilities also include strategic security measures and emergency response procedures. Capacity and location of the source and the user, available and feasible technologies, and economics determine the means of transportation and storage used. The various types and capacities of transportation and storage facilities involved in natural gas handling and value chain are summarized in Table I.

2. NATURAL GAS

Natural gas is a mixture of hydrocarbon and nonhydrocarbon gases; these gases are produced from petroleum and hydrate-bearing formations present in geological sedimentary formations located below the ground surface and the ocean floor. Typical hydrocarbon components of natural gas

TABLE I
Means of Transportation and Storage of Natural Gas^a

Natural gas compression	Direct gas liquefaction	Conversion to hydrate blocks or hydrate-in-oil slurry	Dissolution of petroleum gas in pentane-plus-heavier fractions of crude oil	Chemical conversion to liquid products	Electric power generation
Pipeline gas transportation with intermediate recompression	Liquid storage and loading	Transportation by tanker and pipeline	Crude transportation by tanker	Liquid transportation by tanker	Electric transmission by cable
Gas storage above ground and below ground	Liquefied gas transportation by tanker	Hydrate storage and transport to destination	Crude storage and transport to destination	Liquid storage and transport to destination	Electric distribution
Distribution	Liquefied gas unloading, storage, and transport to destination	Regasification and distribution	Regasification and distribution	Liquid distribution	
Compressed gas tank filling	Regasification and distribution				
Gas tank transportation by truck and other means					
Storage as compressed gas in tanks					
Gas distribution					

^aFrom oil and gas reservoirs, as coal bed methane, as ocean-floor and permafrost hydrates, and gathered from wells.

include a significantly large percentage of methane, and relatively smaller percentages of ethane, propane, isomer and normal butanes and pentanes, hexane, heptane, and other heavier hydrocarbons. The typical nonhydrocarbon components can be classified as acid gases, such as hydrogen sulfide and carbon dioxide; inert gases, such as nitrogen and helium; odorous sulfurous gases, such as hydrogen sulfide and various mercaptans; and other impurities, such as water and mercury vapors and sulfur dissolved in gas.

The composition of natural gas dictates its behavior during the chain of processes involved in gas handling, including its transportation and storage. Its density ρ or specific volume v are given by the real gas equation of state as in Eq. (1):

$$\rho = \frac{1}{v} = \frac{m}{V} = \frac{MP}{ZRT}, \quad (1)$$

where m , V , P , and T denote the mass, volume, pressure, and temperature of the gas, M is the gas molecular mass (16 kg/kmol for methane), R is the universal gas constant (8308 J/kmol $^\circ$ K), and Z is the real gas deviation factor, a function of the pressure, temperature, and composition of gas, usually given by empirical correlations or charts. The density of the methane gas is about 0.65 kg/m³ at 25 $^\circ$ C and 1 atm under ambient conditions. The density of natural gas relative to the density ρ_a of dry air of standard composition, at standard conditions, is referred to as the specific gravity G , given by Eq. (2):

$$G = (\rho/\rho_a)_b = M/M_a, \quad (2)$$

where the molecular weight of air is $M_a \cong 29$ kg/kmol. The standard conditions b are defined as zero or 15.6 $^\circ$ C (60 $^\circ$ F) temperature and 1.0 atmosphere of pressure. The gravity of methane gas is 0.55. The gravity of typical natural gas is about 0.64, hence it is lighter than air. Therefore, the transportation and storage of natural gas in large mass present a challenging engineering problem, requiring sophisticated and expensive technologies. This is because the volume of natural gas per unit mass (referred to as the specific volume) is very large and its density is very low at ambient conditions. For example, 1 kg of methane gas occupies 1.5 m³ at 25 $^\circ$ C and 1.0 atm of pressure. Consequently, natural gas should be compressed or converted to smaller volumes by suitable means in order to be able to handle the large quantities in a convenient manner.

At elevated pressure and sufficiently low temperature conditions, natural gas can liquefy and produce a multiphase system, including a liquid phase,

referred to as condensate, and a solid phase (when combined with water), referred to as a natural gas hydrate. Production of condensates and hydrates of natural gas creates various difficulties during transportation and storage of natural gas. Coping with such problems is an important issue in natural gas transportation and storage, and an economic and technological headache.

3. NATURAL GAS TRANSPORTATION

Transportation facilities of various types are used for collecting gas from distributed sources and for delivering it to end users (the customers), who utilize gas for various purposes. The worldwide trade and utilization of gas greatly depend on the availability and economics of gas transportation and distribution facilities. The primary gas transportation technologies are based on compressed gas transport, direct pipeline transfer, gas-to-liquid conversion, gas-to-solid conversion, and gas-to-power conversion.

3.1 Compressed Gas Transport in High-Pressure Tanks

Natural and petroleum gases can be transported as compressed gas in small quantities inside high-pressure tanks for use at residences, in industry, and in gas-powered vehicles. Based on Eq. (1), the volume of compressed gas at the prevailing pressure P and temperature T can be estimated by Eq. (3):

$$V = V_b Z \frac{T}{T_b} \frac{P_b}{P}, \quad (3)$$

where P_b and T_b denote the base or standard pressure and temperature, respectively, and V_b is the standard gas volume.

3.2 Direct Pipeline Transfer

Pipeline transportation is the oldest and most common technology; pipelines are widely utilized worldwide for transportation of gas. Pipeline transportation of natural gas involves a variety of equipment and instrumentation, such as a pipeline network to carry gas from different sources to the end users; meters to measure the gas quantity and flow; valves to control the gas flow; regulators to control the gas pressure; safety devices, expansion systems, and storage facilities for temporary storage

TABLE II
Worldwide Natural Gas Pipeline Extent^a

Location	Length (km)
North America	405,980
Latin America	35,041
Europe	267,846
Africa	7160
Middle East	6854
Asia–Oceania	145,519
World total	868,400

^aEstimated based on data in *The World Factbook* (available on the Internet through the Central Intelligence Agency Web site at <http://www.cia.gov>).

of gas; dehydrators to remove the moisture and prevent water condensation, hydrate plug formation, and corrosion; compressors to pressurize and “flow” the gas; prime movers and electrical control systems to operate these systems; and remote data acquisition, monitoring, and transmission systems. Table II provides worldwide statistics about the extent of natural gas pipelines based on information provided by *The World Factbook*.

One of the frequently used pipeline flow equations is the Weymouth equation, expressing the gas flow rate q_b at the base pressure and temperature conditions, given by Eq. (4):

$$q_b = \frac{\pi T_b}{4 P_b} \left[\frac{R(P_1^2 - P_2^2) D^5}{M_a G \bar{Z} \bar{f} L} \right]^{1/2}. \quad (4)$$

The symbol L denotes the length of the pipeline, the gas pressures at the inlet and outlet ends of the pipeline are P_1 and P_2 , D is the internal diameter of the pipe, M_a is the molecular weight of standard air, and $\bar{Z}$ and $\bar{f}$ are the average values of the real gas deviation factor and the pipe friction factor over the pipe length.

3.3 Gas-to-Liquid Conversion

Liquids can be more conveniently, safely, and economically transported in large quantities than can gases. Natural and petroleum gases can be transformed into liquids either by direct cryogenic condensation, by mixing, or by chemical conversion. Cryogenic condensation forms liquefied gases, such as liquefied natural gas (LNG), which essentially consists of methane and little ethane, and liquefied petroleum gas (LPG), which consists of the propane and butane cut. Liquefaction occurs by mixing

natural or petroleum gas with the pentane-plus-heavier hydrocarbon fractions separated from an associated gas of crude oil. Chemical conversion yields liquid products such as methanol and gasoline. Chemical conversion of natural gas to liquid products requires the application of a series of complicated and expensive chemical processing methods. Therefore, chemical conversion is not a generally preferred option. Nevertheless, the liquids produced by chemical conversion are stable at ambient conditions and can be conveniently carried and safely stored at moderate near-ambient conditions. Following liquefaction, intercontinental and marine transport of liquids is accomplished by means of specially designed tankers or ships.

Transportation of gas in liquefied form is a broadly exercised option. However, this combines a series of delicate technological issues, including compression, Joule–Thomson expansion turbines, cooling and heating, fractionation, and specially designed storage tanks for tankers. Liquefied natural gas is transported as a cryogen at the normal boiling temperature at atmospheric pressure. This temperature is -161.5°C for pure methane. However, natural gas also contains small quantities of other hydrocarbon gases. Therefore, the liquefaction process undergoes a gradual condensation by cooling from heavier to lighter components at different temperatures, ending with the condensation of methane at -161.5°C . The density of the liquid methane is about 420 kg/m^3 at this condition. This means that 1 kg of liquid methane occupies a volume of only 0.0024 m^3 . Hence, the volume of the liquid methane is 625 times less than the volume of the methane gas at ambient conditions.

The LNG tankers for marine transport are specially designed and constructed to carry large quantities of liquefied natural gas safely. The hull of LNG tankers usually is double, consisting of outer and inner hulls, separated with a space sufficient to protect against damage by contact or collision of tankers with other objects. Essentially, integrated and self-supporting tanks are used as containers in the conventional LNG tankers. The LNG tankers incorporating the integrated-tanks technology contain a series of rectangular tanks containing a membrane to resist the external forces. The LNG tankers incorporating the self-supporting tanks technology contain a series of individually reinforced tanks, usually spherical in shape, designed to resist the hydrostatic force of their fluid content. These tanks are placed at sufficient distances from the neighboring tanks and are insulated

individually. LNG transport by tankers also requires auxiliary services for liquefaction of natural gas, and LNG loading, storage, discharging, and regasification.

3.4 Gas-to-Solid Conversion

Under proper conditions, natural gas combines with water to form a cagelike crystalline solid material similar to ice, called a gas hydrate. The gas and water molecules are held together because of the molecular interaction forces. Hydrates can hold approximately 150 to 180 standard (at 15.6°C or 60°F and 1.0 atm) volumes of natural gas per unit hydrate volume and they are stable in the -10° to -15°C temperature range at atmospheric conditions. Natural gas hydrates exist in ocean floor and permafrost regions. Gas hydrates can also be artificially manufactured in the form of rectangular and cylindrical blocks and transported conveniently via ships and other means. Hydrates can also be transported through pipelines in the form of slurry with oil. For this purpose, the gas and water are placed into contact at conditions near 5°C and 50 atm in stirred-tank or spray-tower reactors to produce gas hydrates. This process generates some heat, which must be removed by appropriate cooling technologies. This rather new technology is under development. Hydrate transportation requires additional processes to form and dissolve the hydrates.

3.5 Gas-to-Power Conversion

Conversion of gas to electric power is accomplished in gas-powered combined-cycle electric power plants. Electricity is the most convenient means for transport, using electric transmission cables of various types. However, electric storage is a challenging technological barrier. Theoretically, 1.0 kg of methane can generate 14 kilowatt-hours (kWh) of electric energy if all the chemical energy can be converted to electric energy with 100% conversion efficiency. However, the actual energy generated by combustion of methane is less than this quantity, because of the loss of energy by friction and other means.

4. NATURAL GAS STORAGE

Temporary natural gas storage capability is essential in order to balance out the supply of natural gas with demand, and as a backup reserve to compensate for shortages and temporary interruptions of gas supply

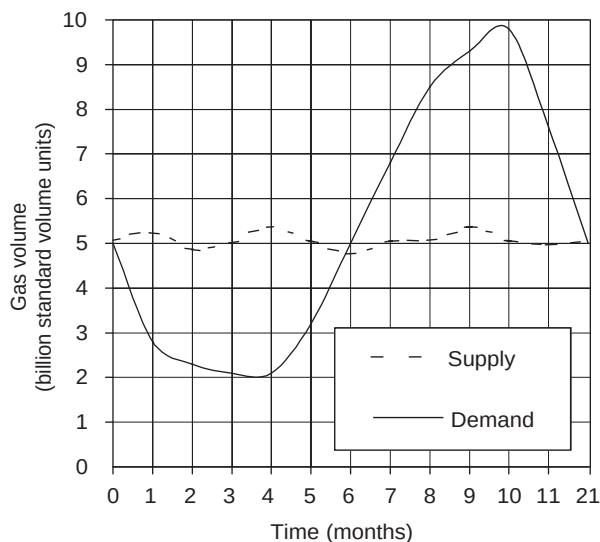


FIGURE 1 Natural gas storage and consumption based on the supply and demand trends.

or demand. Gas storage provides a convenient leverage and advantage in risk management, combating and dealing with peaks in the gas supply-and-demand chain, and accommodating gas price fluctuations. Natural gas is temporarily stored when the market is unfavorable and/or demand is low, and the stored gas is made available under favorable market and demand conditions. For example, as a heating fuel, gas is stored during summers and consumed during winters (Fig. 1). Companies providing various gas utilization services, local distribution, and intra- and interstate pipeline transportation rely on gas storage in order to alleviate the variations in the supply-and-demand chain. When storage cannot meet the demand for natural gas, liquefied natural or petroleum gases are regasified for the market.

Various features, such as capacity, accessibility, and gas injection and production rates, determine the type of gas storage for particular applications and requirements. The conventional gas storage options can be broadly classified as aboveground and underground storage facilities.

4.1 Aboveground Natural Gas Storage

The primary means of aboveground gas storage involve pipelines, cryogenic tanks, and gas hydrates.

4.1.1 Pipeline Storage

Pipeline networks can be used to provide temporary gas storage when they are not needed for gas

transmission purposes. This helps in coping with the rapid surge and peak situations in the supply-and-demand value chain over short time frames. Applying the Weymouth equation given by Eq. (4), the total quantity of gas stored in a pipeline after closing its ends can be estimated as follows:

$$V_b = \frac{\pi D^2 T_b L}{6 P_b T} \left(P_1 + P_2 - \frac{P_1 P_2}{P_1 + P_2} \right). \quad (5)$$

4.1.2 Storage in Cryogenic Tanks

Tank storage affords small capacity and is expensive, but can be facilitated practically anywhere. Natural gas can be stored cryogenically in small quantities in specially designed storage tanks at -155° to -160°C temperatures. These tanks are well insulated against heat influx from the surroundings and therefore cooling is not required. Further, the storage tanks can be placed inside specially designed dikes to confine the liquefied gas in case of accidental spills. Cryogenic storage tanks may be constructed as double-wall or membrane tanks.

Prolonged storage of liquefied gas in storage tanks may create a hazardous condition resulting from external heating at the bottom of the tank over time. Heating can cause a density difference and stratification of the stored liquid, eventually leading to rollover and rapid mixing of the colder and heavier upper layer with the warmer and lighter lower layer of the cryogen. The rollover phenomenon can be avoided by maintaining a uniform-density liquid throughout the tank; this can be accomplished by various methods, including proper loading and discharging procedures, continuous recirculation of the liquid in the tank, and reliquefying the vapors

generated by boil-off and refluxing of the liquefied vapors back into the tank.

4.1.3 Storage as a Gas Hydrate

Gas can be conveniently stored in the form of hydrate blocks. As explained for transporting gas as hydrates, typical hydrates can hold approximately 150 to 180 standard volumes of natural gas per unit hydrate volume, and they are stable in the -10° to -15°C temperature range at atmospheric conditions. This technology enables storing large quantities of natural gas in the hydrate state.

4.2 Underground Gas Storage

There are various options for underground gas storage, varying in capacity and performance, such as in-ground storage tanks, depleted and abandoned oil and gas reservoirs, aquifer storage, salt and hard-rock caverns, and abandoned mine shafts. Table III provides several statistics about the worldwide disposition of underground gas storage capacity.

4.2.1 In-Ground Storage Tanks

In-ground storage tanks and aboveground tanks are constructed similarly, but in-ground tanks are partially or fully buried in the ground.

4.2.2 Depleted and Abandoned Oil and Gas Reservoirs

Oil and gas reservoirs with their existing wells and surface handling facilities are widely used for underground gas storage after their economically recoverable hydrocarbon content has been depleted. Depleted gas reservoirs are preferred because their

TABLE III

Worldwide Underground Gas Storage Capacity^a

Location	Storage sites (number)	Working capacity (10^9 m^3)	Total capacity (10^9 m^3)	Peak withdrawal rate ($10^6 \text{ m}^3/\text{day}$)
United States	386	102	228	1894
Western Europe	66		86	782
Central Europe	17	7.3	—	—
New independent states	46	80.5	—	—
Russia	21	—	—	—
World	554	243	502	

^aBased on data from an international association of natural gas industries and the Institut Français du Pétrole (IFP-CEDIGAZ). (Available at the Web site <http://www.cedigaz.com>.)

structure is more likely to be leakproof compared to oil reservoirs. The injectivity and deliverability of gas storage will vary with the prevailing condition and quantity of stored gas and with several other factors, including the energy or driving mechanisms (such as water drive) of the reservoir storage for gas transport; the gas and storage formation rock compressibility; the amount of base gas; the number, arrangement, and performance of the wells completed into storage; and effectiveness of the surface gas handling facilities.

Reservoir rock is a compacted sedimentary porous formation, often naturally fractured. Fractures form the highly conductive flow paths, and the interconnected pores of the rock blocks (separated by the fractures) provide the essential gas storage capability. Permeability of the storage rock formation determines the rate of gas flow through it. High-permeability reservoirs are preferred for gas storage because these can allow for high rates of injectivity and deliverability. However, over time, the performance of depleted reservoir storage can be reduced by impairment of permeability due to various damage mechanisms, such as clay swelling, fine aggregate ("fines") migration, and precipitation of scales. The performance can be restored by costly special measures, such as acid flushing and hydraulically created fracture stimulation. Depleted reservoir storage requires careful operation and mitigation in order to avoid the possibility of coproduction of water; water production can occur due to the coning phenomenon, because most reservoirs contain an underlying water table. A large quantity of base gas should be retained in underground storage in order to avoid this problem. The possibility of facilitating inert gases is being explored for use as a base or cushion gas. Typically, depleted reservoir storage involves ten to hundreds of wells, 5 million to 10 million standard cubic feet per day deliverability capability, only 33–50% of the total gas content available as working gas, a significant amount of water, and inflexibility, allowing only one to two cycles of injection and production per year with 100 to 150 days of withdrawal periods. Applying the leaky-cylindrical tank model depicted in Fig. 2, the relationship between gas flow rate and pressure is given by Eq. (6):

$$q_b = C(\bar{P}^2 - P_w^2)^n, \quad (6)$$

where $\bar{P}$ and P_w denote the average reservoir gas pressure and the gas pressure at the well bottom, respectively, and C and n denote the performance coefficient and exponent, respectively, which have

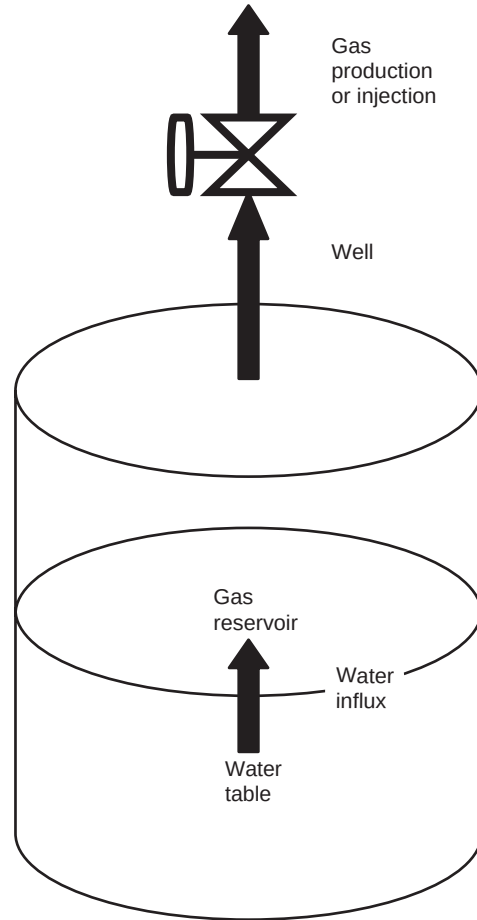


FIGURE 2 Underground gas storage reservoir and water table.

empirically determined specific values for specific storage reservoirs. Assuming constant flow rate, the cumulative production/injection gas volume is given by Eq. (7):

$$G_p = q_b t, \quad (7)$$

where G_p and q_b denote the cumulative volume and volume flow rate of the gas under standard conditions (15.6°C or 60°F and 1.0 atm), and t denotes the total elapsed time.

4.2.3 Storage in Aquifers

Suitable underground water aquifers, leakproof for gas, can be used for gas storage. Aquifer storage must have an impermeable cap rock formation to avoid gas leakage. The injectivity and deliverability of this storage depend on the gas solubility in water and the water-drive potential of the aquifer. Generally, use of aquifer storage requires a large quantity of base gas to be retained in place. Such storage needs careful

operation and mitigation in order to avoid the possibility of coproduction of water with gas. The other features of aquifer storage (including typical inventory) and the injection and production capabilities of aquifers are similar to those of depleted reservoirs.

Based on Fig. 2, the storage gas balance equation can be expressed as follows in order to relate the gas pressure in the storage to the cumulative gas production or injection:

$$\frac{P}{Z} = \frac{P_i}{Z_i} \left(1 - \frac{G_p}{G_i} \right) / \left(1 - \frac{W_e E_i}{G_i} \right), \quad (8)$$

where the subscript i denotes the values at initial or reference condition of the gas storage, and E is the gas expansion factor, given by Eq. (9):

$$E = \frac{T_b}{P_b} \frac{P}{Z} \frac{1}{T}. \quad (9)$$

The water content of the aquifer is denoted by W_e , and G_i and G_p denote the initial volume of gas stored in the storage and the cumulative volume of gas produced from the storage, respectively, both expressed at standard conditions. The water rises when gas is produced from the storage, and recedes when gas is injected. The water influx or efflux rate q_w is given proportionally to the pressure difference between the aquifer water and storage gas as

$$q_w = \frac{dW_e}{dt} = J(\bar{P}_a - P), \quad (10)$$

where W_e is the aquifer water content, t is time, J is an aquifer productivity index, P is the storage reservoir gas pressure, and $\bar{P}_a$ is the average water pressure in the aquifer.

4.2.4 Storage in Salt Caverns or Cavities

Salt caverns are created by dissolution of salt in deep salt beds and domes by hot water circulating through wells drilled into the salt formation. Storage capacity is significantly smaller in these cavities compared to the depleted reservoirs and aquifers, but cavern storage is flexible and allows for multiple cycles of gas injection and production in a year. The construction of salt cavern storage is expensive and is preferred usually only when depleted reservoir or aquifer storage opportunities are unavailable. Cavern storage provides significantly greater gas injectivity and deliverability rates in comparison to the depleted reservoirs and aquifers, and requires holding a much smaller amount of base gas in the storage to be effective. Typically, salt cavern storage involves one to three wells, 8.5 million to 17 million standard cubic meters per day deliverability capability, about

65% of the total gas content available as working gas, a negligible amount of water, and 6 to 12 cycles per year with 10 days of withdrawal periods. Simplifying Eq. (8) by neglecting the water influx effect, the gas pressure and cumulative gas production or injection at isothermal conditions are related by Eq. (11):

$$\frac{P}{Z} = \frac{P_i}{Z_i} \left(1 - \frac{G_p}{G_i} \right). \quad (11)$$

4.2.5 Hard-Rock Caverns and Abandoned Mine Shafts

The possibility of storing natural gas in properly conditioned and prepared hard-rock caverns and abandoned mine shafts is under consideration. Refrigerated mined caverns operating in the -29° to -40°C temperature range and specially lined rock caverns that can store compressed gas at elevated pressures are being explored in order to increase the capacity of flexible gas storage. However, the gas industry has not yet implemented such storage for conventional applications. The operational equation for this type of storage is similar to that for the salt cavern storage.

5. GAS TRANSPORTATION AND STORAGE SERVICES, OPERATION, AND ECONOMIC VALUE

The economic value of gas transportation and storage depends on many factors, including supply and demand trends; means of transportation and storage; measures for deliverability assurance; operational and management strategies; ability of accommodation for seasonal price fluctuations; measures for management of uncertainty, risk, surge, peaks, and pricing options; and conditions and types of available services. Mitigation of the uncertainty in natural gas price, investment, decision-making, and supply and demand is the challenge in the marketing and management issues. Development and implementation of best strategies for effective coordination of various gas sources, flexibility of transportation and storage facilities, measures for absorbing and managing peaks, and dealing with price volatility and seasonal price spreads are among the important problems. Typically, the price of natural gas remained close to \$5.00 per million British thermal units (MMBtu), or \$5.00 per thousand standard cubic feet (Mscf) during the most of the year 2003.

Pipeline operation is a dynamic industry, providing creative services such as gathering, lending, parking, pooling, wheeling, and interruptible and backhaul services. The cost of pipeline transportation depends on the expense, usage, distance, and types of services rendered. While providing valuable service to the suppliers and end users of gas, the gas marketing and trading industry makes a profit from the seasonal buying and selling of gas price differentials, by managing risk and taking advantage of creative services such as options, swaps, and futures.

Gas storage ensures adequate and reliable gas supply with less price volatility. Inventory accountability, deliverability assurance, and gas leakage mitigation are some of the measures of the technical, economic, and environmental performance of natural gas storage. Selection of the gas storage type, location, and operation mode is a delicate and strategic engineering issue. Dictated by the supply, demand, and market price trends, gas storage facilities can be operated at different gas injection and withdrawal modes. Simultaneous injection and production as well as consecutive injection and production modes are often resorted to in this practice. Hence, various gas storage facilities can be broadly classified as having flexible or inflexible storage capabilities with respect to the ability of providing different storage services, gas deliverability, and yearly cycling frequency. Flexible storage (e.g., salt caverns) allows multiple cycles during a year and high injection and deliverability rates, accommodating widely varying intra- and interday gas rates. Gas can be withdrawn quickly from a flexible storage within as short a time frame as a week. This makes it possible to have more than 12 cycles in a year. As a result, flexible storage can capture the price differentials quite effectively over short periods of time according to the trends in the market conditions. Therefore, flexible storage has a beneficial market value when it can take advantage of small price fluctuations, but it would not have an economic value under stable market conditions. In contrast, inflexible storage (e.g., a depleted reservoir and water aquifer) may be cycled only once a year, typically with injection during summer for storage and production during winter for consumption. Consequently, while providing an advantage of storing significantly large quantities of natural gas in the storage place, depleted reservoir and aquifer storages can capture only the long-term seasonal spreads in the gas market. Hence, flexible and inflexible gas storage facilities may have different

response dynamics and economic values under certain market trends.

6. HAZARDS FROM NATURAL GAS TRANSPORTATION AND STORAGE

Natural gas transportation and storage systems can create various pollution and safety problems and should be monitored, inspected, and controlled carefully in order to avoid the escape and emission of dangerous natural gas. Gas storage facilities may undergo significant variation in the gas pressure and storage stress conditions as a result of frequent injection and production under different scenarios associated with various services. Frequent overpressuring and underpressuring of gas storage can induce aging of and various types of damage to the storage facility, creating and increasing the potential for gas leak and hazards.

From the point of gas storage, depleted reservoirs, aquifers, and salt domes often contain natural defects, including natural fractures and faults, which may provide escape routes for gas. Escape, migration, and accumulation near the ground surface of large quantities of gas from storage create potential hazardous situations, such as fire and explosion, and environmental pollution problems, including noxious odors and other health hazards. Nevertheless, gas leaks from underground storage facilities are natural and unavoidable, and should be monitored carefully in order to avoid potential health and safety hazards.

7. NATURAL GAS SECURITY AND RISK ISSUES

Natural gas supply and transportation in parts of the world may be limited and/or interrupted because of political instability and unrest and frequent changes in local governments, customs, and regulations. The security and risk issues in the post-9/11/2001 world are of great concern, and ensuring the accessibility and availability of natural gas reserves and the safe operation and economic viability of natural gas transportation and storage facilities are issues of importance. Liquefied natural gas tankers, facilities for on- and offshore natural gas liquefaction, regasification, and storage, and gas pipelines are particularly vulnerable to sabotage and other man-made damages. Attacks impacting any of the storage

components may lead to rapid release of large gas clouds and to dangerous fires and explosions. Implementation of carefully designed surveillance techniques can avert such problems, and security and emergency response plans and methods for detection and deterrence of unwanted activity inside and around gas transportation and storage facilities have become an integral part of the overall natural gas handling and management task. Effective security measures require the use of highly sophisticated tools and approaches, such as the development and application of vulnerability assessment models, security vulnerability assessment of gas handling facilities, properly selected and installed security equipment, and training of personnel about safety and emergency procedures.

SEE ALSO THE FOLLOWING ARTICLES

Coal Storage and Transportation • Natural Gas, History of • Natural Gas Industry, Energy Policy in • Natural Gas Processing and Products • Natural Gas Resources, Global Distribution of • Natural Gas Resources, Unconventional • Natural Gas Transportation and Storage • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Drilling

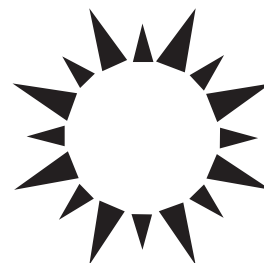
Further Reading

- Ammer, J. R., and Sames, G. P. (2000). "Advances for Improved Storage: Deliverability Enhancement and New Storage Facilities." SPE paper 65638. Proceedings of the Society of Petroleum Engineers (SPE) Eastern Regional Meeting, Morgantown, West Virginia, 17–19 October. SPE, Dallas, Texas.
- Chabrelié, M.-F. (2000). "Natural Gas in the World—2000 Survey." IFP-CEDIGAZ, Institut Français du Pétrole (IFP), Paris.
- Civan, F. (1989). "Review of Methods for Measurement of Natural Gas Specific Gravity." SPE paper 19073. Proceedings of the Society of Petroleum Engineers (SPE) Gas Technology Symposium, Dallas, Texas, 7–9 June. SPE, Dallas, Texas.
- Civan, F. (2000). "Reservoir Formation Damage—Fundamentals, Modeling, Assessment, and Mitigation." Gulf Publ. Company, Houston, Texas.
- Clark, J. (2003). Government, industry forge partnerships for security enhancement. *Oil Gas J.* **101.24**, 20–30.
- Cornot-Gandolphe, S. (1995). "Underground Gas Storage in the World—A New Era of Expansion." IFP-CEDIGAZ, Institut Français du Pétrole (IFP), Paris.
- Energy Information Administration (EIA). (2002). "The Basics of Underground Natural Gas Storage." US Department of Energy, Natural Gas Analysis Publications, Washington, D.C.
- Fitzgerald, A., and Taylor, M. (2001). "Offshore Gas-to-Solids Technology." SPE paper 71805. Proceedings of the Offshore European Conference, Aberdeen, Scotland, 4–7 September. SPE, Dallas, Texas.
- Groves, T. (2001). Terminal LNG tank management system promotes storage safety. *World Refining*, May 2001, 46–48.
- Gudmundsson, J. S., and Mork, M. (2001). "Stranded Gas to Hydrate for Storage and Transport." 2001 International Gas Research Conference, Amsterdam, November 5–8. Gas Technology Institute, Des Plaines, Illinois.
- Ikoku, C. U. (1980). "Natural Gas Engineering: A Systems Approach." PennWell Publ. Company, Tulsa, Oklahoma.
- Javier, M. (2002). "Decision-Making Under Uncertainty With Application to Natural Gas Utilization Investments." Master of Science thesis. University of Oklahoma, Norman, Oklahoma.
- Khilyuk, L. F., Chilingar, G. V., Endres, B., and Robertson, J. O., Jr. (2000). "Gas Migration, Events Preceding Earthquakes." Gulf Publ. Company, Houston, Texas.
- Mathew, J. P. (2003). Liquefied natural gas: What IS it all about? *PMA OnLine Magazine*, January 2003.
- Natural gas hydrate—A viable GTL alternative. (2003). *Chem. Eng. Prog.* **99**(3), 17.
- Perry, W. (2003). Marine transportation facilities face New US Security Regs. *Oil Gas J.* **101.24**, 32–34.
- Rojey, A., Jaffret, C., Cornot-Gandolphe, S., Durand, B., Jullian, S., and Valais, M. (1997). "Natural Gas: Production, Processing, Transport." Institut Français du Pétrole (IFP), Paris.
- Tek, M. R. (1996). "Natural Gas Underground Storage: Inventory and Deliverability." PennWell Publ. Company, Tulsa, Oklahoma.
- Tobin, J. (2001). "Natural Gas Transportation—Infrastructure Issues and Operational Trends." Energy Information Administration, Natural Gas Division, U.S. Department of Energy, Washington, D.C.
- Tobin, J., and Thompson, J. (2001). "Natural Gas Storage in the United States in 2001: A Current Assessment and Near-Term Outlook." Energy Information Administration, U.S. Department of Energy, Washington, D.C.
- U.S. Environmental Protection Agency (EPA). (1999). "National Emission Standards For Hazardous Air Pollutants: Oil and Natural Gas Production and Natural Gas Transmission and Storage, Final Rules." Federal Register, 40 CFR Part 63, Vol. 64, No 116, June 17. EPA, Washington, D.C.
- Williams, D. (2001). "The Value of Gas Storage in Today's Energy Market." Technical paper, Solution Mining Research Institute (SMRI), Fall 2001 Meeting, 7–10 September, Albuquerque, New Mexico. SMRI, Richmond, Texas.

Net Energy Analysis: Concepts and Methods

ROBERT A. HERENDEEN

Illinois Natural History Survey
Champaign, Illinois, United States



-
1. Introduction and Background
 2. Quantitative Energy Balance Indicators
 3. Conceptual and Procedural Issues
 4. Results and Trends
 5. Concluding Comments

Glossary

absolute energy ratio (AER) Energy produced/energy required, including energy from the ground or the sun (dimensionless).

end use Consideration taking into account the efficiency of use of energy produced.

incremental energy ratio (IER) Energy produced/energy required, excluding energy from the ground or the sun (dimensionless).

internal rate of return (IRR) The implied interest rate if energy inputs are considered an investment and energy produced is considered a payoff (units = 1/time).

payback time The time for the accrued energy produced to balance the energy inputs (units = time).

power Energy per unit time.

power curve A graph of power inputs and outputs vs time.

premium fuel An energy type considered especially valuable. For example, oil is usually considered premium relative to coal because it is more flexible in use, less polluting, etc.

Net energy analysis of an energy technology is a comparison of the energy output with the energy needed to supply all inputs—the energy source, materials, and services—to construct, operate, and dispose of the technology.

1. INTRODUCTION AND BACKGROUND

At first glance, net energy analysis (NEA) is a natural, intuitively sensible extension of the idea of energy

cost. Everything requires energy somewhere along the chain of production, including the equipment and services used to find and extract energy. In addition, energy seems to be harder to get with time: efforts to drill, dig, and dam must be in deeper and in more forbidding places for our oil, gas, coal, and hydro-power (offshore for oil and gas, in mountainous terrain for coal, and into the sub-Arctic for hydro-power). This in turn probably implies more energy cost, and shifts the time when this trend will hit the breakeven point. NEA seeks to find out. Some proponents of NEA claim that standard economics will not always indicate this problem (especially if subsidies to energy industries are involved, as they usually are) and that therefore net energy analysis should be carried out largely independent of economic analysis. However, the more common argument is that energy analysis is a supplement to, not a substitute for, economic analysis in decision making.

Net energy concerns peaked in the 1970s and early 1980s following the oil embargo/energy crisis years of 1973 and 1979–1980. Many NEA studies covered oil made from coal or extracted from tar sands and oil shale, geothermal sources, alcohol fuels from grain, biomass plantations, solar electricity from orbiting satellites, and nuclear electricity from ambitious proposed plant-building programs. In 1974, Federal legislation requiring net energy analysis of federally supported energy facilities was enacted. It required that “the potential for production of net energy by the proposed technology at the state of commercial application shall be analyzed and considered in evaluating proposals” [Public Law No. 93-577, Sect. 5(a) (5)]. Since that time, interest in net energy has diminished, but continues sporadically. Adherence to Public Law 93-577 was abandoned early by the U.S. Department of Energy (DOE). In 1982, responding to a criticism from the U.S. General Accounting Office that it was not performing NEAs, DOE said “[it is]

DOE's view that the benefits of [NEA] are not worth the time and effort involved for general application." Net energy analysis is a type of life cycle analysis. The past decade has seen an increase in applications of the latter, but usually these have been applied to consumer products rather than to energy facilities, using toxic and other materials, water, etc., rather than energy, as numeraire.

The DOE had a point; the devil is in the details. NEA is an elusive concept subject to a number of inherent, generic problems that complicate its application. Some of these problems persist, not because they are unstudied, but because they reflect fundamental ambiguities that can be removed only by judgmental decision. The problems include all of those with energy analysis as well as others specific to the "net" question. The difficulty of specifying a system boundary, the question of how to compare energy produced and consumed at different times, and the problem of how to compare energy types of different thermodynamic qualities, density, and ease of storage all make NEA harder to perform and interpret. Some analysts therefore reject net energy analysis but support energy analysis. The latter considers the energy output to be a good or service and attaches no special significance to the fact that it is energy. In this view, it is reasonable and useful to determine how much energy is required to keep a room warm, or to produce a head of lettuce or even a ton of coal, but it is confusing at best and misleading at worst to compare outputs and inputs of a coal mine purely on the basis of energy cost. This fundamental objection attacks the very basis of NEA: NEA assumes that the economic/human life-support system can be separated into the "energy system" and the "rest of the system" and that studying the energy system as a distinct entity is valid. For many applications, this separation, although not perfectly defensible, seems to be acceptable. For others, there are stronger objections.

Little has occurred since 1982 to challenge DOE's dismissive pronouncement, and in this author's opinion NEA has very seldom, if ever, been used as a critical decision criterion. In the Grain Belt, NEA arguments have sometimes been quoted regarding ethanol production from corn, but the sorting out that occurred (with small operators using dry-milled corn shutting down while larger operators using wet milling flourished) can be understood on the basis of the monetary economies of scale, of the role of economic subsidies, and of the details of the flexibility of the two milling processes. In fairness, there is a large debt owed to NEA as a stimulant to proper thinking about resources. Predicting resource avail-

ability on the basis of past trends has been shown to be fallacious by the example of domestic U.S. oil, where increased drilling in "Project Independence" did not reverse the fact that all-time production peak was in 1971. To the extent that prediction by linear extrapolation is an economics technique and that prediction by incorporating the physical aspects of the resource is a net energy analysis technique, NEA deserves credit. But anything can be done poorly, and proper economics should incorporate physical realities. Once that is granted, the question of usefulness lies in the details of NEA.

In advance, it can be anticipated that two specific outcomes of NEA will be least controversial. If the technology appears to be a clear energy loser even after all uncertainties of technique and data are accounted for, the result is useful; the program should be scrapped. Likewise, if the technology appears an unambiguous energy producer, it is likely that NEA can now be deemphasized and the decision to proceed based on other criteria. If the technology is close to the energy breakeven point, NEA seems more appealing. Of course, in this case, a higher degree of accuracy in the result, and hence in the needed data, will be required.

2. QUANTITATIVE ENERGY BALANCE INDICATORS

The statement that an energy technology can be an energy winner or loser implies three assumptions. First, because energy is not created or destroyed (mass-energy in the case of nuclear energy), it would seem that at best an energy (conversion) technology can only break even. NEA usually assumes that the energy in the ground (or coming from the sun) is not to be counted as an input, i.e., is outside the system boundary. This is discussed in detail later. Second, it is usually assumed that high-quality energy (in the thermodynamic sense) is desirable, whereas low-quality energy is not. This would be covered properly if "free energy" were used instead of energy. The common (mal)practice is followed here, i.e., using "energy" actually to mean thermodynamic free energy. Third, it is assumed here that all material and service inputs to an energy facility are expressible in terms of the energy needed to produce them.

Figure 1 shows the basic NEA framework, assuming only one energy type. It is useful to have a normalized indicator of the net energy balance; two are the incremental energy ratio and the absolute

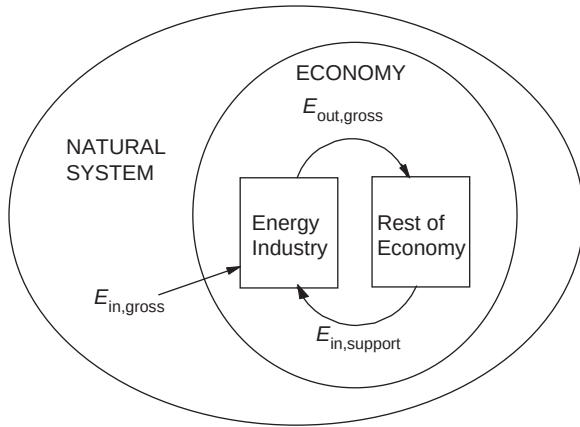


FIGURE 1 Energy flows for net energy analysis.

energy ratio:

$$\text{Incremental energy ratio} = \frac{E_{\text{out,gross}}}{E_{\text{in,support}}}$$

$$\text{Absolute energy ratio} = \frac{E_{\text{out,gross}}}{(E_{\text{in,gross}} + E_{\text{in,support}})}$$

Support energy is that obtained directly from the rest of the economy or used by the economy to provide necessary materials and service inputs. $E_{\text{in,gross}}$ is the energy from the ground or sun, that energy usually referred to as “resource.” $E_{\text{out,gross}}$ is the energy supplied to the rest of the economy. The difference between the two ratios depends on the system boundary. The incremental energy ratio (IER) parallels standard economic practice in that the “cost” of exploiting a resource is the cost of extraction, and includes as inputs only those taken from the rest of the economy. The latter can be thought of as “invested” energy, and the IER is called by some workers “energy return on investment.” An $\text{IER} > 1$ means that we have a net energy producer. The absolute energy ratio (AER) is appropriate to the more global question of the physical efficiency with which a given (say) fossil or renewable energy source can be converted into useful energy. The AER (which never exceeds 1) is useful for determining how much of a given stock of energy resource can be used by the rest of the economy.

Figure 2 illustrates the interplay of the IER and AER. Because of the possibility of energy consumption within the boundary of the energy technology, knowledge of the IER does not uniquely specify the AER, and vice versa. To exhibit how the energy flows vary and cumulate over the facility’s lifetime, we can use a “power curve,” shown in Fig. 3. The power curve tells “everything” (assuming that the system boundary is properly stated, etc.) but is cumbersome. It is desired to have some summary indicator of the whole time path,

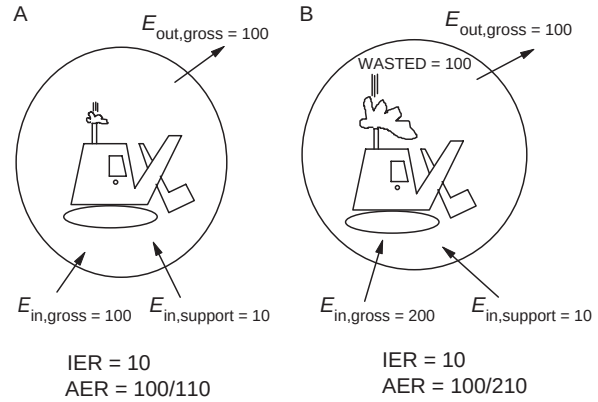


FIGURE 2 Incremental energy ratio (IER) and absolute energy ratio (AER) for two hypothetical coal mines. Both mines have the same IER, but mine B has a lower AER because of energy waste within the system boundary.

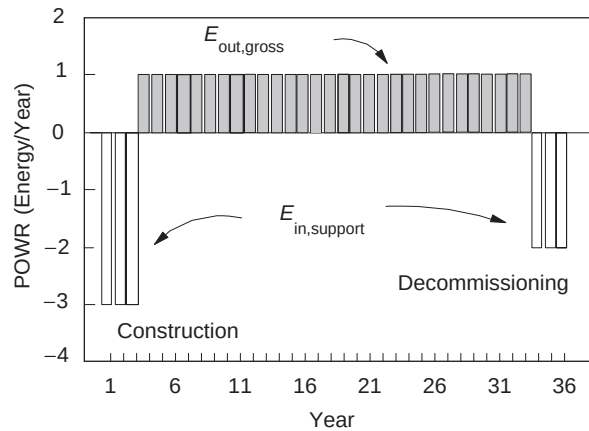


FIGURE 3 Hypothetical power curve for an energy facility through construction, operation, and decommissioning. $\text{IER} = 30 / [(3 \times 3) + (2 \times 3)] = 2$. Payback time = 12 years, but it is not really “simple” because the concept does not apply to a power curve with costs following benefits. $\text{IRR} = 9.1\%/\text{year}$.

such as (1) the energy ratio (already defined), (2) the energy payback time (the time for the accrued energy produced to balance the energy inputs), and (3) the energy internal rate of return (IRR; the implied interest rate if energy inputs are considered an investment and energy produced is considered a payoff). The formal statement is that the IRR is that value of r for which

$$\frac{\sum_{t=1}^N \frac{E_{\text{out,gross},t}}{(1+r)^{t-1}}}{\sum_{t=1}^N \frac{E_{\text{in,support},t}}{(1+r)^{t-1}}} = 1.$$

All three of these indicators have analogues in economics. All require summing energy over many years, and it is necessary to decide how to weigh the present against the future. The IRR by definition

explicitly assumes the standard economic practice of geometric discounting. However, for energy ratio and payback time there is no fixed rule. Most energy analysts do not discount, and that convention will be followed here. Summary indicators will be used for ease of communication but they sacrifice detail and introduce ambiguity. For example, consider simple payback time (i.e., payback time with no discounting), which is defined as that time at which the energy facility has produced as much energy as it has consumed. This definition implicitly assumes a power curve that is negative at first and then becomes, and remains, positive during facility operation. It would be less useful, and even misleading, for more complicated power curves, such as for a facility that requires significant energy inputs late in the lifetime, as for repair, maintenance, or decommissioning (see Fig. 3).

The energy ratio is also subject to difficulties in interpretation. For example, if an electric plant uses some of its own output, should that be added to the

inputs or subtracted from the output? This makes no difference to payback time or to the internal rate of return, but does affect energy ratio. If the power curve is well behaved, i.e., negative at first and then positive over the whole lifetime, then the following statements are equivalent statements of positive net energy production:

- $IER > 1$.
- Simple payback time $<$ facility lifetime.
- Internal rate of return > 0 .

3. CONCEPTUAL AND PROCEDURAL ISSUES

Several of the conceptual and procedural difficulties of NEA are summarized in Table I. It should be noted that there is considerable overlap between issues, and that several also plague economic analysis. Most are

TABLE I
Conceptual and Procedural Difficulties with Net Energy Analysis

Issue	Example	Remedied by careful statement of problem?	Comment
1. Specification of system boundary	Is sunlight "free"?	Yes	Differences of opinion persist among both analysts and users
a. Spatial	Is soil quality maintained in biomass energy production?		
b. Temporal	How much of a passive solar house is functional vs decorative?		
c. Conceptual	Should the energy cost of decommissioning a nuclear plant be included? Should energy costs of labor be included?		
2. End-use consideration	Should NEA of a coal-electric plant be influenced by the efficiency of residential refrigerators? If a car gets better miles per gallon than expected, is the benefit allocatable to the ethanol-from-grain process?	No	An example of the fundamental question of to what extent the energy system is separable from the rest of the economy
3. Opportunity cost	Can energy conservation measures be thought of as producing (saved) energy and then subjected to NEA?	Yes	Difficult-to-set ground rules that are not situation-specific
4. Dynamic problem	Will a rapid buildup of many energy facilities produce net positive energy? The answer can be "no," even if the IER for each facility is > 1	Yes	Raised regarding nuclear and solar energy in the 1970s but shown not to be not significant (rapid buildup did not occur)
5. Existence of more than one kind of energy	Should oil and coal be equally weighted in NEA?	Not completely	Very vexing
6. Average vs. marginal accounting	New oil wells often have a lower IER compared to old wells	Yes	Generic problem
7. Dependence of NEA on economic factors	Energy requirements for steel production depend on price of energy	No	Unavoidable because NEA is based on real technology

problems with energy analysis in general as well as with NEA specifically.

3.1 End-Use Considerations

Should NEA take into account the efficiency of use of $E_{\text{out,gross}}$ to provide an energy service such as cooling a space (e.g., a refrigerator) or moving people (e.g., an automobile)? An example of this problem is the gasohol miles-per-gallon question. It was observed in the late 1970s that cars using gasohol (a mixture of 90% unleaded gasoline and 10% ethanol) got more miles per gallon than would be expected on the basis of the enthalpy of combustion (“energy content”) of the mixture as compared with that of pure gasoline. For the purpose of NEA, this could be viewed as a production of extra energy, i.e., an increase in $E_{\text{out,gross}}$. Should the credit for this increased output be allocated to the ethanol-producing process? If so, the IER increased from 1–1.7 to 1.5–2.4. Or, because the improvement is really due to the existence of the mixture, should it be inadmissible to perform the NEA of ethanol only? And what if the ethanol goes to an end use other than gasohol, for which this issue is moot? Further, this is all dependent on the type of engine. Research provides no definite answer. The user must decide, being attentive to the likelihood that in the event of controversy, different interest groups will tend to select the interpretation that best supports their cases.

3.2 Energy Conservation

NEA can be applied to energy conservation, in which case $E_{\text{out,gross}}$ is the energy saved (“neg-a-watts”), and $E_{\text{in,support}}$ is the energy required to make the insulation, the more efficient motor, etc.

3.3 The Dynamic Problem

This is an age–structure demographic issue, of concern with the rapid buildup of an energy technology. Even though a single facility may have a (lifetime) $\text{IER} > 1$, a program of building many facilities may be a net energy sink for many years. If the program is growing fast enough, the average facility is not even completely built! Concern over this problem was stimulated by exponential program goals for nuclear power plants (in the United States and Great Britain, doubling time = 5 years) and solar heating (United States, doubling time = 3 years). The exponential growth did not occur, and the concern has largely gone away. But it is always a potential problem.

3.4 Energy for Human Labor

There is disagreement on whether to include the energy to support human labor. The two following reasons argue against inclusion:

1. It is difficult to agree on how much of it to include. Is it the energy consequences of a worker’s whole paycheck or just that portion spent for food and lodging?
2. Including labor means that the economy is viewed as producing goods only for government consumption, exports, and capital purchases. Consumption by residences is assumed to be fixed in amount and, especially, in “mix.” Other energy research (for example, energy conservation) exploits the flexibility of personal consumption patterns (to purchase more insulation and less natural gas, for example). This flexibility seems to be preferable.

3.5 Different Energy Types and Premium Fuels

Not all energy types are equally useful. On average, we consider oil and natural gas more useful than coal, and the price reflects this (gasoline retails for about \$10 per million Btu; coal is about \$1.50 per million Btu). Various analysts have accounted for this by giving different weights to different energy types based on criteria such as scarcity or economic price. One example is calculating the IER for “premium” fuels only. Thus, if liquid fuels are defined as premium and an ethanol-from-grain plant can burn coal or crop residues, the IER will be increased over one that burns oil. This explicit introduction of nonenergy criteria again blurs the boundary of NEA.

3.6 Facility Lifetime and Other Sensitivity Issues

In Fig. 3 we see that if the energy facility has a productive life of only 15 years, instead of 30, the IER is reduced to 1, which is the breakeven point. Besides lowered lifetime, lack of reliability could cause diminished output and increased energy inputs for maintenance. Similarly, increased energy costs of decommissioning would reduce the IER; this is a real issue for nuclear power. Of course, the IER could be increased by the opposite of these trends. Explicit treatment of the uncertainty of IER is often performed. An example is an NEA of the proposed solar power satellite (SPS) of the late 1970s. This involved arrays of solar cells in geosynchronous orbit, on which electricity was

converted to microwaves and beamed to terrestrial antennas, for conversion back to electricity. Uncertainties of the lifetime and performance of photovoltaic cells in a space environment, plus relatively high energy requirements for heavy, crystal-grown silicon, which in turn required higher energy costs to place them in orbit, resulted in an IER of 2.1 ± 0.8 for the SPS. The project was not funded, though monetary issues probably dominated net energy questions. Today's solar cells are lighter and more efficient and durable, so that portion of the system would likely be more energetically favorable now.

4. RESULTS AND TRENDS

NEA has changed little since the spate of work in the 1970s and 1980s. A literature search today shows relatively few studies, and these tend to emphasize window coatings to manage solar gain and loss, a continued interest in solid biomass fuels and liquids-

from-biomass, and photovoltaic units. Thus, the data in Table II, showing results for several conventional and unconventional technologies, are dominated by 25-year-old results. Table II indicates that coal tends to have IERs on the order of several 10s. Crude oil and natural gas are on the order of 10. Electricity from a fossil plant is on the order of 3 to 10. Geothermal electricity is of order 4 to 10. For solar technologies, we see that direct combustion of biomass has IERs of 3 to 20, with the energy cost of fertilization a dominant factor at the low end. Biomass liquids have IERs of 1 to 5. One study of ethanol from corn found an IER from 1 to 1.4 for total energy, which increased to 2.3–3.8 when calculated for premium fuel (liquids) only (another study 25 years later similarly found IER = 1.5). Solar space heat has IERs from 2 to 15. The high variation is partly due to climate, partly due to the boundary issue of how much of a solar house is the solar system (especially for passive solar), and partly due to the degree of coverage. The latter is another boundary

TABLE II
Selected Net Energy Results

Technology	Date	IER (unless otherwise stated)	Comment
Coal, U.S. average	1970s	37	
Eastern surface coal	1970s	43	Surface mining tends to have a higher IER compared to deep mining
Crude petroleum delivered to refinery	1970s	7	
Natural gas delivered through gas utility to user	1970s	11	
Coal mine-mouth electric power plant	1977	8	
Solar power satellite	1979	2.1 ± 0.8	Very uncertain; the IER for the photovoltaic cells alone would likely be higher today
Ethanol from corn for gasohol (gasoline:ethanol = 90:10)	1979, 1995	≈ 1.5 to 2.5	Extremely sensitive to issues in Table I
Ceiling insulation as source of saved energy	1970s	136	Best case, previously uninsulated house
Geothermal-electric plant (vapor dominated)	1981	13 ± 4	Geothermal heat is not counted in the IER
Geothermal-electric plant (liquid dominated)	1981	4 ± 1	Geothermal heat is not counted in the IER
Short-rotation (≈ 5 -yr) trees for direct combustion	1980s	10 to 20 if unfertilized; 2 to 4 if fertilized	Unfertilized is likely unsustainable
Short-rotation biomass (trees, ≈ 5 yrs; grass, 1 yr) for direct combustion	1980s	6 if unfertilized; 2 if fertilized	Unfertilized is likely unsustainable
Residential solar space heat	1980s	2–15	Depends on climate and how well insulated the structure is (climate dependent); passive (no moving parts) tends to have higher IER compared to active
Residential solar hot water	1980s	2–5	Climate dependent
Photovoltaic cells	1976	Simple energy payback time ≈ 12 yr	
Photovoltaic cells	2000	Simple energy payback time = 1.1 to 4 yr	Payback time must be compared with device lifetime; this study concludes that the IER ≈ 30 , implying a lifetime of at least 30 + yr

issue. A solar heating system that can supply year-round heat, including that one continuously overcast 5°F week in January, will require more inputs (such as more glass, thermal mass, and storage tanks). This system's IER will be lower than the IER for a system that is not expected to cover the worst periods—for which there is, say, a backup propane furnace that is not considered part of the solar heating system.

A similar comment applies to conservation as an energy source. The $IER > 100$ for ceiling insulation is for the first few inches of insulation in a totally uninsulated house. Putting that same insulation in a house already somewhat insulated would save less energy and give a lower IER. Also, photovoltaic cells have become better net energy producers over time as reliability has increased and material inputs have decreased.

5. CONCLUDING COMMENTS

Interpreting net energy information such as that in Table II almost instantly raises questions of the details and assumptions, which the critical reader must embrace. In spite of these, the concept (and the specter!) of the energy breakeven point and hence exhaustion is compelling enough for continued study. Perhaps the most vivid way to address this is by asking how IERs are changing over time.

SEE ALSO THE FOLLOWING ARTICLES

Aggregation of Energy • Cost-Benefit Analysis Applied to Energy • Emergy Analysis and Environ-

mental Accounting • Exergy Analysis of Energy Systems • Goods and Services: Energy Costs • Input-Output Analysis • Life Cycle Assessment and Energy Systems • Modeling Energy Supply and Demand: A Comparison of Approaches • Multicriteria Analysis of Energy • National Energy Modeling Systems

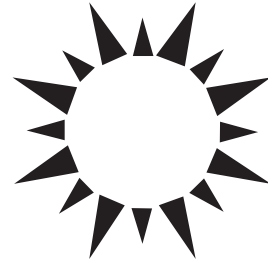
Further Reading

- Bullard, C., Penner, P., and Pilati, D. (1978). Net energy analysis: Handbook for combining process and input-output analysis. *Resourc. Energy* 1, 267–313.
- Chambers, R., Herendeen, R., Joyce, J., and Penner, P. (1979). Gasohol: Does it or doesn't it...produce positive net energy. *Science* 206, 789–795.
- Herendeen, R. (1988). Net energy considerations. In “Economic Analysis of Solar Thermal Energy Systems” (R. West and F. Kreith, Eds.), pp. 255–273. MIT Press, Cambridge.
- Herendeen, R. (1998). “Ecological Numeracy: Quantitative Analysis of Environmental Issues.” John Wiley and Sons, New York. See Chapter 8.
- Herendeen, R., Kary, T., and Rebitzer, J. (1979). Energy analysis of the solar power satellite. *Science* 205, 451–454.
- Knapp, K., Jester, T., Milhalik, G. (2000). “Energy Balances for Photovoltaic Modules; Status and Prospects.” 28th IEEE Photovoltaic Specialists' Conference, Anchorage, AK. Institute of Electrical and Electronic Engineers, New York.
- Leach, G. (1975). Net energy: is it any use. *Energy Policy* 2, 332–344.
- Pilati, D. (1977). Energy analysis of electricity supply and energy conservation options. *Energy* 2, 1–7.
- Roberts, F. (ed.). (1978). “Symposium Papers: Energy Modelling and Net Energy Analysis.” Institute of Gas Technology, Chicago.
- Spreng, D. (1988). “Net Energy Analysis and the Energy Requirements of Energy Systems.” Praeger, New York.
- Whipple, C. (1980). The energy impacts of solar heating. *Science* 208, 262–266.

Neural Network Modeling of Energy Systems

SOTERIS A. KALOGIROU

Higher Technical Institute
Nicosia, Cyprus



-
1. Artificial Intelligence and Neural Networks
 2. Biological and Artificial Neurons
 3. Principles of Artificial Neural Networks
 4. Network Parameters Selection
 5. Modeling of Energy Systems Using Neural Networks
 6. Conclusions

Glossary

artificial intelligence The science devoted to the reproduction of the methods or results of human reasoning and brain activity.

artificial neural network A network of artificial neurons interacting with one another in a concerted manner that achieves “intelligent” results through many parallel computations without employing rules or other logical structures.

artificial neuron Element of a neural network that behaves like a neuron of the brain that possesses a number of input values to produce an output value.

backpropagation A method of supervised learning applied in multilayer feedforward neural networks. The error signal between the actual output and target is fed back through the processing layer, which tries to reduce this error by changing the connection weights.

biological neuron A brain cell that processes data and controls the response of the body. Neurons are connected to the brain and other nerve cells by transmitting connectors (axons) and receiving connectors (dendrites).

epoch A complete pass through the neural network of the entire set of training patterns.

hidden layer A neural network layer of nodes between input and output layers that contains the weights and processing data.

layers A group of neurons in a network. Layers can be input, output, or hidden.

pattern A single record or row of variables, also called an observation, that consists of inputs and outputs.

training In neural networks, training is synonymous to learning. During training, the system adapts in a

specified manner so that some parts of the system suggest a meaningful behavior, projected as output.
weights Parameters for which their values are adapted during training in order to perform a task.

Artificial neural networks, or simply neural networks, are collections of small, individual interconnected processing units. Information is passed between these units along interconnections. An incoming connection has two values associated with it, an input value and a weight. The output of the unit is a function of the summed value. Neural networks, although implemented on computers, are not programmed to perform specific tasks. Instead, they are trained with respect to past history data sets until they learn the patterns presented to them. Once they are trained, new unknown patterns may be presented to them for prediction or classification.

1. ARTIFICIAL INTELLIGENCE AND NEURAL NETWORKS

For the estimation of the flow of energy and the performance of energy systems, analytic computer codes are often used. The algorithms employed are usually complicated, involving the solution of complex differential equations. These analytic programs usually require large computer resources and need a considerable amount of time to achieve convergence and thus to give accurate predictions. Instead of complex rules and mathematical routines, artificial neural networks are able to learn the key information patterns within a multidimensional information domain. In addition, neural networks are fault tolerant, robust, and noise immune. Because they are inherently noisy, data from energy systems represent good candidate problems to be handled with neural networks.

The possibility of developing a machine that would “think” has intrigued humans since ancient times. In 1637, the French philosopher–mathematician Rene Descartes predicted that it would never be possible to make a machine that thinks as humans do. However, in 1950, the British mathematician and computer pioneer Alan Turing declared that one day there would be a machine that could duplicate human intelligence in every way.

Artificial intelligence (AI) is a term that generally indicates the ability of a machine or artifact to perform the same kinds of functions that characterize human thought. The term AI has also been applied to computer systems and programs capable of performing tasks more complex than straightforward programming but still far from the realm of actual thought. Thus, AI is the study of how to make computers do things similar to the way human beings do them. This term is often misinterpreted, and in order to avoid misunderstandings scholars usually employ the term computational intelligence.

It should be noted that solving a computation does not indicate understanding, something a person who solved a problem would have. Human reasoning is not based solely on rules of logic. It also involves perception, awareness, emotional preferences, values, the ability to generalize and weigh many options, evaluating experience, and many more.

Machinery can outperform humans physically. Similarly, computers can outperform mental functions in limited areas, particularly in the speed of mathematical calculations. The fastest computers are able to perform approximately 10 billion calculations per second. However, making more powerful computers is probably not the way to create a machine capable of thinking because computer programs operate according to set procedures, or logic steps, called algorithms. In addition, most computers do serial processing and operations such as those of recognition and computation are performed one at a time. The brain, however, works in a manner called parallel processing—that is, it performs a number of operations simultaneously. To achieve simulated parallel processing, supercomputers have been developed that have multiple processors that follow several algorithms at the same time.

During the past 15 years, there has been a substantial increase in interest in artificial neural networks. Neural networks are suitable for some tasks while lacking in some others. Particularly, they are good for tasks involving incomplete data sets, fuzzy or incomplete information, and for highly complex and ill-defined problems, for which humans

usually decide on an intuitional basis. Additionally, they can learn from examples and are able to deal with nonlinear problems. They also exhibit robustness and fault tolerance. The tasks that neural networks cannot handle effectively are those requiring high accuracy and precision, such as in logic and arithmetic.

Neural networks have been applied successfully in various fields—engineering, mathematics, economics, medicine, meteorology, psychology, neurology, and many others. Some of the most important applications are in pattern, sound, and speech recognition; the analysis of electromyographs, cardiographs, and other medical signatures; the identification of military targets; the identification of explosives and other illegal substances in passenger suitcases; and the prediction of mineral exploration sites. They have also been used in market trends and weather forecasting, electrical and thermal load prediction, adaptive and robotic control, and many others. Neural networks are also used for energy systems process control because they can build predictive models of the process from multidimensional data routinely collected from sensors.

2. BIOLOGICAL AND ARTIFICIAL NEURONS

The basic building blocks of the nervous system are called neurons. A schematic of a biological neuron is shown in Fig. 1. Neurons are excitable cells composed of the cell body or soma (where the cell nucleus is located) and projections branching out from the soma. Each neuron has incoming fibers called dendrites and an outgoing fiber called an axon. Dendrites are input devices through which the neuron receives signals from other neurons. There is usually a single axon per neuron, and it serves as its primary output device. This may be connected to other neurons or may terminate in a muscle tissue, serving as a control signal for the muscle activity. The points of contact between an axon branch and a receiving neuron are called synapses. The change of synaptic transmission efficiency acts as a memory for past experiences. In the brain, there is a flow of coded information (using electrochemical media, the so-called neurotransmitters) from the synapses (having a density of approximately 10^4 synapses per neuron) toward the axon. The axon of each neuron transmits information to a number of other neurons. The neuron receives information at the dendrites from a large number of other neurons. It is estimated that

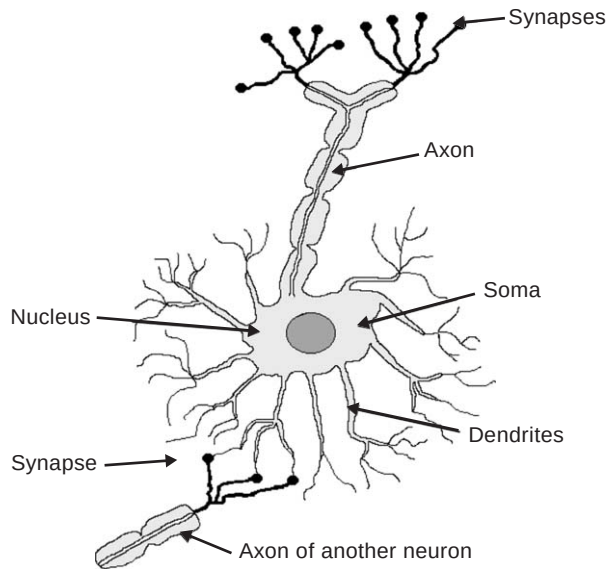


FIGURE 1 A simplified model of a biological neuron showing the soma with nucleus, the axon, dendrites, and synapses. The axon is the neuron output device and the dendrites are the input devices. The point of contact between an axon branch and a receiving neuron is called a synapse.

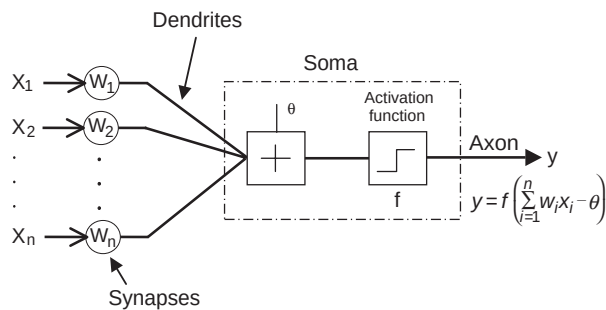


FIGURE 2 A simplified model of an artificial neuron used to model the features of the biological neuron. The synapses are the connection weights, the soma includes the summation and activation units, and the axon represents the output of the artificial neuron as in the biological counterpart. The unit step function is equal to 1 if $x \geq 0$ and is equal to zero otherwise.

each neuron may receive stimuli from as many as 10,000 other neurons. Groups of neurons are organized into subsystems and the integration of these subsystems forms the brain. It is estimated that the human brain has approximately 100 billion interconnected neurons.

Computing networks are far simpler than their biological counterparts. The first simplified neuron model was proposed by McCulloch and Pitts in 1943. Figure 2 shows a highly simplified model of an artificial neuron that may be used to stimulate some important aspects of the real biological neuron. The

artificial neuron has all the features of a biological neuron (i.e., synapses, dendrites, soma, and one output line representing the axon of the neuron). The neuron is considered to be a binary threshold device with n inputs $x_1, \dots, x_n$, one output y , a threshold level θ , and n synaptic weights $w_1, \dots, w_n$. In such a system, excitation is applied to the input of the network. Following some suitable operation, it results in a desired output. At the synapses, there is an accumulation of some potential, which in the case of the artificial neurons is modeled as a connection weight. These weights are continuously modified during training, based on suitable learning rules.

3. PRINCIPLES OF ARTIFICIAL NEURAL NETWORKS

An artificial neural network is a group of interconnected artificial neurons interacting with one another in a concerted manner. It is in fact a massively parallel distributed processor that has a natural propensity for storing experiential knowledge and making it available for use. It resembles the human brain in two respects: The knowledge is acquired by the network through a learning process, and interneuron connection strengths known as synaptic weights are used to store the knowledge.

Artificial neural network models may be used as an alternative method in engineering analysis and predictions. They operate like a “black box” model, requiring no detailed information about the system. They imitate somewhat the learning process of a human brain because they learn the relationship between the input parameters and the controlled and uncontrolled variables by studying previously recorded data. In this way, they operate similarly to nonlinear regression, but they are much more powerful than regression analysis. Neural networks are able to handle large and complex systems with many interrelated parameters. They seem to simply ignore excess input parameters that are of minimal significance and concentrate instead on the more important inputs.

A schematic diagram of a typical multilayer feedforward neural network architecture is shown in Fig. 3. The network usually consists of an input layer, some hidden layers, and an output layer. In its simple form, each single neuron is connected to other neurons of a previous layer through adaptable synaptic weights. Knowledge is usually stored as a set of connection weights (presumably corresponding to synapse efficacy in biological neural systems).

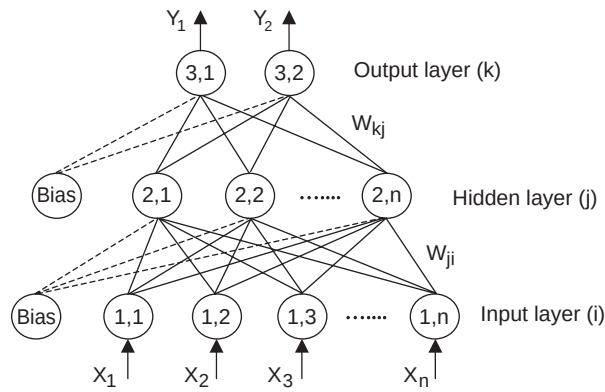


FIGURE 3 Schematic diagram of a multilayer feedforward neural network with an input layer (*i*), a hidden layer (*j*), and an output layer (*k*).

Training is the process of modifying the connection weights in some orderly fashion using a suitable learning method. The network uses a learning mode in which an input is presented to the network along with the desired output and the weights are adjusted so that the network attempts to produce the desired output. Before training, the weights are random and have no meaning, whereas after training they contain meaningful information.

Figure 4 illustrates how information is processed through a single node. The node receives weighted activation of other nodes through its incoming connections. These are first added (summation) and the result is passed through an activation function; the outcome is the activation of the node. For each of the outgoing connections, this activation value is multiplied by the specific weight and transferred to the next node.

A training set is a group of matched input and output patterns used for training the network. The outputs are the dependent variables that the network produces for the corresponding input. When each pattern is read, the network uses the input data to produce an output, which is then compared to the correct or desired output. If there is a difference, the connection weights (usually but not always) are altered in such a direction that the error is decreased. After the network has run through all the input patterns, if the error is still greater than the maximum desired tolerance, the neural network runs through all the input patterns repeatedly until all the errors are within the required tolerance. When the training reaches a satisfactory level, the network holds the weights constant and the trained network can be used to define associations in new input data sets not used to train it.

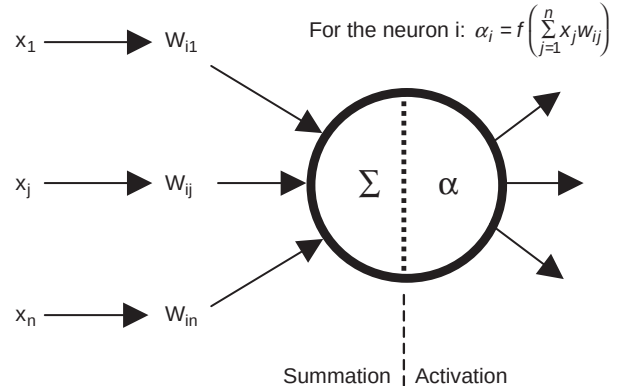


FIGURE 4 Information processing in a neural network unit.

By learning, it is meant that the system adapts (usually by changing suitable controllable parameters) in a specified manner so that some parts of the system suggest a meaningful behavior, projected as output. The controllable parameters have different names, such as synaptic weights, synaptic efficacies, and free parameters.

The classical view of learning is well interpreted and documented in approximation theories. In these, learning may be interpreted as finding a suitable hypersurface that fits known input/output data points in such a manner that the mapping is acceptably accurate. Such a mapping is usually accomplished by employing simple nonlinear functions that are used to compose the required function.

Generally, learning is achieved through any change in any characteristic of a network so that meaningful results are achieved. Thus, learning may be achieved not only through synaptic weight modification but also by modifications in network structure and through appropriate choice of activation functions, number of neurons in hidden layers, number of input parameters used, and others. By meaningful results, it is meant that a desired objective is met with a satisfactory degree of success. The objective is usually quantified by a suitable criterion or cost function. It is usually a process of minimizing an error function or maximizing a benefit function. In this respect, learning resembles optimization.

The information processing at each node site is performed by combining all input numerical information from upstream nodes in a weighted average of the form

$$\beta_i = \sum_j w_{ij} \alpha_{pj} + b_1, \quad (1)$$

where b_1 is a constant term referred to as the bias. The final nodal output is computed via the activation

function. There are many such functions, but the most common one is the logistic

$$\alpha_{(pi)} = \frac{1}{1 + e^{-\beta_i}}. \quad (2)$$

The most popular learning algorithms are backpropagation and its variants. The backpropagation algorithm is one of the most powerful learning algorithms in neural networks. Backpropagation training is a gradient descent algorithm. It tries to improve the performance of the neural network by reducing the total error by changing the weights along its gradient. The error is expressed by the root mean square value (RMS), which can be calculated by

$$E = \frac{1}{2} \left[\sum_p \sum_i |t_{ip} - o_{ip}|^2 \right]^{1/2}, \quad (3)$$

where E is the RMS error, t is the network output (target), and o is the desired output vectors over all pattern (p). An error of zero indicates that all the output patterns computed by the artificial neural network perfectly match the expected values and the network is well trained. In brief, backpropagation training is performed by initially assigning random values to the weight terms (w_{ij}) in all nodes, where j is a summation of all nodes in the previous layer of nodes, and i is the node position in the present layer. Each time a training pattern is presented to the artificial neural network, the activation for each node, α_{pi} , is computed. After the output of the layer is computed, the error term, δ_{pi} , for each node is computed backwards through the network. This error term is the product of the error function, E , and the derivative of the activation function; hence, it is a measure of the change in the network output produced by an incremental change in the node weight values. For the output layer nodes and for the case of the logistic-sigmoid activation, the error term is computed as

$$\delta_{pi} = (t_{pi} - \alpha_{pi}) \alpha_{pi} (1 - \alpha_{pi}). \quad (4)$$

For a node in a hidden layer,

$$\delta_{pi} = \alpha_{pi} (1 - \alpha_{pi}) \sum_k \delta_{pk} w_{kj}, \quad (5)$$

where k is a summation over all nodes in the downstream layer (the layer in the direction of the output layer), and j is the weight position in each node. The parameters δ and α are used to compute an incremental change to each weight term via

$$\Delta w_{ij} = \varepsilon (\delta_{pi} \alpha_{pj}) + m w_{ij(old)}, \quad (6)$$

where ε is the learning rate, which determines the size of the weight adjustments during each training

iteration, and m is the momentum factor. It is applied to the weight change used in the previous training iteration, $w_{ij(old)}$. Both of these constant terms are specified at the start of the training cycle and determine the speed and stability of the network.

In the past few years, a large number and variety of commercial neural network programs have become available. These vary from user-friendly, step-by-step, easy to implement programs to general-purpose programs such as Matlab, in which the user must set up the neural network in the program environment and perform the necessary operations of training and testing of the network. In this case, the neural network component is a special module of the program (toolbox in Matlab) that can use the capabilities of the main program in order to input and manipulate data and produce graphs.

4. NETWORK PARAMETERS SELECTION

Although there are differences between traditional approaches and neural networks, both methods require preparing a model. The classical approach is based on the precise definition of the problem domain as well as the identification of a mathematical function or functions to describe it. However, it is very difficult to identify an accurate mathematical function when the system is nonlinear and there are parameters that vary with time. Neural networks are used to learn the behavior of the system and subsequently to simulate and predict this behavior. In defining the neural network model, first the process and the process control constraints have to be understood and identified. Then the model is defined and validated.

The neural network model and the architecture of a neural network determine how a network transforms its input into an output. This transformation is in fact a computation. The neural network architecture refers to the arrangement of neurons into layers and the connection patterns between layers, activation functions, and learning methods. Often, success depends on a clear understanding of the problem regardless of the network architecture. However, in determining which neural network architecture provides the best prediction, it is necessary to build a good model. It is essential to be able to identify the most important variables in a process and generate best-fit models. How to identify and define the best model is very controversial.

When building the neural network model, the process has to be identified with respect to the input and output variables that characterize the process. The inputs include measurements of the physical dimensions, measurements of the variables specific to the environment, and equipment-controlled variables modified by the operator. Variables that do not have any effect on the variation of the measured output are discarded. These are estimated by the contribution factors of the various input parameters. These factors indicate the contribution of each input parameter to the learning of the neural network and are usually estimated by the network, depending on the software employed.

The selection of training data is vital in the performance and convergence of the neural network model. An analysis of historical data for identification of variables that are important to the process is important. Plotting graphs to determine whether the charts of the various variables reflect what is known about the process from operating experience and for discovery of errors in data is very helpful.

The first step is to collect the required data and prepare them in a spreadsheet format with various columns representing the input and output parameters. If a large number of sequences/patterns are available in the input data file, and to avoid long training times, a smaller training file may be created to select the required parameters and the complete data set can be used for the final training.

Three types of data files are required: a training data file, a test data file, and a validation data file. The former and the latter should contain representative samples of all the cases that the network is required to handle, whereas the test file may contain approximately 10–20% of the cases contained in the training file.

The training of all patterns of a training data set is called an epoch. The training set has to be a representative collection of input/output examples. During training, the network is tested against the test file to determine its prediction accuracy, and training should be stopped when the mean average error remains unchanged for a number of epochs. This is done in order to avoid overtraining, in which case the network learns perfectly the training patterns but is unable to make predictions when an unknown training set is presented to it.

The input and output values need to be normalized. All input and output values are usually scaled individually such that the overall variance in the data set is maximized. This is necessary because it leads to faster learning. The scaling that may be used is either

in the range -1 to 1 or in the range 0 to 1 depending on the type of data and the activation function used.

When using a neural network for modeling and prediction, the following steps are crucial. First, a neural network needs to be built to model the behavior of the process. Second, based on the neural network model built, the output of the model is simulated using different scenarios. Third, the control variables are modified so as to control and optimize the output.

The basic operation that has to be followed to successfully handle a problem with neural networks is to select the appropriate architecture and the suitable learning rate, momentum, number of neurons in each hidden layer, and the activation functions. This is a laborious and time-consuming method, but as experience is gathered some parameters can be predicted easily, thus shortening tremendously the time required.

In backpropagation networks, the number of hidden neurons determines how well a problem can be learned. If too many are used, the network will tend to try to memorize the problem and thus not generalize well later. If too few are used, the network will generalize well but may not have enough “power” to learn the patterns well. Determining the correct number of hidden neurons is a matter of trial and error since there is no science to it.

5. MODELING OF ENERGY SYSTEMS USING NEURAL NETWORKS

Neural networks have been applied successfully in a number of applications. The following are some of the most important:

Function approximation: In this type of problem, mapping of a multiple input to a single output is established. Unlike most statistical techniques, this can be done with adaptive model-free estimation of parameters.

Pattern association and pattern recognition: Neural networks can be used effectively to solve difficult problems in this field, such as in sound, image, or video recognition. This task can even be performed without an a priori definition of the pattern in which the network learns to identify totally new patterns.

Associative memories: This is the problem of recalling a pattern when given only a subset clue. In these applications, the network structures used are

usually complicated, composed of many interacting dynamical neurons.

Generation of new meaningful patterns: This general field of application is relatively new. Some claims have been made that suitable neuronal structures can exhibit rudimentary elements of creativity.

Neural networks have been used by various researchers for the modeling of energy systems. This section presents various such applications in a thematic rather than a chronological or any other order.

5.1 Renewable Energy Systems

In the field of renewable energy, neural networks have been used for the modeling of various components of a concentrating solar collector and for the prediction of the mean monthly steam production of a solar system. They have also been used to model the starting up of a solar steam generator, which is a particularly difficult problem because the system is working in transient conditions. In all these models, experimental data were used to train the neural networks, which were used subsequently to provide predictions for unknown cases.

Neural networks have also been used for the modeling of solar domestic water heating systems and for long-term performance prediction of the systems. In these models, performance data for 30 systems, obtained by testing the systems with the procedure suggested in international standards, together with weather conditions and some physical characteristics of the systems such as the collector area, storage tank volume and heat loss coefficient, tank type (vertical or horizontal), and type of system (open or closed) were used to train the neural network. A large variety of weather conditions were used to enable the network to accept and handle unusual cases. The trained model could be used to provide the useful energy extracted and the increase in storage tank water temperature of other unknown systems using only their physical characteristics and typical weather conditions, which reduces the experimentation required.

Other applications of neural networks in this field include the identification of the time parameters of solar collectors, photovoltaic systems, and solar radiation and wind speed prediction, which are the driving sources of renewable energy systems.

5.2 Heating, Ventilating, and Air-Conditioning Systems

In the field of heating, ventilating, and air conditioning, neural networks have been used for the modeling

of houses and the estimation of the heating and cooling load, for the prediction of the air flow and pressure coefficients of naturally ventilated rooms, for the prediction of the energy consumption of passive solar buildings, and for the energy consumption optimization. The data required for the training of the neural networks were obtained from experimental measurements in actual buildings. Although most of the previously mentioned models were performed for a particular building construction, these can be developed to have a more general usefulness provided that the training databases are enriched with a variety of building constructions, orientations, and weather conditions. In such cases, neural networks need to be retrained to "learn" the new patterns.

Other applications in this field include the use of neural networks for the energy consumption prediction in commercial buildings and modeling of a room storage heater. In these models, the neural networks were used as an expert system.

Neural networks have also been used to control specific systems such as the temperature in a bus air-conditioning system according to the ambient temperature, number of passengers, and time of day. Neural networks can also be used for the modeling of refrigeration systems. In one such application, neural networks were used to model the amount of frost on the coil of a refrigerator, and based on the network results a demand defrost method was proposed.

5.3 Combustion and Incineration

In combustion, neural networks have been used mainly for the modeling and prediction of the combustion emission products. In particular, they have been used for the monitoring of combustion pollutant emissions, for the prediction of the gaseous emissions from a chain grate stocker boiler, and for the estimation of NO_x emissions in thermal power plants. They have also been used for the development of controllers for combustion systems. Applications include the predictive control of unstable combustion systems, process control for fluidized bed combustion, and optimal control of an industrial stoker-fired boiler.

For the latter application, data collected during 300 h of experiments on an actual 0.75-MW chain grate stocker boiler were used to train two neural networks. The control strategy, shown in Fig. 5, is to provide the minimum airflow required for combustion without incurring unacceptably high CO emissions and carbon in ash losses, thus improving the combustion efficiency. The system detects changes in load required and gives the matching coal feed and

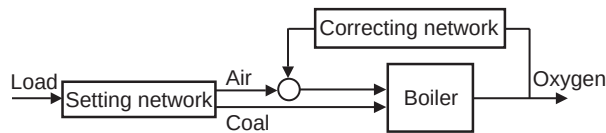


FIGURE 5 Chain grate stoker-fired boiler controller strategy.

airflow. Initially, the “setting network” acts as a look-up table to provide near optimum settings of coal feed and airflow for the desired load. After quasi-steady-state conditions are reached, the “corrective network” is activated to fine tune the airflow rate in order to optimize the combustion process on the grate.

Neural networks have also been used for the modeling of the combustion process in incineration plants, modeling of the temporal evolution of a reduced combustion chemical system, representation of the chemical reactions of combustion flames, and turbulent combustion modeling. In all the previously mentioned models, past history experimental data were used for the training of the neural networks to learn the required behavior and, according to the application, to provide the expected emissions so as to take corrective actions or to control a processes.

5.4 Internal Combustion Engines

Neural networks have been used to design engine controllers, such as a controller for idle speed regulation in internal combustion engines, a knock detector, spark advance control using cylinder pressure, and the control of combustion parameters such as spark ignition timing of petrol engines.

Neural networks have also been used for engine modeling, such as for the nonlinear time series analysis of combustion pressure data, modeling of the emissions and performance of a heavy-duty diesel engine, modeling of the volumetric efficiency and performance prediction of internal combustion engines, and modeling of a gas engine. In these models, experimental data were used for the training of the neural networks and the developed models are applicable to the particular type of engine used to collect the data. More general models can be produced by combining data from a number of engine capacity sizes and retraining the networks in order to produce general-purpose tools for the modeling of these engines.

Neural networks have also been used for the fault diagnosis of small to medium-sized diesel engines and marine diesel engines by providing an early warning of combustion-related faults. In a marine diesel engine model, the neural network was used to classify combustion quality on the basis of simulated data.

5.5 Energy Prediction and Forecasting

In the field of forecasting and prediction, neural networks have been used in power systems load forecasting and in electric power forecasting. They have also been used in tariff forecasting and energy management and in electric load forecasting in smaller capacity units such as supermarkets. The neural network models were trained with a sequence of past data recorded at the utilities and can provide either short-term or long-term predictions. Such systems can be used as effective management tools to regulate the operations of power utilities and other smaller enterprises.

6. CONCLUSIONS

From the applications described previously, it can be seen that neural networks have been applied for modeling and prediction in a wide range of fields of energy systems. Required for setting up such a neural network system are data that represent the past history and performance of the real system and a suitable selection of a neural network model. The selection of the model is usually done empirically after testing various alternative solutions with respect to network architecture and parameters. The performance of the developed neural network model is tested with unknown data collected from a real system that were not used for the training of the network.

The number of applications presented here is neither complete nor exhaustive but merely a sample that demonstrates the usefulness of artificial neural networks for modeling and prediction of energy systems. Artificial neural networks, like all other approximation techniques, have advantages and disadvantages. There are no rules regarding when this particular technique is more or less suitable for an application. Based on work published to date, neural networks offer an alternative modeling method that should not be underestimated.

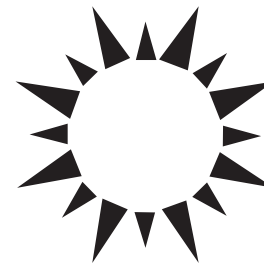
SEE ALSO THE FOLLOWING ARTICLES

Complex Systems and Energy • Computer Modeling of Renewable Power Systems • Fuzzy Logic Modeling of Energy Systems • System Dynamics and the Energy Industry

Further Reading

- Haykin, S. (1994). "Neural Networks: A Comprehensive Foundation." Macmillan, New York.
- Kalogirou, S. (2000). Applications of artificial neural networks for energy systems. *Appl. Energy J.* **67**, 17–35. [Special issue].
- Kalogirou, S. (2001). Artificial neural networks in renewable energy systems: A review. *Renewable Sustainable Energy Rev.* **5**, 373–401.
- Ripley, B. D. (1996). "Pattern Recognition and Neural Networks." Cambridge Univ. Press, Oxford.
- Rumelhart, D. E., and McClelland, J. L. (Eds.). (1986). Learning internal representations by error propagation. *In* "Parallel Distributed Processing: Explorations in the Microstructure of Cognition." MIT Press, Cambridge, MA.
- Warwick, K., Ekwue, A., and Aggarwal, R. (1988). "Artificial Intelligence Techniques in Power Systems." Institution of Electrical Engineers, London.

Nongovernmental Organizations (NGOs) and Energy



BONIZELLA BIAGINI

Global Environment Facility
Washington, D.C., United States

AMBUJ SAGAR

Harvard University
Cambridge, Massachusetts, United States

-
1. Introduction
 2. NGOs and Energy
 3. Conclusion

Glossary

nongovernmental organizations (NGOs) National and international nonprofit organizations as well as local membership organizations, often referred to as grassroots or community-based organizations, that represent the public interest (or the interests of groups of citizens).

The term nongovernmental organization (NGO), as used in this article, refers to national and international nonprofit organizations as well as local membership organizations that are often referred to as grassroots or community-based organizations. NGOs can basically be thought of as being a third sector—separate from the state and the market—that represents the public interest (or the interests of groups of citizens). In this context, NGOs fill gaps in services provided by governments and firms and also enhance the accountability and responsiveness to citizens of these two sectors.

1. INTRODUCTION

Organizational goals of nongovernmental organizations (NGOs) include empowering people socially, economically, or politically; serving underserved or neglected populations; and protecting the local or

global environment. To fulfill these objectives, NGOs are involved in a range of activities, including delivering welfare and development services, monitoring, critiquing and advocating policies, and providing information to citizens (and even to policymakers).

Although such citizens organizations have been a part of the landscape of societies for centuries, there has been a marked and rapid growth of this sector during recent years. The number of NGOs in most countries, as well as those operating transnationally, has increased substantially during the past couple of decades, as have the functions of these organizations. Consequently, these NGOs are playing an increasingly prominent role in many aspects of domestic affairs (within industrialized and developing countries) as well as in the international arena, resulting in what Edwards and Hulme referred to as the “rise and rise of NGOs” and what Salamon referred to as a “revolution.” Nowhere is this transformation more prominent than in the environment and development arena. Whereas fewer than 300 NGOs attended the Stockholm Conference on the Environment in 1972, approximately 1400 NGOs registered for the Earth Summit in Rio de Janeiro, Brazil, in 1992 and roughly 18,000 NGOs attended the parallel NGO forum. More recently, more than 8000 NGOs attended the World Summit on Sustainable Development (WSSD) in Johannesburg, South Africa, and approximately 30,000 NGOs participated in the related Civil Society Forum. By now, NGOs and their networks are central players on the landscape of environmental politics and action.

At the same time, the finances underlying NGO operations have also been bolstered markedly. For example, Edwards and Hulme estimated that in

1975 only 0.7% of the total development aid given by Organization for Economic Cooperation and Development (OECD) countries was channeled to NGOs, whereas by 1985 NGOs received 3.6% of development aid and by 1993–1994 they received 5.0%—with this last figure corresponding to roughly \$5.7 billion. Other sources of NGO funding could have accounted for another \$3 billion by the early 1990s. Of course, these numbers pertain only to funds channeled by aid and development agencies. NGOs in industrialized and developing countries also raise money from other sources, including membership fees, private foundations, individual donors, sales of publications (e.g., *State of the World* by the Worldwatch Institute), and technical services (e.g., organizing conferences and technical consulting). For example, Greenpeace raised more than four-fifths of its \$143 million income in 2000 through personal donations, and the U.S. NGO Natural Resources Defense Council (NRDC) has more than 500,000 dues-paying members (see subsections on Greenpeace and NRDC in Section 2.4).

Another indicator of the growing influence of NGOs is the increasing movement of their personnel into governmental positions. This is particularly true in Western European countries, such as Germany, where members of environmental organizations formed the Green Party that has participated in governing coalitions. There have also been notable examples in the United States under Presidents Clinton and Carter, where individuals from environmental NGOs were nominated for senior governmental positions. This movement is not only into the government, as there are many former government officials now working in NGOs.

Broadly speaking, NGO activities include the following:

- Identification and articulation of needs and concerns that are being overlooked, or that are not receiving sufficient attention, by the state and the market
- Research and analysis to examine specific issues as well as potential solutions
- Advocacy, information, and technical assistance provision (e.g., coalition building) to citizens, policymakers, politicians, media, NGOs, and other groups
- Agenda setting
- Policy development
- Policy change and alternative policy formulation
- Policy implementation
- Delivery and provision of services to citizens or groups of society
- Capacity building in other groups, individuals as well as state and market actors
- Monitoring and critiquing state and market actors and activities.

The scale of NGO operations can also vary from the local (as may be common for many grassroots organizations) to the national and global. Furthermore, to fulfill their function better, NGOs often establish a wide variety of formal and informal links with other NGOs, individuals, governments and intergovernmental organizations (IGOs), development agencies, corporations, and transnational issue networks. For example, NGO relationships with governments and IGOs can be such that the NGOs may assist these latter institutions, take on tasks that have been ignored by them, lobby them to do things differently, evaluate their performance, or even challenge them, and sometimes the same NGOs may exhibit different kinds of linkages on different issues. Another important facet has been the emergence of networks and coalitions of NGOs at the national and transnational levels that have been useful for mobilizing and leveraging resources of a number of groups with common agendas or ideologies.

1.1 Greenpeace

Greenpeace is an environmental nonprofit organization that was founded in 1971 and now has a presence in 41 countries across Europe, the Americas, Asia, and the Pacific. The Greenpeace organization consists of Greenpeace International (Stichting Greenpeace Council) in Amsterdam, Netherlands, and Greenpeace offices around the world. The organization has a large base of individual supporters; as of January 2002, 2.8 million people had taken out or renewed their financial memberships within the previous 18 months.

As with many other major environmental NGOs, Greenpeace is engaged in a number of campaigns, including climate change, through which it draws attention to major environmental problems and possible solutions to them. Given the intimate relationship between energy use and climate change as well as other environmental issues, many of Greenpeace's activities focus on energy issues and, in turn, the organization has had a significant impact on various components of the energy system.

Greenpeace is actively engaged in increasing the public profile of a range of energy-related issues,

including renewable energy, nuclear energy, oil exploration, and climate change. It prepares detailed reports, commissions studies, analyzes policies, and disseminates information to the public on such issues. It also campaigns against, debates, and engages with industry (as needed). For example, whereas Greenpeace has led campaigns to prevent oil exploration in remote areas by firms such as BP or to challenge Shell on Brent Spar (the oil platform in the North Sea), it has also worked with other firms to develop new refrigeration technology. It also seeks to influence governments in international forums, such as those on climate change, as well as in the domestic arena.

Greenpeace was instrumental in the development of Greenfreeze, the world's first climate and ozone safe technology. In 1992, it brought together scientists who had extensively researched the use of propane and butane as refrigerants with an East German company, DKK Scharfenstein (now Foron), to develop this technology. It also gathered tens of thousands of preliminary orders for this new refrigerator from environmentally conscious consumers in Germany. Greenfreeze refrigerators are now sold in Western Europe, China, and Latin America.

Greenpeace does not solicit or accept funding from governments, corporations, or political parties. It relies on the voluntary donations of individual supporters and on grant support from foundations. Greenpeace's net income in 2001 was 112 million Euros (nearly exclusively from grants and donations). It spent approximately 10 million Euros on its climate campaign that year.

1.2 Natural Resources Defense Council

NRDC was founded in 1970 and now has more than 500,000 members in the United States. The organization is active in a number of areas, including clean air, energy, and water; forests and wildlife; toxics; nuclear weapons and waste; and green cities. It uses scientific analysis, legal actions, public education, and targeted grassroots efforts to promote policy changes in all of these areas.

NRDC's efforts to ensure clean air include targeting local, state, and federal governments as well as using legal avenues to promote strict pollution standards and compliance with them. NRDC also uses market-based strategies to promote new, less-polluting technologies. It also aggressively promotes renewable energy sources, such as wind and sun, as well as energy efficiency improvements in appliances, business equipment, and buildings

through new and existing laws, better regulations, and financial incentives. Recently, NRDC has been paying particular attention to deregulation and restructuring of the electric industry to ensure incorporation of incentives for clean energy and efficiency.

NRDC's "earthsmartcars" campaign aims to put cleaner cars on the road by focusing on the demand side (i.e., consumer demand) as well as on the supply side (i.e., new technologies). It advocates for laws and regulations governing vehicle pollution as well as for promotion of public transportation and local planning to reduce vehicle use. NRDC has also been active in targeting diesel-fueled trucks and buses due to the particular health and environmental hazards posed by them. NRDC was instrumental in the efforts to phase out lead from gasoline, reduce acid rain, and help win adoption of national efficiency standards for consumer appliances, all of which have had a major impact on the energy system in the United States.

NRDC also has a major effort on climate change focusing on greenhouse gas (GHG) reduction domestically and internationally.

Nearly three-quarters of NRDC's 2001 operating income of \$40.6 million came from membership and contributions, with foundations accounting for most of the remainder (nearly \$9 million).

2. NGOs AND ENERGY

The main path through which NGOs shape the energy system is by influencing policy processes of governments and IGOs (e.g., the World Bank) that have an impact on the energy system as well as on the pace and direction of technological innovation. These include the following:

- National energy policy frameworks, including regulations, taxes, and publicly funded research
- Local and national environmental policies such as air quality regulation and GHG abatement policies
- Global environmental regimes such as the climate convention
- Contextual policies such as trade policy, export credit, and bilateral and multilateral aid that have indirect influences on global energy supply and demand

NGOs also directly participate in efforts to enhance and improve the delivery of energy services by working on relevant technologies and products as

well as on their deployment and diffusion in energy projects.

2.1 NGO Influence on National Energy Policies

Over the past few decades, NGOs have had a significant influence on shaping national or regional energy policies in most industrialized countries as well as in many developing countries. In industrialized countries, there is by now an established tradition of (mostly environmental) NGOs lobbying for domestic environmental regulations, and energy is often a priority on the agenda. In developing countries, an increasing number of NGOs have played a leading role in promoting sustainable energy technology transfer as well as the recognition of local knowledge and “home-grown” solutions in the South, critiquing existing national or local energy policies and proposing alternative policies, and participating in the design, implementation, and monitoring and evaluation of energy-related development projects. The pathways to influence have also been different in various countries. NGOs, although often with limited resources, are generally quite adroit at working within the institutional, policy, and cultural contexts within their countries and exploiting opportunities to shape energy policies through various mechanisms, as illustrated by the following examples from various countries.

2.1.1 United States

Some of the largest environmental NGOs, such as NRDC, were formed when citizens organized to fight power plants and other energy projects. Energy issues consistently appear high on the agenda for many of the largest NGOs. For example, the Sierra Club, one of the largest U.S. membership NGOs, has for years been a strong opponent of oil drilling in Alaska and a strong proponent of more stringent fuel economy standards.

The ability of NGOs to use courts as a means of influencing energy policy has been particularly important in the United States, where citizen suits to protect the environment were first permitted by judicial decisions and then enshrined in legislation such as the Clean Air Act. Environmental litigation has been used to challenge procedural deficiencies, such as lack of notice and opportunity to participate in decision making, as well as substantive outcomes, such as whether regulations are sufficiently protective of the environment consistent with legislative require-

ments. Litigation has also been an important educational tool in that it often generates publicity and public awareness. For example, environmental groups went to court in 2002 seeking to identify participants in meetings on energy policy with the vice president in an effort to show undue influence by industry.

An example of the high impact of environmental litigation in the United States on energy policy is a lawsuit filed by NRDC in 1983 to contest a decision by President Reagan to oppose the implementation of a law that required energy efficiency standards for refrigerators, air conditioners, and other appliances. NRDC was joined in the lawsuit by several states, including California and New York, as well as other NGOs, including a national consumer organization. The lawsuit was decided in favor of NRDC and resulted in negotiations between the plaintiffs (NGOs and state governments) and appliance manufacturers on national standards that were ultimately adopted through national legislation in 1987. It has been estimated that the standards will save U.S. consumers \$200 billion through reduced energy consumption and avoided investments in power plants. U.S. appliances, which had been among the least energy efficient in the world during the 1970s, are now among the best available.

Although U.S. NGOs receive financial support from members, resources provided by private foundations have been particularly helpful in many cases. This source of funding is generally not available in other industrialized countries or is restricted from supporting litigation. The impact of foundation support for NGO efforts related to energy policy was enhanced in the United States by the founding of the Energy Foundation in 1991. The Energy Foundation was created by the McArthur, Pew, and Rockefeller foundations to provide more strategic support for utility regulatory reform and other technical issues. (Since that time, it has also received support from other foundations.) This more coordinated funding facilitated a network of lawyers and activists, at the state level, who could develop a common strategy; share data, witnesses, and legal materials; and generally work much more effectively. The Energy Foundation was also able to identify gaps in NGO work and support the creation of new organizations such as the Regulatory Assistance Project (RAP) and the Renewable Energy Policy Project (REPP).

2.1.2 Europe

NGOs have influenced domestic energy policies in many Western European countries. In Italy,

Legambiente, a leading grassroots environmental NGO, has worked on energy issues in various spheres. Started with the intellectual support of a group of scientists, many of them associated with the University of Rome "La Sapienza," Legambiente was initially very active in the antinuclear campaign. In 1986, after the accident at Chernobyl, Legambiente and other Italian NGOs promoted a referendum that resulted in a nuclear moratorium. During the 1990s, Legambiente's scientific committee, as well as its legal office, supported the work of green parliamentarians who designed and proposed Laws 9 and 10, subsequently approved by the Italian Parliament in 1991. Laws 9 and 10 introduced the "avoided cost principle" for purchase of power from nonutility generators and created significant economic incentives to produce and distribute renewable energy.

In The Netherlands, the ratio of NGOs to population is the highest in the world. Dutch NGOs have engaged in a number of activities on domestic energy policy, including public awareness, financial and fiscal incentives, and demand-side management programs. For example, in 1997, the Dutch government, the World Wildlife Fund (WWF, a major international NGO), and energy distribution companies decided to coordinate their public information campaigns. In addition, the three stakeholders joined efforts in the "Dutch Green Electricity Project" aimed at introducing a new product: "green electricity." Consumers paid 20% more, and the energy utility, in partnership with WWF, guaranteed that that this amount of electricity came from wind, small-scale hydro, biomass, and solar sources.

In Central and Eastern Europe, energy and climate NGOs have been active since the early 1980s and have significantly increased in number and activities during subsequent decades. Terra Millennium III, engaged at both the domestic and international levels, successfully lobbied the Romanian government toward an early ratification of the Kyoto Protocol. Green Action in Croatia, the Energy Club in Hungary, the Energy Efficiency Program and the Renewable and Energy Savings Center in Czech Republic, and the Polish Foundation for Energy Efficiency in Poland are among many NGOs in the region working to promote energy efficiency, renewable energy, and GHG reduction through policy research, training and education, and joint venture development.

2.1.3 India

NGOs play a prominent role in public discourse on India's development activities, including numerous

aspects of the energy sector. For example, NGO opposition to the country's policy on independent power projects (IPPs), and on specific IPPs (including Enron's Dabhol project) for environmental, social, or economic reasons, has contributed to the government's reexamination of its power policies. The long and ongoing protest by the Narmada Bachao Andolan against a major dam (the Sardar Sarovar), citing reasons of insufficient attention to resettlement and rehabilitation issues as well as overall long-term effectiveness of the project, forced the World Bank to withdraw its support from the project and has rendered impossible a business-as-usual approach on hydroelectric development within the country. The International Rivers Network (IRN) has helped to support and lead a discussion on these kinds of issues in other countries as well (see subsection on IRN in Section 2.4).

Environmental NGOs have also influenced indirectly aspects of energy policy in India. For example, the Center for Science and Environment released a major study on urban air pollution to highlight the urgency of tackling transport-related pollution in Delhi and other major cities in the country. It followed this up with a strong and sustained public campaign that helped to catalyze the transition (with the support and intervention of a determined and activist judiciary) toward cleaner public transport in these cities.

2.1.4 NGOs and Nuclear Energy

No energy technology has been the subject of more NGO focus and influence than has nuclear power. The development of nuclear power has been heavily subsidized, and its further refinement and deployment is supported by a well-financed industry lobby. Nevertheless, since around 1970, nuclear energy's expansion has been blocked and even reversed (in the form of plant closures) in response to public pressure in Europe, the United States, and Japan. As noted in the World Energy Assessment, many official projections of nuclear power in 2020 envision no more, and perhaps less, total output than is the case today. NGO campaigns, coupled with the devastating accident at Chernobyl, have had a major and lasting influence on public attitudes toward nuclear power in Europe. NGOs have engaged in demonstrations (sometimes violent) at plant sites throughout Europe and the United States to highlight concerns about plant safety and waste disposal. They have also used legal means to challenge licensing and construction of plants and have prepared detailed analyses critical of the economics of nuclear power. The result has

been a significant contribution to public skepticism about the technology, ultimately leading to a refusal to accept new plants and, in some cases, pressure to cancel plants under construction or even those already operating.

A further indication of NGO influence is the widespread perception that the future revival of the nuclear industry depends as much on changes in public attitudes as on actual changes in plant operation and safety. As the World Energy Assessment states,

Decisions on future nuclear power will be made largely at the political level rather than on narrow technical and economic grounds. Gaining broad public support ... is not simply a matter of better educating the public on the issues, which is what many in the nuclear industry believe is needed most. The industry should also seek to better understand public concerns. The industry should also consider opening up the nuclear decision-making process to diverse interest groups, so that a well informed public could ensure that its concerns are addressed every step of the way.

2.2 Role of NGOs in Shaping International Environment/Development Policy Frameworks

A major role of NGOs in shaping energy policy is through their influence on international policy frameworks that have implications for energy systems. These include frameworks that are both legally binding, such as the UN Framework Convention on Climate Change (UNFCCC), and nonbinding, such as the Agenda 21 (the sustainable development action plan that includes significant sections about energy policies).

Energy plans and policies, previously seen as domestic issues, now have a transnational dimension with the clear recognition that the way in which energy is consumed and produced affects the regional and global environment through phenomena such as acid rain and greenhouse warming. Hence, agreements have been negotiated at the regional level, such as the Convention on Long-Range Transboundary Air Pollution (LRTAP), as well as at the global level, such as the UNFCCC.

NGOs have played a pivotal role in these negotiations and in the domestic implementation of related energy policies. At the global level, this has occurred thanks to the inclusive policy of the UN system, which allows nongovernmental groups to participate as observers in international forums. Since 1945, NGOs have been involved in consultation arrangements with the Economic and Social

Council (ECOSOC) of the United Nations. Notwithstanding many difficulties and constant debate for decades with some governments that were not in favor of public participation, NGOs have increasingly gained recognition, credibility, and consensus on their involvement in UN processes, to the extent that by the early 1990s ECOSOC already included approximately 1000 NGOs. The UN Conference on Environment and Development (UNCED), or Earth Summit, held in Rio de Janeiro in 1992 was an unprecedented event with respect to NGO participation in global environmental politics. A record number of NGOs participated in the UNCED process (more than 1420 NGOs were accredited) and had the opportunity to influence the outcome of the negotiations. The UNFCCC and Agenda 21 are, among the outcomes of the Rio summit, the most relevant with respect to energy. NGOs are now considered as a major force in environmental politics. At various levels, they have helped governments and other stakeholders to identify issues and set agendas, have provided policy analysis as well as normative and regulations criteria, and have carried out monitoring and implementation activities.

Over the years, the level of expertise of NGOs has been increasingly recognized, thanks also to credible and independent policy analyses and scientific reports. Many governments, especially in developing countries, now rely on NGOs' expert support to help negotiate global environmental treaties and to help translate such agreements into domestic (energy) policies.

Worldwide, most NGOs active in global environmental politics fall into one of three categories: international NGOs (e.g., Friends of the Earth, Greenpeace, WWF, Third World Network), national organizations and think tanks, or (more recently) small and local NGOs. The last ones are particularly helped by joining umbrella organizations that work on thematic issues such as the Climate Action Network (CAN) (see subsection on CAN in Section 2.4).

2.2.1 NGOs' Involvement in the UNFCCC

The UNFCCC was adopted by more than 160 countries at the Earth Summit in Rio de Janeiro, and this agreement forged the first international work plan to address climate change. The ultimate objective of the treaty is the "stabilization of greenhouse gas concentrations in the atmosphere at a level that would prevent dangerous anthropogenic interference with the climate system." In practical terms, this statement implies a significant drop in

GHG emissions that cannot occur without significant changes and reforms in energy policies and technologies worldwide, inevitably leading to complex and contentious negotiations among countries.

NGOs have interfaced with this negotiation process in a number of ways, including directly lobbying government delegates, presenting “position papers” to the climate convention secretariat and parties, making statements during the plenary and organizing thematic “side events,” using the media to exert public pressure on the delegates, and providing technical, policy, and scientific advice to the delegations (as appropriate). Although it is generally difficult to quantify the impact of NGOs, their relevant contributions to the debate, including policy analysis, position papers, fact sheets, scientific research, and economic analysis of mitigation options in the energy sector, among other inputs, have been recognized by the UNFCCC secretariat as well as by governments, which are gradually including a larger number of NGOs in their official delegations.

For example, in 1991, the Center for Science and Environment (CSE), an Indian NGO based in Delhi, published a report called “Global Warming in an Unequal World,” that expressed the frustration of both Southern governments and many NGOs about the lack of attention given to the equity issue in the climate deliberations and that forcefully injected the topic into the international discussions. CSE was also asked to advise the Indian government on climate change, and it continues to have a significant influence on the country’s climate policies. On the other end of the policy process, ENDA Energy, the energy branch of ENDA Tier Monde from Africa, is helping to implement the UNFCCC in African countries (see subsection on ENDA Energy in Section 2.4).

During the climate negotiations, NGOs have found numerous ways in which to exert influence. CAN NGOs publish a daily newspaper, *ECO*, that reports and comments on the details of the negotiations and is considered by many negotiators and observers to be a key source of information, policy analysis, and humor. CAN representatives also organize a daily press conference and distribute press releases commenting on the status of the climate negotiations. A Canada-based NGO, the International Institute of Sustainable Development (IISD), also publishes a more journalistic report on the negotiations, the *Earth Negotiations Bulletin*.

Over time, NGOs have gradually become more effective at networking by working together with their counterparts from various parts of the world.

Increased funding from foundations and from bilateral, multilateral, and private sources, combined with North–South joint fund-raising, has increased NGOs’ ability to travel and ensure a continued and consistent presence during more than a decade of climate negotiations.

2.2.2 NGOs and Agenda 21

The influence of NGOs was even more relevant in the Agenda 21 than in the climate convention due to its non-legally binding nature. Agenda 21 is intended to set out an international program of action for achieving sustainable development during the 21st century. Agenda 21 is a report of more than 500 pages and comprising 40 chapters. Energy issues are explicitly spelled out in many chapters, although there is no single energy overview or focus reflecting political differences on this topic. In the section on changing consumption patterns, NGOs provided language proposing to improve production efficiency to reduce energy and material use and to minimize the generation of waste concepts reflected in the final text of Agenda 21. The text of Agenda 21 is a particularly positive example of a collaborative approach and mutual tolerance among governments and civil society, represented mostly by NGOs.

Agenda 21 also explicitly recognizes, in its chapter 27, the role of NGOs in implementing the agenda:

Nongovernmental organizations play a vital role in shaping and implementing participatory democracy. Their credibility lies in the responsible and constructive role they play in society. Formal and informal organizations, as well as grassroots movements, should be recognized as partners in the implementation of the Agenda 21. ... Nongovernmental organizations ... possess well-established and diverse experience, expertise, and capacity in fields which will be of particular of particular importance to the implementation and review of environmentally sound and socially responsible sustainable development.

2.2.3 NGOs and Multilateral Development Banks

During the 1980s, NGOs launched a major campaign to reform the environmental policies of the multilateral development banks (MDBs). Such a targeted reform campaign on MDBs was based not only on the size of their loan portfolios but also on the fact that MDBs act as a catalyst for additional funding from the private sector and other banks.

Energy was a key issue in this campaign. The lending decisions of the World Bank and other MDBs often favored large-scale, capital-intensive, centralized projects (e.g., large dams, fossil fuel power plants) as opposed to more decentralized,

small-scale, renewable energy or energy efficiency technologies.

Because of the significant influence of the U.S. government on the World Bank and other MDBs, a coalition of U.S.-based NGOs launched a strategy based on lobbying the U.S. Congress to influence the MDBs' lending policies through an evaluation of their environmental practices and policies. The NGOs started their campaign by convincing members of the U.S. House and Senate and their staff to hold hearings on the environmental performance of the MDBs. The U.S.-based NGOs testified, together with colleagues from developing countries, on the environmental and social impacts of the projects, especially in the energy and agriculture sectors. Recommendations on new procedures for MDB lending decisions drafted jointly by NGOs and staff of the House Banking Subcommittee were proposed to, and accepted by, the Department of the Treasury.

During the late 1980s, the campaign expanded to include an increasing number of NGOs from developing countries, economies in transition, and Western Europe. Although the campaign had a tangible influence on environmental policies of the World Bank, within the energy sector it resisted shifting its lending policy from conventional energy sources (e.g., fossil fuels, large dams). In 1989, only 2% of the World Bank's energy portfolio included end-use energy efficiency components.

Many developing country NGOs gradually increased their participation in areas of action traditionally handled by Northern NGOs. Protest against international aid agencies has demonstrated to be a very successful means to raise media, public, and political attention. In the MDB campaign, monitoring and denouncing project activities in the field became a key element of the campaign, giving developing country NGOs a pivotal role to play. In some cases, thanks to the testimony of developing-country NGOs about negative local environmental and social impacts, Northern NGOs managed to stop or at least question the project funding.

The NGO initiative calling for a World Bank reform further expanded during the 1990s, especially in 1994 when Friends of the Earth International took the lead with the "50 Years Is Enough" campaign.

The relationship between MDBs and NGOs has also changed dramatically during recent years as a result of such campaigns. The World Bank and other aid agencies have included public participation and disclosure guidelines to give more access to information and participation at all stages of project design,

implementation, monitoring, and evaluation, and formalized consultations are now sometimes done on the Internet. The ongoing NGO campaign that embraces a number of MDBs' initiatives is called the "Campaign for the Reform of the World Bank," where development groups, such as Both Ends, have a leading coordinating role.

In the context of aid agencies, it is also relevant to mention the interaction between the Global Environmental Facility (GEF) and NGOs. GEF has allocated \$580 million in renewable energy during the past decade or so and has engaged in bilateral consultations with NGOs since its inception. GEF public participation policy consists of information dissemination, consultation, and stakeholder participation, including those potentially affected by GEF-financed projects. The GEF Small Grant Program also includes a project portfolio that involves NGOs as central proponents, designers, and implementing stakeholders.

2.2.4 NGOs and International Finance/Aid Flows

NGOs have recently started to examine the activities of export credit and investment insurance agencies that play a major role in financing infrastructure projects in developing countries. Many of these infrastructure projects are in the energy area and, hence, have important long-term climate as well as other environmental and social implications. Many export credit agencies (ECAs) have no formal policies to evaluate such implications of the projects they fund; hence, many NGOs are building on their past work on World Bank projects to direct attention to filling this gap.

Recent examples include advocacy (so far unsuccessful) in the United States to include renewable energy targets in legislation reauthorizing the Export-Import Bank's charter and budget, establishment by the Export-Import Bank of a renewable energy advisory committee (REAC) that was pushed by advocacy groups, and greater integration of the ECA issue into climate advocacy group efforts both within the context of climate negotiations and in the runup to the WSSD.

2.3 NGOs and Energy Service Delivery

Given the centrality of energy services in improving the human condition, numerous development NGOs have, not surprisingly, been involved in efforts to increase and improve the delivery of essential energy services to poor people in many developing countries.

Thus, the main thrusts of these activities have been on improving the availability and use of energy for basic human needs, such as lighting and cooking, as well as for enhancing health services and opportunities for basic economic development. Hence, NGO efforts have aimed to enhance, and make more reliable, supplies of electricity using renewable resources (biomass, microhydro, and solar photovoltaic [PV]) as well as to make available technologies for economically or socially productive uses of energy (e.g., biomass gasifiers for small-scale industrial applications, refrigeration for health clinics) and improved cooking stoves for reducing health-adverse impacts of traditional stoves.

The presence and sophistication of NGOs involved in such activities have grown significantly over the past few decades and have often compensated for the inadequacies of the public sector in many developing countries. As mentioned earlier, given the local nature of many NGOs, development agencies interested in improving energy service delivery have also recognized their value, as have many actors in the private sector. Thus, NGOs continue to play an increasing role in this arena.

NGO involvement in energy service delivery takes many forms, but these organizations have been particularly useful at overcoming some of the major barriers—technology, finance, information—that often impede efforts to increase access to energy in rural areas.

Effective implementation of small-scale energy projects in rural areas often hinges on the availability of appropriate energy technologies. Although specific components of these technologies may be available in the commercial markets, often the development of the overall system is dependent on local resources and constraints. This necessitates making careful choices about technology design, possibly carrying out modifications to existing designs or developing new designs or components altogether. NGOs can play a useful role in such technology modification or development processes.

This is well illustrated by the case of improved cooking stove programs where, to be successful, the designs of the new stoves have had to take into account the locally available energy sources, the preferences of the users in terms of stove characteristics (not only functional but sometimes also aesthetic), and the skills and materials available locally for stove manufacture. For example, the successful Kenyan *jiko* stove was based on a design from Thailand that was modified to make it useful

and appealing to local users. The training and demonstration work that underlay the development of the *jiko* was carried out with the assistance of the Kenya Energy and Environment Organization (KENGO), an indigenous NGO. Similarly, in the case of the improved cooking stove program in Sri Lanka, the Intermediate Technology Development Group (ITDG), a U.K.-based NGO that has been involved in energy technology projects for many decades (see subsection on ITDG in Section 2.4), played an important role in working with local firms to identify appropriate production materials and processes for indigenous production of cooking stoves. As another example, The Energy and Resources Institute (TERI, formerly known as Tata Energy Research Institute), a major NGO in India (see subsection on TERI in Section 2.4), has been involved in the development of biomass-gasifier-based technologies for thermally productive uses such as silk dyeing and cardamom drying where the design of the system, to be successful, is very much dependent on the specific needs of the small-scale enterprises that are its eventual users.

The success of the Grameen Bank in Bangladesh highlighted both the importance of microcredit in rural areas and the feasibility of NGOs and community organizations in bridging the financing gap for poor people who are deemed credit risks by banks and other traditional lending agencies. The resulting microcredit revolution has had an impact on the energy sector also where rural users were often unable to purchase energy systems for lack of financing options. NGOs, often with the help of development agencies, have begun to play a role by helping to provide microfinance to individual users or local entrepreneurs for promoting energy access. For example, Grameen Shakti (GS), a subsidiary of the Grameen Bank, has been quite successful at promoting energy service delivery through the provision of microcredit and other services (see subsection on GS in Section 2.4).

NGOs also provide other kinds of assistance to entrepreneurs and other actors in the energy service delivery arena by providing training and building capacity in technical as well as business issues. They can also play a critical role through public education and dissemination of information about energy options. In some cases, NGOs can be involved in the actual delivery of energy services. And in some of these cases, they may have functions similar to those of private energy service companies (ESCOs), although presumably maintaining their nonprofit character.

2.4 Illustrative NGOs

2.4.1 *The Climate Action Network*

CAN is a global network of more than 330 environment and development NGOs from some 80 countries working to promote equitable measures to limit human-induced climate change to ecologically sustainable levels. CAN is the largest recognized environmental nongovernmental network through which environmental groups coordinate on climate and energy policy at the national and international levels. CAN members work to achieve this goal through the coordination of information exchange and NGO strategy on international, regional, and national climate issues. CAN has seven regional coordinating offices that coordinate these efforts in Africa, Central and Eastern Europe, Europe, Latin America, North America, South Asia, and Southeast Asia. Diverse environmental organizations from around the globe, ranging from large international groups, such as WWF, Greenpeace, and Friends of the Earth, to small local groups in developing countries, such as Terre Vivante in Mauritania and the Green Coalition in The Philippines, work collaboratively within CAN. Its members place a high priority on both a healthy environment and development that “meets the needs of the present without compromising the ability of future generations to meet their own needs.” CAN’s vision is to protect the atmosphere while allowing for sustainable and equitable development worldwide.

Established in March 1989, CAN has been actively monitoring and seeking to influence the climate negotiations as well as climate-related policies and measures at the national and international levels. Member organizations include noted experts on issues of climate change science and policy. CAN is the recognized umbrella NGO in the international negotiations through which environmental groups work. CAN’s credibility has been built over years of involvement at both the international and national levels. The organization’s flexible structure provides an NGO forum to get together and share ideas, advocacy strategies, and information on climate change. Membership is open to nonprofit public interest organizations that are active on the global warming issue and that subscribe to the mission of CAN.

CAN’s major objectives include assisting the global community in reaching the goals of the climate convention, promoting awareness and capacity building of governments and citizens worldwide on the climate issue, developing an objective understanding

of climate change and its causes, sharing this information with all people (especially all members), and organizing, supporting, inspiring, and coordinating its members to take effective action on climate change.

2.4.2 *ENDA Energy*

ENDA Energy is a branch of the organization ENDA Tiers Monde (Environnement et Développement du tiers-monde), an association of autonomous entities coordinated by an executive secretariat, and collaborates with grassroots groups in search of alternative development models on the basis of the experience, expectations, and objectives of marginalized peoples.

ENDA Energy’s work focuses on energy use and management in the African context, with an emphasis on the linkages between energy and development. ENDA Energy is also helping with the implementation in Africa of the United Nations Conventions on Climate Change and Desertification, with both areas being intimately related to the energy issue in Africa. It is also involved in the development of alternative energy technology, collecting information on energy, and implementing projects on both local and regional levels. ENDA Energy often works in partnership with other groups and through networks and jointly led projects.

ENDA Energy’s team has approximately 10 permanent members, engineers, economists, and environmentalists focusing on the knowledge and planning of energy systems, the promotion of alternative energy systems suitable for the African context, and the broader challenge of integrating energy with environment for sustainable development.

Although the program originally focused mainly on energy use by grassroots communities, over time it has become involved in maintaining databases, undertaking direct fieldwork and training, and lobbying African policymakers. ENDA Energy has also helped to set up an African network, RABEDE (African Network on Bioresources–Energy–Development–Environment), which circulates information, advice, and experience on these topics through a variety of mechanisms, including the *Bulletin Africain*, the journal of the RABEDE network.

2.4.3 *Grameen Shakti*

GS is a subsidiary of Bangladesh’s Grameen Bank, a member-owned organization founded in 1976 that provides microcredit to its members. GS was founded in 1996, when only 16% of the 21 million households were covered by the national electric

grid, to help provide renewables-based energy to rural households in Bangladesh.

By mid-2001, GS had installed more than 5800 PV home systems, with an installed capacity of 290 kWp, and planned to install 8000 systems within the subsequent 3 years. GS provides soft financing for customers as well as maintenance of its system for a year after sales. GS provides training to customers so that they can take care of minor problems themselves, and its intention is bring all customers under this program. GS also provides training to develop skilled technician-cum-retailers in the rural areas who will be able to install solar systems, provide accessories and after-sales services to PV systems buyers, and educate the rural people in renewable energy technologies. GS has trained more than 300 PV technicians under this program. GS also engages in promotion and demonstration efforts to build awareness of solar PV technology and also encourages customers to develop income generation schemes.

Under its wind program, GS is developing micro-enterprise zones using wind-based electricity generation systems; it has installed at least six turbines so far. Under its biogas program, GS is promoting the use of biodigesters to produce clean-burning biogas from cow dung and other waste biomass. GS has installed 30 biogas plants in northern Bangladesh and also provided training to customers for producing biogas from their own resources.

GS also has a research program that explores ways in which to develop appropriate renewable energy technologies (RETs) and their uses, develop methods to popularize renewable energy systems and make them accessible to large number of people, innovate financial services to facilitate rapid expansion of RETs, and develop and fabricate solar home system accessories and components to reduce total system cost.

A number of organizations have provided financial support to GS, including the International Finance Corporation, the U.S. Agency for International Development, the Swedish International Development Agency, Grameen Trust, and Grameen Fund.

2.4.4 International Rivers Network

IRN was established in 1985 in the United States as a nonprofit, all-volunteer organization of activists experienced in fighting economically, environmentally, and socially unsound river intervention projects. IRN collaborates and coordinates with local river activists worldwide. It sees itself as part of a

larger citizen-based movement working on these issues and believes that its legitimacy is to be derived from the respect of individuals groups all over the world with whom it works.

IRN has a two-pronged approach that combines work on changing global policies with campaigning on specific key projects around the world. Through this approach, its campaigns for policy change are informed by specific project examples while also recognizing that a project-only approach does not address the root causes and policies that underlie unsound river interventions.

IRN works with groups from around the world on cooperative campaigns for community-based river development. This has included campaigns on the Sardar Sarovar project in India, the Three Gorges project in China, and the Lesotho Highlands Water project.

IRN also undertakes in-depth research and provides project critiques, analyses of alternatives, and activist briefings. It monitors and critiques the policies of financial institutions, including the World Bank, and provides analyses and recommendations for reforming their practices.

IRN helped to establish and continues to coordinate the International Committee on Dams, Rivers, and People (ICDRP), an informal network of NGOs and peoples movements from 13 countries. ICDRP monitored the World Commission on Dams (WCD) and facilitated civil society input into the process, and it continues to advocate for the implementation of the WCD guidelines. IRN has also supported the formation of regional networks to promote progressive change in water and energy policies.

IRN's support and revenue for the year ended December 2000 amounted to \$1.2 million, of which nearly 80% came from grants and approximately 14% came from contributions. Nearly 70% of its expenses were directed toward campaigns, and just over 7% of its expenses were directed toward information and media.

2.4.5 The Intermediate Technology Development Group

ITDG, an NGO registered in the United Kingdom, was founded in 1966 by E. F. Schumacher. It has a presence in Latin America, East Africa, Southern Africa, and South Asia and focuses particularly on Peru, Kenya, Sudan, Zimbabwe, Sri Lanka, Bangladesh, and Nepal.

In these countries, ITDG works with poor communities to develop and disseminate appropriate technologies in a number of sectors, including energy,

transport, and small enterprise development. In the energy area, ITDG's efforts focus on increasing poor people's access to energy technology options. It has done this through the development and commercialization of low-cost cooking stoves as well as small-scale, off-grid, sustainable energy supply technologies such as microhydro plants, small-scale wind generators, affordable solar lanterns, and biogas plants.

As an example, in Peru, ITDG has been promoting the dissemination of microhydro energy for the past 20 years. Its integrated approach has included technology development, training, pilot projects, research on institutional issues, and advocacy work. This approach allowed ITDG to overcome the main barriers to rural electrification in Peru, namely, lack of appropriate technology, financial mechanisms, local capacity, and legal frameworks. Through a program with the Inter-American Development Bank (IDB), ITDG has implemented a small hydro revolving fund consisting of soft loans and technical assistance for isolated communities. Between 1994 and 2000, there were 21 small hydro schemes (more than 1 MW total) benefiting 15,000 people with an allocation of \$700,000 that, in turn, leveraged \$2.5 million from local, regional, and central governments as well as other NGOs. ITDG also helped to open a training center to build local capacity and actively disseminated the results of its effort. This program won the CTI World Climate Technology Award in 2000.

Lessons from ITDG's grassroots experience are spread through consultancy services, publishing activities, education, policy, research, and an international technical inquiries service. Thus, dissemination is an explicit part of ITDG's focus.

ITDG's income for the financial year ended March 2000 was £12.9 million (U.S. \$20.8 million at then prevailing exchange rates), of which £3.6 million was from donations by individuals, companies, and trusts; £6.5 million from grants from the U.K. government, multilateral development agencies, NGOs, companies, and trusts; and the remainder from its in-house consulting, publishing, and manufacturing activities. More than half of ITDG's expenditures were devoted to its technology programs.

2.4.6 The Energy and Resources Institute

TERI was established in 1974. Although during the initial period its focus was mainly on documentation and information dissemination activities, research activities in the fields of energy, environment, and sustainable development were initiated toward the end of 1982. With a staff of some 500 drawn from multidisciplinary and highly specialized fields, TERI

is the largest developing country institution focusing on issues relating to energy and environment and is involved in an enormous range of activities in these areas. Not only does TERI play a prominent role in energy, environment, and other related issues in India, it also has an international presence through its international offices and affiliate institutes.

TERI engages in research and analysis, technology development, consultancy, training and capacity building, and information collection and dissemination. It also organizes numerous seminars and workshops that bring together various stakeholders and experts on assorted topics. It also set up the TERI Institute of Advanced Studies, a degree-granting institution focusing on energy and environment technology, regulation, and governance, among other things.

TERI is also involved in analyzing policy issues on various aspects relating to energy such as global environmental issues, oil and gas resources, urban and transport systems, and economic modeling and statistical analyses.

Through its energy-environment technology division, TERI is involved in the development and dissemination of energy technologies as well as efforts to promote energy conservation, biomass energy technologies, rural energy (development and implementation of renewable energy-based programs as well as building capacity for doing so), and chemical and hydrogen energy (with a particular focus on fuel cells).

TERI's revenues have grown rapidly during the past decade or so. During the year 2000–2001, it received approximately Rs. 170 million (~U.S. \$3.7 million at then prevailing exchange rates) from international sources and Rs. 110 million (~U.S. \$2.4 million) from domestic sources.

3. CONCLUSION

The energy sector is often a high priority on NGOs' agendas due to the increasing demand for energy services in both developed and developing countries and its high environmental and social impacts. Consequently, NGOs from many developed and developing countries have played a substantial and generally effective role in shaping international and domestic energy agendas.

Importantly, NGOs' influence on the energy system comes through their engagement in a range of activities—from grassroots activism, to policy research, to technology development and dissemination.

TABLE I

Some Examples of NGOs That Have Had an Influence on the Energy System and/or Its Potential Trajectory within Their Home Countries or Globally and Their Main Modes of Influence

	CAN	ENDA	GS	Greenpeace	IRN	ITDG	NRDC	TERI
Research	X	X	X	X	X	X	X	X
Advocacy	X			X	X		X	
Agenda setting	X	X		X			X	X
Policy development		X						X
Policy reorientation/alternative policy formulation	X			X			X	
Policy implementation		X						
Service delivery			X			X		X
Information provision		X		X	X		X	
Technical assistance			X			X		X
Capacity building		X			X	X		
Watchdog	X			X	X		X	

Through their bottom-up approach, NGOs have shown increasing expertise, technical knowledge, and capability to monitor the implementation of energy policies, highlight problems, and suggest solutions to governments, UN agencies, multilateral aid agencies, and other energy stakeholders. And on the other end of the spectrum, many NGOs have helped local and national governments, as well as other organizations, to design improved energy policies.

Although most NGOs are involved in a number of activities at the same time (as Table I illustrates), the emphasis of their efforts is generally of a limited set of core activities that are in line with the organizations' foci. However, there are significant synergies and complementarities between the efforts of NGOs that aim to influence the energy sector within their home countries and those that aim to do so internationally. Although NGOs may collaborate and network with other like-minded NGOs in the North and South (e.g., in the MDB reform campaign), they also form partnerships with government agencies, donor organizations, and academic organizations.

Given the range of energy-related activities in which NGOs are involved, their creative strategies for influencing the energy system, and their responsiveness to the needs of citizens and marginalized stakeholders, these organizations occupy an important niche in the energy field. Through their advocating, opposing, negotiating, and consulting activities, NGOs have contributed to strengthening the links among the general public and governments,

aid agencies, and other key stakeholders that act as decision makers in the energy sector and to catalyzing change. These organizations, in turn, have begun to realize the value of NGOs as allies in the design, implementation, monitoring, and evaluation of energy projects. If anything, over time, the roles played by NGOs in contributing to an evolution toward an improved, effective, and sustainable energy system will only increase.

SEE ALSO THE FOLLOWING ARTICLES

European Union Energy Policy • Global Energy Use: Status and Trends • International Energy Law and Policy • Labels and Standards for Energy • Renewable Energy Policies and Barriers • Sustainable Development: Basic Concepts and Application to Energy • United Nations Energy Agreements • World Environmental Summits: The Role of Energy

Further Reading

- Anheier, H., Glasius, M., and Kaldor, M. (eds.). (2001). "Global Civil Society 2001." Oxford University Press, Oxford, UK.
- Edwards, M., and Hulme, D. (1992). "Making a Difference: NGOs and Development in a Changing World." Earthscan, London.
- Lin, G. (2000). Energy development and environmental NGOs: The Asian perspective. In "The Global Environment in the Twenty-first Century: Prospects for International Cooperation" (Pamela S. Chasek, Ed.). United Nations University Press, New York.

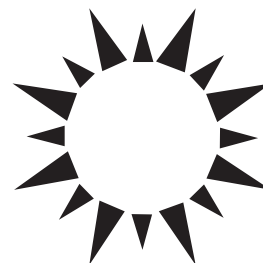
- Hurrell, A., and Kingsbury, B. (Eds.). (1992). "The International Politics of the Environment: Actors, Interests, and Institutions." Oxford University Press, Oxford, UK.
- Keck, M. E., and Sikkink, K. (1998). "Activists beyond Borders: Advocacy Networks in International Politics." Cornell University Press, Ithaca, NY.
- Princen, T., and Finger, M. (1994). "Environmental NGOs in World Politics." Routledge, London.
- Salamon, L. M. (1994). The rise of the nonprofit sector. *Foreign Affairs* 73(4), 109–122.
- Smillie, I., Helmich, H., German, T., and Randel, J. (Eds.). (1999). "NGOs and Governments: Stakeholders for Development." Earthscan, London.
- Willetts, P. (Ed.). (1996). "The Conscience of the World: The Influence of Nongovernmental Organizations at the United Nations." Brookings Institution, Washington, DC.

Nuclear Engineering

JOSÉ N. REYES, JR. and JOHN B. KING, JR.

Oregon State University

Corvallis, Oregon, United States



-
1. Nuclear Power and World Electricity Generation
 2. Some Basics of Nuclear Science
 3. Radiation Protection
 4. Design of Nuclear Power Reactors
 5. Commercial Nuclear Power Plants
 6. Nuclear Power Plant Safety Systems
 7. Next-Generation Nuclear Power Plants
 8. Nuclear Waste Disposal
 9. Conclusion

Glossary

boiling water reactor A direct-cycle power system that produces steam for a turbine-generator set by boiling water in a nuclear reactor core.

breeder reactor A nuclear reactor that breeds fissile fuel through neutron absorption in nonfissile isotopes.

containment building A large concrete- and steel-reinforced building that houses a commercial nuclear steam supply system.

passive safety systems A system that provides nuclear core or containment cooling by natural processes such as buoyancy, condensation, and evaporation or by a system of prepressurized tanks filled with coolant.

pressurized water reactor A power system that produces steam for a turbine-generator set by transferring heat from a nuclear core filled with high-pressure, subcooled water to an intermediate heat exchanger, known as a steam generator.

The methods developed in 1942 by Enrico Fermi to design, construct, operate, and control the first nuclear reactor formed the basis of a new discipline in engineering, nuclear engineering. It is the purpose of this article to introduce the reader to this exciting field. Today, nuclear engineers are active throughout the world, designing a new generation of nuclear reactors that can be safely used to produce electricity or medical isotopes or even to provide the power

needed for deep-space exploration. However, the guiding principle for the nuclear engineer has remained the same, to find creative ways of safely harnessing and using the energy of the atom for the benefit of mankind.

1. NUCLEAR POWER AND WORLD ELECTRICITY GENERATION

I am going to tantalize you with a vision of vast supplies of energy far surpassing the wildest desires of the engineer-resources so illimitable that the idea of fuel economy is not to be thought of.

—Sir Arthur Eddington, Berlin, 1930

Sir Arthur Eddington's general address on subatomic energy at the 1930 World Power Conference in Berlin stirred the imagination of every scientist and engineer present. The challenge was clear: find a practical means of accessing, controlling, and using the enormous energy locked in the atom as predicted by Einstein's remarkable mass-energy relation, $E=mc^2$. On December 2, 1942, Enrico Fermi transformed Eddington's visionary challenge into reality by producing the world's first controlled, self-sustaining nuclear reactor, Chicago Pile 1. Six decades later, nuclear energy now produces 16% of the world's electrical power. As shown in Fig. 1, this amount compares to 40% for coal, 15% for natural gas, 10% for oil, and 19% for renewables such as hydro, solar, wind, and geothermal power. Although the percentages for nuclear power are relatively high in countries such as France (78%), Japan (35%), Germany (32%) and the United Kingdom (27%), the United States and Canadian nuclear shares are only 20% and 15%, respectively. Although the percentage is somewhat lower in the United States, it actually boasts the largest number of nuclear power reactors in the world. That is, 104 of the 441 nuclear power reactors in the world are operating in the United

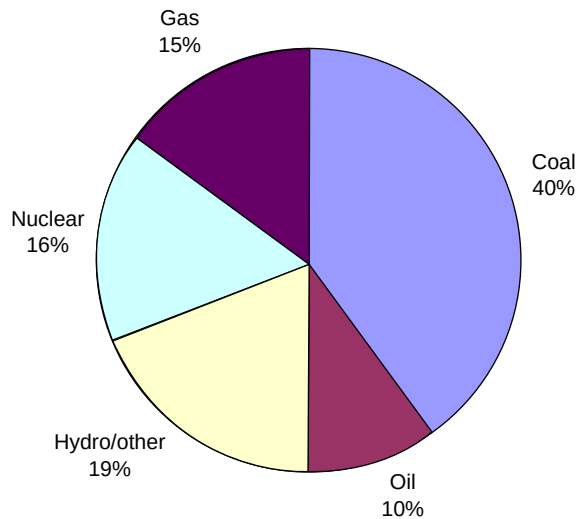


FIGURE 1 World electricity generation (statistics from the World Nuclear Association, World Energy Needs, and Nuclear Power, July 2002).

States. These reactors have a combined power output of 98.2 gigawatts (GW) of electricity, more than any other nation in the world.

2. SOME BASICS OF NUCLEAR SCIENCE

Before discussing nuclear engineering and the technology it has created, it is necessary to explain some basics of nuclear science. To this end, an atom is composed of negatively charged electrons that orbit around a positively charged nucleus. The nucleus, in turn, consists of positively charged protons and electrically neutral neutrons, both of which have masses approximately 1800 times greater than that of an electron. Thus, the preponderance of the atomic mass lies in the nucleus, which forms the tiny, central core of the atom.

Because the atom is electrically neutral, the number of electrons is identical to the number of protons, but the number of neutrons can and does vary. However, for any given chemical element, the number of protons is constant and serves to identify the element in question. For instance, all hydrogen atoms have 1 proton; all helium atoms have 2 protons, and all uranium atoms have 92 protons. Because the number of protons is uniquely linked to each element, this quantity is referred to as the atomic number Z . As already mentioned, however, for any given chemical element (i.e., for constant Z), the number of neutrons may vary. Thus, whereas all hydrogen atoms have a single proton and a single

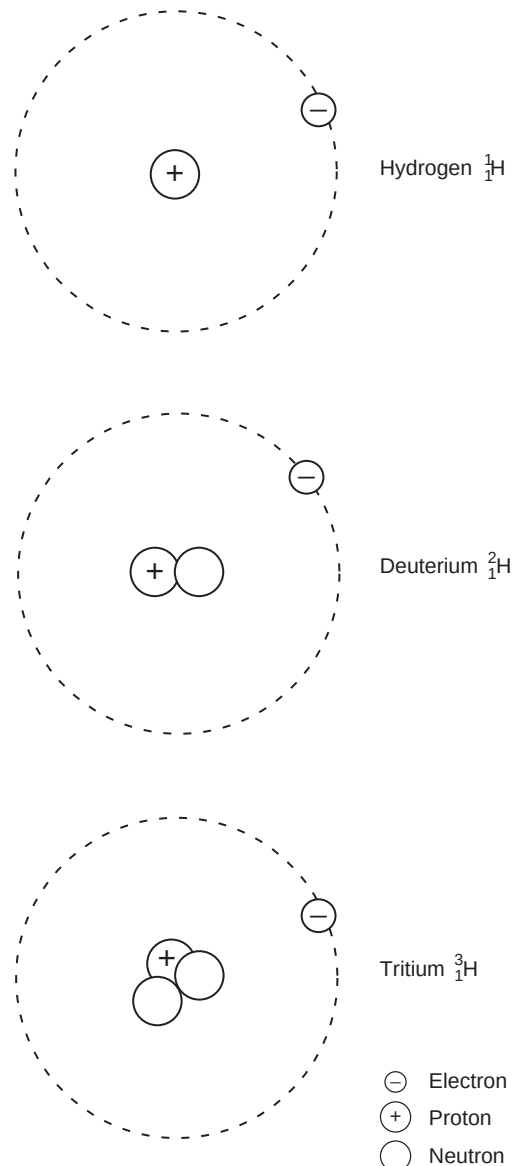


FIGURE 2 Isotopes of hydrogen.

electron, the number of neutrons varies from 0 to 2, as shown in Fig. 2. Given this fact, to uniquely specify a specific nucleus, it is necessary to specify both the number of protons (and hence the element) and the number of neutrons. Alternately, because the nuclear mass number A is specified by the total number of protons and neutrons, the nucleus can be uniquely specified by the combination of the atomic number Z and the mass number A . For instance, because all hydrogen atoms have a single proton ($Z=1$), but can have 0, 1, or 2 neutrons, the mass number A will assume values of 1, 2, or 3, respectively. Accordingly, using the Z value for the subscript and the A value for the superscript, the

three types of hydrogen nuclei can be represented as ${}^1_1\text{H}$, ${}^2_1\text{H}$, and ${}^3_1\text{H}$. Moreover, because the Z value is implicit in the chemical symbol H , these nuclei may be represented more compactly as ${}^1\text{H}$, ${}^2\text{H}$, and ${}^3\text{H}$, known as hydrogen-1, hydrogen-2 (or deuterium), and hydrogen-3 (or tritium), respectively. Finally, because these various nuclei all have the same Z value, they are all forms of hydrogen and will therefore occupy the same place in a periodic table of the elements. For this reason, they are referred to as isotopes, a word derived from the Greek terms *isos* (same or equal) and *topos* (place), which literally means “same place.”

If all the isotopes of the various elements are plotted on a graph of N (the number of neutrons) vs. Z , a chart of the nuclides is obtained, a portion of which is shown in Fig. 3. The chart identifies the stable isotopes and lists essential data for each radioisotope, such as its decay constant, the type of radiation emitted, and the energy of the emitted particle. For nuclear engineers and physicists, the chart of the nuclides provides a direct means of visualizing nuclear reactions and therefore functions as a nuclear analogue of the chemists’ periodic table. The known nuclides fall within a narrow band, which becomes more horizontal as N increases. Accordingly, only certain combinations of N and Z are possible, and the ratio of N to Z increases for the larger nuclei. The reason for this latter effect is that the number of protons must remain sufficiently small

relative to the total number of nuclear particles, to prevent the electrostatic repulsion from exceeding the nuclear binding force.

As mentioned previously, the mass of the atom is approximately equal to the nuclear mass due to the relative lightness of the electrons. However, if the nuclear mass is measured, it is found that the actual nuclear mass is less than the total mass of its individual particles. The reason for this mass defect is that, in forming a nucleus, the particles achieve a lower energy state together than they would separately. The absolute value of this energy difference is called the binding energy of the nucleus and it is related to the mass defect through Einstein’s famous formula: $E = mc^2$ (or in this case, $\Delta E = \Delta mc^2$).

To put the nuclear binding energy on a specific particle basis, the binding energy of every isotope may be obtained and divided by its mass number A to obtain the binding energy per nucleon (i.e., per nuclear particle), as shown in Fig. 4. From Fig. 4 it is seen that the binding energy per nucleon is maximized for nuclei around $A = 60$. Thus, for light nuclei, energy can be liberated by bringing the nuclei together. On the other hand, for heavy nuclei, energy can be released by splitting them apart. The former process, known as fusion, is operative in the sun and in the explosion of hydrogen bombs, but cannot yet be controlled and sustained at the moderate rates needed for power generation. The second process, by contrast, known as fission, was utilized in the first

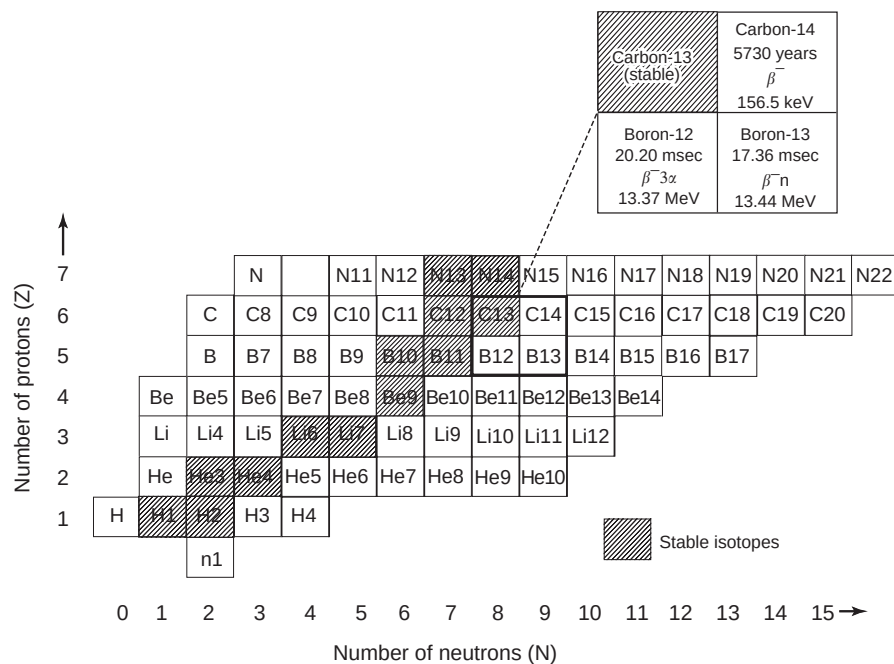


FIGURE 3 Chart of the nuclides from hydrogen to carbon.

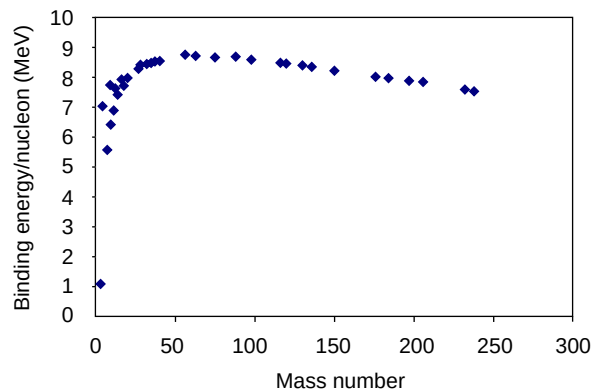


FIGURE 4 Binding energy of isotopes.

atomic bombs. Unlike fusion, fission can be controlled at sustained, moderate rates and is therefore currently utilized in nuclear power plants. For this reason, the subsequent discussion is restricted to fission technology.

With respect to fission, all nuclei are theoretically fissionable because they can be made to undergo fission by high-energy neutrons. However, some nuclei are said to be fissile, because they will undergo fission through the absorption of even slow-moving neutrons. Such slow neutrons are referred to as thermal neutrons because they have reached thermal equilibrium with their surroundings and are therefore moving as slowly as the local temperature will allow. The kinetic energy of particles is typically expressed in units of electron volts (eV), where 1 eV is the energy obtained in passing an electron or a proton through a potential difference of 1 V. Natural uranium is composed of 0.7% uranium-235, which is fissile, and 99.3% uranium-238, which is merely fissionable. Given this low percentage of the fissile isotope, natural uranium must be enriched to approximately 3% uranium-235 to serve as fuel in most reactor types. However, although the uranium-238 is not fissile, it may be transformed into fissile plutonium-239 by the absorption of a neutron. Thus, uranium-238 is said to be fertile because it can be used to breed fissile material. Given this fact, both isotopes of uranium may be utilized through special reactor designs known as breeder reactors (see later).

In addition to man-made transformations, some nuclei disintegrate spontaneously through a process known as radioactivity. As a result of radioactive decay, such nuclei emit three main types of radiation, denoted by the Greek letters α , β , and γ . In α decay, a nucleus emits an α particle, which consists of 2 protons and 2 neutrons and is identical to a helium-4 nucleus. As a result of α decay, the nuclear charge

TABLE I

Biological Quality Factors for Various Types of Radiation

Type of radiation	Biological quality factor
X rays and γ rays	1
β particles	1–1.7
Neutrons	2–11
α particles	10
Heavy nuclei	20

and mass decrease by 2 and 4 units, respectively. In β decay, the nucleus emits either an electron (β^-) or a positron (β^+). In β^- decay, a neutron transforms itself into a proton by emitting an electron and an antineutrino. In β^+ decay, a proton becomes a neutron by emitting a positron and a neutrino. Thus, in β decay, the nuclear charge is either increased (β^-) or reduced (β^+), but the nuclear mass remains constant. By contrast, neither the charge nor the mass of the nucleus is changed in γ decay. Rather, the nucleus gives off a high-energy photon (i.e., electromagnetic radiation) known as a γ ray. Finally, it should be mentioned that some nuclei decay by neutron emission, although this form of decay is least common. Obviously, in neutron emission, the nuclear mass is reduced without affecting the nuclear charge. As a result, the nucleus changes to a different isotope of the same element.

With respect to radiation, a major concern is its biological impact. In this regard, the effects are found to depend on both the energy and the type of radiation absorbed. With regard to energy, the unit of absorbed dose is the “radiation absorbed dose” (rad), which is equivalent to the absorption of 0.01 joule/kilogram (J/kg) of material. However, beyond the quantity of energy absorbed, the biological effect also depends on the quality of the energy. For instance, due to their high penetrating power, γ rays deposit less energy per unit length and therefore have a smaller linear energy transfer (LET). By contrast, α particles have far less penetrating power and therefore a higher LET. Given this fact, an absorbed dose of α particles may be more dangerous than the same dose of γ rays because the energy deposition is more localized. To account for this difference, quality factors for various types of radiation have been devised, as shown in Table I. Multiplying the absorbed dose by the relevant quality factor yields the dose equivalent in “roentgen equivalent man” (rem), a measure of biological impact. Thus, based on Table I, an absorbed α dose of 1 rad would give a dose equivalent of 10 rem. The need to gain a greater

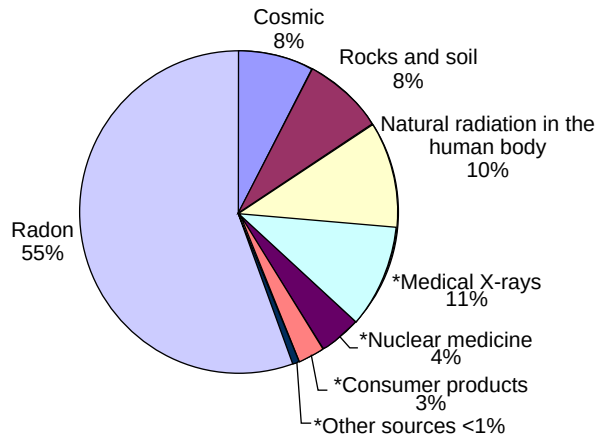


FIGURE 5 Average annual exposure to natural and man-made (asterisks) sources of radiation. The average annual dose in the United States is 360 mrem (National Council on Radiation Protection Measurements Report No. 93, 1988).

understanding of the biological effects of radiation exposure and to provide effective radiation protection methods has led to the development of a discipline known as radiation health physics.

3. RADIATION PROTECTION

Nuclear engineers work in conjunction with health physicists to assure that all activities involving radiation exposure to nuclear plant workers or to the public are kept well below the U.S. requirements stated in Title 10, Part 20 of the Code of Federal Regulations (10 CFR 20). In fact, the current industry practice is to apply the ALARA (“as low as reasonably achievable”) principle to every exposure-related activity. To this end, nuclear engineers have been widely successful in designing nuclear plants that limit the dose to the public. As shown in Fig. 5, the average American receives a radiation dose of about 360 mrem (or 0.36 rem) per year. Of this exposure, the inhalation of radon gas accounts for 200 mrem per year, with another 28 mrem coming from the rocks and soil. (Radon is a decay product of the uranium naturally present in rocks and soil, and, as a noble gas, radon easily seeps into buildings.) Because both of these sources of radiation are soil based, the actual dose equivalent will vary with location. In addition to this soil-based radiation, another 27 mrem results from cosmic rays, which consist of high-energy radiation emanating from outer space. Because Earth’s atmosphere acts as a filter, however, the actual exposure will vary with elevation. Beyond these external sources, the human

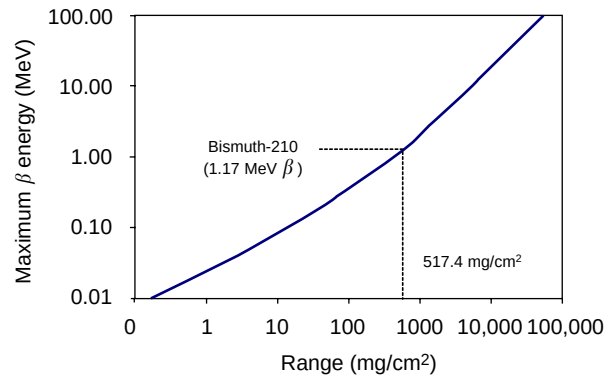


FIGURE 6 The range–energy curve for β particles.

body is also radioactive, exposing itself to about 40 mrem per year. Finally, man-made sources of radiation (such as X rays) add about 65 mrem. Nuclear power provides a mere 0.003 mrem per year, thus nuclear plants are far safer than the environment! Thus, the routine emissions of nuclear reactors are not a primary safety concern.

The three governing principles of radiation protection are time, distance, and shielding. Doses can be minimized by limiting the time of exposure, by maximizing the distance to the source, and by providing shielding to attenuate the radiation. The α particles are the easiest form of radiation to shield against. For example, an α particle with the very high energy of 4 MeV would travel only ~ 2.3 cm in air. Hence, relatively short distances from an α source would serve as adequate protection. However, a β particle of the same energy can travel considerably farther in air, in excess of 1 m. Both α and β particles have finite ranges of travel in different materials and a significant effort has been made to document their range–energy curves. Figure 6 presents a graph of the range–energy curve for β particles. The range is defined as the shielding material density (milligrams/cubic centimeters) times the shielding material thickness (centimeters) needed to stop a β particle with a given energy. Hence, the units for the range are expressed in milligrams/square centimeter. Suppose we wish to design an aluminum shield to block the 1.17-MeV β particles emitted from a bismuth-210 source. The graph in Fig. 6 indicates that at 1.17 MeV, β particles have a range of 517.4 mg/cm². Dividing the range by the known density of aluminum, 2700 mg/cm³, yields a shield thickness of approximately 0.2 cm.

Unlike α and β particles, which have finite ranges in materials, γ rays and X rays are exponentially absorbed in materials. That is, they obey the following equation, as illustrated in Fig. 7, for a

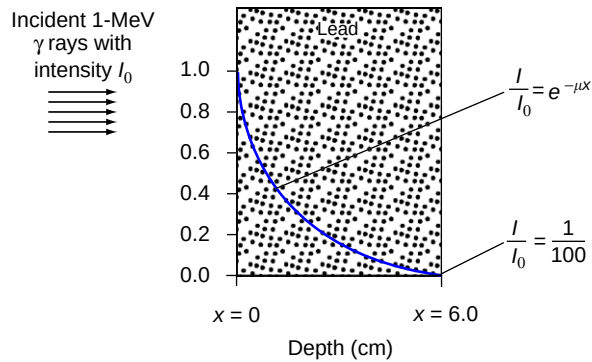


FIGURE 7 Exponential attenuation of 1.0-MeV γ rays in a lead slab.

beam of monoenergetic (single-energy) γ rays impinging on a slab:

$$I/I_0 = e^{-\mu x}.$$

This equation states that the gamma intensity, I , at a depth, x , in a shielding material is equal to the γ intensity at the front face times an exponential attenuation factor, $e^{-\mu x}$. The exponential attenuation factor is dependent on an attenuation coefficient, μ , which in turn depends on the energy of the γ ray. Values for the attenuation coefficient have been tabulated in the literature for a wide range of γ energies and materials. For example, the linear attenuation coefficient for 1.0-MeV γ particles in lead is 0.177 cm^{-1} . Thus, as shown in Fig. 7, 1.0-MeV γ rays would require 6 cm of lead to be attenuated by a factor of 100.

Many additional factors must be considered when designing shielding. For example, secondary radiations may be caused by the interaction of energetic γ rays with the shield material nuclei, thus requiring a greater material thickness than that predicted by the exponential attenuation equation. When shielding a radioactive isotope, consideration must also be given to its daughter products, which may have higher decay energies than the parent has. For example, strontium-90 emits a 0.54-MeV β particle, but its daughter product, yttrium-90, emits a higher energy 2.27-MeV β particle. Therefore, the shielding should be designed relative to the highest energy emission expected in the decay chain.

4. DESIGN OF NUCLEAR POWER REACTORS

Having discussed the basics of nuclear science and radiation protection, it is now possible to consider the

practical application of nuclear fission as a power source. As mentioned previously, uranium-235 fissions through the absorption of thermal neutrons. Each fission, in turn, releases approximately 200 MeV of energy and two to three additional neutrons. Of these newly created neutrons, some will be lost through leakage and absorption in nonfissile material. Others, however, will be absorbed in uranium-235 nuclei to generate additional fissions and neutrons. Accordingly, it is possible to sustain the fission process via a chain reaction. If one fission induces more than one fission on average, the system is supercritical, and the fission rate will increase. On the other hand, if one fission induces less than one fission on average, the system is subcritical, and the fission rate will decrease. Finally, if each fission induces exactly one fission on average, the system is critical, and the fission rate will remain constant. In a nuclear reactor, therefore, the trick is to control the chain reaction to maintain a constant power level. In practice, this is done by the adjustment of control rods, which contain a neutron-absorbing material such as boron. By adjusting the position of the control rods, the fission rate and the corresponding heat generation rate may be increased, decreased, or held constant.

Although understanding the basic operation of a nuclear reactor is relatively straightforward, the design process is quite challenging. Nuclear engineers are tasked with calculating the time-dependent population of neutrons at each position within a nuclear reactor. The conserved quantity of interest is the neutron flux, ϕ . The neutron flux represents the number of neutrons per unit area, per unit time, at a given position and time inside the reactor. One method of calculation employs the neutron transport equation. This integro-differential equation is mathematically formidable and therefore not presented here. The equation, however, represents a balance of the various mechanisms by which neutrons can be lost or gained from an arbitrary volume in the system. As shown in Fig. 8, neutrons having an energy within a specified energy range and traveling in a direction within a specified range of directions can enter or leave the volume by a variety of mechanisms. The gain mechanisms include sources of neutrons, such as fission neutrons, generated within the volume, external neutrons with just the right energies and directions entering through the surface that bounds the volume, and neutrons with a different energy and direction within the volume that experience a scattering collision such that their new energies and directions fall within the specified ranges. The loss mechanisms include neutrons leaking

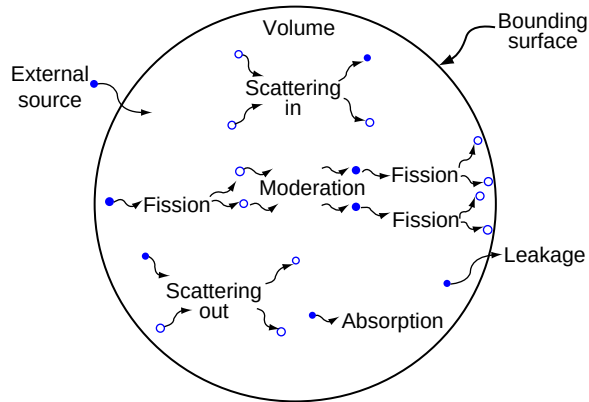


FIGURE 8 Various mechanisms by which neutrons can be lost or gained from an arbitrary volume in a reactor. Solid circles denote neutrons with energies and directions within the specified ranges. Open circles denote neutrons with energies and directions outside the specified ranges.

out through the bounding surface, neutrons captured by nuclei within the volume, and neutrons within the volume that experience a scattering collision such that their new energies and directions fall outside of the specified ranges. Neutron-reflective materials such as graphite, beryllium, and heavy water (D_2O) are often used to improve fuel economy in thermal reactors. Neutrons leaking out of the reactor core interact with the reflector material nuclei and some are scattered back into the reactor. The use of reflectors allows the designer to reduce the overall dimensions of the core.

The rates at which a neutron scattering, neutron capture, or nuclei fission event occurs in a material is calculated by the reaction rate, $\Sigma\phi$. Here ϕ is the neutron flux as defined previously. The reaction rate represents the frequency with which a particular reaction occurs between single-energy neutrons and the nuclei of the interacting material. It is expressed in units of reactions per unit volume, per unit time. The macroscopic cross section, Σ , consists of the nuclei density, N , of the interacting materials, in units of nuclei/cubic centimeter, times the microscopic cross section, σ , expressed in units of square centimeters. The value of the microscopic cross section varies significantly with the properties of the interacting material, the neutron kinetic energy, and the type of reaction being considered. Fission, capture, and scattering cross sections have been experimentally determined and tabulated for a wide range of materials and neutron kinetic energies. Graphs of reaction specific cross sections versus neutron kinetic energies reveal the energy ranges at which a reaction would preferentially occur in a

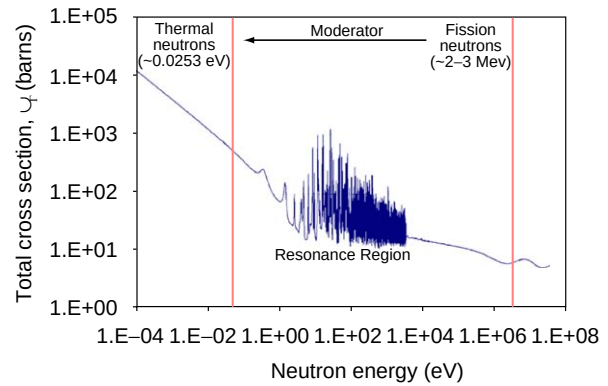


FIGURE 9 Values of total cross section for uranium-235 as a function of neutron kinetic energies.

material. For example, Fig. 9, shows the values of total cross section for uranium-235 as a function of neutron kinetic energies. The total cross section is relatively close to that of the fission cross section for this isotope. The cross sections in this graph are expressed in units called barns, where 1 barn is equivalent to 10^{-24} cm^2 . The neutron kinetic energy is expressed in electron volts. Figure 9 shows that neutrons have an increased likelihood of interacting with uranium-235 nuclei at lower neutron energies. Neutrons emerge from a fission event with kinetic energies in the range of 2–3 million electron volts (2–3 MeV). By contrast, thermal neutrons at room temperature have average kinetic energies of only 0.0253 eV. Thus, to take advantage of the higher fission cross sections that arise at the lower neutron energies, thermal neutron reactors use the scattering properties of a moderator, such as graphite or water, to reduce the kinetic energy of the neutrons by eight orders of magnitude.

In summary, the neutron transport equation is a balance equation used to calculate the neutron flux as a function of position and time in a reactor. The solution to this equation is not easily obtained because its source and sink terms are given by reaction rate equations that also depend on neutron flux. In addition, the reaction rates vary with material nuclei densities and the energy-dependent neutron cross sections for fission, scattering, and capture events. A variety of numerical techniques have been developed to obtain accurate computer solutions of the three-dimensional energy-dependent neutron transport equation.

Nuclear engineers often employ an approximation to the neutron transport equation—a diffusion theory model—to predict the temporal and spatial neutron population in a reactor. Diffusion theory is much

TABLE II
Distribution of Fission Energy in a Nuclear Core^a

Process	Percent of total reactor energy	Range
Kinetic energy of fission fragments	80.5	Very short
Kinetic energy of fast neutrons born of fission event	2.5	Medium
γ energy released at fission	2.5	Long
Kinetic energy of delayed neutrons	0.02	Medium
β decay energy of fission products	3.0	Short
Neutrinos associated with β decay Nonrecoverable	5.0	
γ energy of fission products	3.0	Long
Nonfission reactions due to excess neutrons plus β and γ decay energy of neutron absorption and γ emission (n, γ) products	3.5	Short and long

^aData from El-Wakil (1981).

simpler and can provide excellent results when the region of interest is not directly adjacent to a strongly absorbing medium or boundary. However, accurately predicting the fission rate of the fissionable fuels (uranium-235 and plutonium-239 produced from neutron capture in uranium-238) in the core is essential to good fuel economy and requires sophisticated modeling using benchmarked computer codes.

The amount of energy produced by fission in the core is directly related to the neutron flux profile, ϕ , the average energy per fission event, the number of fissionable nuclei per volume, and the effective fission microscopic cross section. Table II shows how this energy is distributed in the reactor. Most of the energy of fission, 80.5%, is associated with the kinetic energy of the fission fragments and is deposited directly in the fuel. A portion of the heat generated in a nuclear core is due to the decay of fission products. Even after the fission events in a nuclear reactor have been terminated, by the action of control rods, for example, a significant amount of decay heat continues to be generated in the core. This unique feature of nuclear heat generation imposes some very important safety constraints on the design of a nuclear power reactor. In particular, decay heat

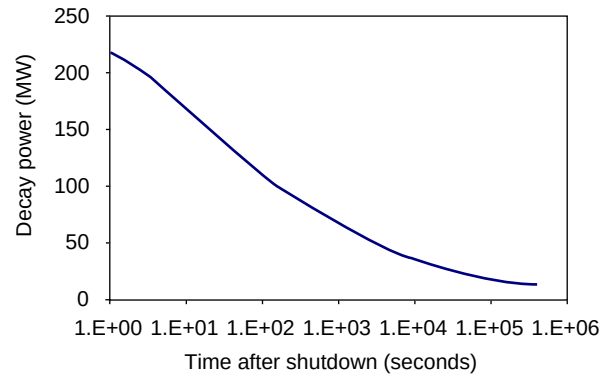


FIGURE 10 Decay power for a uranium-based thermal reactor with an initial core power of 3400 MW (thermal) and shutdown after 1 year of continual operation.

removal systems must be capable of cooling the nuclear core even after shutdown. The 1979 American National Standard for Decay Heat Power in Light Water Reactors provides the standard models useful to predict the amount of decay heat generated after reactor shutdown. The decay power is a function of the reactor power at the time of shutdown and the length of time the reactor has been at power. Figure 10 shows how core decay power varies with time after shutdown for the case of a 3400-MW thermal reactor that has been operating continuously for 1 year. One second after the reactor is shut down, the core power has dropped to $\sim 6.5\%$ of its initial value. However, this still represents ~ 220 MW of decay power, and significant core cooling would still be required to prevent damaging the fuel. One hour after shutdown, the core decay power is roughly 50 MW, which represents 1.5% of the original core power. Nuclear fuel is typically permitted to decay for several days before it is removed from the core as part of a refueling operation.

Having explored the fundamental concepts of nuclear engineering design as applied to nuclear reactor cores, the following section uses those concepts to describe the nuclear power plant designs that are currently employed to produce electricity.

5. COMMERCIAL NUCLEAR POWER PLANTS

The general design criteria for commercial light water reactors (LWRs) in the United States are presented in Title 10, Part 50, Appendix A of the Code of Federal Regulations (10 CFR 50 Appendix

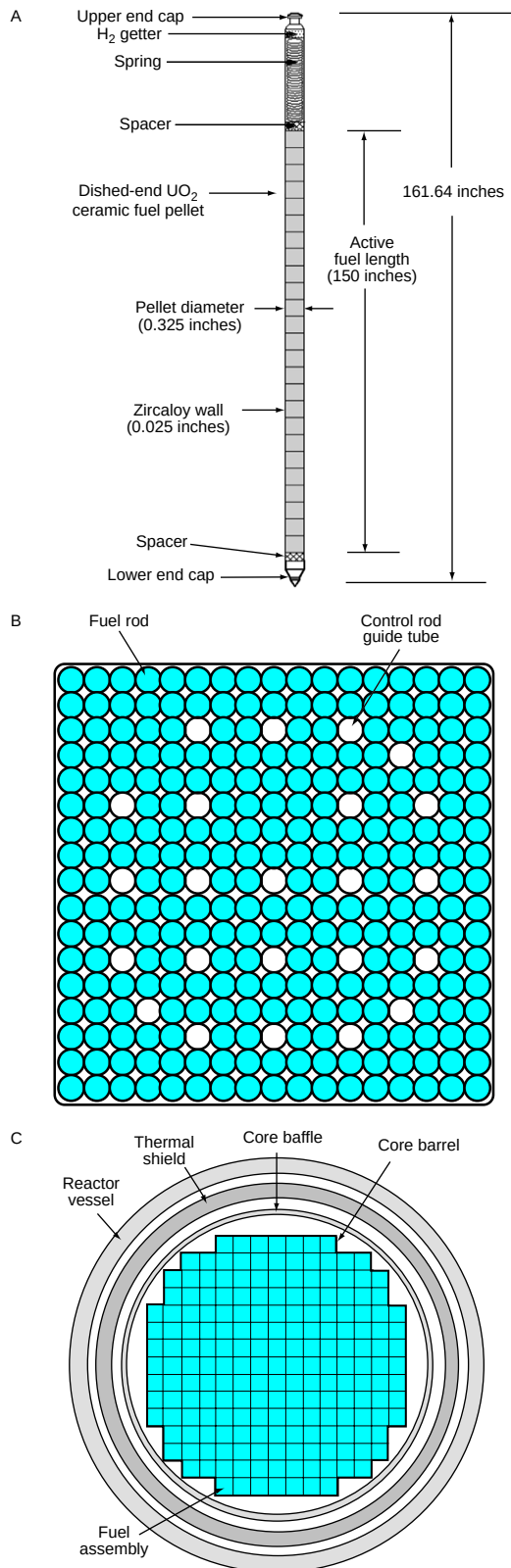


FIGURE 11 Reactor fuel geometry. (A) Fuel rod, (B) a 17 × 17 fuel assembly, and (C) the reactor core.

A). Inherent to these criteria is a safety philosophy known as defense-in-depth. The defense-in-depth philosophy requires that all of the safety systems be redundant (multiples of the same kind of safety system), diverse (different types of systems performing the same safety functions), and independent (a single failure would not cause the loss of multiple safety systems). The outcome of this approach is that nuclear plants are designed with multiple barriers that prevent the release of fission products to the environment. Because the source of heat and fission products is the fuel, the discussion begins there. A typical reactor uses cylindrical fuel pellets of uranium dioxide (UO₂) enriched to about 3–5% uranium-235. Ceramic fuel is selected because of its ability to withstand very high temperatures. Each pellet is dished on the ends to allow thermal expansion and is slightly porous to trap fission gases. The pellets are stacked end to end in a long metal tube and held in place by a compressed spring. This tube, known as the cladding, provides structural support to the fuel and prevents the radioactive fission products from entering the reactor coolant. To fulfill these demands, the cladding must be a strong corrosion- and heat-resistant material as well as a poor neutron absorber. For all these reasons, a zirconium alloy called zircaloy is commonly used for cladding. As shown in Fig. 11A, taken together, the fuel pellets and the cladding form what is known as a fuel rod. The fuel rods are assembled and aligned with high precision using spacer grids to form rectangular arrays known as fuel assemblies (Fig. 11B). These assemblies, in turn, are grouped together to form a reactor core that approximates a right circular cylinder (Fig. 11C). In light water reactors, water flows through the core to remove the heat from the fuel rods. Moreover, in addition to serving as the coolant, the water also functions as the moderator, in that it acts to slow down (or moderate) the high-speed neutrons released from fission. In this regard, the moderating capacity of water stems from the fact that the average kinetic energy lost by a particle in an elastic collision is maximized when it collides with a second particle of the same mass. Accordingly, because the hydrogen nuclei have the same mass as the neutrons, light water is a very effective moderator.

With respect to LWRs, there are two major designs: the pressurized water reactor (PWR) and the boiling water reactor (BWR). As shown in Fig. 12A, the PWR utilizes three types of flow loops. In the primary loop, which passes through the reactor core, the water is highly pressurized [15.5 megapascals (MPa) = 2250 pounds per square inch

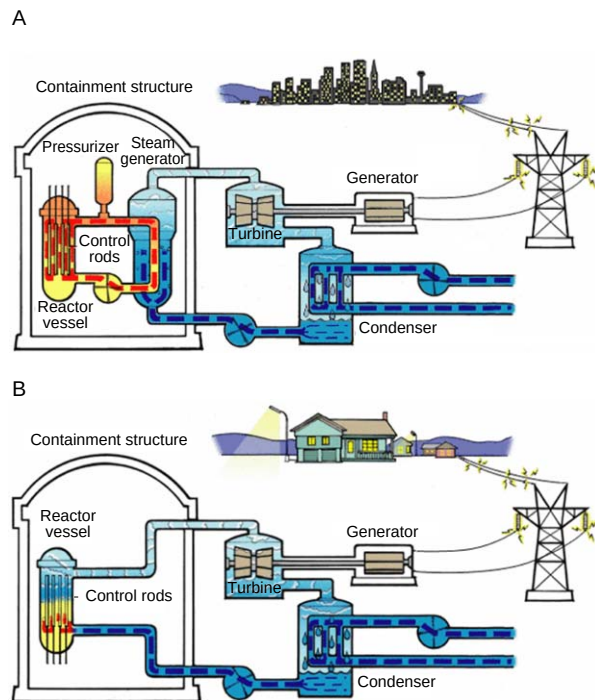


FIGURE 12 Commercial light water reactors. (A) Overall schematic for a pressurized water reactor and (B) a boiling water reactor.

gauge at atmospheric pressure (psia)] to prevent boiling. The primary coolant is heated in the core and flows through the tubes of a steam generator. The hot primary fluid gives off heat to the secondary fluid (with which it does not mix) in the steam generator, and is pumped back to the reactor to complete its circuit. The secondary fluid boils within the steam generator shell and all of the moisture is removed via separators and dryers to produce a very dry, high-pressure, steam. The steam expands to a high-velocity vapor that subsequently turns a turbine shaft connected to a generator to produce electrical power. From the turbine, the secondary fluid flows as a liquid-vapor mixture to the condenser, where it condenses to liquid by means of heat transfer to a third liquid stream. From the condenser, the secondary fluid is pumped back to the steam generator to complete its circuit. The third stream, which receives the waste heat in the condenser, generally dissipates this heat in a cooling tower or a nearby body of water such as a river or a lake. Because the PWR was the first commercially available reactor design, it is the most commonly used reactor concept worldwide.

In the BWR shown in Fig. 12B, the pressure is somewhat lower [6.9 MPa (1000 psia)], allowing the water to boil directly in the core. The moisture from

the steam is removed using dryers and separators to produce a dry, high-pressure steam that goes directly to the turbine. Here the steam expands to a high-velocity vapor, which turns the turbine shaft connected to a generator to produce electrical power. From the turbine, the water flows as a liquid-vapor mixture to the condenser, where it condenses back to liquid and is then pumped back to the core. As before, the secondary stream passing through the condenser connects to a nearby body of water or a cooling tower. Although not as prevalent as the PWR, the BWR is nevertheless employed on a worldwide scale. In this regard, an improved version of the BWR, known as the advanced boiling water reactor (ABWR), is currently utilized in Japan and Taiwan.

From this description of reactors, it should be evident that many features of a nuclear power plant are identical to those of conventional power plants with the exception of the power source. Because of this exception, however, precautions must be taken to contain the radioactive fission products in the event of an accident.

6. NUCLEAR POWER PLANT SAFETY SYSTEMS

Because a nuclear explosion in a nuclear power plant is impossible due to the low fuel enrichment, the worst conceivable accident is a severe loss-of-coolant accident (LOCA), leading to a core meltdown. Although the fission process would be immediately stopped by a control rod insertion, the radioactive fission products would continue to generate decay heat in the fuel. Thus, a LOCA producing core uncovering could cause the fuel to melt. In the most extreme case, a molten mass would fall to the bottom of the reactor, melting through the reactor pressure vessel and the underlying concrete, and eventually coming to rest about 6.1 m (20 ft) underground. In the process, airborne radioactivity would likely be generated as a result of flashing steam and the gaseous products of chemical reactions. To preclude such a scenario, numerous engineered safety features have been incorporated into the design of commercial nuclear power plants. Lists of typical engineered safety features for PWRs and BWRs and brief descriptions of their functions are provided in Tables III and IV, respectively. The first four systems in each table comprise the emergency core cooling system for each design.

The zircaloy-encapsulated fuel and the primary loop can be considered the first line of defense. The reactor pressure vessel and the pipes are designed

TABLE III
Some Engineered Safety Features of Pressurized Water Reactors

Engineered safety feature	Description of function
High-pressure injection system	This system consists of multiple pumps connected to a large supply of borated water (neutron absorbing). It is capable of injecting coolant into the core at moderate flow rates while at full system pressure
Low-pressure injection system	This system consists of multiple pumps connected to a large supply of borated water capable of injecting coolant into the core at very high flow rates while at moderate system pressures
Accumulators	This system consists of a set of prepressurized tanks filled with borated water. No external power is required to inject large volumes of cold borated water into the core
Residual heat removal system	This system consists of multiple pumps capable of providing very large quantities of borated water to the core at low system pressures
Control rod shutdown margin	This system consists of an additional bank of control rods that are available to shut down the fission reaction
Auxiliary feedwater system	This system consists of multiple pumps connected to the condensate storage tank, to provide backup water for the steam generators
Containment building	This is a large-volume prestressed, posttensioned concrete building that houses the nuclear steam supply system. It is reinforced with a steel liner and steel reinforcement bars and is capable of withstanding the maximum pressures for worst-case accident conditions
Containment ventilation system	This system provides containment air cooling and uses high-efficiency filters to remove any radioactive particles inside containment
Containment sprays	This system sprays cold water into the containment to reduce containment pressure in the event of a steam release
Containment sump	This system consists of a large concrete sump at the base of the containment. It collects all of the water leaking from the reactor. It serves as the water supply for the safety injection systems during the long-term cooling phase of an accident
Redundant offsite power	This system consists of two independent sources of external AC power for all of the safety systems
Diesel generators	This system consists of multiple diesel generators (4180 V), each capable of powering all of the safety systems. In the event of a loss of offsite power, they are able to provide full power within 10 sec

with a thickness sufficient to withstand the high operating pressures. The reactor pressure vessel is made of stainless-steel-clad carbon steel with a thickness of about 25.4 cm (10 inches). However, assuming that a leak does occur, as shown in Tables III and IV, nuclear reactors have safety injection systems with independent piping, pumps, diesel generators, and water supplies. Moreover, these systems are designed to inject emergency coolant into the reactor at flow rates sufficient to meet very pessimistic conditions. For instance, the safety injection systems are designed to keep the core covered with water even if the incoming coolant pipe (the cold leg) suffers a double-ended guillotine break, with reactor coolant flowing out from both ends!

Nevertheless, assuming that such a break does occur and that the safety injection systems fail, then it would be necessary to contain the airborne radioactivity that the melting fuel would generate. For this reason, the reactor is contained in a large, dome-shaped building called the containment. The

containment is a 1.2-m (4-ft)-thick, heavily reinforced and steel-lined concrete structure that is designed to withstand the internal pressures generated by flashing steam. Moreover, it is also designed to protect the reactor from outside disturbances such as earthquakes, tornadoes, hurricanes, and jet airline crashes. Consequently, when the combination of the vessel and piping thicknesses, the safety injection systems, and the rugged containment is considered, the probability of a significant release of radioactivity is nil. Although the containment design is utilized by nuclear reactors throughout most of the world, there is a notable exception. The Russian RBMK reactors (so named for the acronym in Russian, *reactor bolsoi mochnosti kipyashiy*, meaning “large power boiling reactor”), such as the one that exploded at Chernobyl in 1986, do not have a containment. Because of this difference (among others), RBMK reactors are not as safe as conventional designs and therefore do not provide an appropriate benchmark to assess worldwide reactor safety.

TABLE IV

Some Engineered Safety Features of Boiling Water Reactors

Engineered safety feature	Description of function
High-pressure injection system	This system consists of multiple pumps connected to a large supply of borated water (neutron absorbing). It is capable of injecting coolant into the core at moderate flow rates while at full system pressure
Low-pressure injection system	This system consists of multiple pumps connected to a large supply of borated water capable of injecting coolant into the core at very high flow rates while at moderate system pressures
Automatic depressurization system	This system consists of a set of valves that automatically open to reduce reactor pressure in a controlled manner
Core spray	This system consists of a set of pumps that spray water directly into the core
Control rod shutdown margin	This system consists of an additional bank of control rods that are available to shut down the fission reaction
Reactor core isolation cooling system	This system isolates the reactor from the turbines and condenser. Steam from the reactor is used to drive a turbine pump connected to the condensate storage tank to provide backup water for the core
Standby liquid control system	This system provides an alternate means of shutting down the fission reaction by injecting highly concentrated borated water into the core
Containment building	This is a large-volume prestressed, posttensioned concrete building that houses the nuclear steam supply system. It is reinforced with a steel liner and steel reinforcement bars and is capable of withstanding the maximum pressures for worst-case accident conditions
Containment sprays	This system sprays cold water into the containment to reduce containment pressure in the event of a steam release
Containment suppression pool	This system consists of a large concrete pool, or torus, at the base of the containment. It serves as a very large supply of the water for all of the safety injection systems. It also serves as a pressure suppression pool to reduce containment pressure
Redundant offsite power	This system consists of two independent sources of external AC power for all of the safety systems

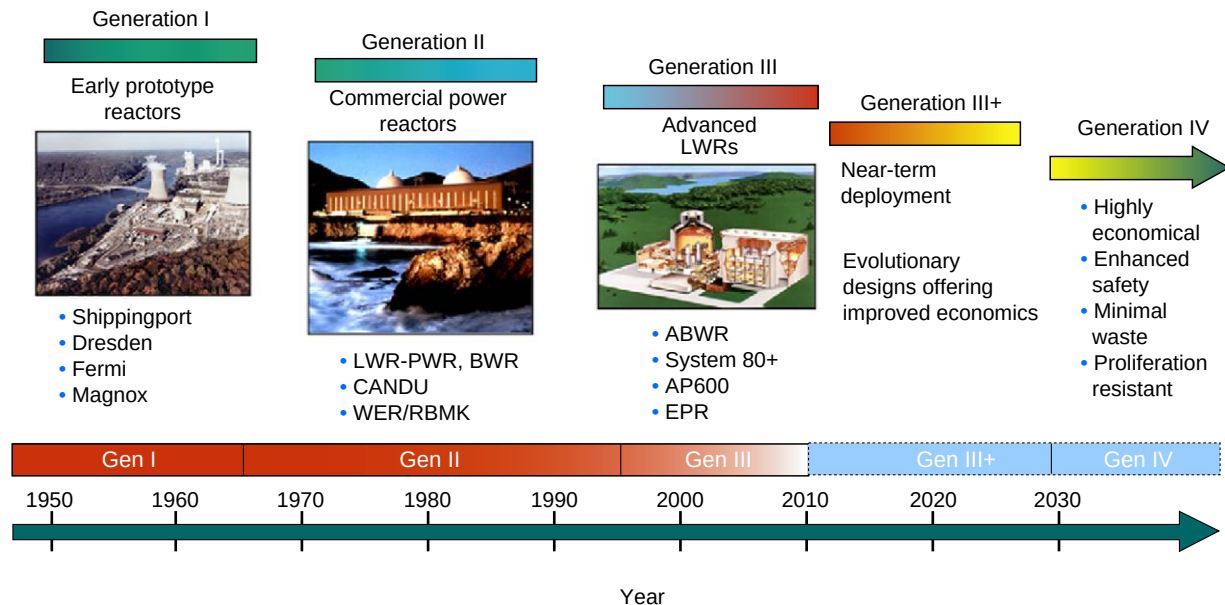


FIGURE 13 Next-generation nuclear plants, building on the experience of previous generations. LWRs, Light water reactors; PWR, pressurized water reactor; BWR, boiling water reactor; CANDU, Canadian deuterium uranium reactor; WER/RBMK, Russian light water and boiling water reactors; ABWR, advanced boiling water reactor; EPR, European pressurized water reactor. Courtesy of the U.S. Department of Energy.

7. NEXT-GENERATION NUCLEAR POWER PLANTS

Nuclear engineers are using the vast experience of two generations of reactor operation to design the nuclear power plants of the future. Figure 13 shows that a third generation of reactors will soon transition

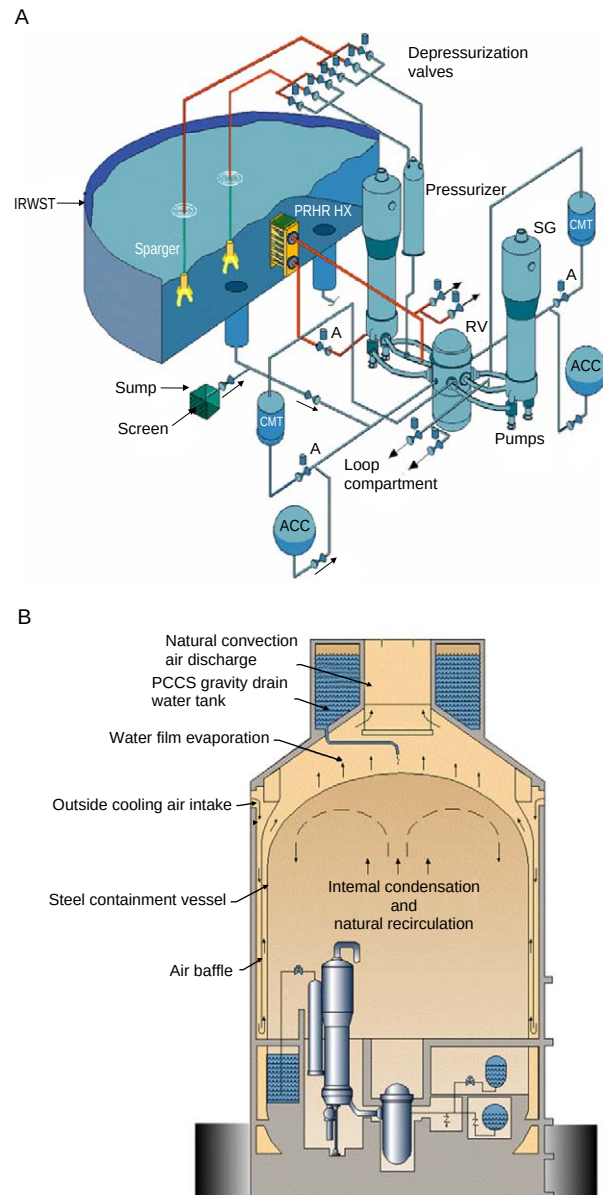


FIGURE 14 The AP600 reactor. (A) Passive safety systems. IRWST, In-containment refueling water storage tank; CMT, core makeup tank; ACC, accumulator; SG, steam generator; RV, reactor vessel; PRHR HX, passive residual heat removal heat exchanger; A, isolation valves. (B) Containment design. PCCS, Passive containment cooling system. Reproduced with permission from Westinghouse Electric Company.

from the drawing board to reality, ready for deployment within the next 10 years. Even greater advancements are expected for the fourth-generation designs that will be ready for deployment in the next 20 to 25 years. Third- and fourth-generation nuclear reactors will be safer and will exhibit a greater degree of reliability. A common design feature of these new reactors is the use of passive safety systems. Although the comprehensive systems used to safeguard today's nuclear power plants offer a very high degree of protection, even greater reliability and safety can be achieved by replacing all of the active safety systems with passive safety systems, which are driven by natural forces such as pressure and gravity. After all, whereas active pumps requiring external power can fail, gravity cannot. An example of such a system, shown in Fig. 14A, is an advanced design known as the AP600. The AP600, a 600-MWe plant designed by the Westinghouse Electric Corporation, is the first passively safe nuclear power plant certified by the U.S. Nuclear Regulatory Commission (USNRC). All of the externally powered safety systems have been replaced by elevated tanks filled with cold borated water. In the event of a LOCA, while the pressure is still high, the core makeup tanks (CMTs) begin to circulate cold water through the core by virtue of negative buoyancy. At intermediate pressures, water injection occurs via compressed nitrogen in the accumulators. Finally, at low pressures, injection from the in-containment refueling water storage tank (IRWST) occurs by virtue of gravity. Moreover, due to the containment design shown in Fig. 14B, the flashing steam is condensed on the containment walls and is recycled back to the IRWST and the lower containment compartments, to guarantee long-term cooling of the core. The heat generated by the core serves to drive all of the natural convection cooling processes. Therefore, cooling will continue as long as cooling is needed!

Because of the novelty of the design, licensing tests to demonstrate the safety of the AP600 were conducted at various experimental facilities in Italy, Japan, and the United States. With regard to the latter, 68 tests were conducted at the Advanced Plant Experiment (APEX) test facility at Oregon State University (OSU). As shown in Fig. 15, APEX is a one-quarter-scale integral system test facility designed to simulate the passive safety systems of the AP600. The nuclear core is simulated by electric heater rods, and all the components of the passive safety system are modeled. After conducting 28 accident simulations for Westinghouse and the U.S. Department of Energy and an additional 40 confirmatory tests for

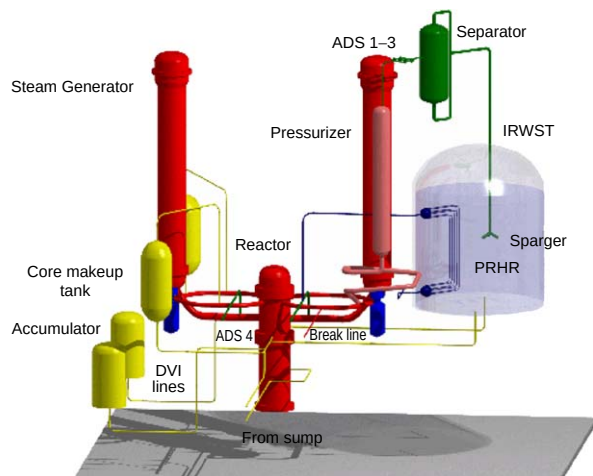


FIGURE 15 The one-quarter-scale advanced plant experiment (APEX) testing facility. IRWST, In-containment refueling water storage tank; PRHR, passive residual heat removal; ADS, automatic depressurization system; DVI, direct vessel injection. Drawing courtesy of Scott C. Franz.

the USNRC, it was concluded that even when some safety systems fail, the AP600 has sufficient redundancy to keep the core covered with water. The USNRC certified the AP600 on 15 December 1999. A larger version of the AP600, the AP1000, is currently in the licensing stage. The AP1000 is a 1000-MWe PWR with the same passive safety systems used in the AP600. To certify this design, the APEX facility at OSU has been modified by the U.S. Department of Energy (USDOE). The new facility will also be utilized to perform the confirmatory tests for the USNRC.

So far, the discussion has concerned light water reactors. In Canada, however, a reactor system has been developed that uses heavy water (i.e., water molecules with hydrogen-2, or deuterium, atoms, rather than hydrogen-1 atoms). This system is known as the Canadian deuterium uranium (CANDU) reactor. Because heavy water, compared to light water, is a much poorer neutron absorber, parasitic absorption is reduced. As a result, it is possible to use natural uranium fuel, thereby avoiding the costs of enrichment. A unique feature of the CANDU is its online refueling capability. The Canadian ACR-700 reactor is a potential candidate for near-term deployment.

To obtain thermodynamic efficiencies approaching 40%, some reactors utilize a gas cycle rather than steam. In these reactors, the gas circulates in a loop that includes both the reactor and the turbine. Two variations on this theme are the American high-temperature gas reactor (HTGR) and the British

advanced gas reactor (AGR), which use helium and carbon dioxide coolants, respectively. Both systems utilize a solid graphite moderator due to the poor moderating ability of low-density gasses. The fuel rods are placed within holes in graphite blocks, with additional holes made for coolant flow. The pebble bed modular reactor (PBMR) has received significant attention because of its inherently safe fuel. The fuel consists of 0.5-mm uranium oxide particles that are triple coated with ceramic materials that will not fail even at extremely high temperatures. Approximately 15,000 particles are embedded in a single 50-mm sphere that is coated with high-purity carbon. The spheres are arranged in a critical geometry to create a nuclear core. The PBMR is a direct-cycle system that uses helium to cool the core while driving a high-efficiency Brayton gas cycle to produce electricity.

Another variation is seen in the liquid metal fast breeder reactor (LMFBR), which uses liquid sodium as the coolant. Because liquid sodium is a poor moderator of neutrons, the high-energy neutrons produced by fission stay fast to enhance breeding. For a reactor to breed, it is necessary that more fuel be produced from fertile materials than is consumed in running the reactor. Thus, for every fission, one or more of the neutrons produced must go on to breed new fuel, because exactly one of these neutrons is needed to sustain the chain reaction. Consequently, because two or more neutrons must be utilized per fission, it is necessary to maximize neutron production while minimizing leakage and parasitic absorption. Because neutron production increases, and parasitic absorption decreases, with increasing neutron energy, having a fast neutron spectrum enhances breeding. Moreover, by placing a blanket of fertile uranium-238 around the core, the leaking neutrons are utilized to breed plutonium-239 in the blanket. In this way, the LMFBR produces fissile plutonium from the nonfissile uranium-238, which constitutes 99.3% of the naturally occurring uranium. Consequently, because breeder reactors do not waste this abundant isotope, they will be able to supply world energy needs for millennia.

After an extensive review, the USDOE has narrowed their research and development efforts to focus on six generation IV reactor concepts. Each design was specifically selected because it offers a high degree of safety, reliability, proliferation resistance, economy, and efficiency. The six designs selected are the gas-cooled fast reactor (GFR), the lead-cooled fast reactor (LFR), the sodium-cooled fast reactor (SFR), the very-high-temperature reactor (VHTR), the supercritical water-cooled reactor

(SWCR), and the molten salt reactor (MSR). The VHTR design is of particular interest because its 1000°C (1832°F) outlet temperature can be used to crack water molecules to produce hydrogen for a wide range of energy-related applications, such as automobile fuel cells. All of these designs represent a significant leap forward in the development of nuclear technology and high-temperature, high-strength, materials.

8. NUCLEAR WASTE DISPOSAL

The timely deployment of the next generation of nuclear power plants will undoubtedly depend on having in place a fully functional nuclear waste repository. As shown in Fig. 16, the nuclear fuel cycle represents the life cycle of the nuclear fuel from its initial mining to its final burial. In this figure, the various waste streams are also indicated. The uranium is initially mined and separated from its ore in the milling process to produce yellowcake (U_3O_8). The yellowcake is then sent to a chemical conversion plant to produce uranium hexafluoride (UF_6) gas. Through a process of gaseous diffusion, the UF_6 is next enriched in its uranium-235 content at the enrichment plant. Essentially, because lighter

particles diffuse more rapidly through permeable membranes, the gaseous streams crossing such membranes are progressively enriched in their uranium-235 content. Once the resulting stream is sufficiently enriched (about 3%), the gas is taken to the fuel fabrication plant to produce the uranium dioxide fuel pellets and thence fuel rods and assemblies. This fuel, in turn, is then taken to a nuclear reactor and burned for about 3 years to produce electrical power.

After the spent fuel is taken from the reactor, one of two options will be used. In the current U.S. approach, the fuel rods are stored in a water storage pool at the reactor site for several years, to allow the most intensely radioactive fission products to decay. Once these rods have sufficiently cooled, they are then removed to dry storage to await burial 300 m (1000 ft) underground in a dry geologic medium. In the second approach, the fuel rods will be taken from storage to a reprocessing plant where the uranium and plutonium will be separated out and reprocessed back into new fuel. Moreover, certain other isotopes having medical or industrial value will also be separated out. Finally, certain long-lived fission products known as actinides will also be separated out and recycled back into the fuel to be fissioned in a reactor, and thereby transformed into stable or rapidly decaying isotopes. This process of eliminating long-lived fission products through recycling and fission is known as transmutation. Because the uranium and plutonium constitute 97.1% of the spent fuel, and because the industrial isotopes and actinides constitute another 2.6%, reprocessing will utilize valuable resources within the waste while producing a 300-fold reduction in the waste volume. As a result, the annual per capita high-level waste will have the volume of an aspirin tablet. Such waste will be incorporated into a glass matrix, placed in a corrosion-resistant metal container, buried deep 300 m (1000 ft) underground in a dry geologic medium, and packed into place with bentonite clay. Moreover, because the long-lived fission products will have been separated out, the buried waste will decay more quickly. As a result, its toxicity will approach that of natural uranium within 10,000 years.

To preserve the waste intact for this amount of time, a defense-in-depth strategy will be employed. Essentially, the wastes will be mixed with molten glass, poured into a corrosion-resistant steel or titanium canister, and allowed to harden. The canisters, in turn, will be buried deep 300 m (1000 ft) in a geological medium that is both dry

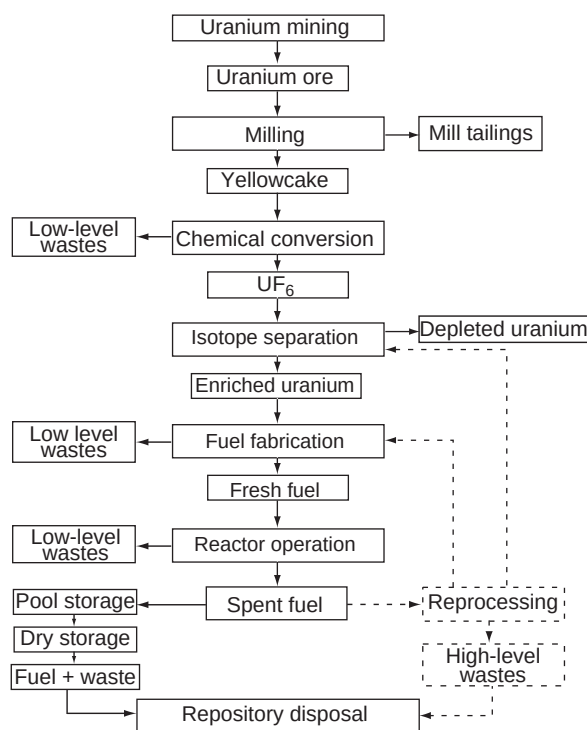


FIGURE 16 Nuclear fuel cycle.

and impermeable to water. In the United States, the currently envisioned site is Yucca Mountain, Nevada, where the water table is 550 m (1800 ft) below the surface and the annual rainfall is a mere 15 cm (6 inches). Thus, at a depth of 300 m (1000 ft), the waste will be safe from both rainfall and upwelling groundwater. The proposed waste disposal facility is known as a repository and will consist of a central vertical shaft that connects with horizontal tunnels. In these tunnels, the waste canisters will be placed in holes drilled into the tunnel floors and packed in with bentonite clay. Finally, when the repository becomes full, the tunnels and the central shaft will be filled in with bentonite clay to prevent the intrusion of water.

As should be apparent, a repository employs many barriers to isolate the waste. First, water must travel some distance through an impermeable geological medium. Second, the groundwater must penetrate a layer of bentonite clay, which will swell to prevent its further passage. Third, the groundwater must next penetrate a corrosion-resistant titanium or steel canister. Fourth, the water must then leach the wastes out of a solid glass matrix. Fifth, the groundwater must then carry the wastes long distances through an impermeable geological medium to reach major population centers. Sixth, the progress of some of the waste species will be delayed relative to the groundwater due to chemical reactions with the geological medium. Thus, by using multiple barriers, it will be possible to contain the waste for a long time and to delay its spread thereafter. As a result, the wastes can be kept from the public until their radioactivity has largely decayed away.

Having examined the nuclear fuel cycle, it remains to consider the transportation of radioactive materials. In this regard, the most critical issue is the possibility of radioactive releases resulting from accidents involving spent fuel shipments (for reprocessing or burial). To prevent such an occurrence, U.S. shipping casks are designed to withstand a series of four extreme tests without leaking. In order of occurrence, such tests include a 9.1-m (30-ft) fall onto an unyielding surface (i.e., from an overpass), a 1.02-m (40-inch) fall onto a 15.2-cm (6-inch) diameter pin (i.e., onto a bridge abutment), a 30-minute fire at 801.7°C (1475°F) (i.e., a crash with flammable material), and an 8-hour submersion in water (i.e., fall into a roadside creek). Moreover, such casks have been deliberately slammed into concrete barriers at 37.6 m/sec (84 miles/hour), hit broadside by locomotives going 35.8 m/sec (80 miles/hour), and then subjected to fire tests without leaking.

As a close to this discussion on waste disposal, it is fitting to consider briefly the implications of the current U.S. waste disposal policy that would result in the burial of usable fuel, including large amounts of plutonium and potentially useful medical isotopes, as opposed to spent fuel reprocessing. It is important to note that the technical feasibility of reprocessing nuclear fuel is not at issue. France has been doing so for many years. Reprocessed fuel would consist primarily of a mixture of uranium and plutonium in an oxide form, the remaining isotopes having been removed. This fuel is known as a mixed oxide fuel (MOX) fuel. Reprocessing spent fuel solely to obtain medical isotopes would not likely be cost-effective because the isotopes of interest represent only a small fraction of the total inventory. Indeed, because of their very special purity requirements, medical isotopes might still be produced more economically in a nuclear reactor dedicated solely to isotope production, even if the isotopes arise as a by-product of producing mixed oxide nuclear fuel. However, this argument should be revisited as technology advances and as the use of different medical isotopes broadens. The chief argument against reprocessing has been the concern that the plutonium-239 separated from spent nuclear fuel might be diverted to produce nuclear weapons. Although burying the spent fuel in its entirety reduces the risk of proliferation, it does not eliminate the long-term concern. It merely passes the concern to the future generations tasked with maintaining and protecting such a repository. Plutonium-239 has two viable uses: it can be used to produce a nuclear weapon or it can be fissioned in a nuclear reactor to produce electricity. Once fissioned, that quantity of plutonium is gone forever. It is interesting to note that a spent fuel reprocessing facility could also be used to convert weapons-grade plutonium to low-enriched mixed oxide nuclear fuel for electricity production, such as has been proposed for the dismantled nuclear weapons produced in the former Soviet Union. It is an encouraging commentary on society whenever swords are turned to plowshares.

9. CONCLUSION

The field of nuclear engineering encompasses nuclear science, radiation protection, and reactor core design; to understand the current state of commercial nuclear power and the role played by nuclear engineers in its development requires an understanding of these subjects. The current generation

of commercial nuclear power reactors has served humankind exceedingly well, providing 16% of the world's total electric power production. Nuclear engineers are now designing and testing innovative designs that will be deployed worldwide over the next 20 years. In particular, the fourth-generation reactors will be safer, more efficient and reliable, cost competitive, and proliferation resistant. The use of passive safety systems will reduce the complexity of the plants and will provide a new level of protection while promoting public confidence. Plans to store nuclear wastes in a secured national repository have been approved and have effectively closed the nuclear fuel cycle in the United States. Other countries have adopted a combined strategy of spent fuel reprocessing and a waste repository. The concept of using highly enriched plutonium from dismantled nuclear weapons to produce limited quantities of mixed oxide fuel for electricity production has significant merit and may lead to new discussions on fuel reprocessing.

Although the field of nuclear engineering is young relative to other engineering disciplines, few other disciplines have such a prestigious heritage. Born of Einstein's remarkable theory, fueled by Eddington's visionary challenge, and ushered into reality by Fermi's experimental genius, nuclear engineering

remains faithful to its original guiding principle—to find creative ways of safely harnessing and using the energy of the atom for the benefit of mankind.

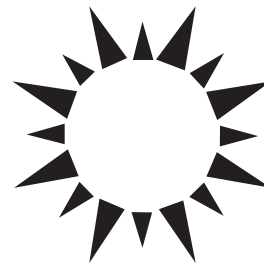
SEE ALSO THE FOLLOWING ARTICLES

Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

- Beckmann, P. (1985). "The Health Hazards of Not Going Nuclear." The Golem Press, Boulder, Colorado.
- Binney, S. E. (1987). Nuclear energy. In "The Encyclopedia Americana," Vol. 23, pp. 511h–516.
- El-Wakil, M. M. (1981). "Nuclear Heat Transport." American Nuclear Society, La Grange Park, Illinois.
- Murray, R. L. (1994). "Understanding Radioactive Waste." Battelle Press, Columbus, Ohio.

Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors



DOUGLAS K. VOGT

Lawrence Livermore National Laboratory
Livermore, California, United States

-
1. Electricity Production from Light Water Reactors
 2. Pressurized Water Reactors
 3. Boiling Water Reactors
 4. Ship Propulsion

Glossary

boiling water reactor (BWR) A light-water-cooled and -moderated reactor that operates at a pressure of ~ 1000 pounds per square inch at atmospheric pressure (psia). At this pressure, reactor coolant is converted from liquid water to steam by thermal energy from nuclear fission. The boiling occurs within the reactor core. Steam from the core of a BWR flows to a conventional turbine-generator system, where it is converted to electricity.

decommissioning At the end-of-life stage of a nuclear power plant, the facility contains highly radioactive components. As part of decommissioning, these components, including the reactor and other plant areas, are decontaminated to reduce the amount of radioactivity. Some components will be removed from the reactor site and buried in an engineered disposal site. Pressurized water, boiling water, and naval reactors are decommissioned.

high-level radioactive waste repository An engineered facility that isolates spent fuel or the radioactive waste from spent fuel reprocessing, safeguarding the environment.

light water Hydrogen has two stable isotopes, (1) protium (99.985% natural abundance), which has a nucleus consisting of one proton, and (2) deuterium (0.015% natural abundance), which has a nucleus consisting of one proton and one neutron. Water molecules formed from water with natural abundance hydrogen (99.985% protium) are called light water. Water molecules formed

from hydrogen enriched in deuterium are called heavy water.

light water reactor (LWR) Any reactor that is cooled and moderated by ordinary or light water. Both pressurized water and boiling water reactors are light water reactors.

pressurized water reactor (PWR) A reactor that operates at high pressure (typically > 2000 psia). At this pressure, PWR coolant water does not experience significant boiling. Reactor coolant is heated by thermal energy from nuclear fission, resulting in a temperature increase. Heated coolant flows to a heat exchanger with lower pressure water (~ 1000 psia) on the “nonreactor side.” The lower pressure water is converted to steam. The steam from the steam generator flows to a conventional turbine-generator system, where it is converted to electricity.

Scientists in the former Soviet Union and in the West who developed nuclear weapons in World War II and in the early days of the Cold War realized that the tremendous energy produced by nuclear fission could be tapped either for direct use or for generating electricity. It was also clear that nuclear fission energy would allow development of compact, long-lasting power sources that could have various applications, including powering ships, especially submarines. From their roots in naval nuclear power programs, light water reactors have become an important source of electric power in many countries. Light water reactors built over the past 30 years continue to generate electricity, and new reactor designs, the advanced boiling water reactor and the advanced pressurized water reactor, are planned for operation in the 21st century.

1. ELECTRICITY PRODUCTION FROM LIGHT WATER REACTORS

The first nuclear reactor to produce electricity (albeit a trivial amount) was the small experimental breeder reactor (EBR-1) in Idaho, in the United States. EBR-1 started operating in December 1951. In 1953, President Eisenhower proposed his “Atoms for Peace” program, setting the course for civilian nuclear energy development in the United States. However, it was the U.S. effort under Admiral Hyman Rickover that led to the development of the pressurized water reactor (PWR) for submarine use. As part of the PWR development, Westinghouse Electric Corporation built a land-based submarine reactor prototype at the National Reactor Testing Station near Arco, Idaho. The prototype started operation in March 1953, and in 1954, the first nuclear-powered submarine, the *USS Nautilus*, was launched. In 1959, both the United States and the former Soviet Union launched their first nuclear-powered surface vessels.

The success of the submarine prototype reactor led to the U.S. Atomic Energy Commission decision to build the 90-megawatt electric (MWe) output Shippingport demonstration PWR near Pittsburgh, Pennsylvania. Shippingport started operation in 1957 and operated as a commercial nuclear power plant until 1982. In the United States, Westinghouse Electric Corporation designed the first fully commercial 250-MWe PWR, Yankee Rowe, which started operation in 1960 and operated until 1992. As the PWR was being developed, Argonne National

Laboratory and General Electric Company were developing the boiling water reactor (BWR). A small prototype BWR, Vallecitos, generated electricity in Pleasanton, California from 1957 to 1963. Dresden-1, which was designed by General Electric, was the first large commercial BWR (200 MWe). Dresden-1 was constructed 50 miles southwest of Chicago, Illinois and operated from 1960 until 1978.

In 1964, the first Soviet land-based PWR power reactor began power operations in Novovoronezh (Volga region). This reactor was a new 210-MWe pressurized water reactor known as a VVER (an acronym in Russian for *veda-vodyanoi energetichesky reaktor*, meaning “water-cooled power reactor”). In the Murmansk region of the Arctic Northwest, a slightly bigger VVER with a capacity of 440 MWe was constructed at the Kola Nuclear Power Plant. This became a standard Russian PWR design.

By the end of the 1960s, electric utilities worldwide were placing orders for PWR and BWR units of 400 MWe to more than 1000 MWe. Other countries have built primarily PWR and BWR types, so that today, 65% of the world nuclear electric generating capacity is PWR and 23% is BWR. Overall, 16% of the world’s electricity now comes from nuclear energy. Throughout the world, there are 438 commercial nuclear generating units with a total capacity of about 351,000 MWe. Worldwide, LWRs generated over 2,000,000,000 megawatt-hours (MWh) of electricity in 2001. Figure 1 shows the location of nuclear power plants worldwide that are under construction, in operation, or undergoing decommissioning. Figure 2 summarizes the

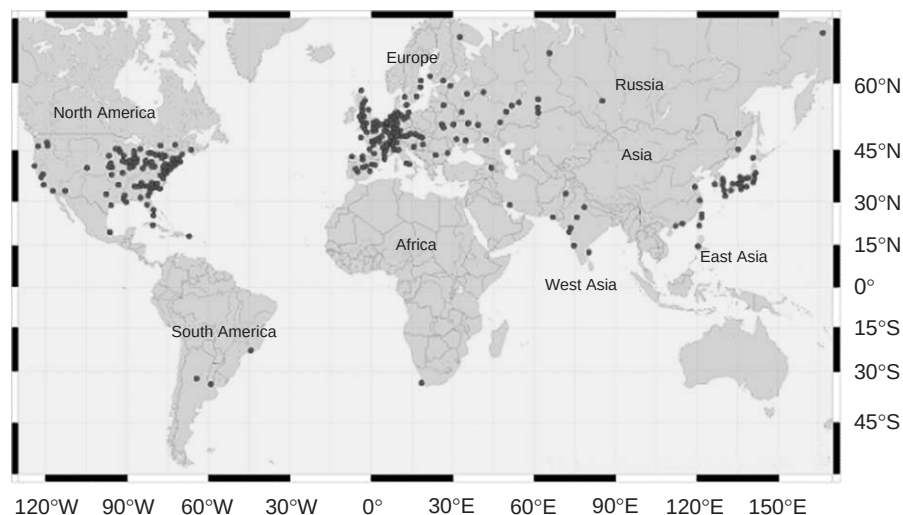


FIGURE 1 Location of nuclear power plants worldwide. Map from the International Nuclear Safety Center (Argonne National Laboratory).

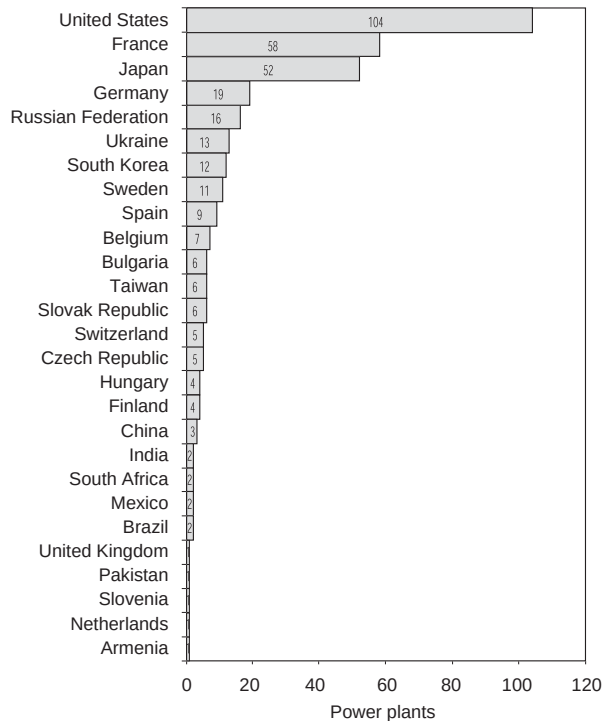


FIGURE 2 Number of light water reactor nuclear power plants worldwide.

number of operating LWR nuclear power plants in each country.

To minimize dependence on fossil fuel imports, several countries have placed significant dependence on nuclear energy from LWRs. The following countries generate more than 30% of electricity output using LWRs: France, Belgium, Slovak Republic, Ukraine, Sweden, Bulgaria, South Korea, Hungary, Slovenia, Switzerland, Armenia, Japan, Finland, and Germany. Eight countries have manufactured large LWRs for electricity generation: the United States, France, Japan, Germany, Russia, South Korea, Sweden, and the United Kingdom. Figure 3 shows the 32 states with operating LWRs in the United States.

2. PRESSURIZED WATER REACTORS

2.1 Plant Design Characteristics

Figure 4 provides a schematic view of a pressurized water reactor power plant. The power plant systems include a nuclear reactor and steam generator (called the nuclear steam supply system), which produce steam using energy from nuclear fission; a reactor “containment system,” which limits release of radio-

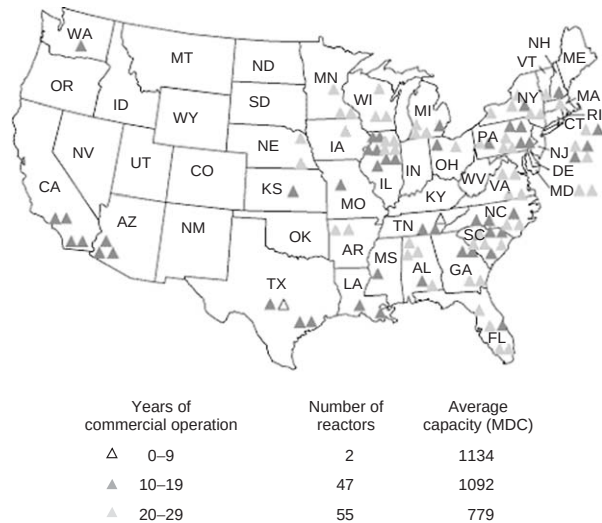


FIGURE 3 Locations of operating light water reactors in the United States as of December 2000 (there are no commercial reactors in Alaska or Hawaii). MDC, Maximum dependable capacity. This map and additional data are available on the Internet at the Web site of the U.S. Nuclear Regulatory Commission (<http://www.nrc.gov>).

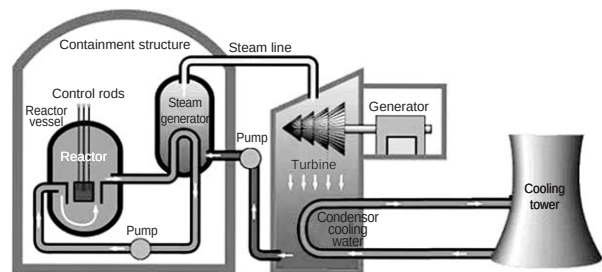


FIGURE 4 Schematic view of a pressurized water reactor.

active material from the reactor under accident conditions; a steam-powered turbine-generator that produces electricity; and a condenser and cooling towers for heat rejection.

The nuclear steam supply system includes nuclear fuel that is configured as the “reactor core” in a reactor pressure vessel. Reactor coolant water is pumped through the reactor core and heated. Heated water passes through tube-and-shell heat exchangers called “steam generators.” On the reactor side of the steam generator, the water enters the inside, or tube side, of the steam generator, passes through, and eventually returns to the reactor through a closed piping system, to be reheated. On the turbine-generator side, the water passes over the outside, or shell side, of the steam generator tubes and is heated. Heat transfer through the tubes provides the energy to convert liquid water to steam on the shell side of

the steam generator. Moisture (entrained liquid water) is removed from the steam and the steam flows to a conventional steam turbine. The steam turbine is coupled to a conventional electric generator. After steam passes through the steam turbine, it passes through another heat exchanger called a “condenser,” where the steam is condensed to liquid water. The condensed water from the steam turbine is then pumped through heat exchangers called “reheaters” and is returned to the steam generators in a closed piping loop. The rejected heat from the condenser system is discharged to the environment by evaporative cooling from cooling towers. The electric generator output is supplied to a local transformer yard that is connected to the local electrical grid. The reactor containment building, turbine-generator, and cooling towers are the prominent buildings visible at most nuclear power plant sites.

2.2 Fuel Design Characteristics

There are numerous PWR nuclear fuel assembly designs. However, most PWR power reactor fuel designs will be conceptually similar to the fuel assembly schematic shown in Fig. 5. The PWR uranium dioxide (UO_2) fuel pellets are nominally 0.375 inch in diameter and 0.5-inch-long solid cylinders. From 200 to 300 fuel pellets are loaded into hollow zirconium fuel tubes called fuel rods. A column of fuel pellets, typically 12 feet (144 inches) long, is loaded into a fuel rod. The PWR fuel tube is typically 150 inches in length. The nominal 6-inch length of the fuel rod that does not contain fuel pellets is called the “plenum.” The plenum provides a space to accommodate fuel pellet axial thermal expansion during operation and the fission product gases that are released during power operations. A spring is inserted into the top of the fuel rod to hold the fuel pellets down in the fuel tube during transportation to the reactor and end caps are welded onto the upper and lower ends of the fuel rod. Fuel rods are inserted into spacer grids that are fabricated from a zirconium or inconel alloy. Fuel rods are held in place in the spacer grids by small leaf springs. Control rod guide tubes replace fuel rods typically at 20 to 25 fuel rod locations. The spacer grids are welded to control rod guide tubes. Some fuel assemblies may have a central fuel rod location used for in-core instrumentation. Depending on fuel assembly design, from 196 to 289 fuel rod locations are available in a PWR fuel assembly. Fuel assembly designs normally arrange fuel rods in a square array with 14, 15, 16, or 17 fuel rod locations on a side. At

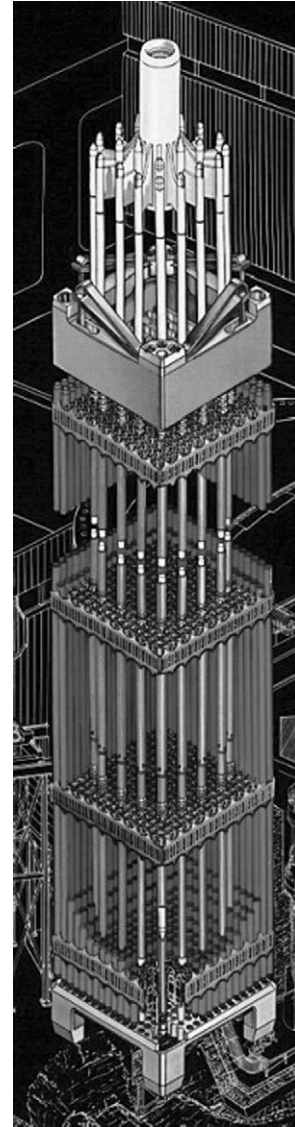


FIGURE 5 Schematic view of a pressurized water reactor fuel assembly.

the top and bottom of the fuel assembly are end fittings. The end fittings are typically machined from stainless steel and provide structural strength to the fuel assembly and help control inlet and outlet flow from the reactor core during operation. In the reactor pressure vessel, 120 to 193 fuel assemblies are arranged into an approximately cylindrical geometry.

Fission of uranium dioxide fuel in the reactor core is the heat source in a nuclear reactor. Fuel typically operates in a reactor for 3–5 years. During this time, the equivalent of 3–5% of the uranium atoms initially loaded into the reactor fission, leaving fission products in the fuel. Plutonium and other actinides are also produced by neutron capture. At the end of its

power-generating life, fuel is discharged from the reactor and replaced with fresh fuel. The discharged fuel is called “spent nuclear fuel” or “spent fuel.” Though some fission products in spent fuel are stable (nonradioactive) or decay to nonradioactive isotopes, spent fuel remains radioactive for very long periods of time following discharge from the reactor. Spent fuel generates decay heat and penetrating radiation (high-energy γ rays and β particles). Spent fuel is stored at reactor sites for several years after discharge, typically 5 or more years. There are two methods of disposition of spent fuel. It is either shipped (in shipping packages designed for radioactive material) to a nuclear fuel reprocessing facility for recovery of residual uranium and plutonium or shipped to a high-level radioactive waste repository for disposal. Table I provides a summary of key design characteristics of a typical pressurized water reactor.

3. BOILING WATER REACTORS

3.1 Reactor Design

Figure 6 provides a schematic view of a boiling water reactor power plant. The power plant systems include a nuclear steam supply system, which produces steam using energy from nuclear fission; a reactor “containment system,” which limits release of radioactive material from the reactor under accident conditions; a steam-powered turbine-generator that produces electricity; and a condenser and cooling towers for heat rejection.

The nuclear steam supply system includes nuclear fuel (discussed later) that is configured as the “reactor core” in a reactor pressure vessel. Reactor coolant water is pumped through the reactor core and heated. At the BWR system operating pressure, liquid water is converted to steam as the water is heated. Moisture is removed from the steam in a moisture separator located in the reactor pressure vessel above the reactor core. The steam flows out of the reactor building to a conventional steam turbine. The turbine-generator, condenser, and cooling towers for a BWR are similar to those for a PWR.

3.2 Fuel Design Characteristics

There are numerous BWR nuclear fuel assembly designs, but most BWR power reactor fuel designs will be conceptually similar to the fuel assembly schematic shown in Fig. 7. The BWR uranium dioxide (UO_2) fuel pellets are nominally 0.4 inch in diameter and 0.4-inch-long solid cylinders. From 200

TABLE I

Pressurized Water Reactor and Fuel Design Parameters

Design parameter	Typical value
Thermal power	3411 MWt
Electrical power	1100 MWe
Reactor pressure	2250 psia
Reactor coolant inlet temperature	560°F
Reactor coolant outlet temperature	620°F
Reactor coolant flow rate	140×10^6 lb m/hr
Number of fuel assemblies in core	193
Number of steam generators	4
Steam turbine inlet temperature	541°F
Steam turbine inlet pressure	970 psia
Fuel rod array	17×17
Fuel rods per assembly	264
Control rod locations per assembly	25
Fuel assembly cross section	8.426×8.426 inches
Fuel assembly length	160 inches
Fuel rod outer diameter	0.36 inch
Fuel rod clad thickness	0.0225 inch
Fuel pellet outer diameter	0.3088 inch
Fuel pellet length	0.53 inch
Fuel assembly weight	535 kg
Fuel assembly uranium weight	423 kg

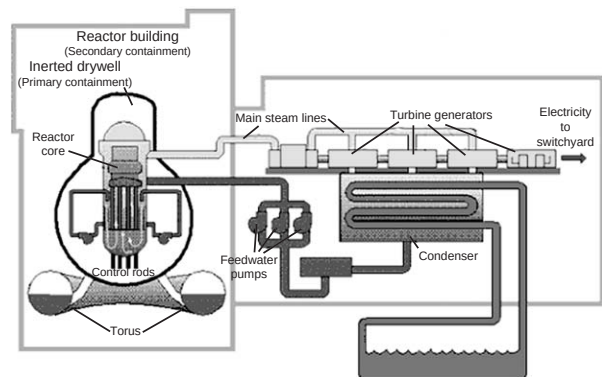


FIGURE 6 Schematic view of a boiling water reactor system.

to 300 fuel pellets are loaded into hollow zirconium fuel tubes called fuel rods. A column of fuel pellets, typically 150 inches long, is loaded into a fuel rod. The fuel tube is nominally 160 inches in length. The nominal unfilled length of the fuel rod that does not contain fuel pellets is called the “plenum.” The plenum provides a space to accommodate fuel pellet axial thermal expansion during operation and fission product gases that are released during power operations. A spring is inserted into the top of the

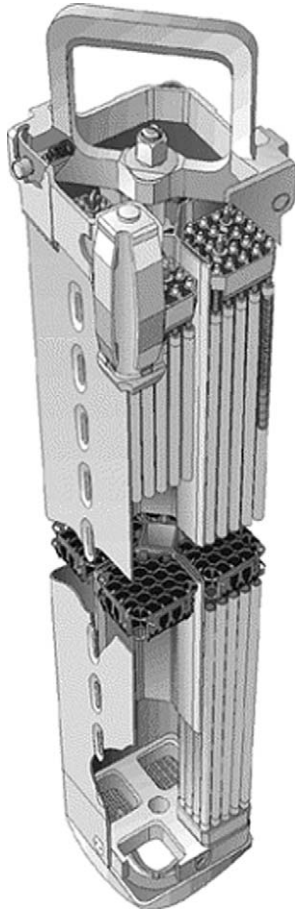


FIGURE 7 Schematic view of a boiling water reactor fuel assembly.

fuel rod to hold the fuel pellets down in the fuel tube during transportation to the reactor and end caps are welded onto the upper and lower ends of the fuel rod. Fuel rods are inserted into spacer grids that are fabricated from a zirconium or inconel alloy. Fuel rods are held in place by small leaf springs in the spacer grids. Depending on the fuel assembly design, from 36 to 81 fuel rod locations are available in a BWR fuel assembly. Most fuel assembly designs arrange fuel rods in a square array, with 7, 8, or 9 fuel rod locations on a side. Depending on the fuel assembly design, several fuel rod locations contain an open-ended zirconium alloy rod to allow water to flow the length of the fuel assembly as water boils in the rest of the reactor core. This open-ended rod is known as a “water rod.” At the top and bottom of the fuel assembly are end fittings. The end fittings are typically machined from stainless steel and provide structural strength to the fuel assembly and help control inlet and outlet flow from the reactor core

TABLE II

Boiling Water Reactor and Fuel Design Parameters

Design parameter	Typical value
Thermal power	3579 MWt
Electrical power	1150 MWe
Reactor pressure	1050 psia
Reactor coolant inlet temperature	533°F
Reactor coolant outlet temperature	551°F
Reactor coolant flow rate	104×10^6 lb m/hr
Number of fuel assemblies in core	800
Steam turbine inlet temperature	540°F
Steam turbine inlet pressure	965 psia
Fuel rod array	8×8
Fuel rods per assembly	62
Water rod locations per assembly	2
Fuel assembly cross section	5.4×5.4 inches
Fuel assembly length	176 inches
Fuel channel thickness	0.12 inch
Fuel rod outer diameter	0.483 inch
Fuel rod clad thickness	0.032 inch
Fuel pellet outer diameter	0.41 inch
Fuel pellet length	0.41 inch
Fuel assembly weight	220 kg
Fuel assembly uranium weight	183 kg

during operation. A BWR fuel assembly has a square cross section fuel channel that encloses the fuel assembly on all four sides. The fuel channel limits the flow of water from one fuel assembly to adjacent fuel assemblies in the reactor core. In the reactor pressure vessel, 400–800 fuel assemblies are arranged into an approximately cylindrical geometry.

As with the PWR, BWR fuel typically operates in a reactor for 3–5 years. The general characteristics of BWR and PWR spent fuels are similar. BWR spent fuel is stored on-site and is eventually shipped off-site for reprocessing or to a high-level radioactive waste repository for disposal. Table II provides a summary of key design characteristics of a typical boiling water reactor.

4. SHIP PROPULSION

Nuclear power is particularly suitable for ships, which need to be at sea for long periods without refueling, or for powerful submarine propulsion. Today, over 150 ships are powered by small nuclear reactors. The United States Navy operates some 100

nuclear ships. Though most nuclear-powered vessels are submarines, they range in type from icebreakers to aircraft carriers.

The development of nuclear propulsion marked the transition of submarines from slow underwater vessels to warships capable of sustaining 20–25 knots while submerged for months. The *Nautilus* led to the parallel development of further submarines, powered by single pressurized water reactors, and an aircraft carrier, the *USS Enterprise*, powered by eight reactor units in 1960. A cruiser, the *USS Long Beach*, followed in 1961 and was powered by two of these early reactor units. Remarkably, the *Enterprise* remains in service. Over time, standardized reactor designs for naval propulsion were developed and built by both Westinghouse and General Electric. The U.S. technology was shared with Britain; nuclear-powered ship development in France, the former Soviet Union, and China proceeded separately.

The Soviet Union built 245 nuclear submarines between 1950 and 1994. The largest submarines are the *Typhoon* class, powered by twin 190-MWt PWR reactors. The Soviet Union developed both PWR and lead-bismuth cooled reactor designs, the latter using highly enriched (>90%) fuel. Eventually, three generations of Russian submarine PWRs were utilized, the last entering service in 1987. At the end of the Cold War, in 1989, there were over 400 nuclear-powered submarines operational or being built. Some 250 of these submarines have now been scrapped and some on order were canceled, due to weapons reduction programs. Russia and the United States had over 100 each; the United Kingdom and France had less than 20 each and China had 6. The United States is the main navy with nuclear-powered aircraft carriers; both the United States and Russia have had nuclear-powered cruisers. Russia has eight nuclear icebreakers in service or being built.

Development of nuclear merchant ships began in the 1950s but has not been commercially successful. The U.S.-built *NS Savannah* was commissioned in 1962 and decommissioned 8 years later. It was a technical success, but not economically viable. The German-built *Otto Hahn* cargo ship and research facility sailed some 650,000 nautical miles on 126 voyages in 10 years without any technical problems. However, it proved too expensive to operate and was converted to diesel. The Japanese *Mutsu* was the third civil vessel. It faced technical and political problems. These three vessels used reactors with low-enriched uranium fuel.

In contrast, nuclear propulsion has proved both technically and economically feasible in the Soviet Arctic. The power levels and energy required for icebreaking, coupled with refueling difficulties for other types of vessels, are significant factors. The icebreaker *Lenin* was the world's first nuclear-powered surface vessel and remained in service for 30 years, though new reactors were fitted in 1970. It led to a series of larger icebreakers, the *Arktika* class, launched beginning in 1975. These vessels have two reactors and are used in deep Arctic waters. The *Arktika* was the first surface vessel to reach the North Pole. For use in shallow waters such as estuaries and rivers, shallow-draft *Taymyr*-class icebreakers with one reactor are being built in Finland and fitted with their nuclear steam supply system in Russia. They are built to conform to international safety standards for nuclear vessels.

Naval reactors are pressurized water types, which differ from commercial reactors producing electricity in that they have high power density in a small volume and therefore run on highly enriched uranium (>20% uranium-235) and long core lives, so that refueling is needed only after 10 or more years. New cores are designed to last 50 years in carriers and 30–40 years in submarines. Decommissioning nuclear-powered submarines has become a major task for both the United States and Russia. After removing the spent fuel, normal practice is to cut the reactor section from the vessel for disposal in shallow land burial as low-level waste.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

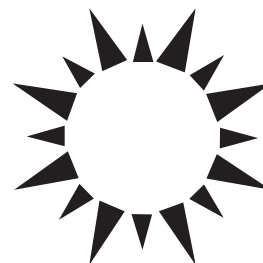
Duke Power Company (1996). "McGuire Nuclear Station Updated Final Safety Analysis Report, May 1996." Duke Power Company, Charlotte, North Carolina.
International Atomic Energy Agency (IAEA). (2001). "Energy, Electricity and Nuclear Power Estimates for the Period up to

- 2020.” Reference Data Series No. 1, July 2001. IAEA, Vienna, Austria.
- Mississippi Power and Light Company (1995). “Grand Gulf Final Safety Analysis Report, December 1995.” Mississippi Power and Light Company (Entergy Corporation), New Orleans.
- Uranium Information Center (2002). “Outline History of Nuclear Energy.” Nuclear Issues Briefing Paper # 50, February 2002. Uranium Information Center, Melbourne, Australia.
- Uranium Information Center (2002). “Nuclear-Powered Ships.” Nuclear Issues Briefing Paper # 32, June 2002. Uranium Information Center, Melbourne, Australia.

Nuclear Fuel: Design and Fabrication

LEON C. WALTERS

Argonne National Laboratory
Idaho Falls, Idaho, United States



-
1. Introduction
 2. Design Considerations
 3. Fabrication
 4. Performance
 5. Summary

Glossary

blanket fuel Nuclear reactor fuel that contains the fertile isotopes that are bred into fissionable isotopes.

driver fuel Nuclear reactor fuel that contains the fissionable isotopes along with fertile isotopes that are bred into fissionable isotopes.

fertile isotopes Common fertile isotopes are ^{238}U and ^{232}Th .

fissionable isotopes Common fissionable isotopes are ^{235}U , ^{233}U , and ^{239}Pu .

natural uranium Uranium as it is mined from the earth. The composition is 99.3% ^{238}U and 0.7% ^{235}U .

Successful nuclear fuel must meet the requirements set forth by the nuclear reactor design team. These requirements include considerations of safety, economy, and ease of reprocessing or disposal. The environment within which nuclear fuel must perform, design considerations, fabrication, and performance are reviewed for nuclear fuel designs that have survived beyond the experimental stage.

1. INTRODUCTION

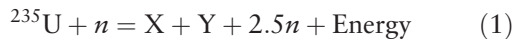
Economically recoverable uranium in the world is estimated to be 4 million tons. This uranium is

composed of two isotopic forms, ^{235}U and ^{238}U . The two forms differ, with ^{235}U having three fewer neutrons in the nucleus. Otherwise, the two forms are identical chemically, meaning that they have the same number of electrons and protons. The uranium that is mined from the earth is called natural uranium. Natural uranium contains 0.7% ^{235}U , with the balance being ^{238}U . Some nuclear reactors were designed to operate with fuel made of natural uranium, whereas other reactor designs required fuel with a concentration of ^{235}U that is higher than 0.7%. For these latter reactor designs, the fuel had to be enriched in ^{235}U , over a range of concentrations, to as high as 93% ^{235}U for certain reactors.

Uranium that remains after the enrichment process is called depleted uranium because the ^{235}U concentration has been reduced to 0.3% or less. This depleted uranium can be converted to a fissionable isotope, ^{239}Pu , by a process called breeding. In reactors designed to optimize this breeding process, neutrons are absorbed by the nucleus of ^{238}U and the ^{238}U is converted to ^{239}Pu . It also should be pointed out that thorium can be converted to the fissionable isotope ^{233}U by the same neutron capture process. The reserves of thorium in the earth are far larger than those for uranium. Thus, if all the recoverable uranium and thorium in the world were converted to fissionable isotopes, there would be an ample energy supply for humankind for several millennia.

Reactor driver fuel is composed of the element or compounds of the fissionable isotopes ^{235}U , ^{233}U , and ^{239}Pu along with the fertile isotopes ^{238}U and ^{232}Th . Some fuels contain mixtures of more than one fissionable isotope. Reactor blanket fuel is composed of the element or compounds of the fertile isotopes ^{238}U and ^{232}Th . Fertile is a term used to indicate that the isotope can be bred to a fissionable isotope.

The important nuclear processes that take place in the fuel are the following:



Most nuclear reactors deployed are fueled with uranium that either has the natural enrichment of 0.7% ^{235}U or is somewhat enriched in ^{235}U , to approximately 5%. These reactors, called thermal reactors, generate their power with new fuel according to fission reaction 1. The energy released in the fission reaction 1 is equivalent to 1 MW day for each gram fissioned. As the ^{235}U is fissioned, some of the ^{238}U in the fuel is converted to ^{239}Pu according to neutron capture reaction 2. The ^{239}Pu in turn is fissioned with approximately the same energy as for reaction 1. A fact that seems to go unnoticed is that approximately 40% of the power from a thermal reactor originates from plutonium at the end of the life of the fuel, in approximately 4 years. Thorium is converted to ^{233}U by the same process of neutron capture, reaction 3.

The X and Y of reaction 1 are called fission products. They represent most of the elements of the periodic chart, they are all radioactive, with relatively short half-lives, and they can exist in the fuel as a solid, as a gas, or, under extreme conditions, as a liquid. Their presence in the fuel and their chemical and physical behavior are critical to the performance of the fuel.

Another series of successive neutron capture reactions occurs, such as reactions 2 and 3. The products of these neutron capture reactions are the so-called minor actinides, americium, curium, and neptunium. They are also radioactive but their half-lives are extremely long. Their presence in the spent fuel is the reason for high-level waste repositories that must maintain their integrity for millennia.

The coolant in a thermal reactor is generally water or occasionally helium gas. The excess neutrons from fission reaction 1 are slowed down or moderated by nuclear collisions in the water or, in the case of the gas-cooled reactor, with graphite designed for the purpose of slowing down the neutrons. The fission reaction produces highly energetic neutrons that are slowed down to near equilibrium with the thermal motion of the surrounding material. The name “thermal” reactor thus arises by virtue of this slowing-down process. These low-energy neutrons

have a higher likelihood of producing fission according to reaction 1.

The ^{238}U that is converted to the fissile ^{239}Pu is called fertile blanket fuel. The ^{232}Th that is converted to the fissile ^{233}U is also called fertile blanket material. A “fast” breeder reactor is specifically designed to optimize the conversion of fertile material to fissile material by elimination of as much moderating material as possible to keep the neutrons as energetic as possible. When designed properly, the fast breeder reactor can easily produce more fissile material from the fertile blanket material than is fissioned to produce energy. The energy density in a fast reactor is much higher than that of a thermal reactor and thus liquid metals, such as sodium, which has a much higher heat capacity than water, are used as the flowing coolant.

Chicago Pile 1 was the first nuclear reactor that achieved criticality on December 2, 1942. It was designed, built, and brought to the world’s first chain reaction by a team led by Enrico Fermi. The fuel for this reactor was natural uranium dioxide pellets placed in a matrix of graphite blocks that served as the neutron moderator. The reactor produced little power and thus the fuel design was of the most rudimentary nature, being simply unclad pellets. Since that time there have been a wide variety of nuclear fuels developed, the designs for which were dependent on the mission of the nuclear reactor. Nuclear reactors have been constructed for electricity generation, breeding of uranium and thorium for the production of fissile isotopes, production of medical isotopes, tumor irradiation, and ship and space propulsion.

This article discusses the phenomena that must be considered for the design of any fuel type and then shows how the effects of these phenomena have been accommodated in the design of specific fuel types. This discussion is limited mostly to fuel types that have survived the early experimental ventures and have gone on to be candidates for commercial deployment.

2. DESIGN CONSIDERATIONS

The geometry of fuel has many variations. The radial dimension, the dimension toward the coolant, is always small and is dictated by the necessity that the energy density of the fuel during the fissioning process is enormous and a great deal of heat must be extracted to prevent the melting of the fuel. Figure 1 is a simplified drawing of a generic fuel pin, shown only

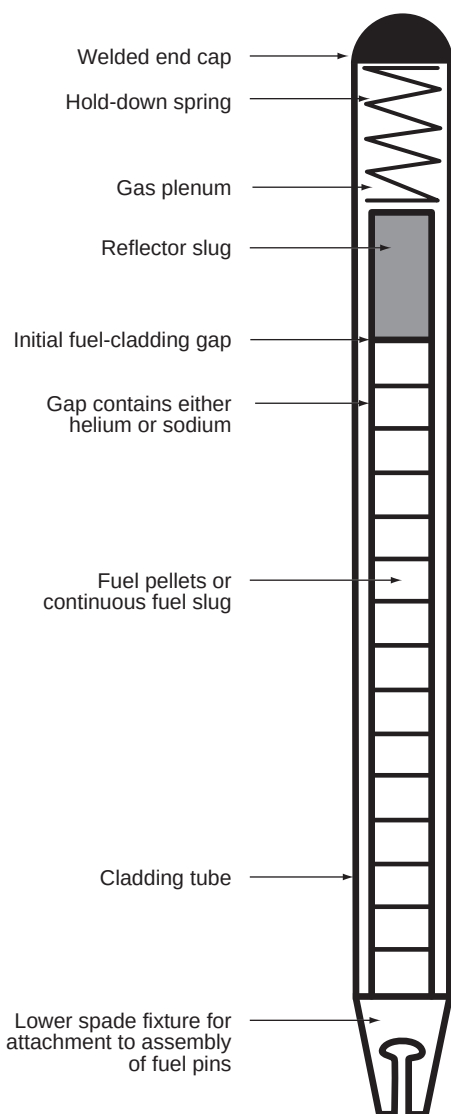


FIGURE 1 Schematic of a generic fuel pin illustrating important components.

to illustrate particular design features. Because of the small dimension between the center of the fuel and coolant, the temperature gradient is large, and this large temperature gradient causes cracking of the fuel, drives mass flow, and creates stress between the fuel and cladding due to thermal expansion differences.

The cladding is generally a metal tube that is sealed around the fuel. The cladding serves several purposes. It is usually referred to as the first barrier against the release of radioactive elements to the environment. As such, its integrity is a very important consideration. For some fuel designs, the cladding is necessary to support the fuel column and to preserve the geometry and heat transfer path of a bundle of

fuel pins. Furthermore, the cladding, for many fuel pin designs, is necessary to contain a fuel transfer medium in the gap between the fuel column and the cladding. This fuel transfer medium could be either a gas, such as helium, or a liquid, such as sodium.

The environment within which the fuel pin must perform is severe. When the reactor is brought to power, a number of events occur simultaneously. High temperatures and high temperature gradients are imposed on the fuel pins. A number of high-energy particles begin disrupting the matrices of the fuel and cladding. The fission product elements cause the fuel to swell, eventually reaching the cladding. The stress generated can eventually breach the cladding. This is called the fuel-cladding mechanical interaction. Some also call this phenomenon the pellet-cladding interaction. Fission products that form gases escape from the fuel and create additional stress on the cladding as the pressure builds with time. Some of the fission products, such as iodine and cesium, may attack the cladding by corrosion reactions.

Cladding surrounding the fuel column experiences a similar set of challenges. Not only is the cladding subject to high-temperature corrosion processes from the inside but the coolant flowing on the outside of the cladding, whether it be water, liquid metal, or gas, may be corrosive to the cladding. Furthermore, the physical properties of the cladding may change dramatically under prolonged exposure to neutron irradiation. Neutron irradiation tends to embrittle most metals, some metals swell under neutron bombardment, and processes that depend on diffusion of material, such as carbide precipitation and creep, may be greatly accelerated by neutron irradiation.

Within the environment just described, the reactor designer will lay forth a set of requirements that he or she anticipates can be met by the fuel system. The inlet and outlet temperatures of the coolant will be specified, which fixes the best efficiency of the system. The reactor designer designates the exposure time of the fuel in the reactor, which is usually measured in terms of the percentage of uranium and plutonium fissioned in the fuel or the energy generated per ton of fuel. Exposure time of the fuel is important as an economic measure for the supply of fuel that must be manufactured as well as the amount of spent fuel that must be stored or reprocessed. Should the fuel be reprocessed by a particular reprocessing scheme, then the fuel must be compatible with the scheme.

The regulatory requirements demand that the fuel system survive possible off-normal events without impact to the public. Off-normal events could

involve loss-of-coolant flow conditions where the temperature of the fuel gradually increases above normal conditions over an extended period of time or more rapid events where a control rod might be inadvertently removed.

The reactor designer and regulatory system impose a set of expectations for the nuclear reactor fuel system and the environment in the reactor imposes its own set of constraints. A final qualified fuel design is always the result of many iterations and compromises. The fuel development team, usually composed of materials scientists, is supported by experts in neutronics, thermal hydraulics, core design, cladding and duct development, nuclear safety, chemistry, remote examination technology, experiment vehicle design, analytical chemistry, microscopy, and more.

3. FABRICATION

For a new fuel type, the means of fabricating the fuel is always an iterative and challenging process. Generally, small versions of the envisioned full-scale fuel pin are fabricated, irradiated in a reactor under prototypic conditions, and examined. The range of fabrication variables may be quite extensive at this experimental stage as the search proceeds for the best combination. Through a series of irradiations and examinations, the variable range is narrowed until an interim fuel fabrication specification is agreed on. At this point, a number, perhaps six, of full-size assemblies of fuel pins are irradiated under prototypic conditions to their end of life, which generally means to the point of cladding failure. Along the way, during this test period, a few of the fuel pins would be removed from the assemblies for testing under off-normal conditions in a reactor specifically designed for that purpose. If the results of these tests did not meet expectations, then there may be additional tests on improved versions. Should the results meet expectations, the fabrication specifications would be frozen. Strict quality assurance programs would be invoked in the fabrication process to ensure that the fuel performance would be repeatable.

The process of moving from a nuclear fuel concept, dictated by a reactor design, to a fully qualified and licensed fuel system takes 10 years or more. The schedule for conducting such a program is controlled by the time it takes to complete the fabrication, irradiation, and examination of a novel fuel type. Irradiation of nuclear fuel takes 4 to 5

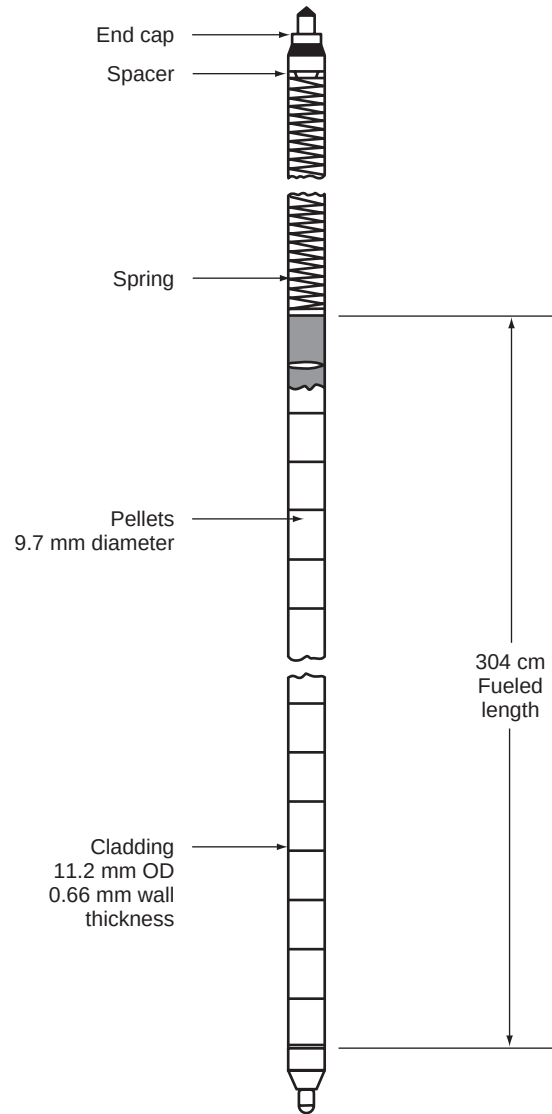


FIGURE 2 Fuel pin of a pressurized water reactor.

years to complete and generally more than one iteration is required.

Figures 2, 3, and 4 show three of the many fuel types that have been developed. They were chosen because they represent classes of fuel pins that are commercially deployed or they are under active investigation for the next generation of reactors. Figure 2 shows the fuel pin utilized in a typical pressurized light water reactor. The fuel pin has evolved from nearly 50 years of utilization, first in experimental reactors and for the past 40 years in commercial reactors. The fuel column is a stack of uranium oxide or uranium and plutonium oxide fuel pellets that is held in place with springs. Prior to assembly, the cladding tube that surrounds the pellet

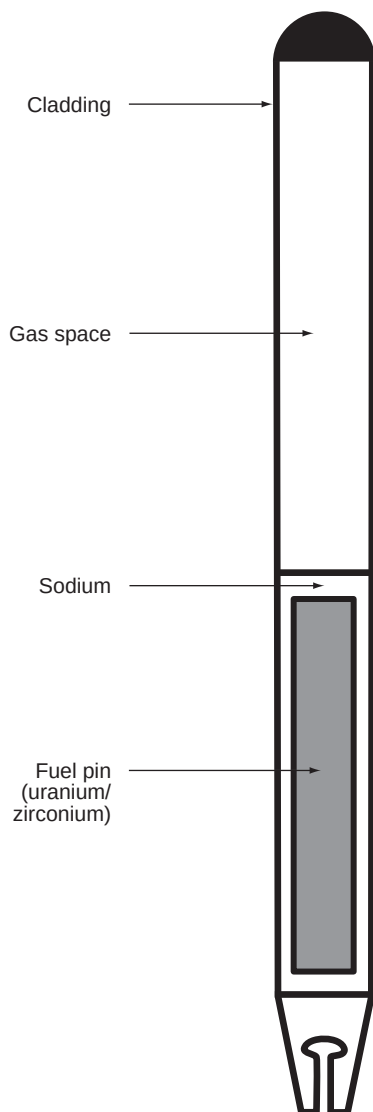


FIGURE 3 Fast reactor metal-fuel pin.

stack is filled with helium at a somewhat elevated pressure. The helium gas serves as a good heat transfer medium and the elevated pressure of the helium prevents the cladding from collapsing around the fuel column from the external pressure of the water in the pressurized water reactor. The technology for hot-pressing the oxide pellets has been well perfected to achieve high and uniform density as well as consistency in stoichiometry. The zirconium and zirconium alloy cladding has evolved over the years to a composition and heat treatment that are resistant to internal and external corrosion effects.

Figure 3 shows a fast reactor fuel pin with a metallic fuel column. The fuel column for a fast reactor fuel pin can be uranium and plutonium

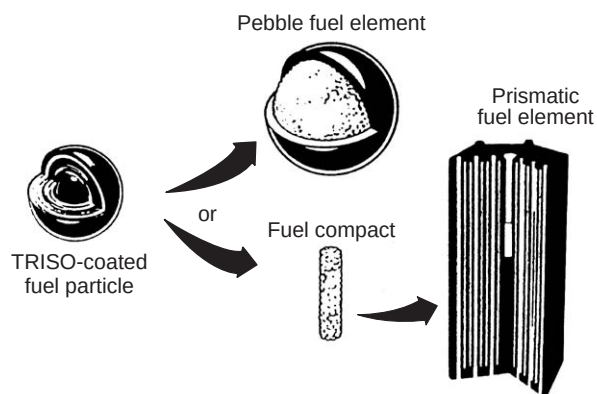


FIGURE 4 Particle fuel element.

oxide, carbide, nitride, or an alloy of uranium and plutonium. The cladding is steel. No clear choice exists for the most preferable fuel column type. Each type has its attributes. Uranium–plutonium oxide is by far the most commonly used fast reactor fuel. The PUREX (plutonium uranium extraction) fuel reprocessing technology, used for light water oxide fuel, also is acceptable for the oxide fast reactor fuel. Alternatives, but less mature reprocessing schemes, have been developed for the other fast reactor fuel types. The first cladding material for fast reactor fuel was commercial grades of austenitic stainless steel. These steels tended to swell when subjected to a high level of neutron irradiation. The steel alloys were improved to the point where little swelling occurs and the fast reactor fuels are capable of achieving very high fissile burn-up.

Most fuel pin designs are cylindrical in shape, for both fast reactor and light water reactors. Figure 4 shows the fuel design for a type of high-temperature gas-cooled reactor (HTGR) that is a departure from the cylindrical designs. These fuel types have undergone successful development in several countries; e.g., South Africa is implementing a program for commercial deployment of these HTGRs. The diameter of the fuel particle in Fig. 4 is dependent on the particular design. The central sphere, called the kernel, contains the fissile material and ranges from 150 to 500 μm , depending on the design. The kernel is composed of uranium oxide, uranium carbide, as well as thorium included with the uranium. The kernel is embedded in two layers, the BISO concept, or the now accepted three-layer version, called the TRISO concept. BISO and TRISO stand for bi-structural isotropic and tri-structural isotropic, respectively. The isotropic refers to the property of the pyrolytic layers, the bi and tri to the

number of layers. The innermost layer is porous pyrolytic carbon and is called the buffer layer. This layer supplies the void space to accommodate fission-induced swelling and gas release from the kernel. The next layer is silicon carbide, which is the load-bearing layer that contains the pressure of the fission gases. The outer pyrolytic layer protects the silicon carbide layer from reaction with the coolant impurities. Each of the layers is 30 to 50 μm thick. These small TRISO particles are further embedded either in graphite spheres for the pebble bed reactor or in cylindrical fuel compacts that are located in vertical channels in hexagonal graphite blocks in the prismatic-compact reactor. Approximately 10,000 coated particles are contained in each 6 cm diameter graphite sphere.

4. PERFORMANCE

4.1 Metal Fuel

Metal fuel was among the first fuels to be used in nuclear reactors. The Experimental Breeder Reactor I, the first nuclear reactor to generate useful electricity, on December 20, 1951, used metal uranium and plutonium fuels. Subsequently, several fast reactors used metal fuel. Metal fuel was not adopted for use in water-cooled reactors because, in the event of a cladding breach, the metal fuel was not compatible with the water coolant. The fuel will form metal hydrides and oxides when in contact with water at elevated temperature. However, metal fuel continued to be of interest for use in fast reactors.

In the mid-1960s, when the large commercial prototype fast reactors were being designed, metal fuels were largely abandoned in favor of oxide fuel. Metal fuel at that time was unable to achieve high burn-up and there was a concern that metal fuel would be unable to reach the high temperatures that reactor designers desired without the formation of a liquid phase between the fuel and the cladding. Even though the performance information on oxide was limited, it was thought that oxide fuel would not be limited in these respects.

Metal fuel possesses certain unique attributes and thus research continued to improve its performance. Metal fuel has a much higher fissile element density than all other fuels, which allows a metal fuel reactor to convert more uranium to plutonium. Metal fuel is easier and cheaper to fabricate, especially in a hot-cell environment, on reprocessed fuel. The thermal conductivity of metal fuel is higher than that of other

fuels and this brings with it the possibility of designing large commercial reactors that are inherently safe, that is, reactors that will shut themselves down during an upset event, without any human intervention.

A design modification that allowed metal fuel to achieve very high burn-up was discovered. Early metal fuel was designed such that there was no gap or only a small gap between fuel and cladding. During irradiation, the accumulation of fission products would cause the fuel to swell. The gaseous fission products would nucleate into small bubbles, and when these bubbles grew into a critical size, the gas pressure in the bubble would overcome the surface tension of the uranium matrix and the fuel would swell rapidly. Compounding the swelling from the gaseous fission products was the swelling component due to the accumulation of solid fission products in the uranium matrix. The fuel swelling would cause the cladding to breach after just a few months at a burn-up of approximately 1% of the heavy metal.

A simple design change avoided the problem of early cladding breaches due to fuel swelling. The gap between the fuel column and the cladding was widened to allow the fuel to swell freely until the volume change of the fuel reached approximately 30%. When the volume change reached approximately 30%, the porosity in the fuel, caused by the fission gas bubbles, would interconnect and the fission gas would be released to a plenum volume above the fuel column. The plenum volume above the fuel was sized such that the stress on the cladding during the life of the fuel pin would remain reasonably low. Figure 5 shows the fractional gas release as a function of volume change for several metal fuel compositions. With this design change, metal fuel pins reached burn-ups in excess of 20% and were capable of remaining in the reactor for 3 to 4 years.

As with all fuel types, the fuel restructures substantially during irradiation. Figure 6 shows the cross section of a uranium-plutonium-zirconium alloy fuel pin at various elevations after 17% burn-up. The ring formation is a result of diffusion, driven by the temperature gradient, where the ring boundaries are phase boundaries. The interconnected porosity is evident throughout the cross sections. The centerline temperature of the fuel is approximately 750°C, whereas the cladding temperature is approximately 600°C.

The other perceived issue with metal fuels, that of a liquid phase forming before the design temperature is reached, never turned out to be a problem.

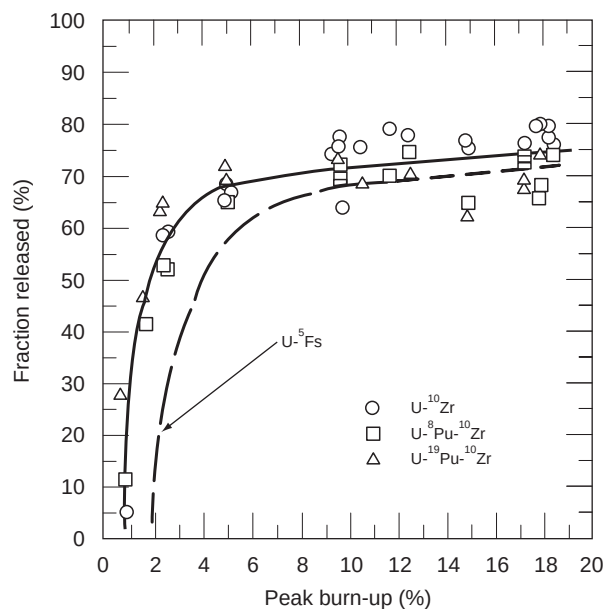


FIGURE 5 Fission gas released to the plenum, above the fuel, for various metal fuels as a function of burn-up.

Cladding steels had a peak useful temperature of no more than 600°C, even though reactor designers desired higher temperatures. At approximately this temperature, the steels used experience a decrease in yield strength such that the stress on the cladding from the fission gas and fuel swelling would cause the cladding to deform and fail. The temperature of 600°C was below the temperature where a liquid phase would form between the fuel and the cladding.

A phenomenon that persisted to limit the capability of fast reactor fuels, whether the fuel was metal or ceramic, was the swelling of the cladding and duct materials due to the neutron damage. Figure 7 shows the progressive improvement in cladding deformation. 304L SA and 316 SA are common solution-annealed austenitic stainless steels. D9 20% CW is a titanium modified 316 stainless steel that has been cold-worked, while HT-9 20% CW is a ferritic steel that has been cold-worked and has, by far, the best performance. This issue is no longer considered a limiting factor with the utilization of ferritic steels for cladding that show very little irradiation-induced swelling.

4.2 Ceramic Fuels

Uranium oxide is by far the most utilized nuclear fuel for commercial power reactors in the world. Mixed uranium and plutonium oxide fuel was the primary choice for sodium-cooled fast reactors and in

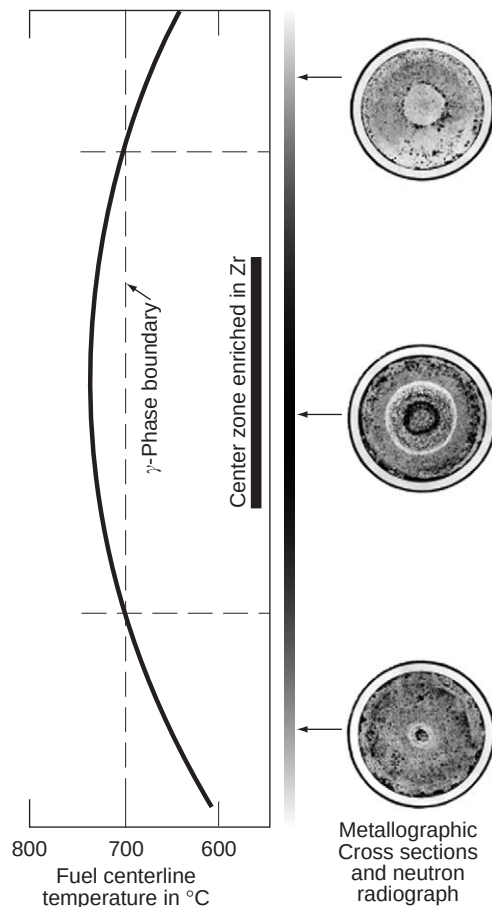


FIGURE 6 Axial temperature profile along the centerline of a U-Pu-Zr pin at 17 at.% burn-up and its relation to radial zone patterns shown in a neutron radiograph and three cross sections.

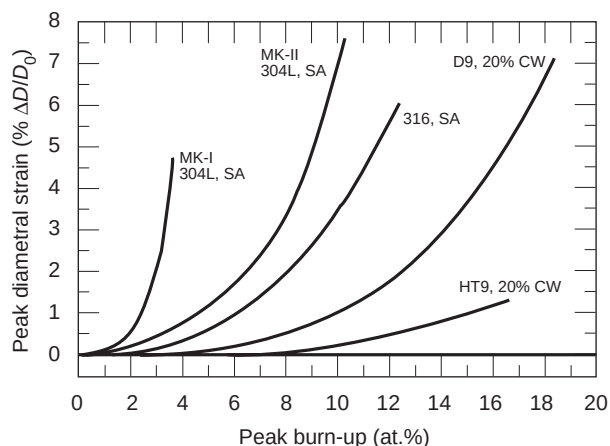


FIGURE 7 Progressive improvement in cladding deformation (swelling and creep) of metal fuel pins.

addition is being increasingly used in thermal reactors as a means of reducing the increasing stockpile of reprocessed plutonium. The fuel in fast

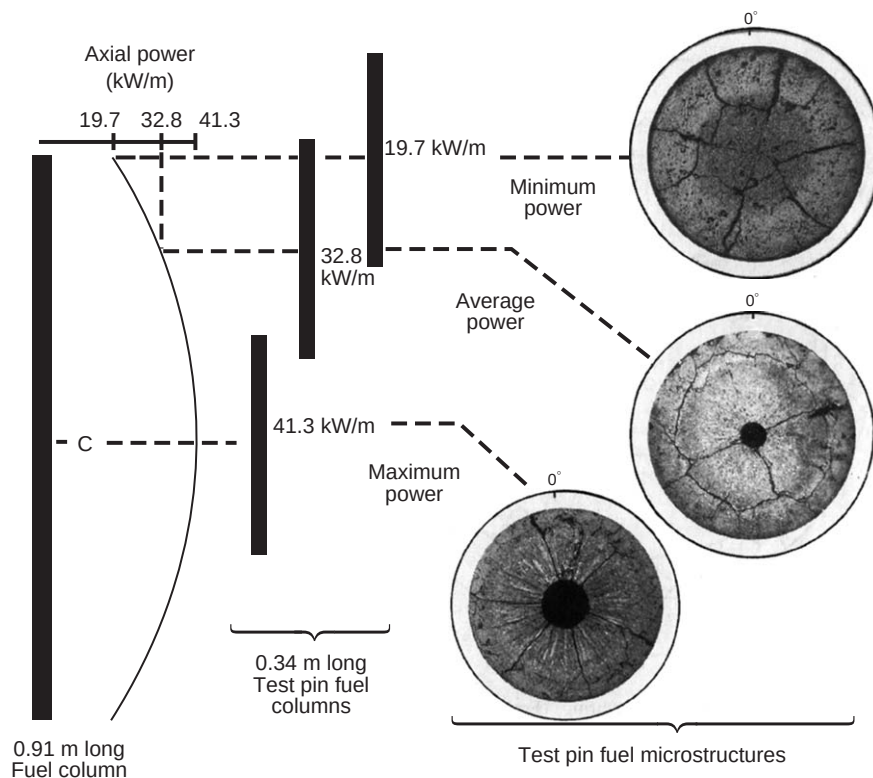


FIGURE 8 Mixed-oxide fuel restructuring versus linear heat rate.

reactors operates at a much higher fissile and energy density and thus the peak fuel and cladding temperatures are much higher. Oxide fuel columns restructure to a great extent because of the low thermal conductivity of the fuel and the high temperature gradient from the center of the fuel to the fuel-cladding interface. The peak centerline temperature of fast reactor fuel is approximately 2100°C , whereas the cladding temperatures reach 600°C in normal operation. The corresponding temperatures for a pressurized water reactor oxide fuel are 1300 and 350°C , respectively. Thus, the restructuring for fast reactor fuel is more severe. Figure 8 shows the axial variation in fuel restructuring for a fast reactor, a uranium-plutonium oxide fuel pin. At the core midplane location, the restructuring leads to a central void. The void results from the transport of the initial porosity up the thermal gradient. The high temperature results in significant grain growth. In addition to grain growth, the large thermal gradient causes a radial redistribution of oxygen and plutonium.

The fission gas that is generated behaves quite differently in thermal and fast reactor fuels. Because of the difference in fuel temperatures, the thermal reactor fuel retains most of the fission gas, whereas

the fast reactor fuel releases most of the fission gas to the gas plenum in the fuel pin. The margins to failure, as a result of the gas retention or release, are considered in the design of the fuel. For thermal reactor fuel, the concern is that the fission gas will be suddenly released during an upset event and overpressurize the cladding. This eventuality has been modeled, understood, and taken into account in the allowable exposure of the fuel element. With the fast reactor fuel, the plenum must be large enough to hold the continuous generation of fission gas without overpressurizing the cladding.

Cladding attack from the volatile fission products, cesium, tellurium, and iodine, is of concern for both fast reactor and thermal reactor fuels although the controlling phenomena are different. With the zirconium and zirconium alloy claddings, early failures were due to stress corrosion cracking from the iodine. In addition, zirconium hydride formation on the outside of the thermal reactor zirconium cladding led to cladding breach. However, these problems have become well understood and were solved, so that modern thermal reactor fuel is highly reliable. For fast reactor cladding, which is steel, the volatile fission products, along with excess oxygen, tend to oxidize the cladding components, such as the

chromium. Control of the initial oxygen-to-metal ratio in the fuel minimized the extent of the oxidation to the point that the oxidation was not a limiting factor in the operation of the fuel.

Development of uranium oxide fuel for thermal reactors and uranium-plutonium fuel for both fast and thermal reactors has matured to the point where they are considered highly reliable and economic fuels. The only disadvantage of oxide fuel for fast reactors is that the fissile density is low, which makes an oxide-fueled fast reactor an inefficient breeder of plutonium.

4.3 Particle Fuels

Particle fuels have a long history of development. The Dragon reactor in the United Kingdom, the Peach Bottom and Fort St. Vrain in the United States, and the AVR and THTR-300 in Germany all operated with particle fuel. The HTTR in Japan and the HTR-10 in China are still operating with particle fuel.

Failure of a fuel pin is generally regarded as the point where the cladding fails and fission products are released to the coolant. In the case of particle fuel, failure is the point where the protective coatings around the particle fail and fission products are released to the coolant. The “ameba effect” was a failure mechanism recognized in early irradiations of particle fuel. When the particles were subjected to steep temperature gradients at high temperature or when fuel fabrication defects yielded asymmetrical kernel geometry, carbon transport from the high-temperature side of the kernel would tend to migrate to the low-temperature side. The kernel would then tend to move to the high-temperature side, ultimately breaching the carbon and silicon carbide layers. This phenomenon is well understood and is avoided by a combination of the proper kernel composition and the reduction of the imposed temperature gradients across the particle during irradiation.

Excessive anisotropy of the pyrolytic graphite layers, which surround the kernel, has led to failure and fission gas release. The need to define the exact coating conditions to avoid the critical anisotropy levels makes fabrication specifications important, where the recipes must be passed to industry and other laboratories.

Fission products react with the silicon carbide layer. In particular, the palladium group metals diffuse into the silicon carbide layer and form intermetallic compounds that can lead to failure of the layer. By

keeping the temperature and time-in-reactor within known limits, failure of the layer is avoided.

Although in many respects particle fuel is complex, in both fabrication and performance the fuel has been perfected and has been shown to be highly reliable. In Germany, the fuel has been irradiated to 15% burn-up at a temperature of 1250°C. Statistical analysis has shown a failure rate for this fuel of 2×10^{-5} at the 95% confidence level. However, the German experience is better than that experienced in the development of particle fuel by other countries.

4.4 New Fuels

For completeness, new fuels that are under development are briefly discussed in this section. Proliferation issues and concerns surrounding nuclear waste repositories have given rise to a generation of special-purpose nuclear fuels. Relative to the proliferation of nuclear weapons, an issue existed regarding the diversion of highly enriched fuel, which is used in research reactors around the world. A large number of these reactors were converted to low-enriched fuel that consisted of a moderately high density of uranium silicide particles, embedded in an aluminum matrix. The fuel was fabricated in the form of plates. However, the fissile density of this fuel was not high enough to allow the conversion of the larger research reactors. To solve this problem, an effort was initiated to develop a much higher density low-enriched fuel such that these larger research reactors could be converted. This effort has been successful with the development of a fuel plate design that has metal uranium-molybdenum particles embedded in an aluminum matrix.

As discussed, the elements that dictate the longevity of a nuclear spent fuel repository are the actinides—plutonium, americium, neptunium, and curium. With fast reactors and the proper spent fuel reprocessing cycle, these elements are recycled into the reactor and fissioned to achieve useful energy. With the recent and perhaps temporary rejection of fast reactors and the associated fuel cycle due to proliferation concerns, it was proposed that neutron-generating accelerators could fission the actinides that are in the commercial thermal-reactor spent fuels and thus reduce the radioactive burden to nuclear waste repositories. New reactor fuels were required that would contain a high concentration of actinides in a matrix that was inert in the sense that the matrix would contain no uranium that could be converted to higher actinides. A number of promising options are being explored, such as the coated

particle option, where the particle kernel would be an actinide oxide or a carbide. Metal fuels are also being proposed as an option where the actinides would alloy in a matrix of zirconium. A similar option is a mixture of actinide nitrides in a matrix of zirconium nitride. Finally, a series of fuels is being studied where an oxide or other nonmetallic compound of the actinides is embedded in an inert ceramic matrix.

5. SUMMARY

Nuclear fuel development has progressed remarkably over the past 50 years. The fuel residence times in the reactors have increased by more than a factor of 20 and the reliability has increased by several orders of magnitude. Fuel is a small cost of generating power from a nuclear reactor as opposed to fuel costs from other fossil fuel energy technologies.

Nuclear fuel design and fabrication techniques will continue to evolve, with their evolution driven by economics and new uses for nuclear energy. A particular fuel design must not only meet the needs of the reactor designer but also must be economical to fabricate. Furthermore, if the fuel is to be reprocessed, it must be amenable to existing reprocessing schemes or new reprocessing technology must be developed to accommodate the fuel. Some fuel types are subjected only to once-through cycles with no reprocessing. In this case, the spent fuel and secondary containment must meet the requirements of the repository.

Looking to the future, high-temperature nuclear reactors for space travel will surely be commonplace. High-temperature fuels, such as nitride fuel, will continue to be developed for this purpose. Ground transportation will move toward hydrogen as a fuel not only to reduce carbon emissions for environmental benefits but also for many countries to reduce their dependence on foreign oil. Nuclear reactors

with high outlet temperatures when coupled with thermochemical cycles will be used for hydrogen production. Here, particle fuels or yet to be developed fuels will be used in these reactors. Water resources will diminish in this century. Nuclear reactors used for seawater desalination purposes will be developed that require long-life fuel. These reactors could be deployed in developing countries, with no need for refueling for two or three decades. All of these future uses will continue to challenge fuel developers for decades.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

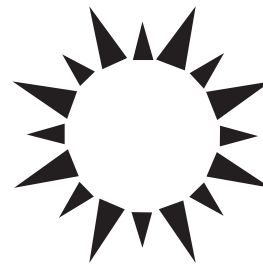
Further Reading

- Cahn, R., Haasen, P., and Kramer, E. (Eds.). (1994). "Materials Science and Technology, Volumes 10A and 10B, Nuclear Materials," Part I. VCH, New York.
- Gulden, T., and Nickel, H. (1977). Coated particle fuels. *Nucl. Technol.* 35, 206–213.
- Leggett, R., and Walters, L. (1993). Status of LMR fuel development in the United States of America. *J. Nucl. Mater.* 204, 23–32.
- Olander, D. (1976). "Fundamental Aspects of Nuclear Reactor Fuel Elements." Technical Information Center, Energy Research and Development Administration, Oak Ridge, TN.
- Weisman, J., and Eckart, R. (1981). Basic elements of light water reactor fuel rod design. *Nucl. Technol.* 53, 326–343.

Nuclear Fuel Reprocessing

HAROLD F. MCFARLANE

Argonne National Laboratory
Argonne, Illinois, United States



-
1. Role of Reprocessing
 2. Spent Nuclear Fuel
 3. Reprocessing
 4. Reprocessing Facilities
 5. Reprocessing Research and Development

Glossary

boiling water reactor A thermal-spectrum nuclear power reactor in which the cooling water is allowed to boil as it rises through the core.

burn-up The fraction of heavy metal atoms fissioned in irradiated nuclear fuel, often expressed as average energy produced per unit mass of initial uranium.

centrifugal contactor The solvent extraction equipment in which separation of the agitated phases is accelerated by centrifugal action.

cladding A metal tube that separates the nuclear material in a fuel element from the reactor coolant.

decontamination The removal of an undesirable component from a product.

depleted uranium The by-product of the uranium enrichment process; uranium containing less than 0.7% ^{235}U .

electrolytic reduction The use of an electric current in an electrolyte to promote the reduction of a ceramic fuel to a metal.

electrorefining The process of decontamination by selectively oxidizing a metal at an anode in an electrolyte while simultaneously reducing it in purer metallic form on a cathode.

enrichment The process of partial isotopic separation in order to increase the fraction of one isotope relative to another, typically ^{235}U relative to ^{238}U .

fast-spectrum reactor A nuclear reactor with a highly energetic neutron spectrum.

fissile An isotope that readily undergoes fission by thermal neutrons.

irradiated nuclear fuel The portion of nuclear fuel that has been exposed to sufficient neutron flux within a reactor to cause measurable nuclear transmutation.

melt refining The partial decontamination of metallic nuclear fuel by a process of melting and allowing fission products to volatilize or react with a substrate.

mixer-settler The equipment for solvent extraction, consisting of a region that thoroughly mixes two liquid phases and a settling region that allows them to separate again.

partitioning The separation of spent fuel into multiple like groups.

plutonium–uranium extraction (PUREX) A solvent extraction process that has been commercialized for reprocessing spent nuclear fuel.

pressurized water reactor A thermal-spectrum nuclear power reactor in which the system pressure is sufficiently high to prevent boiling of the cooling water.

pulsed column The equipment for solvent extraction, consisting of a tall pipe in which the two liquid phases flow in opposite directions, mixed by pressure pulses that force them through perforated disks inside the column.

pyrochemical A high-temperature reprocessing technique that uses selective reduction and oxidation in fused salts or metals to recover nuclear materials.

pyrometallurgical A high-temperature reprocessing technique that processes the materials in metal form.

raffinate The residue of contaminants removed from a product stream.

rare earths The lanthanide group of 14 elements.

reduction–oxidation (REDOX) The first countercurrent solvent extraction process chosen for reprocessing spent nuclear fuel.

reprocessing The separation of irradiated nuclear fuel into potentially useful product materials and waste.

sodium-cooled reactors Fast-spectrum nuclear reactors cooled with molten sodium.

solvent extraction A technique used in chemical separations involving two immiscible solvents in which one or more components are transferred between the solvents.

spent nuclear fuel Assemblies containing nuclear fuel that has reached the end of its economic lifetime.

thermal-spectrum reactor A nuclear reactor in which the neutron spectrum approaches thermal equilibrium with the core.

tonne A metric ton (= 1000 kilograms).

transmutation The act of changing one element into another through nuclear processes.

transuranic extraction (TRUEX) A solvent extraction process for transuranic elements.

tritium The radioactive hydrogen isotope with mass number 3.

vitrification The stabilization of waste in a glass matrix.

zeolite A mineral with a cage-like molecular structure; used as a catalyst or an ion exchanger.

Nuclear fuel reprocessing is the separation of irradiated nuclear fuel into potentially useful product materials and waste. The separation is accomplished by a combination of mechanical, chemical, and physical processes. Conservation of uranium resources and improved management of radioactive waste are the two primary motivations for reprocessing. In principle, almost all of the constituents of irradiated nuclear fuel can be recycled for some further use. In practice, only the recycling of plutonium and uranium has been of commercial interest.

1. ROLE OF REPROCESSING

Reprocessing is essential to closing the nuclear fuel cycle. Natural uranium contains only 0.7% ^{235}U , the fissile isotope that produces most of the fission energy in a nuclear power plant. Prior to being used in commercial nuclear fuel, uranium is typically enriched to 3–5% in ^{235}U . If the enrichment process discards depleted uranium at 0.2% ^{235}U , it takes more than 7 tonnes of uranium feed to produce 1 tonne of 4%-enriched uranium. Nuclear fuel discharged at the end of its economic lifetime contains less than 1% ^{235}U , but still more than the natural ore contains. Less than 1% of the uranium that enters the fuel cycle is actually used in a single pass through the reactor.

The other naturally occurring isotope, ^{238}U , directly contributes in a minor way to power generation. However, its main role is to transmute into plutonium by neutron capture and subsequent radioactive decay of unstable uranium and neptunium isotopes. ^{239}Pu and ^{241}Pu are fissile isotopes of plutonium that produce more than 40% of the fission energy in commercially deployed reactors. It is recovery of the plutonium (and to a lesser extent the uranium) for use in recycled nuclear fuel that has been the primary focus of commercial reprocessing. Uranium targets irradiated in special-purpose reactors are also reprocessed to obtain the fission product ^{99}Mo , the parent isotope of technetium, which is widely used in medical procedures. Among the

fission products, recovery of expensive metals such as platinum and rhodium is technically achievable but is not economically viable in current market and regulatory conditions.

During the past 60 years, many different techniques for reprocessing spent nuclear fuel have been proposed and tested in the laboratory. However, commercial reprocessing has been implemented using a single type of aqueous solvent extraction technology, plutonium–uranium extraction (PUREX). Similarly, hundreds of types of reactor fuels have been irradiated for different purposes, but the vast majority of commercial fuel is uranium oxide clad in zirconium alloy tubing. As a result, commercial reprocessing plants have relatively narrow technical requirements for spent nuclear fuel that is accepted for processing.

2. SPENT NUCLEAR FUEL

The majority of commercially reprocessed nuclear fuel is discharged from pressurized water reactors, the type of reactor that accounts for two-thirds of the world's nuclear generating capacity. Most of the balance is made up of fuel from boiling water reactors, which comprise more than 20% of the capacity. Though there are many variations in fuel design for specific reactors, the fuel assemblies for these two types of nuclear reactors are similar in physical and chemical composition. (Fig. 1).

The composition of spent nuclear fuel depends on the fission energy generated within the fuel assembly during irradiation in the nuclear reactor. Burn-up, usually expressed in megawatt-days/tonne of initial uranium, is a measure of the total energy produced by a fuel assembly. Average burn-up has increased with improvements in fuel technology; 40,000 MWd/tonne (MWd/tonne) being typical of spent fuel discharged at the beginning of the century. The mass of uranium contained in two fresh fuel assemblies for a pressurized water reactor is approximately 1 tonne. Each spent fuel assembly includes approximately 150 kg of radioactive hardware that holds the fuel pellets in place. At 40,000 MWd/tonne burn-up, the spent fuel assembly contains approximately 94.5% unused uranium, 1% plutonium, 0.2% minor actinides (neptunium, americium, and curium), and 4.3% fission products by weight. The fission products, which span the range from zinc to holmium on the periodic chart, include metals, rare earths, and gases. A small amount of tritium is also produced during ternary fission.

3. REPROCESSING

The status of large-scale commercial reprocessing is defined by the thermal oxide reprocessing plant



FIGURE 1 Boiling water reactor fuel assembly. Some fuel elements have been removed to show more structural detail. Photograph courtesy of Framatome.

(THORP) in Sellafield, United Kingdom, operated by British Nuclear Fuels Plc (BNFL), and the UP2 and UP3 plants in France, operated by Compagnie G n rales des Mat ri s Nucl aires (COGEMA). These plants, which are similar in design and operating principles, represent more than 90% of the world market in reprocessing services. They recover high-purity plutonium and uranium for recycle and stabilize radioactive waste in forms suitable for disposal. The fundamental process steps are storage of the spent fuel, size reduction, dissolution, chemical separations, product purification, and waste treatment. The process steps are typical of modern practices, although there are usually several practical alternatives for each step.

3.1 Head End Processes

Storage is important to reprocessing in that it allows the spent fuel to cool sufficiently to meet the input decay heat specification. Spent fuel is typically stored in canisters that are about 6 m in length, each weighing about 100 tonnes. The canisters are placed within a storage pool in which the water quality is carefully monitored to prevent corrosion. This storage capacity, which typically provides adequate buffer for at least several months of throughput, represents a significant capital investment. With a length of 150 m, the THORP storage facility (Fig. 2) holds enough fuel for 4 years of reprocessing. When removed from the canisters, individual fuel assemblies are subjected to nondestructive examination to



FIGURE 2 The THORP spent fuel storage pool in Sellafield, United Kingdom. Photograph courtesy of British Nuclear Fuels Plc.

verify that key nuclear characteristics match documentation that has been provided by the owner.

The first major step is mechanical disassembly to reduce the fuel to manageable dimensions and to expose the uranium inside of the metal cladding. In modern plants, the entire fuel assembly is chopped into small pieces as it moves through a shearing machine. A typical assembly contains an array of 17×17 fuel elements. Each stroke of the shearing blade cuts a few centimeters from the entire array. The end pieces, which contain no uranium, are collected separately. Shearing fuel is a dusty process in which the equipment is subjected to high stresses. Some gaseous fission products, such as krypton and xenon, are released during this operation.

The chopped fuel is transferred to the dissolver, a large tank containing hot nitric acid. Dust from the shearing operation is also delivered to the dissolver through the ventilation system. Uranium and most of the fission products dissolve readily, whereas dissolution of the plutonium proceeds somewhat more slowly. Plutonium-rich fuel grains do not always fully dissolve. For fuel irradiated to greater than 30,000 MWd/tonne, a portion of the metal fission products (e.g., ruthenium, rhodium, technetium, palladium, and molybdenum) form an insoluble alloy. The residual fission product gases and volatile fission products (e.g., tritium and iodine) are released into the off-gas system. After removal from the dissolver, the leached cladding hulls are washed with nitric acid and water, checked for residual fuel, and stored prior to processing into disposable waste. Solid particles of a few micrometers in diameter that are suspended in the dissolver liquid are removed with a centrifuge. The particles are subjected to an additional leaching step, contributing to the approximately 99.5% recovery of the nuclear material from the fuel.

The dissolver product liquor is conditioned prior to further processing. It is sparged with air at a higher temperature to remove residual iodine, then is cooled and sparged with nitrogen dioxide to adjust the plutonium from its hexavalent state to the tetravalent state, which is needed for extraction. The acidity is adjusted by the addition of HNO_3 . With the final clarification of the liquor, a materials accountancy sample is taken to establish the quantity of fissile material entering the process.

3.2 Chemical Separations

The purpose of chemical separations is to partition the elements into three different product streams: a

pure plutonium stream, a pure uranium stream, and a waste stream that contains the fission products and minor actinides. PUREX, with its many possible variants, is the separations process that has been universally selected for industrial reprocessing. It is a solvent extraction technique that relies on the selective transfer of components between two immiscible liquids—an aqueous phase and an organic phase. The organic phase contains tributyl phosphate (TBP) diluted to $\leq 30\%$ by volume with a hydrocarbon such as kerosene. The aqueous phase is a nitric acid solution. The two liquids are agitated into intimate contact to increase the surface area across which the species exchange can take place. They separate naturally due the difference in density, like oil floating to the top of water. Uranium and plutonium are first extracted together from the highly radioactive dissolver liquor, which simplifies the downstream separations steps. After partitioning, the separate uranium and plutonium streams are purified to the required level.

Three types of equipment are used to bring the organic and the liquid phases into contact: mixer-settlers, pulsed columns, and centrifugal contactors. Because the organic solvent is damaged by radiation and high temperature, the holdup time needs to be minimized. Multistage countercurrent pulsed columns or centrifugal contactors are preferred to mixer-settlers for high-activity separations. Mixer-settlers are used in scrubbing and stripping stages, during which the radioactivity is less intense. Pulsed columns (Fig. 3) can be several meters in height, containing dozens of mixing stages characterized by perforated disks or an alternating series of disks and rings within a tall pipe. Each pressure pulse pushes the organic phase up the column, breaking it into droplets as it passes through the perforations in the disks. These droplets contact the continuous aqueous phase as it is fed by gravity down the column. This heavier phase is removed from the bottom of the column, while the organic solvent containing the extracted product is removed from the top.

3.2.1 Initial Decontamination

Plutonium and uranium are extracted from the aqueous solution by forming nitrate complexes with TBP, leaving most of the fission products behind:



Some fission products are carried over in the solvent along with the uranium and plutonium. These



FIGURE 3 Pulsed column. The boxes around the column contain sensors that measure the progress of the extraction. Photograph courtesy of COGEMA.

contaminants are eliminated in a scrubbing operation in which the solvent is washed with aqueous nitric acid. After all traces of solvent are removed, the aqueous raffinate is held in a tank pending further treatment as high-level radioactive waste.

3.2.2 Separation of Plutonium from Uranium

If the decontamination in the first extraction cycle is sufficient, uranium/plutonium partitioning can be done in either a mixer-settler or a pulsed column. The uranium and plutonium pass into the contactor in the solvent stream, where uranous nitrate stabilized by hydrazine is introduced as a reducing agent for the plutonium. Plutonium in the Pu^{3+} state passes into a fresh acidic aqueous stream. The uranium, which remains in the organic solvent, goes into a mixer-settler (Fig. 4) for additional decontamination. Uranium is then removed from the solvent by dilute nitric acid. The uranium and plutonium undergo separate purification processes.

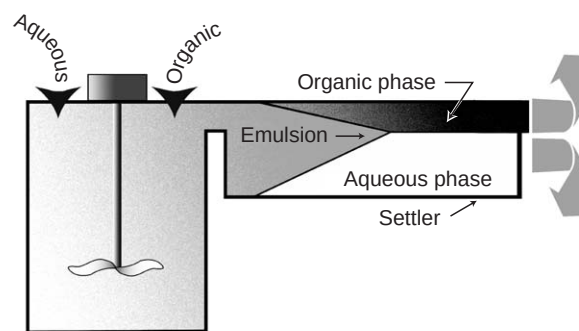


FIGURE 4 Mixer-settler concept.

3.2.3 Uranium Purification

The uranium stream passes into additional mixer-settlers, where the uranium is first extracted into the solvent and then transferred back to the aqueous phase. Through careful control of the chemistry, this series of steps removes most of the residual contaminants, principally neptunium, plutonium, technetium, and ruthenium. The uranium solution is stored in a buffer tank prior to finishing. An evaporator concentrates the uranyl nitrate after removal of any residual solvent. The concentrated nitrate solution is put through a thermal denitration process, which releases nitric acid and forms uranium trioxide. The UO_3 is packaged in stainless-steel drums for storage.

3.2.4 Plutonium Purification

The plutonium-bearing stream is washed with kerosene to remove residual TBP, then evaporated to concentrate the solution. Using pulsed columns, the plutonium is passed between the aqueous and organic phases to remove contaminants. The resulting plutonium nitrate is taken through a final kerosene wash prior to temporary storage. During the finishing process, plutonium nitrate is converted to plutonium dioxide. The properties of the PuO_2 product are adjusted to yield a powder that meets the input specifications for fuel production. The powder is packaged in a stainless-steel canister that is encased in two outer cans and transferred to a highly secure storage vault.

3.3 Waste

Radioactive waste—in gaseous, aqueous, organic, and solid forms—is generated throughout the reprocessing flow. The early experience with reprocessing for nuclear weapons production in the United States and the former Soviet Union left a legacy of inadequately managed waste and unfortunate environmental releases. However, in modern commercial plants, waste

stream management is a priority, leaving relatively small volumes for treatment and disposal.

3.3.1 Gaseous Effluents

Radioactive gases are evolved from various operations in the plant. Shearing the fuel assemblies releases some of the fission gases; most of the remaining fraction is released to the off-gas system when the fuel is dissolved. In order to avoid downstream chemistry problems, temperature and acidity are adjusted to ensure that at least 98% of the iodine is volatilized from the dissolver liquor. Some volatile ruthenium can also be released from high-temperature processes. A high-activity metal fission product that will condense on cool surfaces, ruthenium released into the ventilation system proved to be a problem in the first reprocessing plant. Ruthenium has since been controlled through process chemistry.

The noble gases krypton and xenon are released to the environment. They contribute insignificantly to the radiation dose of the workers or the public. Because noble gases are not reactive, they are difficult to trap and sequester safely. The perceived risk from a large-scale accidental release of stored krypton has outweighed that of a low-intensity continuous release. However, krypton capture and storage has been demonstrated at the small Tokai Reprocessing Plant in Japan.

Caustic scrubbers are typically used to remove the other radioactive gases and noxious gases, such as nitrogen oxides. These scrubbers are packed columns in which the gases come into contact with a countercurrent of sodium hydroxide solution. The gases are usually filtered to remove dust and small radioactive particles prior to the caustic scrub.

3.3.2 Solid Wastes

The solid wastes include primarily the cladding hulls and end pieces of the fuel as well as some plutonium-contaminated materials such as filters or process residues. The solids are compacted and stored in canisters, or they may be immobilized in concrete prior to storage. In Europe, these solids are categorized as intermediate-level wastes that require deep geologic disposal.

3.3.3 Liquid Wastes

The primary liquid waste is the first-cycle raffinate containing most of the fission products. In the United States, high-level waste is legally defined as this raffinate or any solids derived from it. The liquid is concentrated by evaporation and temporarily stored. Because the concentrated liquid is highly self-heating

from the decay of ^{137}Cs , ^{90}Sr , and other fission products, the storage tanks must be cooled.

The high-level liquid waste is pumped out of storage into a treatment plant, where it is calcined and mixed with glass frit. The mixture is melted and drained into a stainless-steel canister, where it solidifies into a vitrified waste form. Lids are welded on the canisters before they are placed in cooled storage, where they remain for decades prior to final disposition. The modern French plants produce less than 0.5 m^3 of high-activity waste for every tonne of uranium processed, or less than 25% of the estimated volume required to store an equivalent amount of spent fuel. To the extent possible, the liquid chemical streams are cleaned and recycled in modern reprocessing plants. Diluent and solvent are both recovered during organic waste treatment. The principal wastes from this process are various salt residues from the separations steps and the decomposition products of TBP.

3.4 Environmental Effects

The modern British and French plants are able to contain their environmental effluents within strict regulatory limits. The plants are located near the coast; the principal environmental effluent is lightly contaminated water that is discharged into the ocean. Trace amounts of radionuclides are contained in the discharged water. Most of the tritium is released to the sea, because it is chemically inseparable from ordinary water. Because of its volatility, iodine is unsuitable for high-temperature waste processing. It has traditionally been diluted with the natural iodine in the seawater, as this is the safest known practice, although the search for a suitable alternative has been underway for decades.

The UP2 and UP3 plants (Fig. 5) are located at La Hague, France, where the average dose rate to the public from the environmental discharges is less than $0.01\text{ millisievert (mSv)/year}$, compared to a dose from background radiation of 2.4 mSv/year . In the United Kingdom, the modern THORP plant is collocated at the Sellafield Site with an older reprocessing plant, and the peak dose to critical seafood consumers reached almost 2 mSv at that location in 1976. Today, the peak dose rate has dropped to approximately 0.1 mSv/year , a level that is due primarily to a legacy of previous seabed contamination. At both the French and the British sites, exceptionally low exposure levels were first achieved in the late 1980s.

Tight control of the radioactive inventory at plants was not the main priority in the early years of reprocessing. In the United States, reprocessing of



FIGURE 5 COGEMA reprocessing facilities at La Hague, France. Photograph courtesy of COGEMA.

short-cooled fuel resulted in release of radioactive ^{131}I between 1944 and 1947. Subsequently, the fuel was cooled for a longer time, which allowed decay of most of this isotope with its 8-day half-life. Filters and scrubbers were also added to the exhaust systems of plants to trap iodine and other radionuclides. More than $350,000\text{ m}^3$ of liquid high-level waste was generated and stored in steel tanks, creating an expensive legacy that is now being addressed.

The largest environmental problems were experienced in the former Soviet Union as a result of plutonium production during the Cold War. These operations at the Mayak, Tomsk-7, and Krasnoyarsk-26 sites resulted in record releases of radioactive materials into the environment. The towns at these Siberian sites are now known as Ozersk, Seversk, and Zheleznogorsk, respectively. At Mayak, liquid wastes were discharged into Lake Karachai, the Techa River, and a series of small reservoirs on the river. At Tomsk and Krasnoyarsk, liquid wastes were discharged primarily to deep underground geologic formations, but also to surface waters.

Though the Cold War legacies from weapons plutonium production in the United States and the former Soviet Union bear scant resemblance to modern commercial reprocessing practices, their influence on public perception continues. COGEMA and BNFL prominently display up-to-date environmental information on their Internet Web sites to help allay public concern about radioactive releases.

4. REPROCESSING FACILITIES

4.1 United States

Large-scale reprocessing had its start in the Manhattan Project. During World War II, three pluto-

onium production reactors and two reprocessing plants (the T plant and the B plant) were built at Hanford, Washington. In operation by December 1944, the T plant was the world's first reprocessing facility. Plutonium was separated from metallic uranium fuel that had been lightly irradiated in the B reactor. Though there are few similarities to today's commercial plants, some features of the early building designs by the DuPont engineers have been retained in modern facilities. The T plant was a canyonlike 245 m long, 31 m high, and only 26 m wide. Constructed of steel-reinforced concrete, the process cells were heavily shielded. Access to the process equipment was by crane through the roof of the concrete cells. The T plant used a bismuth phosphate process, selected and designed by DuPont from several separations flow sheets developed by the Metallurgical Laboratory at the University of Chicago. The B plant, the second facility based on this precipitation process, operated at Hanford before the end of the war.

Two other large reprocessing plants were also built at Hanford. Completed in 1951, the REDOX (so named because of the reduction and oxidation process involved) plant was the first reprocessing plant based on countercurrent, continuous-flow separation of plutonium and uranium. The process was developed by Argonne National Laboratory and implemented by the General Electric Company at Hanford. The PUREX (plutonium/uranium extraction) plant went into operation in 1956, applying the technology that has since been commercialized in several countries. The PUREX flow sheet was developed by the Knolls Atomic Power Laboratory and implemented at Hanford by General Electric.

The first large PUREX reprocessing plant built by the U.S. government, called the F-Canyon, was designed and constructed in 1954 by DuPont at the Savannah River site in South Carolina. PUREX was also used in specialized plants for reprocessing highly enriched uranium fuels at Savannah River (H-Canyon) and in Idaho (Idaho Chemical Processing Plant). None of these plants was designed or operated for processing commercial fuel. Notwithstanding environmental legacies, the government-owned, contractor-operated plants had an impressive production record, particularly considering the first-of-a-kind engineering that went into each plant. The REDOX plant processed more than 19,000 tonnes of spent fuel during its 16 years of operation. The capacity of the PUREX and F-Canyon plants was considerably higher, in the range of 2000–3000 tonnes of initial uranium per year of operation.

By contrast, attempts to commercialize reprocessing in the United States were unsuccessful. Of the three plants that were constructed, only one ever operated—the Nuclear Fuel Services plant at West Valley, New York. From 1966 through 1972, the West Valley plant processed fuel from several utility-owned reactors as well as some fuel from the N reactor in Hanford, Washington. With a nominal capacity of 300 tonnes/year, the plant's PUREX process worked well. However, when the plant was shut down due to changing regulatory requirements, modifications were assessed to be too expensive and the plant never restarted.

A novel 300-tonne/year plant was designed and built by General Electric in Morris, Illinois in the early 1970s. The Midwest Fuel Recovery Plant was designed to use a combined solvent extraction and fluoride volatility process rather than the complete PUREX flow sheet. However, nonradioactive testing revealed that the compact design had not incorporated sufficient buffer storage between processes. Realizing that the plant would not be capable of reliable operation, General Electric simply used the fuel storage facility, and never started the plant. Between 1970 and 1975, Allied General Services built a large 1500-tonne/year reprocessing plant in Barnwell, South Carolina. The plant underwent startup testing with fresh uranium in 1976 while the rules on plutonium conversion and high-level waste vitrification were being resolved by the Nuclear Regulatory Commission. However, in 1977, the Carter administration canceled the licensing proceedings and the plant had to be mothballed.

4.2 Western Europe

Several European reprocessing plants have been built and successfully operated, including the Eurochemic 300-tonne/year plant at Mol, Belgium, a 35-tonne/year German pilot plant, and several modern British and French commercial reprocessing plants (described in the following sections).

4.2.1 France

More than any other nation, France has successfully implemented its commercial reprocessing ambitions. As a part of the international nuclear services conglomerate AREVA, COGEMA operates the two 800-tonne/year UP2 and UP3 plants at La Hague on the Brittany coast. In addition to the French national utility Electricité de France (EDF), COGEMA's reprocessing customers include Japan, Belgium, Germany, Spain, and Switzerland. Prior to 1998,

COGEMA also operated a 400-tonne/year plant in Marcoule to reprocess fuel from the French gas-cooled reactors.

4.2.2 United Kingdom

The highest capacity reprocessing plant in the United Kingdom is the 1500-tonne/year Magnox plant, which has processed fuel from the nation's gas-cooled Magnox reactors since 1964. However, these relatively old reactors are scheduled to be shut down by 2010. Once the magnesium-clad fuel has been processed, the Magnox plant will also be closed.

Completed in 1994 near Sellafield on the Irish Sea, THORP is one of the newest reprocessing plants in the world. The plant is 500 m long, 48 m high, and 120 m wide, with a chimney that stands 125 m above the landscape. The expected throughput was approximately 1200 tonnes/year, but operational difficulties, primarily with the waste vitrification plant, have not allowed it to reach that potential. The nominal capacity is now considered to be more in the range of 850 tonnes/year, though individual pieces of equipment are scaled for higher throughput. Because the capacity of the plant far exceeds the domestic need for its reprocessing services, THORP has relied on foreign contracts, particularly with Japan and Germany, for most of its business.

4.3 Asia

4.3.1 Japan

Lacking indigenous uranium resources, Japan has a long-standing policy of supporting reprocessing. Japan has both imported foreign technology and conducted a vigorous national research and development program. The Tokai Reprocessing Plant has operated with an approximately 90-tonne/year capacity since 1977. In 1993, Japan Nuclear Fuel Limited (JNFL) began constructing a large reprocessing plant near Rokkasho on Japan's northern coast (Fig. 6). The 800-tonne/year plant is scheduled to begin full operation in 2005.

The Rokkasho plant incorporates much of the French commercial technology. One notable exception is that plutonium will be coextracted with 50% uranium as an additional nuclear safeguard measure. The site includes an interim high-level waste storage facility, which in 1995 began to receive canisters of waste from Japanese fuel that had been reprocessed in Europe. The site also includes a low-level waste disposal facility, a large spent fuel storage capacity, and a uranium enrichment plant, all currently operating. An icon of modern technology, the

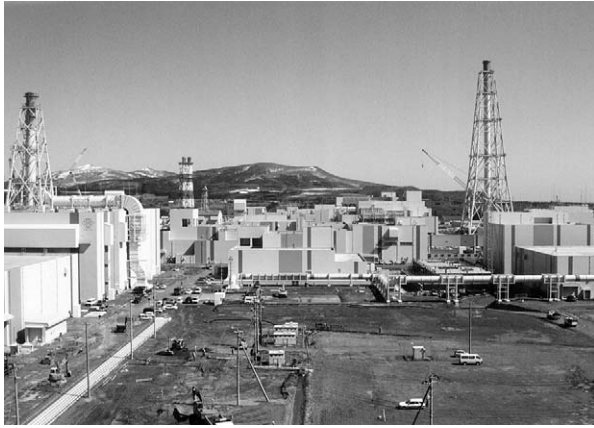


FIGURE 6 Rokkasho reprocessing plant on Japan's northern coast. Photograph courtesy of Japan Nuclear Fuel Limited.

Rokkasho plant provides a benchmark (Table I) for the scale of commercial reprocessing facilities.

4.3.2 Russia

In Russia, as in the United States, most reprocessing experience has been with plutonium production reactor fuel. At the Mayak, Tomsk-7, and Krasnoyarsk-26 sites, more than 300,000 tonnes of uranium fuel were reprocessed between 1948 and 1996. This is an estimable production record, considering that it would take 100 years of modern commercial reprocessing to match it. The Russian production reprocessing used a somewhat unique aqueous technology based on the precipitation of sodium uranyl acetate and sodium plutonyl acetate from nitric acid solutions containing the dissolved fuel. Plutonium was separated from the uranium by reducing plutonium to the Pu^{4+} or Pu^{3+} state, which remained in solution. The acetate process created a high volume of sodium nitrate and sodium acetate waste.

A reprocessing facility for power reactor fuel was completed at Mayak in 1976. Called RT-1, the plant uses a more conventional PUREX process. It was built with a storage capacity of 400 tonnes of spent fuel. Russia is building a much larger RT-2 plant at Zheleznogorsk. To be constructed in stages, the initial operation is intended to have a capacity of 1500 tonnes/year. However, funding has been insufficient for timely completion of the plant, which is less than half finished. It has a storage capacity of 6000 tonnes, with a planned increase to 9000 tonnes.

Russia has conducted extensive research on nuclear waste management and reprocessing at many of its universities and laboratories. Since the mid-1990s, Russia has been developing plans to open an

TABLE I

Dimensions of Main Buildings at the Rokkasho Reprocessing Plant^a

Building	Approximate floor area (m ²)	Height (stories)	
		Aboveground	Belowground
Spent fuel receiving and storage	9400	3	3
Head end	6000	5	4
Separations	5700	4	3
Purification	6500	6	3
Uranium denitration	1500	5	1
Uranium-plutonium codenitration	2700	2	2
Uranium oxide storage	2700	2	2
Uranium-plutonium mixed oxide storage	2700	1	4
High-activity liquid waste vitrification	5100	2	4
Vitrified package storage	5700	1	2
Low-activity liquid waste treatment	2600	3	2
Low-activity waste treatment	9500	4	2
Control	2900	3	2
Analytical laboratory	4900	3	3

^aData courtesy of Japan Nuclear Fuel Limited.

international spent fuel reprocessing center. Because the reprocessing capacity is not yet adequate, the center would initially provide spent fuel storage.

4.3.3 China and India

China has operated a reprocessing pilot plant since 1970 and has plans for constructing an 800-tonne/year facility by 2010. India is operating two small plants, each capable of about 60 tonnes/year throughput. India also has planned to install some 1000 tonnes/year of reprocessing capacity to manage the spent nuclear fuel from its growing nuclear energy infrastructure.

4.4 National Policies

The rationale for nuclear fuel reprocessing, or alternatively for deferring the practice, has been the subject of intense international discussion since the late 1970s. The crux of the debate lies in the military origin of the technology and the increasing global concern about

weapons of mass destruction. Although many refinements have been made, the technology that has been implemented commercially is fundamentally the same technology that was used to develop a military nuclear capability in the United States, the former Soviet Union, France, and the United Kingdom, and it can produce very pure plutonium.

Weapons-grade plutonium has been specified to have at least 92% of the isotope ^{239}Pu . Because of spontaneous neutron emission, radiation, and heat generation, the other isotopes of plutonium cause significant difficulties for the design, fabrication, storage, and transportation of nuclear weapons. Typical commercial spent fuel contains plutonium that has less than 80% of the ^{239}Pu isotope, so reactor-grade plutonium was initially not considered weapons usable. However, in the 1970s, weapons experts informed U.S. policymakers that it was in fact possible to make a nuclear explosive from reactor-grade plutonium, if not a military weapon. In light of this information and an unexpected nuclear test explosion by India in 1974, the U.S. position on reprocessing was reexamined during the Gerald Ford administration. In 1977, President Jimmy Carter canceled the licensing of the Barnwell plant, stranding half a billion dollars of private investment. When the reprocessing ban was quietly lifted in 1981 during the Ronald Reagan administration, there was no longer viable commercial interest in reprocessing within the United States. In 1993, President Bill Clinton reinstated the national policy against reprocessing, but in 2001 the Bush administration began to encourage research into new reprocessing technologies that would not produce weapons-suitable material.

Unfavorable economics has further fueled the debate over recycling nuclear materials. In an era of cheap uranium, fuel produced from recycled plutonium is more expensive than is standard fuel produced from enriched uranium. Though estimates of the cost premium for recycled fuel span a wide range, its actual impact on the cost of electricity production is minimal. Unlike fossil-fueled power plants, nuclear fuel accounts for no more than 20% of the total generating cost.

The original assumption in the 1960s was that the recovered plutonium would be used to fuel sodium-cooled reactors, which would improve uranium resource utilization by two orders of magnitude. The actual practice has been to recycle plutonium into existing water-cooled reactors. Primarily for safety reasons, not all of the existing reactor fleet can accept plutonium fuel. Most of those reactors that do

accept the mixed plutonium/uranium oxide fuel (MOX) can use it in only one-third of the core loading. Further, the capacity of the MOX fabrication facilities has not caught up with the reprocessing capacity. As a result, there is a growing accumulation of excess civilian plutonium, which stood at more than 200 tonnes at the beginning of the 21st century.

The reprocessing debate has been predominantly a U.S. phenomenon. Approximately one-half of the spent nuclear fuel discharged annually around the world is slated for reprocessing, and the other half is slated for direct disposal. A practical argument in favor of reprocessing is that though uranium may be plentiful, land suitable for the disposal of nuclear waste is not. No nation has yet opened a deep geologic repository for high-level waste. The United States is moving toward that goal, having initiated a licensing process for Yucca Mountain, Nevada, after 20 years of study and legal challenges. If the schedule is maintained, the first fuel assemblies would be shipped in 2010.

France, Japan, and Russia have firm pro-reprocessing policies. With limited land area, more than 50 nuclear power plants, and no significant indigenous energy resources, Japan's reprocessing policy is based on a desire to have some measure of energy independence. Requiring reprocessing as a part of its national energy strategy, France has a complex spent fuel management plan that calls for leveling the total accumulation of plutonium in the entire fuel cycle before a national geologic repository is developed. Russia has expressed keen interest in providing full nuclear fuel cycle services to international clients in addition to domestic reprocessing.

4.5 International Safeguards

Commercial reprocessing plants are subject to nuclear materials safeguards verification by the International Atomic Energy Agency (IAEA) under the Treaty on the Nonproliferation of Nuclear Weapons. Entered into force in 1970, the treaty now has 188 member states. Only Israel, Pakistan, and India have not signed, although in January 2003, North Korea announced its intention to withdraw. The technical objectives of safeguards agreements established under the treaty are the timely detection of diversion of significant quantities of nuclear material from peaceful uses to manufacture nuclear weapons, and to deter such diversion by early detection. The objectives are based on the principle that a certain quantity of nuclear material is needed to manufacture a nuclear explosive device and that a

certain amount of time is needed to convert the material into a weapons-usable form.

The IAEA uses a variety of physical and chemical measurements and other sources of information to determine whether a diversion has occurred. The agency has implemented safeguards in 70 countries at more than 900 facilities, including six reprocessing plants. The IAEA also has safeguards jurisdiction over a rapidly growing inventory of plutonium in spent nuclear fuel, almost 600 tonnes in 2003.

5. REPROCESSING RESEARCH AND DEVELOPMENT

5.1 Cost Reduction

After 50 years of PUREX implementation, reprocessing is still not a mature industry. With changing objectives, evolving input fuel specifications, more stringent environmental regulations, and feedback from operating plants, the opportunities for developing process, equipment, or plant improvements are ample. Established reprocessing corporations are focused on driving down costs. Options include process modifications to make them move compact, faster, safer, less chemical intensive, more reliable, more transparent to safeguards, or easier to maintain. There is considerable incentive to reduce capital cost through simplification and smaller facilities, which could be accomplished by fewer cycles or less buffer storage. Existing plants were designed to process fuel that had cooled for only several months, whereas in practice the actual cooling time has been years, if not a decade or more. This greatly reduces the radiation and heat load requirements on the facility, which opens the door to cost-saving design changes.

As new technologies are developed, they are implemented in operating plants to the extent practicable, rather than saved for the next-generation facility. For example, utilities are routinely discharging fuel that has an average of 48,000 MWd/tonne burn-up, with 60,000 MWd/tonne expected to be achievable within a few years. Since some plants were designed with lower burn-up limits and initial enrichment limits, modifications must be made just to stay in business.

5.2 Alternative Processes

From hundreds of potential chemical separation schemes, PUREX has been adopted for commercial

implementation. However, as new types of reactors and fuels are considered or as the objectives for reprocessing evolve, research is being conducted on alternative separation technologies as well as on PUREX and its aqueous siblings. Nonaqueous processes have been studied for decades because of potential advantages that include much less susceptibility to radiation and heat damage, fewer chemical steps, more compact equipment, and larger batch sizes. They have not found their way into commercial application because of disadvantages that include an inability to meet product specifications for current-generation nuclear fuel and required operation at high temperature with corrosive or reactive materials. Most nonaqueous processes can be classified as pyrometallurgical, pyrochemical, or fluoride volatility processes.

Pyrometallurgical processes include volatilization, distillation, fractional crystallization, liquid metal extraction, fused salt extraction, fused salt electrolysis, and partial oxidation. Many fission products are more volatile than uranium. Volatilization typically works by vacuum distillation of crushed, molten, or dissolved fuel to separate most of the volatile elements from the uranium. It is used to remove noble gases and iodine during conventional aqueous processing. Fractional crystallization has been tested for partitioning fission products from metallic fuel using a low-melting-temperature liquid-metal extractant and for removing bulk uranium from the dissolver liquor at very low temperature. Partial oxidation is a process in which volatile fission products are vaporized from molten metal fuel, whereas the rare earths and the more reactive fission product metals are oxidized by a sacrificial crucible liner. Argonne National Laboratory successfully conducted a pilot-scale demonstration of this melt refining process on more than 30,000 fuel pins during the late 1960s, but decontamination was considered to be too incomplete for further development.

Pyrochemical processes use selective oxidation and reduction reactions in fused salts or metals to recover uranium and plutonium. Russia has an active program that has demonstrated pyrochemical processing of uranium oxide fuels. There is also an active Japanese program to develop pyrochemical processes for removing transplutonium actinides from waste streams and for processing uranium nitride fuels.

Fluoride volatility processes take advantage of the unusual property of uranium, neptunium, and plutonium to form volatile hexafluorides. Unlike other nonaqueous processes, fluoride volatility is

more readily capable of thorough decontamination of the product materials. Fluoride volatility processes can be combined with aqueous processes, as was done in the General Electric plant at Morris, Illinois.

5.3 Partitioning and Transmutation

During the 1990s, the concept of much more aggressive nuclear waste management began to take hold. This line of research developed out of concern about leaving a legacy of radiotoxic waste that would persist for hundreds of thousands of years, the uncertainty about the future availability of deep geologic repositories, and proliferation potential of thousands of tonnes of plutonium accumulating in spent nuclear fuel in more than 30 countries. In its ultimate application, partitioning and transmutation would destroy all actinides by fission and would transmute long-lived fission products into stable elements or short-lived radioisotopes. In order to achieve transmutation of the minor actinides and some plutonium isotopes, a fast-spectrum reactor would be required.

5.3.1 Objectives of Partitioning and Transmutation

The principle behind partitioning and transmutation is that there is an optimum means of managing each group of elements in spent nuclear fuel. Within practical limits, the elements are separated into like groups as defined by hazard, transmutation potential, chemistry, heat-generation rate, proliferation risk, half-life, and energy value. Each group is managed to meet appropriate specific objectives: Uranium, which comprises 95% of the mass, can be stored until needed sometime in the future. Plutonium can be used in thermal-spectrum reactor fuel. Plutonium and minor actinides can be used in fast-spectrum reactor fuel. Highly soluble, long-lived fission products such as ^{129}I and ^{99}Tc can be fabricated into transmutation targets or can be stabilized in custom waste forms. Short-lived waste products can be stabilized in compact waste forms. The high-heat-load fission products cesium and strontium can be separately managed until they are sufficiently cool after several decades.

Temperature is the limiting design parameter for the proposed Yucca Mountain repository. Intended to store unprocessed spent fuel, the peak mountain temperature is predicted to occur hundreds of years after closure. The transuranic elements, in particular neptunium and americium, are the principal sources of long-term heat generation (Fig. 7). A second

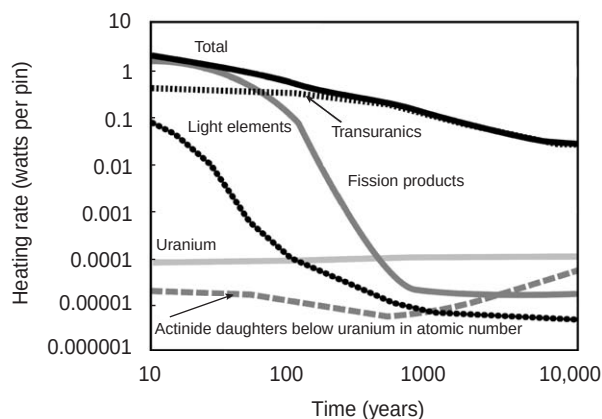


FIGURE 7 Components of the heat generation rate for spent nuclear fuel with an average burn-up of 33,000 MWd/tonne. Courtesy of Argonne National Laboratory.

repository, operated in a partitioning-transmutation scenario with active ventilation, could safely store the waste from five to a hundred times as many reactor-years of power production. The key is recycling the actinides, which would remove most of the volume, long-term heat generation, and long-term radiological hazard. Without actinides, the radiotoxicity of the stored nuclear waste also decays relatively quickly, becoming less toxic than uranium ore within a few hundred years.

5.3.2 Separations Research Programs

Japan, Russia, the United Kingdom, France, and the United States all sponsor research on advanced separations methods. Japan is pursuing both advanced aqueous and pyro processes. Russia has led pyrochemical process development. The United Kingdom is focused on improving conventional reprocessing while supporting a modest effort in pyroprocess development. France has a comprehensive program that supports a broad spectrum of research, ranging from molecular modeling to hardware improvements. The U.S. Department of Energy has consolidated all separations research into a single advanced fuel cycle program.

The U.S. aqueous process development work is aimed at determining the practicality of multiple separations. At least five different partitioning steps are being considered. After conventional nitric acid dissolution, uranium and technetium would be separately extracted, thereby removing most of the material from the main process stream. Cesium and strontium would be extracted in a second process. Neptunium and plutonium would then be coextracted to avoid clean plutonium and its

attendant proliferation risk. Transuranic extraction (TRUEX), an established process, would separate the bulk fission products from the remaining actinides. A fifth process would separate the americium and curium from the acid stream, which after that point would contain principally rare earth fission products.

This approach to multiple separations presents many technical challenges, including managing several different solvent streams and minimizing radiation damage to the organic phases while retaining fission products and high-activity α emitters through several cycles. Radiation damage to the solvent is controlled by using banks of centrifugal contactors, which minimizes the irradiation time of the solvent phases.

A second line of separations development is aimed at using a fused salt pyroprocess to partition the fuel into two product streams and two waste streams. The spent fuel would be reduced to metal by direct electrolytic reduction in a vessel containing molten LiCl. The metal product would be electrorefined (Fig. 8) to recover uranium in second vessel containing a LiCl/KCl electrolyte, in much the same way that aluminum is produced. The salt and uranium phases would be separated by vacuum distillation, followed by melt consolidation of the uranium. The cladding hulls and metals such as technetium and rhodium that are not oxidized in the salt would be withdrawn, separated from adhering salt, and processed into a metallic waste form. Argonne National Laboratory has successfully demonstrated

electrorefining by processing more than 2.5 tonnes of metal fuel since 1996.

Salt containing uranium and transuranic chlorides would be treated by electrolysis or an electrometallurgical process to extract a second product containing uranium, plutonium, neptunium, americium, curium, and some rare earth fission products. This alloy could be used directly to manufacture fast-spectrum reactor fuel or it could be sent to a small purification line to be adequately decontaminated for thermal-spectrum reactor fuel. After the salt was sufficiently laden with fission products and stripped of actinides, it would be treated in a zeolite column to remove the fission product chlorides. The zeolite would be processed into a glass-bonded ceramic waste form.

5.3.3 Future Research and Development

Reprocessing research and development are limited only by the imagination of the chemist, available funding, and national policy restrictions. Ultimately, fuels, reactors, separation processes, waste forms, and geologic repositories will benefit from being designed as a secure, compatible system, rather than following independent development trajectories as they did in the 20th century.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel: Design and Fabrication • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

- Benedict, M., Pigford, T. H., and Levi, H. W. (1981). "Nuclear Chemical Engineering." McGraw-Hill, Boston.
- Bradley, D. J. (1997). "Behind the Iron Curtain: Radioactive Waste Management in the Former Soviet Union." Batelle Press, Columbus, Ohio.
- Cochran, R. G., and Tsoulfanidis, N. (1999). "The Nuclear Fuel Cycle: Analysis and Management." American Nuclear Society, La Grange Park, Illinois.
- Feiveson, H. A., von Hippel, F., and Williams, R. H. (1979). Fission power: An evolutionary strategy. *Science* 203, 330–337.

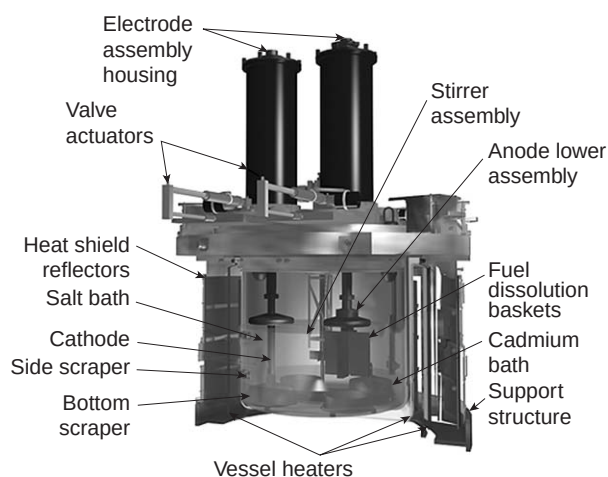


FIGURE 8 Drawing of an electrorefiner used in pyroprocess demonstrations. Courtesy of Argonne National Laboratory.

- Garwin, R. L., and Charpak, G. (2001). "Megawatts and Megatons." Knopf, New York.
- Hoffman, D. C. (ed.). (2002). "Advances in Plutonium Chemistry 1967–2000." American Nuclear Society, La Grange Park, Illinois, and University Research Alliance, Amarillo, Texas.
- Office of Environmental Management. (1997). "Linking Legacies: Connecting the Cold War Nuclear Weapons Production Processes to Their Environmental Consequences." DOE/EM-0319. U.S. Department of Energy, Washington, D.C.
- Rasmussen, N. C. (Chair), Committee on Separations Technology and Transmutation Systems. (1996). "Nuclear Wastes: Technologies for Separations and Transmutation." National Research Council. National Academy Press, Washington, D.C.
- Seaborg, E., and Seaborg, G. T. (2001). "Adventures in the Atomic Age: From Watts to Washington." Ferrar, Straus & Girox, New York.
- Wilson, P. D. (ed.). (1996). "The Nuclear Fuel Cycle: From Ore to Waste." Oxford Univ Press, Oxford, United Kingdom.

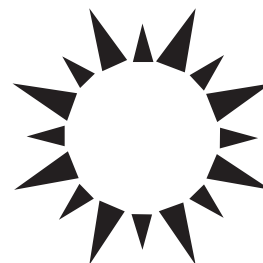
Nuclear Fusion Reactors

PHILIPPE MAGAUD and G. MARBACH

Euratom–Commissariat à l’Energie Atomique
Cadarache, France

IAN COOK

Euratom–U.K. Atomic Energy Authority
Abingdon, United Kingdom



-
1. Introduction
 2. The Fusion Reactor
 3. Conclusion

Glossary

International Thermonuclear Experimental Reactor (ITER) Project whose main goal is to demonstrate the scientific and technological feasibility of fusion energy for peaceful purposes; ITER will study burning plasmas (i.e., plasmas where heating is provided mainly by the alpha particles created from fusion reactions) and will also be the first machine incorporating most of the technologies crucial for the preparation of a future fusion reactor (e.g., plasma facing components, tritium management, robotics, tests of breeding blankets).

Joint European Torus (JET) Flagship of the European Community Fusion Program; JET is the largest and most powerful fusion experiment in the world.

keV Abbreviation for kiloelectronvolt; in nuclear physics, energy is often expressed in electronvolts and multiples thereof; as a result of popular misuse of language, temperature is sometimes also indicated in kiloelectronvolts.

At the core of the sun and stars, light nuclei combine—or fuse—to create heavier nuclei. This process releases a significant quantity of energy and is the source of the heat and light that we receive. Our sun has been shining for more than 4 billion years, and according to astrophysics, it will shine for this long again before entering the phase that will lead to its extinction. Harnessing this type of reaction on Earth for the purpose of generating energy would open the way to nearly unlimited resources. This is the aim of fusion research undertaken by the leading industrial nations. After a reminder of the main

principles, this article focuses on the final objective of controlled fusion research: the energy-generating fusion reactor. Many concepts are already available, whereas some would require so much progress in terms of physics and technology that they could only be implemented well into the future. An exhaustive description is not possible. Therefore, this article focuses only on the most developed: the magnetic fusion-based concepts. Worldwide studies have been done, or are ongoing, with the main goals to prove their credibility, reliability, and economic viability and that the advantages of this type of reactor in terms of safety and the environment are respected.

1. INTRODUCTION

1.1 Fusion Reactions

To obtain a fusion reaction, two nuclei that have a natural tendency to repel each other, both being positively charged, must be brought close enough together. Therefore, a certain amount of energy is necessary to overcome this natural barrier and reach the zone close to the nucleus where nuclear forces are capable of overcoming the barrier. The probability of crossing this barrier may be quantified by the “cross section” of the reaction. The variation of the effective cross section of several fusion reactions as a function of interaction energy is provided for light elements in Fig. 1. This gives rise to several points:

- The cross sections of fusion reactions are on the order of 1 barn ($1 \text{ barn} = 10^{-28} \text{ m}^2$), compared with 600 barns for the fission reaction of ^{235}U obtained by thermal or slow neutron bombardment. Therefore, fusion reactions feature extremely small cross sections, even compared with other competitive processes

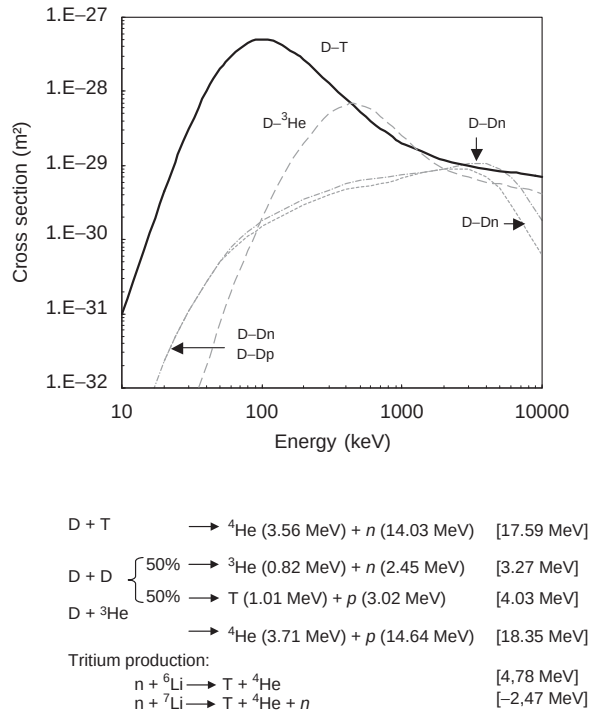


FIGURE 1 Cross section of main fusion reactions.

such as ionization and coulomb scattering (collision of charged particles). This constitutes one of the inherent difficulties of fusion.

- Fusion demands high temperatures, typically more than 100 million degrees (10 keV). At such temperatures, electrons are separated from nuclei; hence, a plasma, the fourth state of matter, is formed. Plasma exists within the universe in extremely diverse forms and with extremely variable temperature and density characteristics. Plasma is the most common state of matter in the universe and has already found numerous applications in everyday life (e.g., neon tube, plasma torch).

- The most accessible fusion reaction is that involving deuterium and tritium, the two isotopes of hydrogen. Research into controlled fusion focuses on this reaction. Although deuterium is abundant, tritium exists in only trace amounts. It needs to be produced by means of nuclear reactions with lithium (Fig. 1).

1.2 Conditions Necessary for Fusion Reactions

As explained, fusion requires high temperatures. Other conditions must also be met if fusion is to be used as an energy source.

As a hot gas, plasma is never totally isolated; thus, it is subject to various losses through radiation but

also through convection and conduction. In general, a time is defined synthesizing all said losses, referred to as the energy confinement time (τ_E); this is the time the plasma requires to lose its energy content if the energy sources feeding it are cut suddenly. In some ways, τ_E characterizes the quality of plasma insulation.

For the fusion reaction to be energetically viable, the energy generated by fusion reactions must at least compensate for these losses. This condition imposes a limit less than the product of energy confinement time and particle density (n) given by the Lawson criterion:

$$n \cdot \tau_E > g(T) \cdot f(Q),$$

where $g(T)$ accounts for the variation of the reaction rate with temperature T and Q corresponds to the relationship between the fusion power generated and the external power supplied to the plasma. The Q factor is frequently referred to as the energy amplification factor. There are two characteristic values of Q :

1. $Q = 1$ indicates that power produced by the plasma is equal to the power coupled to it from the exterior. This is referred to as the “break-even” point and is approached in the most efficient of the current experimental devices.

2. $Q = \infty$ indicates that external power contributed to the plasma is zero. The plasma is self-sustaining and is said to be in ignition.

For a deuterium and tritium plasma, the function $f(Q)$ is equivalent to approximately 1 for $Q = 1$ and tends rapidly toward 5 for higher values of Q . Under these conditions, and for a temperature of 10 keV, the Lawson criterion may be written

$$n \cdot \tau_E \approx 10^{20} (m^{-3} \cdot s).$$

1.3 Fusion in the Stars, Fusion on Earth

In the sun and the stars, the conditions necessary for fusion regarding temperature, density, and confinement time are maintained by gravity. This solution is impossible to implement on Earth. The fact that the Lawson criterion sets a limit, not for density and confinement time separately but rather for the product of these characteristics, allows two completely different approaches:

1. Bring a small volume of matter to a high temperature under high pressure for an extremely short time period. This is referred to as inertial confinement. The aim is to obtain the greatest possible number of fusion reactions before the

plasma disperses. Confinement time characteristics are on the order of a few picoseconds.

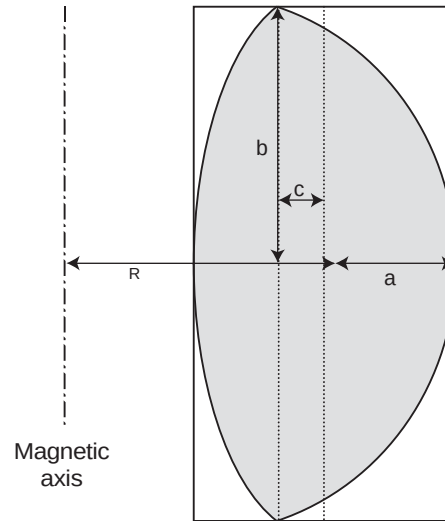
2. Trap and hold plasma at a very high temperature. This plasma is confined in an intangible box created by magnetic fields. This is referred to as magnetic confinement. In this case, the aim is to “allow time” for particles to fuse. Confinement time characteristics are on the order of a few seconds.

Even if inertial confinement energy-oriented concepts exist, research on inertial fusion is supported mainly for simulation of nuclear weapons. Magnetic fusion is being developed in an extremely open international framework due to the fact that its applications are focused solely on the production of energy. What follows is focused on this second approach.

1.4 Magnetic Boxes to Contain Plasma

Plasma is an electrically conducting fluid electrically neutral from the outside in which ions and electrons move practically independently of each other. Submerged in a magnetic field, ions and electrons will follow helical trajectories winding around field lines and will be forced to move along the field. This is the magnetic confinement principle. Initially, straight (or cylindrical) topologies were examined. However, these featured a drawback in that they let the plasma escape at extremities. To avoid this, the cylinder was closed on itself, resulting in a toric configuration, frequently characterized by the ratio between its major radius and its minor radius, referred to as the aspect ratio (Fig. 2). However, in this type of configuration, the curve (and therefore centrifugal force) and lack of homogeneity of the field (the magnetic field is stronger on the inner surface of the torus than on the outer surface) give rise to a drift of charged particles. Ions and electrons tend to separate; some move to the top and the others move to the bottom and end up leaving the magnetic trap. To compensate for this effect, the field lines were modified and made helicoidal. Thus, particles successively cross at the top and then at the bottom of the magnetic configuration. Therefore, the drift effect, which is always in the same direction, is on average compensated. This is achieved by adding another magnetic field (“poloidal” field) perpendicular to the toric field (“toroidal” field). The method implemented to generate these helicoidal field lines gives rise to two types of machines:

1. In a “tokamak,” an axial current circulating in the plasma itself creates the poloidal magnetic field.



R : major radius
 a : minor radius
 $A = R/a$ = aspect ratio
 b/a = elongation
 c/a = triangularity

FIGURE 2 Cross section of a toric fusion plasma (main geometrical characteristics).

This field becomes the secondary circuit of an enormous transformer. Thus, the tokamak is an impulse device (Fig. 3).

2. In a “stellarator,” the magnetic configuration is based entirely on currents circulating in external helicoil coils. This configuration, because of its greater complexity regarding principles and geometry, is lagging behind development of the tokamak configuration. However, it does feature some inherent qualities that have motivated continuation of work in this area.

Plasma behaves like a gas and exerts a (kinetic) pressure toward the exterior, the value of which increases with temperature and density. To confine the plasma, this pressure must be balanced by pressure toward the inside. This is the function of the (magnetic) pressure exerted by the magnetic field. In practice, to avoid the appearance of instabilities, kinetic pressure must be much less than magnetic pressure (a factor of 10). Therefore, for a given magnetic field and temperature, this places a limit on density. With standard temperature (10–20 keV) and magnetic field (5–10 T) values, this limit density is on the order of 10^{20} m^{-3} . As a consequence, this also sets the confinement time value required for magnetic fusion: on the order of a few seconds (refer to the Lawson criterion).

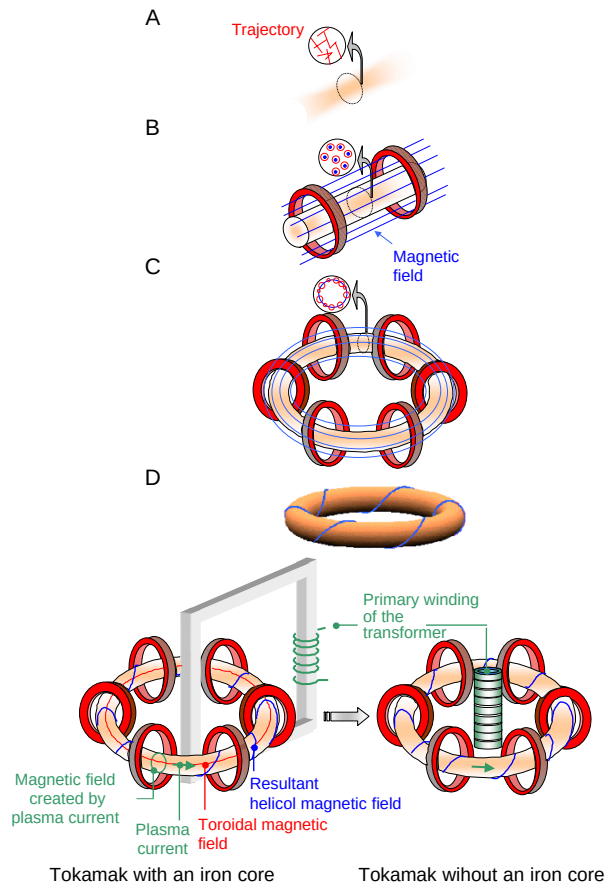


FIGURE 3 Main confinement and tokamak principle.

The stability limit is frequently characterized by the parameter β_n that is a function of plasma topology (e.g., aspect ratio, shape of the poloidal section) as well as current and pressure profiles. This parameter is on the order of 2.5 to 3.0% in current configurations but can achieve 5.0% in “advanced” configurations.

1.5 How Is Plasma Heated?

Irrespective of the manner in which the plasma is created within a confinement structure, it does not start at a temperature necessary for fusion reactions. There are three methods available to heat a plasma:

1. The current circulating in the tokamak plasma heats the plasma by the Joule effect. The latter is efficient up to a temperature on the order of 1 keV. Beyond this temperature, plasma resistivity becomes too weak and this method becomes less efficient. In a stellarator, there is no central current and, therefore, no “ohmic” heating.

2. Heating by injection of high-energy neutral particles (neutral beam heating) involves creating and accelerating an ion beam outside the confinement containment. The beam is then neutralized by charge exchange before penetrating into the plasma, where the neutral particles become ionized again and are confined by the magnetic field. Energy is then redistributed through collisions and the plasma temperature rises.

3. Plasma can absorb energy from electromagnetic waves at the characteristic frequencies of the medium. The energy of the electromagnetic waves is transferred to the plasma by antenna covering part of the confinement area. By choosing the frequency, it is possible to define the particle species (ions or electrons) that will be heated and, to some extent, the area where the wave and heating will be absorbed.

In a magnetic confinement-type thermonuclear fusion reactor, plasma temperature may be raised to a suitable level through a combination of the methods described previously. Once the number of fusion reactions becomes significant, energy contributed by helium nuclei produced that remains confined within the plasma constitutes the main heating method ($Q > 5$).

1.6 Toward Continuous Operation in Tokamak Machine

In addition to heating the plasma, the heating methods described previously allow current to be generated and, thus, allow continuous operation of the tokamak to be envisioned (despite its inherent impulse nature).

Under some conditions, the plasma may also give rise to a plasma-generated toroidal current referred to as the bootstrap current resulting from collisions between circulating electrons and locally trapped electrons. The fraction of this current with respect to total plasma current is highly dependent on temperature and density profiles. Development of reliable plasma scenarios favoring this type of current (e.g., how to maintain good profiles, where and how heat should be applied and current should be generated) is of utmost importance from the standpoint of steady-state operation, which will be the common state of a commercial fusion reactor.

2. THE FUSION REACTOR

A schematic diagram of a fusion tokamak-based reactor is provided in Fig. 4. A general description is

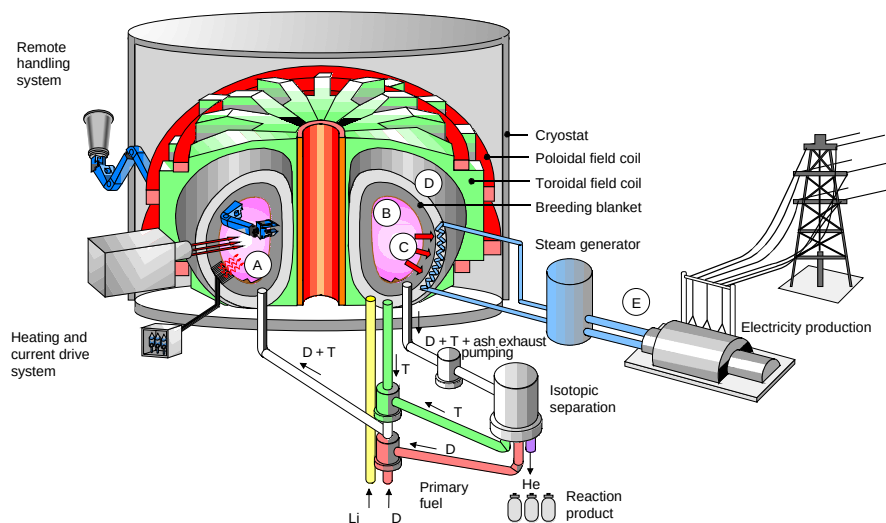


FIGURE 4 Schematic view of a fusion reactor based on a magnetic confinement: (A) vacuum chamber; (B) plasma; (C) plasma radiation; (D) blanket; (E) electricity production.

dealt with briefly here before examining the main reactor elements in detail. The deuterium/tritium fuel mixture is injected in gaseous or frozen ice pellets form into a vacuum chamber (Fig. 4A) where, by means of a confinement and heating system, it passes to the plasma state and burns continuously (Fig. 4B). The plasma produces ash (helium atoms for a D–T plasma) and energy in the form of particles and radiation (Fig. 4C). Ashes are extracted in gaseous form and are processed outside the reactor. Charged particles and radiation are absorbed, losing their energy in the first physical element surrounding the plasma (the so-called “first wall”). Neutron kinetic energy ($\sim 80\%$ of the total energy for a D–T plasma) is converted (or even amplified through suitable nuclear reactions) into heat in the so-called “blanket” (Fig. 4D) component located after the first wall but still within the vacuum chamber. The vacuum chamber encloses the space where the fusion reaction takes place. The first wall, blanket, and vacuum chamber are cooled down by a heat extraction system. The heat is used to generate steam that supplies a conventional turbine and alternator electricity generating system (Fig. 4E).

2.1 Reactor Fuels

As already mentioned, the fusion reaction involving deuterium and tritium is the most accessible and, obviously, is the reaction on which the most credible concepts are based. Tritium is a radioactive element with a relatively short half-life (12.3 years) that will be produced on-site by neutron capture in lithium.

Therefore, the primary fusion fuels are deuterium and lithium (Fig. 1), two nonradioactive elements.

Deuterium is a stable isotope of hydrogen. It is very abundant and may be extracted from seawater (33 g of deuterium per m^3 seawater) by conventional industrial processes (e.g., distillation, electrolysis, isotopic exchange). The estimated resource in oceans is on the order of 4.5×10^{13} metric tons, corresponding to an energy potential of 5×10^{11} TW per year. Bearing in mind that current global energy consumption is on the order of approximately 12 TW per year (year 1990), deuterium energy resources would exceed the lifetime of the sun (~ 5 billion years).

The lithium content in Earth’s crust is approximately 50 ppm (0.05 g/kg). It is more abundant than tin or lead and is even 10 times more abundant than uranium (3–4 ppm). Lithium occurs naturally in a mixture of two isotopes: lithium-6 (7.5% of naturally occurring lithium) and lithium-7 (92.5% of naturally occurring lithium). Although it would be possible to use naturally occurring lithium in a fusion reactor, mixture enriched with lithium-6 (40–90% according to the design) would be preferable because, in this case, reactions are exothermic and feature higher cross sections. The enrichment processes are well known and have been validated (lithium-6 is used to produce tritium for nuclear weapons). Lithium may also be extracted from seawater that contains 0.17 g/m^3 , constituting a potential reserve of 230,000 million metric tons. Proposed extraction methods are based on conventional ionic exchange, solvent extraction, or coprecipitation-type chemical processes. In a fusion reactor generating 1 GWe per year,

between 0.5 and 3.5 metric tons of naturally occurring lithium would be consumed depending on the enrichment implemented. Therefore, the use of telluric lithium would ensure reserves for several thousand years, and use of lithium extracted from seawater would push this limit to several million years.

Use of “advanced” fuels (D–D or even D– ^3He reactions) features the dual advantage of avoiding use of tritium and production of high-energy neutrons. However, conditions regarding temperature, density, and confinement time necessary to compensate the small cross sections render their implementation extremely hypothetical and, in any case, impossible in the near future. It should also be emphasized that in a reactor based on these reactions, nearly all fusion energy produced is transferred to the first wall elements, accentuating the constraints on these components that were already very high (in the case of D–T reactions, 80% of fusion energy is carried by neutrons that give up their energy in the tritium breeding blanket).

2.2 Reactor Physics

Since the beginning of the tokamak era at the end of the 1960s, considerable advances have been made, both regarding understanding of physical phenomena and regarding technologies implemented in the construction of experimental tools. These results were obtained for numerous facilities of various sizes designed during the 1970s and operated since the beginning of the 1980s. Two highlights should be retained:

1. One is the 16 MW of fusion power obtained in the European device, JET, over approximately 1 s in 1997, with a power amplification factor, Q , of approximately 0.65, close to break-even conditions ($Q=1$).

2. Another is long-duration plasmas produced in the Tore Supra device (Cadarche, France), fitted with superconducting magnets and plasma facing components actively cooled by water circulation. These technologies enabled several discharges of a few minutes such as the record discharge on September 18, 2002, of $4\frac{1}{2}$ min, sustained by a heating power of 3 MW and requiring to exhaust more than 750 MJ of thermal energy during the experiment (Fig. 5).

Reliable responses to most questions raised by the studies concerning controlled fusion have already been provided, albeit independently. The next step should involve integration of all these results within a single facility. This is one of the goals of the

International Thermonuclear Experimental Reactor (ITER) project.

The objectives of current devices target mainly qualification of confinement scenarios selected for next generation machines. Thus, it was discovered that confinement improves with increasing machine size deteriorates with increasing additional heating power coupled to the plasma and that, under some conditions, there is a power threshold where confinement is suddenly improved. This improved confinement system is referred to as H mode (for “high confinement”), as opposed to the confinement mode obtained below the power limit, referred to as L mode (for “low confinement”). This allows the confinement time to be improved by a factor of nearly 2 with respect to L mode. Discovery of this improved confinement mode was critical, and H mode currently constitutes the reference scenario for next generation devices, notably for ITER. Stabilizing mechanisms that allow reach into H mode are not yet entirely understood but are based on stabilization of turbulence, which is the source of confinement deterioration. A transport barrier is created at the edge of the plasma, holding heat and particles in the plasma core. The most characteristic point of this scenario is the appearance of strong gradients in the edge zone. These extremely steep gradients at the edge give rise to instabilities specific to H mode, referred to as ELM (for edge localized modes). The plasma pressure profile relaxes periodically toward shallower slopes, releasing bursts of particles and heat that escape from the plasma at each ELM, thereby imposing significant constraints on vacuum chamber components.

Results acquired on a global scale from tokamak experiences were gathered in a database from which an empirical scale law was established expressing the confinement time based on the main parameters of machines and plasma. Although this method is empirical, it is extremely important to be able to extrapolate current performance levels to those of next generation machines.

Physics options adopted for reactor operation are obviously based on strong or weak extrapolations of current scenarios. They target continuous operation or at least operation with extremely high discharge durations (>12 h) and, therefore, are based on implementation of scenarios favoring a significant fraction of noninductive current. These scenarios guarantee good confinement and control of turbulence phenomena that could damage first wall elements. Reactor size is highly dependent on these extrapolations. Diagrammatically, two major extrapolation classes limit the domain (Table I):

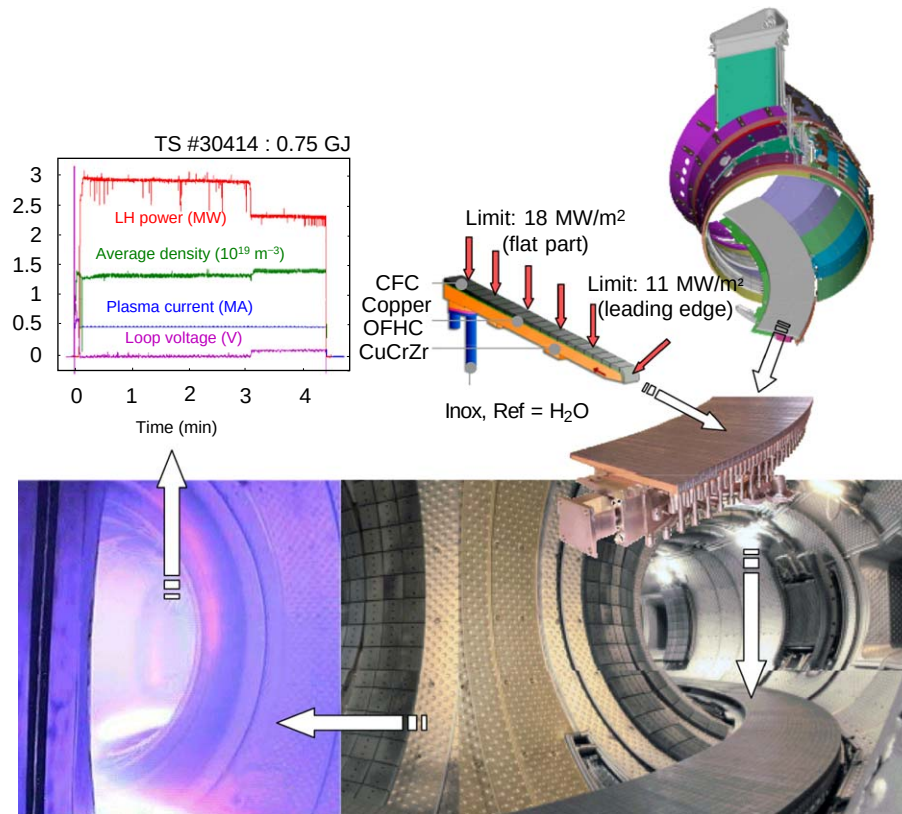


FIGURE 5 Record discharge reached in Tore Supra on September 18, 2002: a plasma discharge lasting 4 minutes 25 seconds requiring the exhaust of more than 750 MJ of thermal energy. This remarkable result was obtained due to the actively cooled plasma facing components, ensuring that every square centimeter of the wall in front of the plasma is actively cooled.

1. *Reactors involving moderate extrapolations from those adopted for ITER.* However, they imply suppression of H mode ELM and control of power transmitted to first wall components by controlling edge plasma radiation (e.g., injection of impurity). The bootstrap current fraction is generally below 50%, leading to a heating power that is quite high and, therefore, a recirculating power (power taken from the system to supply heating sources) on the order of 20%. The reactor efficiencies are similar to those acquired with conventional nuclear power plants (30–40%). Regarding size, the major radius is typically approximately 9 m (1500 MWe class).

2. *Reactors with large extrapolations from current knowledge but still credible.* These are generally accompanied by the same extrapolations at a technological level. Plasma parameters and topology are selected to maximize high-bootstrap current fraction, close to 75%, minimizing recirculating power. Instabilities are assumed to be completely controlled, as is plasma radiation at the edge where a radiative “mantle” is established without any reper-

cussions on confinement and performance levels of the central plasma. This leads to heat fluxes on the first wall components lower than those of previous designs (factor 2) and a reactor size close to that adopted for ITER, that is, typically 6 m (1500 MWe class).

2.3 Plasma Facing Components

The plasma facing components constitute the first physical surface encountered by the plasma. Excluding accidental situations, the plasma does not come into contact with this surface. In normal operation, plasma facing components are exposed to neutronic, thermal, radiative, and thermomechanic stress loading. Thus, the management of the interface between the plasma and this wall is critical. Furthermore, the main results obtained concerning confinement, like H mode, were based on efficient management of this interface. Several types of phenomena may enter into play at the wall, for example, absorption of large quantities of gas issued by the plasma, erosion due to

TABLE I

Main Parameters of Two Generic Fusion Reactors

Parameter	ITER	Reactor type 1 (small extrapolation)		Reactor type 2 (advanced)	
Fusion power (MW)	500	3500		2500	
Electrical power (MW)	—	1500		1500	
Net reactor efficiency (percentage)	—	35–45		60–65	
Q	10	20		35	
Plasma duration	400 s	Continuous		Continuous	
Major radius (m)	6.2	9–10		6	
Minor radius	2.0	3.0–3.3		2	
Aspect ratio	3.1	3.0		3	
Plasma current (MA)	15	30		15	
Toroidal field on axis (T)	5.3	7		5.6	
Toroidal field on coil conductor (T)	12	13		13.5	
Elongation (95%/separatrix)	1.7/1.85	1.7/1.9		1.9/2.1	
Triangularity (95%/separatrix)	0.33/0.48	0.27/0.4		0.47/0.7	
Normalised β	2.0	3.3		4.5	
Average temperature (keV)	10	20		12	
Average density ($10^{20}/\text{m}^3$)	1	1.1		1.4	
Bootstrap fraction (percentage)	Experiment	35		75	
P_{ext} : heating power (MW)	70–110	250		70	
Average neutron load (MW/m^2)	0.6	~ 2		~ 2.5	
Average divertor heat load (MW/m^2)	10	10–15		5	
Armor material	Be, CFC, W	W		W	
Breeder	None	Pb-17Li	Li_2TiO_3 (pebble bed)	Li	Pb-17Li
Multiplier	None	None	Be (pebble bed)	None	None
Coolant	Water	Water	He	Self	Self
Typical outlet temperature ($^{\circ}\text{C}$)	150	325	500	600	1100
Structure	316	Ferritic	Steel	V alloy	SiC/SiC
Reactor efficiency (percentage)	—	33	37	45	60

Note. Type 1 is a “small extrapolation” front current physics and technology. Type 2 is more advanced and is based on large (but credible) extrapolations.

the action of fast particles, heating (the heat flux can reach several megawatts/square meter). Furthermore, plasma radiation losses vary significantly with the quantity and type of impurities that plasma contains (radiation increases with the atomic number of the relevant impurity). Therefore, the plasma facing wall must be designed, on the one hand, to minimize this type of pollution and, on the other, to withstand particles and radiation from the plasma. To illustrate the problems and issues associated with these components, we propose focusing on the so-called divertor.

All charged particles present in the plasma are subject to a confinement effect caused by the magnetic field. Consequently, if no action is taken, Helium ion concentration within the plasma in-

creases with the combustion rate. This effect must be limited because not only is the fuel diluted, but losses through radiation also increase. A specific device referred to as the divertor (Fig. 6) manages continuous extraction of reaction ash and, thus, maintains its concentration in the plasma at an acceptable level. It involves creating, by means of the poloidal magnet system, a final closed magnetic surface that stretches in two sections to the divertor. Beyond this separatrix, all magnetic surfaces are open and reach the wall. A particle first confined close to the magnetic axis moves slowly away by diffusion to reach the closed magnetic surface and will then encounter the divertor wall. Here, it is neutralized and, thus, may be extracted by pumps that continuously empty this zone. The divertor fulfills three functions that

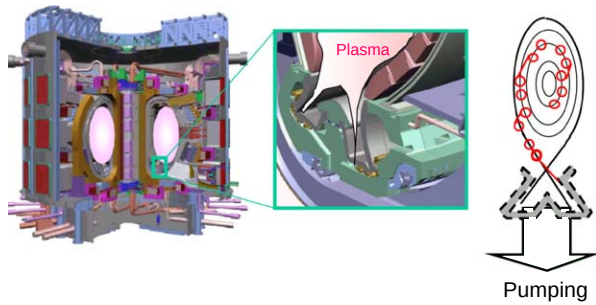


FIGURE 6 Principle of the divertor.

have extremely significant repercussions for the plasma:

- Extraction of particles during deconfinement (e.g., helium, deuterium, tritium, impurities) in a precise and controllable location
- Extraction of heat (particles striking the divertor lose their energy here)
- Control of plasma density through the divertor pumping capacity

From a technological point of view, heat extraction at the level of several megawatts per squared meter is rather challenging. Creation of a divertor presents technological difficulties associated with extraction of heat. Particle flows along the two lines of the separatrix create thermal fluxes of several megawatts per squared meter. These fluxes can be attenuated by locally injecting impurities (e.g., neon, argon) to increase radiation in this zone and by sweeping the separatrix over the entire surface of the divertor.

Components constituting the divertor are made of armor tiles mounted on a cooled structure. Numerous designs were developed through judicious selection of materials, including the armor tiles, structural materials, joining methods, and coolant.

Designs validated on current machines include a stainless steel supporting structure to which carbon fiber composite (CFC) and hardened copper (CuCrZr) armor tiles are attached and are capable of withstanding continuous thermal fluxes on the order of several megawatts per squared meter (Fig. 5).

The ITER plasma facing components design was derived from these concepts. The armor tiles will be made either of a low Z material, such as CFC or beryllium, or a refractory material, such as tungsten, for the most exposed sections of the divertor. These technologies were tested successfully in laboratories in specific installations under thermal fluxes reaching 20 MW/m^2 over thousands of cycles, but they have not yet been qualified at the industrial level.

From a reactor standpoint, objectives regarding lifetime, reliability, and safety must be satisfied, as must economic criteria, along with the constraints already imposed by plasma physics. Notably, the energy transferred to the first wall and the divertor (20% of energy released by a D-T plasma) must be recovered to be used to generate electricity. Therefore, the coolant outlet temperature must be as high as possible. Safety criteria, notably those involving authorized tritium inventories, lead to the exclusion of carbon fiber composites because these materials trap and easily retain tritium. The mechanical strength of beryllium and its erosion rate render implementation of this material in a reactor environment difficult. Tungsten is the most promising material, although its mechanical strength under irradiation is still being studied. Its implementation is the basis of plasma facing components used in several reactor designs (Fig. 7). Schematically, there are four divertor types:

1. *Designs based on extrapolations that are limited with respect to technologies implemented in ITER.* They use water under pressurized water reactor (PWR) conditions as coolant. This type of design allows the withstanding of heat fluxes of 15 MW/m^2 . The structural material constituting the cooling tube is a hardened copper alloy (CuCrZr) featuring thermal conductivity properties 30 times greater than that of the stainless steel used in ITER. The armor material is tungsten jointed to a copper tube through of a compliance layer. The operation constraints of these materials, as well as deterioration of CuCrZr thermomechanical performance levels under thermal cycling, are taken into consideration. Uncertainties are also clearly identified, as is the behavior of CuCrZr under irradiation. Use of CuCrZr imposes a coolant temperature that is quite low (150°C). Optimized versions using steel and thermal barriers allow a higher coolant temperature (325°C) and improved efficiency.

2. *Designs involving helium coolant, tungsten as the armor material, and steel as the structural material.* This type of design allows resistance to heat fluxes on the order of 10 MW/m^2 while using a high-temperature coolant (500°C or even 800°C in some optimized versions).

3. *Designs involving liquid metal as the coolant and tungsten as the armor material.* These designs generally use Li-17Pb as the coolant and SiC-SiC-type ceramics as the structural materials. This would require highly significant technological advances, notably concerning improvement of thermal-mechanical properties and the resistance of these materials

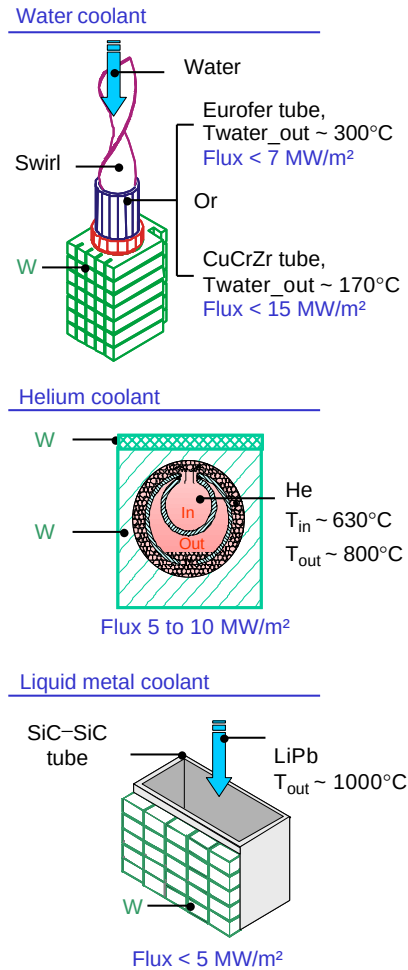


FIGURE 7 Different types of high heat flux components for divertor concepts.

when subject to irradiation. They would allow coolant temperatures of approximately 1000°C. These designs are generally intended for reactor versions that also use significant extrapolation regarding physics and adequate control of the plasma edge.

4. *Alternative designs.* There are other “nonconventional” designs in which the protective material is replaced by liquid metal (e.g., lithium, gallium) in the form of a continuously flowing film, drops, or in pool. The idea is to overcome problems of erosion, lifetime, and thermal load associated with solid materials (this type of design supports fluxes of several tens of megawatts/squared meter). However, some issues are also under investigation regarding, for example, the effects of these liquid surfaces on the plasma edge or the pumping capacity of helium and impurities.

Regarding the reactor, plasma facing components and the divertor specifically remain an open issue.

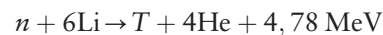
The main difficulty concerns integrating complicated phenomena that control the plasma edge and technological constraints due to materials. ITER will provide some crucial answers in this area regarding feasibility and operating procedures for a reactor.

2.4 The Tritium Breeding Blanket

The element located after the plasma facing components is called the blanket (Fig. 4). This component performs several functions:

- First, the blanket provides shielding by significantly reducing neutron/gamma radiation energy to protect the outer following components (e.g., vacuum chamber, magnetic system). Its second function is to recover the energy left by the neutrons by heating materials. A coolant circulates in the structure and evacuates the heat generated to conventional equipment for electricity production.

- In the case of a reactor based on the D–T reaction, the blanket must also produce tritium necessary for fusion reactions by neutron bombardment on lithium:



One should note that the first of these two reactions is exothermic and, thus, is preferable. This requires 6Li-enriched lithium to be used. This enrichment operation is simple and already mastered in industry (for military purposes). It may also be noted that the blanket is the site of reactions generating energy that altogether account for 20% of the reactor balance. Lithium may be in a solid form (ceramic) or a liquid form (metallic alloy) depending on the blanket design implemented. The presence of a tritium breeding material does not suffice to ensure suitable production of tritium. Ended, the reaction of a tritium and deuterium atom produces only one neutron, and the reaction of this neutron on lithium produces only one tritium atom. Because neutron losses are inevitable, an adequate amount of tritium cannot be produced under these conditions. To remedy this drawback, a neutron multiplier is introduced into the blanket, allowing the appropriate balance. Lead or beryllium may be used as a multiplier. Blanket tritium-generating capacity is characterized by the self-sufficiency ratio (or tritium breeding ratio). If the tritium breeding ratio is greater than 1, the blanket produces more tritium than is consumed by the reactor.

Numerous tritium breeding blanket designs are available that fulfill the three functions just

described. They differ in terms of tritium breeding material type, coolant type, and structural material types implemented. The combinations adopted are the result of trade-offs regarding compatibility of materials with each other and the allowable operation windows (e.g., operation temperature, swelling resistance). The choice of coolant will also depend on solutions adopted for plasma facing components, with a single coolant being preferable for obvious reasons of simplicity. The choice of structural material is also vital because, in general, it is the structural material that dictates performance levels in terms of efficiency (via its maximum allowable operation temperature) and regarding lifetime of blanket elements. The main criterion is resistance under 14-MeV neutron irradiation, where a maximum of 150 dpa is usually adopted, corresponding to approximately 5 years operation at full power. Therefore, periodic replacement of blanket components is provided for in the design. The long-term induced activation of components can be tailored by proper selection of materials to avoid generation of waste that would require deep geological disposal. These materials (so-called “low-activation” materials) could be steel in which elements that are penalizing regarding activation are replaced with other elements that are more interesting but compatible from a metallurgical standpoint (e.g., nickel, molybdenum) or other material families. These are presented in order of development risk:

- Low-activation ferritic–martensitic steels have been developed to the point where their use in a nuclear component may be possible in the relatively short term. They are notably developed in Japan and in Europe, with the latter having focused its studies on said materials. A particular grade was selected (EUROFER steels) and has already given rise to significant melts (a few metric tons), allowing the manufacture of samples that were subject to numerous tests (behavior under irradiation, corrosion, welding, etc.). The operation temperature is approximately 550°C, or even 600 to 650°C, through specific optimizations (e.g., ODS steel, composition optimization). Operation temperature windows lead to efficiency of approximately 35 to 40%. The main question that still remains open concerns the effect of 14-MeV neutrons and, specifically, the effect that the formation of hydrogen and helium produced by transmutation has on embrittlement.

- Vanadium alloys are interesting in that they can be used in up to 700°C. This would allow yields of up to 45%.

- Silicone carbide composites (SiC–SiC) could operate at very high temperatures (1000°C) and, therefore, allow access to efficiency close to 60%. However, their status of development would not allow them to be used in the short term.

Schematically, there are two main types of blankets.

2.4.1 Solid Tritium Breeding Material Designs

In this type of blanket, the tritium breeding material is a lithiated ceramic (Li_2O , Li_4SiO_4 , Li_2TiO_3 , or Li_2ZrO_3) and beryllium is used as the neutron multiplier material. These materials are generally in the form of pebble beds (diameter on the order of millimeters). Tritium is extracted by gas circulating (He) that is in contact with the tritium breeding material and transports tritium outside the blanket. The structure is made of low-activation ferritic–martensitic steel, possibly hardened (ODS-type steel) to allow a higher coolant temperature. Coolant may be either pressurized water (typically 320°C and 15 MPa) or helium (typically 500°C and 8 MPa), allowing efficiency of between 35 and 40%. The main difficulties taken into consideration in this type of blanket design are as follows:

- Chemical compatibility of beryllium and water that may induce production of hydrogen and a risk of explosion
- Tritium permeation in water
- Evolution of pebble beds (ceramic and beryllium) when subject to irradiation
- Lower shielding capability if helium is used as a coolant
- Fuel manufacturing costs, the renewal obligation due to consumption of lithium, and fuel reprocessing to recover lithium that has not been consumed

2.4.2 Liquid Tritium Breeding Material Designs

In this type of blanket, the tritium breeding material is a liquid metal, either metal lithium, an eutectic of lithium and lead (Pb–17Li), or a molten salt composed of lithium, beryllium, and fluorine called Flibe (Li_2BeF_4). Usually, the multiplier material is incorporated into the tritium breeding material. Liquid metal circulates in the reactor and, outside the reactor, undergoes processing designed to extract tritium present in dissolved form. This also allows online control of the lithium-6 titer, thereby avoiding depletion of the fuel in lithium and the requirement to replace lithium periodically. The Pb–17Li eutectic

is frequently used due to its significant tritium production capacities, high thermal conductivity, ability to withstand irradiation damage, and inertness with air. Pb–17Li compatibility with ferritic–martensitic-type steels has been proven in up to 500°C. Use of helium as a coolant leads to difficulties in achieving self-sufficiency in tritium, and water (or even the liquid metal itself) is preferred. This gives rise to two types of liquid tritium breeding blankets:

1. *Water-cooled liquid blankets (typically 325°C and 15 MPa).* These are generally associated with steels as the structural material and give rise to efficiency of approximately 35%.

2. *Self-cooled blankets where the tritium breeding material itself acts as the coolant.* This is implemented at high temperatures and is associated with a compatible structural material (vanadium alloy or SiC–SiC composite). Efficiency can be as high as 60%.

The economic performance and technological maturity of the various designs is provided in Fig. 8.

2.5 The Magnetic System

Coils, in which an electrical current circulates, are used to create the magnetic fields necessary for reactor operation; current is constant in the toroidal coils and variable in the poloidal coils. Toroidal coils are typically more than 10 m in height, and the largest of the poloidal coils measure more than 20 m. Currents passing through said coils are high intensity

(a few tens of kiloamps) and, therefore, give rise to an electrical energy consumption problem. For a reactor under steady-state operations, implementation of conventional coils (water-cooled copper coils) severely depletes the total energy balance; therefore, superconducting coils should be used. For these coils, electrical consumption is limited to the cryogenic system that maintains coils at very low temperatures (~ 4 K). This technology has been implemented since 1989 to generate the toroidal magnetic field of the Tore Supra tokamak and will be applied to the ITER magnetic system.

Regarding the toroidal fields, the highest values possible must be achieved due to the relationship between kinetic pressure and magnetic pressure. Limits are imposed by the superconducting materials selected that, to preserve superconductor status, must fulfill some criteria: temperature must be less than the critical temperature, T_c ; the magnetic field applied must be less than a certain value, H_c ; and the current density applied must be less than the critical current density, J_c . Furthermore, the superconducting material is fragile and loses its superconducting properties when subject to high-stress loading. Reactor designs provide for fields on the conductor in the neighborhood of 16 T at a standard operating temperature of 4 K, values that are not far from those adopted for ITER (13 T and 4 K). Within these field ranges, the operation limits of Nb₃Sn, selected for ITER, are achieved and Nb₃Al could be used. For higher fields, implementation of immature

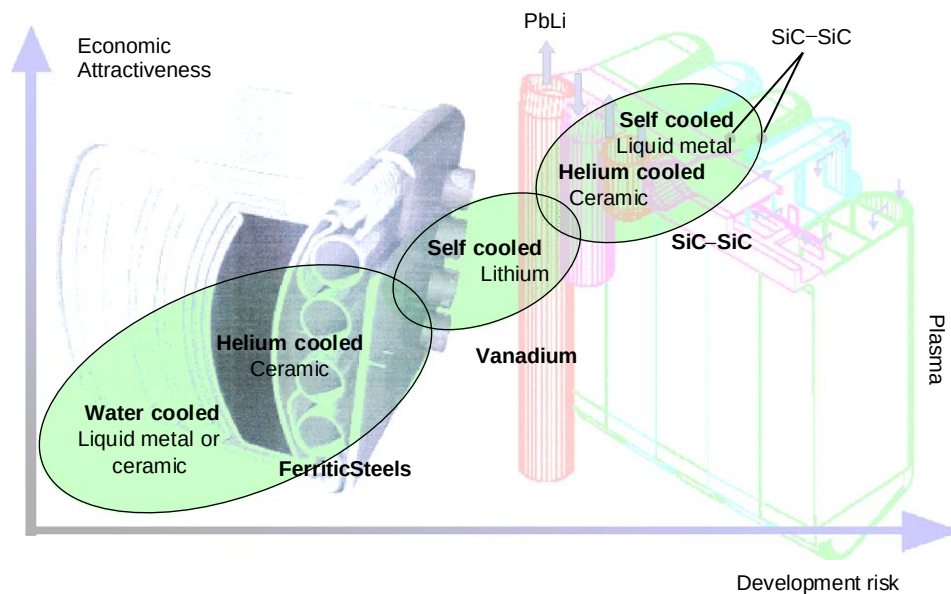


FIGURE 8 Economic attractiveness and development risk of various kinds of breeding blankets.

technologies may be considered, for example, use of Nb_3AlGe (24 T on the conductor) or even implementation of superconducting materials referred to as critical high-temperature materials.

There are several superconducting cable designs available. However, recall that the strand consists of filaments of superconducting material (NbTi or Nb_3Sn , diameter of a few microns) embedded in a copper matrix and that the total strand diameter is slightly less than 1 mm. Using this strand, a cable is made composed of more than 1000 strands positioned according to a precise architecture. The cable obtained in this manner is placed within a jacket (e.g., steel, incoloy) intended to withstand mechanical stress loading (Fig. 9). Liquid helium circulates and bathes the cable, and the entire magnetic system is placed within a huge cryostat. Manufacture of the conductor is quite sophisticated and implies other steps that are not described here, intended to ensure that superconducting properties (heat treatment) and good stability (introduction of copper strands) are achieved.

2.6 Robotized Maintenance

The operation of fusion reactors will require maintenance phases. These outage periods (scheduled or not) must be minimized to avoid penalizing facility availability (80% availability is targeted). Implementation of remote interventions is absolutely mandatory, notably to perform interventions (e.g., inspec-

tion, replacement of components) in the vacuum chamber where the radioactive environment does not allow for other solutions. Therefore, all of the objects already mentioned (e.g., plasma facing components, tritium breeding blankets) are designed for remote interventions. Even the reactor environment will be different; complex maintenance operations have already been performed successfully, for example, the fully remote exchange of the JET divertor.

Therefore, maintenance will be taken into consideration as of design, in particular for the divertor and tritium breeding blanket modules that must be replaced periodically. For example, a trade-off between segment size (limited by carrier capacity) and segment number (limited by intervention rapidity) controls the segmentation of tritium breeding blanket segmentation. Several maintenance schemes have been proposed. Their repercussions on the general design of the reactor are extremely significant. The three main schemes are as follows (Fig. 10):

1. The blanket and divertor are segmented into modules (Fig. 10A), where each module (blanket plus first wall and divertor) is removed through specific ports, the number and size of which are optimized so as to minimize intervention time.
2. The torus is divided into sectors (typically 16 sectors), with each sector including a toroidal coil, the entirety of the first wall, and the blanket (Fig. 10B). Connection interfaces are minimized, and the number of objects to be handled is reduced but leads to a sophisticated manipulation.

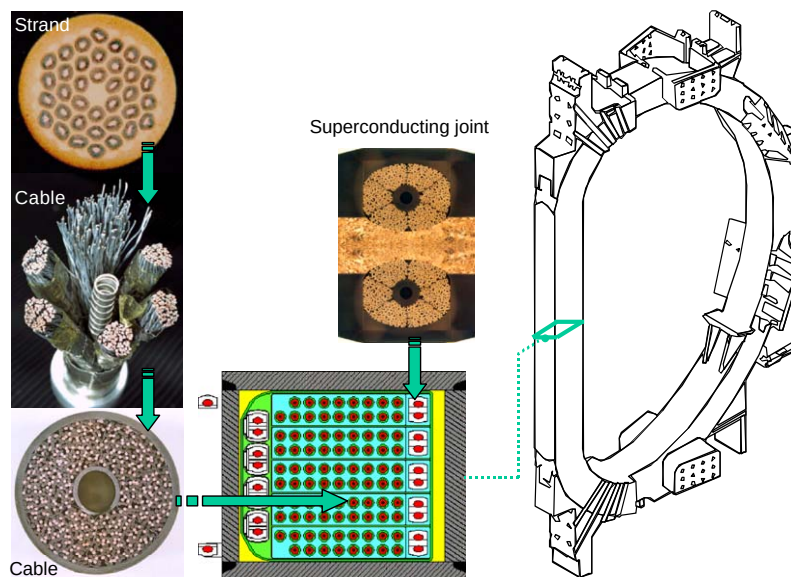


FIGURE 9 Example of superconducting cable (ITER toroidal field coil cable).

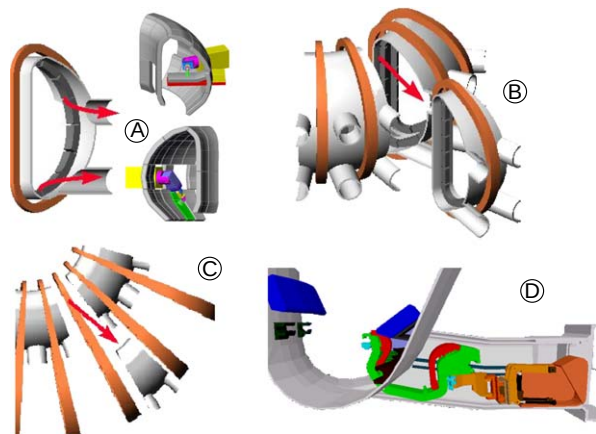


FIGURE 10 Possible maintenance schemes for divertor and blanket modularized blanket (A), large sector with toroidal coil (B) or without toroidal coil (C), and divertor maintenance (D).

3. The torus is divided into sectors that may be handled whole with the exception of toroidal coils (Fig. 10C). This assumes that toroidal coils are large. Poloidal coils are trapped and are manufactured in place. This configuration is interesting because it frees the radial space, thereby facilitating maintenance operations because a sector may be replaced in a few tens of hours.

2.7 Safety

The viability of fusion as a future energy source will be partially determined by factors associated with safety and the environment. Notably, two fundamental objectives need to be verified:

1. It must be demonstrated that no accident would require evacuation of the population.
2. Waste should not constitute a burden on future generations.

To fulfill both of these conditions, the fusion reactor falls back on its inherent advantages (e.g., the absence of any risk of a runaway reaction) and on the integration of solutions offered by implementation of structural materials referred to as low-activation materials in solution design.

2.7.1 Advantages Regarding Safety

The primary fuels of the reactor are deuterium and lithium, two nonradioactive and nontoxic bodies. Small quantities of these elements will be used (<1 kg of deuterium and <10 kg of naturally occurring lithium are enough to supply a day's generation of 1000 MWe). Therefore, management

of procurement, transport, and storage of fuel would appear to be relatively simple.

Fusion reaction conditions require the use of a plasma that is not dense (a few grams of fuel in a volume of several tens of cubic meters), is extremely pure, and is at a high temperature. Therefore, the quantity of fuel in the combustion chamber during the reaction is always very low. The slightest uncontrolled interference with this environment causes the rapid cooling and automatic shutdown of the fusion reaction. There is no possibility of uncontrolled power runaway because inherent processes limit reactivity excursions of the plasma. After the plasma shutdown, the residual energy is low and the structures cannot suffer any major damage due to accidents. Thus, the main safety function to ensure is confinement (for a fission reactor, control of the reaction and removal of residual power must also be ensured). Particular attention is paid to the confinement strategy due to the presence of tritium characterized by high diffusion in most materials. Thus, design engineers have focused on minimizing the tritium inventory, both in the torus and in the fuel cycle. To minimize emissions in the event of an accident, the "multibarrier" approach is generally adopted. This provides for three confinement barriers: the vacuum vessel, the cryostat, and the reactor building itself. Analyses that were as detailed as possible concerning energy inventories, accidents likely to occur, and expected repercussions were undertaken. These works revealed that there is no accidental sequence that could impugn reactor integrity. Even in the event of a severe accident, it would not be necessary to evacuate the population around the site.

2.7.2 The Absence of Long-Term Radioactive Waste

With the exception of first start-up that requires an initial tritium load, a fusion reactor does not require transport of radioactive material.

At the end of the fusion reactor life, material surrounding the plasma comprising the reactor structure will be activated. Implementation of low-activation materials (or, more precisely, rapid decay time materials) will allow the induced radioactivity to be minimized.

The quantity of materials to be considered varies according to the reactor design, generally between 60,000 and 100,000 metric tons. After a period of 100 years following final decommissioning of the reactor, the majority (or even entirety) of materials may be considered as very low-activity waste (meeting nuclear waste declassification standards defined

by the International Atomic Energy Agency) or may be recycled in the nuclear industry. This quality may be illustrated by a strong image (Fig. 11); after 100 years of decay, the average radioactivity of fusion reactor materials is lower than that of ash from an amount of coal that would have generated the same quantity of energy (coal contains trace amounts of uranium and thorium). Elimination of fusion waste by the generation that resulted in its creation constitutes a realistic objective. Specifically, there is no requirement to provide for deep storage.

2.8 Economy

It seems premature to speak of kilowatt-hour production costs for an energy source that is still several tens of years away from commercialization. The first objective is to illustrate the repercussions of various physical variables or technological production assumptions on costs. These results, which set the relative orders of magnitude and variation directions, indeed have a direct impact on development strategy.

The second point of interest constitutes verification that the proposed device complies with market demand. Economic models used correspond to a direct extension of models used for the design, optimization, and cost calculations of current machines or projects such as ITER. From many standpoints, ITER is similar to the reactor, but above all, construction costs were

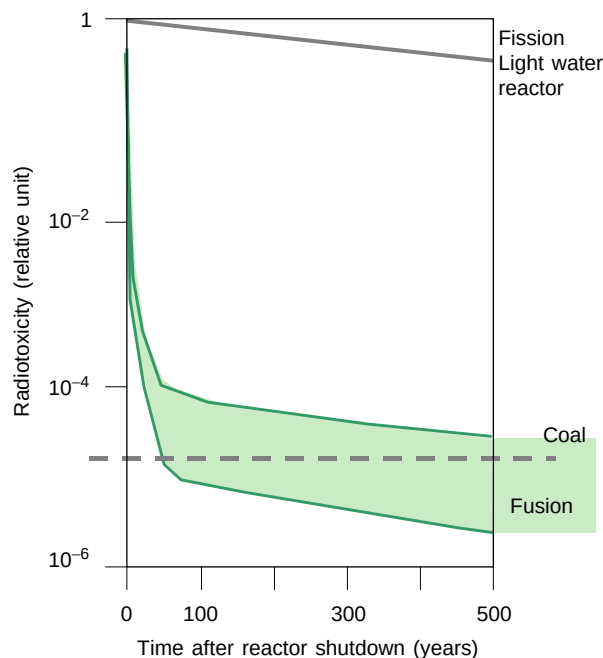


FIGURE 11 Radiotoxicity of fusion waste after reactor shutdown.

calculated directly by the involved industry in Europe, Japan, Russia, and the United States. Thus, a solid evaluation basis is available and, to return to kilowatt-hours, outstanding uncertainties are significant but concern reactor availability more than the direct cost of its components.

2.8.1 Direct Costs (or Internal Costs)

In terms of energy production, direct costs (or internal costs) correspond to costs invoiced to consumers. Put simply, they include plant construction costs, general operation costs, costs associated with fuel purchases, and costs of waste storage and plant dismantlement (regarding the nuclear industry).

The standard direct cost structure for various energy sources is provided in Fig. 12. It may be noted that for a fusion reactor, fuel cost accounts for less

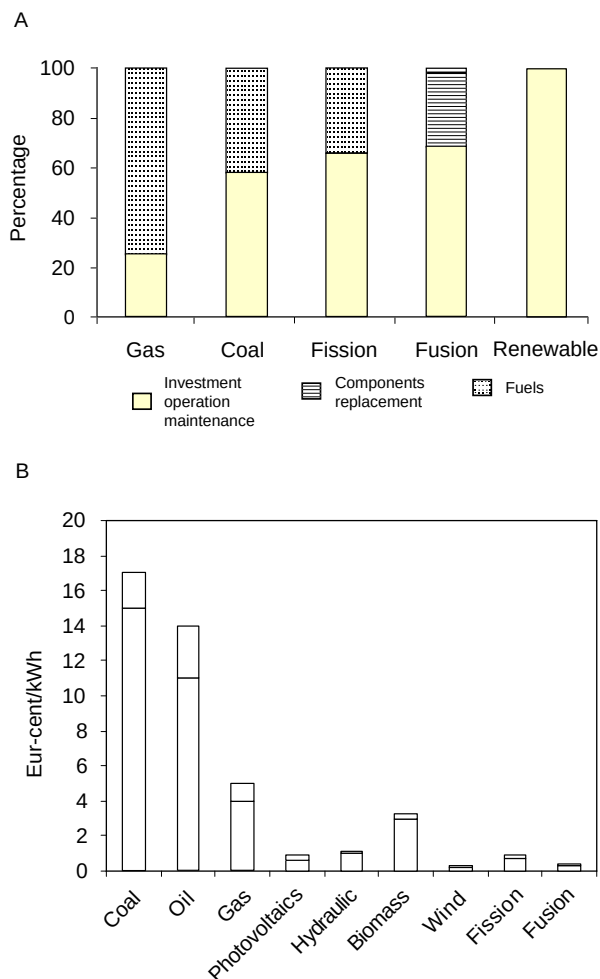


FIGURE 12 Typical breakdown of direct costs for various sources of energy (A) and comparison of externalities of several energy sectors (B).

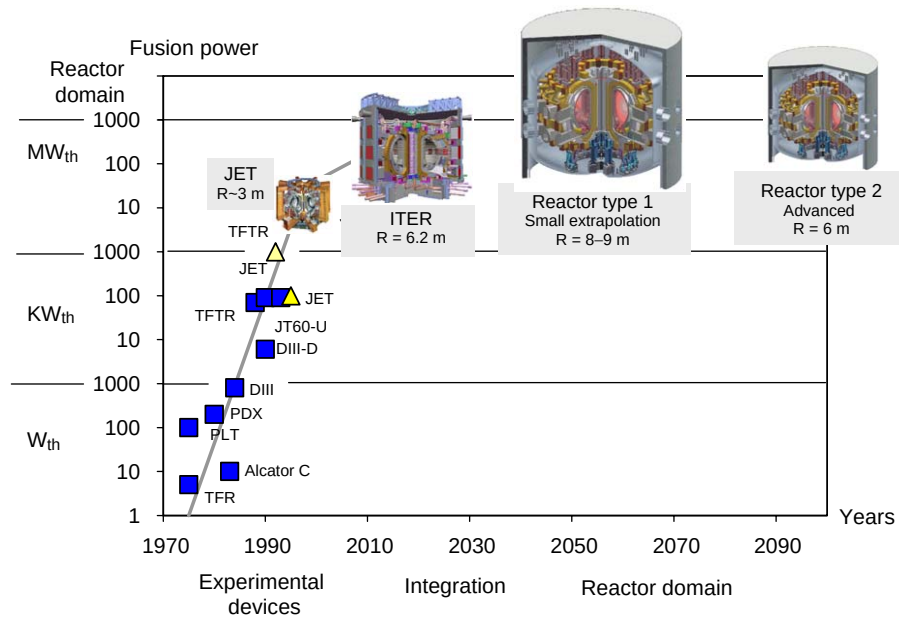


FIGURE 13 Path to the reactor.

than 1% of the total kilowatt-hour cost. Thus, the cost of energy generated by a fusion reactor is calculated by the volume of initial investment to which the cost of regular replacement of aged components is added.

The standard volume of the fusion reactor combustion chamber is on the order of 1000 m^3 . For comparison purposes, the reactor vessel volume of a fission reactor (1400 MWe) is less than 300 m^3 . Consequently, fission investment requirements are high simply due to facility size. The absence of fuel cost partially compensates for this drawback with respect to conventional energy sources (e.g., coal, gas, fission), but only slightly. Currently, economic calculations place the cost of a fusion kilowatt-hour at between U.S. \$45 million and \$90 million, with a probable value of approximately \$65 million, that is, twice the cost of conventional energy production (e.g., coal, gas, fission) and between the cost of offshore wind production and photovoltaic production (assuming that in the case of these two latter sources, evaluations do not take energy storage cost into consideration). Therefore, two results should be retained. First, and contrary to what is occasionally stated, fusion reactor production costs would not automatically jeopardize the future of fusion energy. Second, the fusion energy source is characterized by a high capital cost.

2.8.2 Indirect Costs (or External Costs)

The concept of externality or external costs corresponds to a method enabling the environmental

impact of an activity to be measured. The European Union has developed a method to assess externalities of an energy production system (ExternE studies). It takes into consideration identification of the emissions attributable to this system, the transfer of pollutants in the environment, and finally determination and quantification in terms of the cost of repercussions for the environment and health. This analysis is performed at all stages of the reactor type in question (e.g., fuel extraction, plant construction, operation, accident, dismantlement). For example, it allows consideration of harmful effects that mining operations have on health or of pollution associated with using fossil-based energy sources (e.g., respiratory problems). Applied to fusion, this method illustrates that the external costs of generating electrical energy by means of the fusion reactor are approximately the same as for renewables and are well below those obtained for fossil-fueled power (Fig. 12).

3. CONCLUSION

Fusion energy features numerous qualities. It is a nearly unlimited energy source that does not generate any greenhouse gas effect or environmental pollution, and it features undeniable advantages in terms of safety. With adequate design, radioactive wastes from the operation of a fusion plant should not constitute a burden for future generations.

Harnessing this energy is on a par with the advantages it features. Despite still very challenging open issues (e.g., physics of a plasma in combustion, heat extraction, manufacturing of sophisticated components such as the blanket, effect of 14-MeV neutrons on materials), scientific bases today are solid enough to realize a device demonstrating the scientific and technological feasibility of fusion energy. This is the main goal of the international ITER project.

Insertion into the energy offer could take place in the second part of the 21st century (Fig. 13), at a time when the exhaustion of conventional resources and climatic consequences of our consumption will start to have serious repercussions. It is absolutely clear that it is the duty of this generation to prepare the knowledge base and know-how that will allow tomorrow's decision makers to serenely consider all of the possible energy solutions. This is the objective of current research into fusion.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors •

Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Power Economics • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

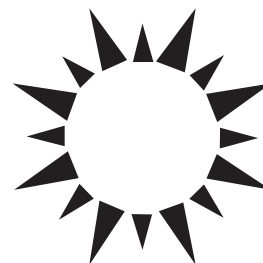
Further Reading

- Borrelli, G., *et al.* (2001). "Socio-Economic Research on Fusion (SERF): Summary of EU Research 1997–2000," European Fusion Development Agreement report. European Union, Paris.
- Cook, I., *et al.* (2001). "Safety and Environmental Impact of Fusion (SEIF)," European Fusion Development Agreement report. European Union, Paris.
- Giancarli, L., *et al.* (2002). Candidate blanket concepts for a European fusion power plant study. *Fusion Eng. Design* 49/50, 445–456.
- International Atomic Energy Agency. (2001). "Summary of the ITER Final Design Report." ITER EDA Documentation Series No. 22. AIEA, Vienna, Austria.
- International Atomic Energy Agency. (2002). "ITER Technical Basis Details." ITER EDA Documentation Series No. 24. AIEA, Vienna, Austria.
- Marbach, G., *et al.* (2002). The EU power plant conceptual study. *Fusion Eng. Design* 63/64, 1–9.

Nuclear Power Economics

GEOFFREY ROTHWELL

Stanford University
Stanford, California, United States



-
1. The Nuclear Fuel Cycle
 2. Nuclear Power Plant Technologies
 3. Nuclear Generation Costs
 4. Price and Net Present Value
 5. Nuclear Power Economic Uncertainties

Glossary

capacity factor The ratio of actual to potential energy output of a generating plant.

cost of capital A weighted average of the interest rate charged by lenders and the rate of return on investment required by investors.

decommissioning The release of a nuclear power plant from jurisdiction of the safety regulator by removing radioactive materials, such as spent nuclear fuel and the reactor core; can also involve the decontamination of all hazardous materials and the dismantling of all structures.

discounting The process of evaluating money of one period in an earlier or later period, reflecting the “time value of money.” A firm’s discount rate is usually equal to its cost of capital.

enrichment The process of increasing the percentage of the isotope uranium-235 (only 0.7% abundance in nature) in mined uranium (for example, to 4% for use as nuclear fuel in light-water reactors).

interest during construction (IDC) The cost of borrowing funds from lenders and investors during construction. The IDC is compounded, so it grows exponentially during construction.

levelize Transforming a dollar amount in one period into an annuity (a series of equal annual payments) over many periods.

nuclear steam supply system (NSSS) The nuclear reactor plus equipment required to transform heat from the reactor into steam (in water-cooled reactors), to drive the turbine-generator. The NSSS and the turbine-generator make up a generating unit. One or more units constitute the nuclear power plant.

Nuclear Waste Trust Fund A U.S. government fund to finance the management and disposal of spent nuclear

fuel; U.S. nuclear power plant owners contribute \$1/megawatt-hour of electricity produced.

reprocessing The process of removing uranium and plutonium from spent nuclear fuel, either to reduce the volume for geologic disposal or to recycle in fresh nuclear fuel.

Nuclear power economics analyzes the costs and uncertainties of generating energy with radioactive fuel in a nuclear reactor at a nuclear power plant (NPP) and compares these costs and uncertainties with other forms of energy generation. Although NPPs have been used to produce heat and to desalinate water, and might be used to produce hydrogen, commercial NPPs produce electricity, measured in megawatt-hours. As in any enterprise, generating electricity with nuclear power involves fixed and variable costs and cost uncertainties, all of which are influenced by political realities, economic conditions and policies, and social concerns.

1. THE NUCLEAR FUEL CYCLE

The generation of nuclear energy begins with uranium mining and milling. These activities are at the “front end” of the nuclear fuel cycle (Fig. 1). The cycle continues through conversion and enrichment and fuel fabrication for light water reactors (natural uranium is fuel for heavy water reactors), with each of these operations having associated costs (see Section 3.3.1). The “back end” of the cycle includes spent nuclear fuel (SNF) management, reprocessing, and high-level radioactive waste disposal. There are two types of back ends: once-through and closed cycle. Under the once-through fuel cycle, SNF is cooled at the NPP, stored or transported to a centralized facility, and finally shipped to a geologic repository, such as the one being planned in the United States at Yucca

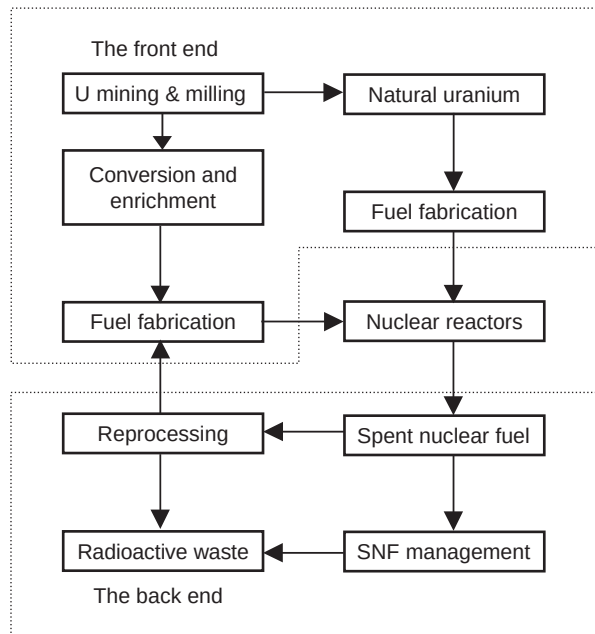


FIGURE 1 The nuclear fuel cycle. Converted and enriched uranium is used in fuel fabrication processes for light water reactors; natural uranium is used for heavy water reactors. SNF, Spent nuclear fuel.

Mountain, Nevada. The closed-cycle back end involves reprocessing SNF to recycle uranium and plutonium for fuel. SNF can only be recycled in current NPPs a few times before the buildup of unstable plutonium isotopes renders the fuel unusable. Eventually, recycled SNF is separated into waste streams, some of which are processed for geologic disposition.

2. NUCLEAR POWER PLANT TECHNOLOGIES

The different types of nuclear power plant technologies are depicted in Fig. 2, which details the numbers of each type of technology (in 2002) under construction (“Build”), in operation (“On”), and permanently retired (“Off”). (Commercial fast reactors, also known as breeder reactors or liquid metal reactors, are not discussed here.) NPPs are generally classified by the moderator (to control the chain reaction) and the coolant (to remove heat). Water (light or heavy; heavy water is also known as deuterium oxide, D_2O) and graphite are used as moderators. Water (light or heavy) and gas (air, carbon dioxide, or helium) are used as coolants. Also, NPPs use either direct or indirect cycles to

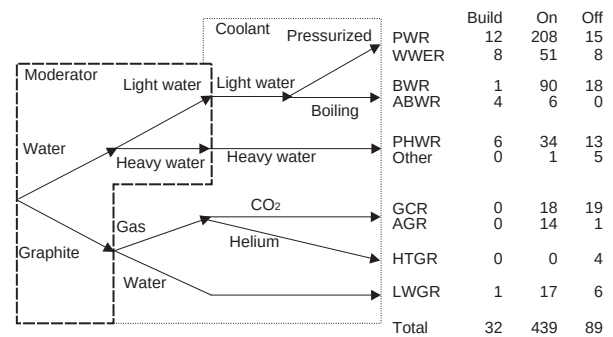


FIGURE 2 Commercial thermal reactor technologies. Key: Build, under construction; On, in operation; Off, retired; PWR, pressurized light water reactor; WWER, pressurized water electricity reactor; BWR, boiling light water reactor; PHWR, pressurized heavy water reactor; GCR, gas-cooled, graphite-moderated reactor; AGR, advanced GCR; HTGR, high-temperature, gas-cooled, graphite-moderated reactor; LWGR, light-water-cooled, graphite-moderated reactor. Data from IAEA (2002).

transfer energy to the turbine-generator to produce electricity. The direct approach allows the coolant to boil in the reactor core, creating steam to drive a turbine. The indirect approach transfers heat from a pressurized primary cooling circuit to a secondary circuit (e.g., through a steam generator) to drive the turbine.

Although many experimental reactors have combined these options, only four types of commercial thermal NPPs have been built:

1. Light water reactors (LWRs), including pressurized light-water-moderated and -cooled reactors (PWRs), Soviet/Russian pressurized water-moderated, water-cooled electricity reactors (WWERs), boiling light-water-moderated and -cooled reactors (BWRs), and advanced BWRs (ABWRs).

2. Heavy water reactors (HWRs), including pressurized heavy-water-moderated and -cooled reactors (PHWRs), such as the Canadian deuterium uranium (CANDU) reactor.

3. Gas-cooled, graphite-moderated reactors (GCRs), including early British and French GCRs and British advanced GCRs (AGRs) and high-temperature, gas-cooled, graphite-moderated reactors (HTGRs).

4. Soviet light-water-cooled, graphite-moderated reactors (LWGRs), such as the type of reactor at Chernobyl.

Because most of the currently operating NPPs are LWRs, the remainder of this article focuses on the cost of generating electricity from light water technology.

3. NUCLEAR GENERATION COSTS

Nuclear energy systems compete with other technologies in the electricity market (see Section 4). When evaluating whether to build NPPs, owner-operators consider (1) the total plant cost in millions of dollars (\$M) (because capital is at risk while the plant is under construction and there are no revenues), (2) the time to build a plant (because financing costs accumulate while the plant is under construction and demand conditions can change), (3) the ratio of average fixed (capital) costs to the market price of electricity (because cash flow problems can arise with large payments to debt and equity holders), (4) the ratio of average variable (production) costs to electricity price (because this will drive dispatch under regulation or bidding strategies under deregulation), and (5) the difference between average total generation cost and electricity price (because this determines profitability and the net present value of the project). Here, a distinction is made between “fixed costs,” which are sunk when the plant is completed, and “variable costs,” which can be avoided if the plant were to cease production. Fixed costs (FCs) include construction costs, financing costs, and decommissioning costs. Variable costs (VCs) include fuel costs, operation and maintenance costs, and capital additions. (Throughout the following discussions, costs and prices are given in real dollars, deflated using the U.S. gross domestic product implicit price deflator, or inflated to future years assuming an appropriate inflation rate, such as 3%.)

3.1 Plant (Fixed) Costs

The electric utility industry distinguishes between base (overnight) construction cost and financing costs. Because electricity-generating plants cannot be built instantaneously, lenders require payments during construction. These payments are known as “interest during construction” (IDC). IDC is equal to base costs discounted (into the future) to the start of commercial operation. Total plant costs (PCs) are equal to base costs plus IDC. Estimated total plant costs are equal to base costs plus IDC plus contingency (discussed later). Fixed costs (or total capital costs) are equal to total plant costs plus discounted decommissioning costs.

Cost estimates also distinguish between those for the first commercial plant of a specific technology (“first-of-a-kind” plants) and later plants, known as “*n*th-of-a-kind” plants. First-of-a-kind costs include

costs for research, development, demonstration, and deployment. These costs can be assigned to the first plant or amortized over a series of “transitional” plants. (However, some research, development, demonstration, and deployment costs could be subsidized with public funding, so these would not be amortized.) The *n*th-of-a-kind estimates also reflect learning during the construction of the first and transitional plants. The following discussion assumes “*n*th-of-a-kind” costs.

3.1.1 Base Construction Costs

Construction costs account for the largest portion of the cost of generating electricity from the current generation of NPPs. Cost-engineering estimates divide the construction project into a series of accounts. For example, Table I presents estimated total plant costs for a 1400-MW ABWR at a two-unit site in the United States. The accounts in this table are defined in the U.S. Department of Energy’s *Energy Economic Data Base*. Direct costs include land, structures, and equipment purchased from vendors, including the nuclear steam supply system (NSSS) manufacturer. Indirect costs include construction and engineering services provided by the architect-engineer, construction manager (who can be the architect-engineer), and the owner-operator (who can be both architect-engineer and construction manager). Direct plus indirect costs equal the base construction cost.

Also included in estimated total plant costs is “contingency,” which accounts for uncertainty in the cost estimate, usually as a percentage of estimated total plant costs. For example, for a simplified estimate, contingency is 30–50% of base construction cost; for a preliminary estimate, it is 15–30%; for a detailed estimate, it is 10–20%; and for a finalized estimate, it is 5–10%. Contingency can also be defined in association with the probability distribution of base construction cost. Under this interpretation, contingency is a function of the accuracy of the estimate, the confidence level of the estimate, the variance of the estimate, and the risk aversion of the estimate evaluator.

3.1.2 Financing Costs: Construction Duration and Interest During Construction

Translating base construction costs into total plant costs depends on spending in each period *t* during construction (CX_t , construction expenditures); construction duration, known as the lead time (LT), in years; and the cost of capital (*r*). A simple function for interest during construction is given by the

TABLE I
Example of a Nuclear Plant Cost Breakdown^a

Account	Cost	Amount
Direct costs		
21	Structures and improvements	\$430
22	Reactor plant	\$520
23	Turbine plant	\$230
24	Electrical plant	\$150
25	Miscellaneous plant	\$45
26	Main heat rejection system	\$45
	Total direct costs	\$1420
Indirect costs		
91	Construction services	\$250
92	Engineering home office	\$70
93	Field office services	\$190
94	Owner's cost	\$200
	Total indirect costs	\$710
	Base (overnight) construction cost	\$2130
	Contingency	\$125
	Interest during construction at 5% (or 10%) (5-year lead time)	\$282 (\$564)
	One-unit cost at 5% (or 10%) ^c	\$2537 (\$2819)
	Plant cost at 5% (or 10%) per kilowatt (gross) ^b	\$1812 (\$2013)

^aTotal plant cost for a 1400-MW advanced boiling water reactor at a two-unit site built in the United States (in millions of 1998 dollars).

^bKilowatt (gross) does not account for electricity consumed by the plant; a 1400-MW plant typically produces 1350 MW (net).

^cBased on data from the Nuclear Energy Agency's *Reduction of Capital Costs of Nuclear Power Plants*.

following equation,

$$IDC = \sum_{t=LT}^1 CX_t \left[(1 + r_t)^{(t-0.5)} - 1 \right] \quad (1)$$

assuming expenditures are made in the middle of the period. This formula discounts construction expenditures to the start of operation. This formula can be simplified by assuming that the cost of capital is equal in each period, continuous discounting, and a function for the expenditure rate, i.e., the percentage of the base cost spent in each period. The cost of capital is generally taken as a weighted average cost of capital (debt and equity), which can be adjusted

for taxes, although taxes are not discussed here. The expenditure rate can be assumed to follow a probability density function, such as the uniform or normal. Under a uniform expenditure rate with continuous compounding and constant cost of capital, the IDC can be approximated as $r \times LT/2$; e.g., with $r = 10\%$ and $LT = 5$ years, $IDC \approx 25\%$. (If LT is expressed in months, r must be converted to the monthly cost of capital.) Of course, if the construction expenditures and the cost of capital are known *ex post*, the IDC can be calculated precisely.

This analysis shows that the cost of financing depends heavily on the length of the construction period. The start of construction can be either the date of the issuance of the construction permit or the date of the first pouring of concrete. The difference between these two can involve months of planning and site preparation. Although site preparation can be lengthy, construction expenditures and IDC are low before the pouring of concrete. The end of construction is either when the plant is connected to the transmission grid ("grid connection") or when plant operation is transferred from the builder to the commercial operator ("commercial operation") and generation near rated capacity begins. The difference between these dates can involve testing, including low-power testing and power-ascension tests; regulatory requirements, such as the completion of a final safety analysis report; and the issuance of an operating permit. This difference can be less than a month to over a year. The commercial operation date is more appropriate for the calculation of IDC. Therefore, LT in Eq. (1) is usually measured as the time from the first concrete pour to commercial operation.

Construction time increased dramatically in the United States after the accident at Three Mile Island in 1979 due to changes in safety requirements for plants still under construction. In the United States, construction time from the first pouring of concrete to the connection of the unit to the grid was 4.8 years for units completed between 1966 and 1971, 6.1 years for units completed between 1972 and 1977, 9.4 years for units completed between 1978 and 1983, and 12 years for units completed between 1984 and 1989. On the other hand, in Japan, construction times were 4 years for units completed between 1972 and 1977, 4.6 years for units completed between 1978 and 1983, 3.9 years for units completed between 1984 and 1989, and 4.6 years for units completed between 1990 and 1995. Therefore, based on the Japanese experience (where the ABWR has been built), the nuclear power

industry assumes that new (large) NPPs can be built in 4 to 5 years.

3.1.3 Decommissioning Costs

Although decommissioning costs are not spent until after the NPP has been retired, once nuclear fuel has been loaded into the reactor and the reactor experiences a self-sustaining nuclear reaction, regulatory authorities, such as the Nuclear Regulatory Commission (NRC), require that the owner-operator set aside funds to pay for the decommissioning. Officially, decommissioning refers to the release of the NPP from regulatory jurisdiction. In the United States, this means the removal of the reactor vessel, the spent nuclear fuel, and any radioactive contamination under NRC regulation. In practice, there are various levels of decontamination and dismantling after NPP retirement. These are defined by other regulators (in particular, national and state environmental regulators) or by laws defining liability for residual radioactive, hazardous, or mixed (radioactive and hazardous) contamination. These levels of decontamination can be defined as follows:

1. Black field (also known as “safe storage”). NRC regulations have not been completely satisfied; for example, spent nuclear fuel has been placed in storage on site, but the reactor has not been dismantled. The site can be used for other nuclear activities.
2. Gray field (includes the U.S. Environmental Protection Agency Superfund sites). NRC regulations have been satisfied, but hazardous waste remains; ownership transfer is therefore difficult, but the same owner can use the site for other electricity-generating activities.
3. Brown field. The site has been decontaminated and can be sold for restricted (e.g., industrial) use, but the original owner remains liable for any residual (possibly unknown) contamination.
4. Green field. There is no contamination; the site can be used for any activity.

Because of these different levels of cleanup and because the requirements for certification of decontamination vary from jurisdiction to jurisdiction (country to country and state to state), the cost of decommissioning varies widely. Decommissioning costs have been estimated from experience at completed decommissioning projects, with the application of unit-cost factors or with a cost-engineering model. The NRC relies on cost-engineering models to specify the minimum amount that should be set aside by the owner during operation for eventual

decommissioning, including costs during “safe storage.” These minimum amounts are \$105 million (1986 dollars) for a PWR and \$135 million (1986 dollars) for a BWR.

Since 1990, U.S. owners have been required to establish a nuclear decommissioning trust (NDT) fund, outside the financial control of the owner. These funds have been collected from ratepayers (like other capital expenses) and are invested at rates approximated by the interest rate on low-risk corporate bonds (e.g., an interest rate of 5%). However, the NRC does not require NDT funds to cover “the cost of removal and disposal of spent fuel or of nonradioactive structures and materials beyond that necessary to terminate the license” [U.S. Code of Federal Regulations (CFR), parts 50–75]. To account for these costs, state regulatory commissions have required higher amounts (collected from ratepayers) in the NDT than are required by the NRC.

Because of the wide variety of cost estimates, the cost of decommissioning is assumed to be one-third of the direct construction cost, discounted 40 years to the start of operations. For example, at \$1420/kW (see Table I, total direct costs), decommissioning would be \$473/kW. Discounted at 5% for 40 years to the start of commercial operation, the present value would be \$67/kW or \$94 million for a 1400-MW NPP. Thus, decommissioning costs represent the present value of decommissioning (decontamination and dismantlement) costs at the time of commercial operation.

3.1.4 Plant Costs: Empirical Evidence and Projections

The econometric analysis of U.S. NPP plant costs began in the mid-1970s. This literature focused on the issue of whether there were economies of scale in NPP generating capacity, controlling for the year of commercial operation, construction duration, and plant characteristics, such as location, number of units at the site, and NSSS manufacturer. Although early analyses found increasing returns to scale, analyses of later data could not reject constant returns to scale, primarily due to the imprecision of the estimate of scale parameters when large, expensive units came into production after the accident at Three Mile Island in 1979. For example, plant costs at Watts Bar, the last unit to attain commercial operation in the United States, were nearly \$7 billion (in mixed-nominal dollars), or about \$6000/kW with a construction duration of 24 years from 1972 to 1996.

Estimated costs for advanced light water reactors, based on more recent experience (primarily in Asia), are much lower. In NEA (1998), plants are either commercially available or expected to be commercially available between 2005 and 2010. Base construction cost varies between \$1020 in China (using Chinese technology) and \$2521 in Japan (for an ABWR).

3.2 Productivity and Average Fixed Costs

To understand fixed costs, total fixed costs can be converted into a cost per megawatt-hour (MWh) by determining an annual, uniform series of payments equal to the present value of total fixed costs at the time of commercial operation and then dividing this annual payment by the annual electricity output. Annual output is discussed first, then average fixed costs.

3.2.1 Productivity and the Capacity Factor

There are three measures of NPP performance. First, productivity refers to the ability of the NPP to produce electricity. In the nuclear power industry, productivity is usually measured by the annual capacity factor, CF:

$$\text{MWh} = \text{CF} \cdot \text{MAX} \cdot Y \quad (2a)$$

$$\text{CF} = \text{MWh}/(\text{MAX} \cdot Y), \quad (2b)$$

where MWh is annual output in megawatt-hours, MAX is the NPP's annual net maximum dependable capacity (net capacity equals gross capacity minus electricity consumed by the plant), and Y is the total hours in a year. For example, a plant with a net maximum dependable capacity of 1350 MW (equivalent to a gross capacity of 1400 MW) operating at a 90% capacity factor for 8760 hours/year generates about 10,600,000 MWh for sale per year. Second, availability measures the readiness of the NPP to generate electricity. It can be measured with time or energy units. Associated with the time definition is the service factor, SF, measured as the ratio of generating hours (H) to total hours in a year: H/Y . Third, reliability measures the certainty of NPP generation. The reliability factor, RF, is equal to $(1-\text{FR})$, where FR is the forced outage rate, i.e., the percentage of time devoted to recovery from a forced outage.

To relate these performance indicators to the CF, consider the following equation:

$$\begin{aligned} \text{CF} &= [\text{MWh}/(\text{MAX} \cdot H)] \cdot (H/Y) \\ &= \text{CU} \cdot \text{SF}. \end{aligned} \quad (3)$$

The first term, $\text{MWh}/(\text{MAX} \cdot H)$, is the capacity utilization rate, CU, a measure of how close the NPP

is to potential output when it is running. The second term, H/Y , is the service factor (SF), the percentage of the time the NPP is running. So, the capacity factor is equal to the product of capacity utilization and the service factor. Further, the service factor is equal to the product of the "scheduled availability" rate, $\text{SA} = [(Y-S)/Y]$, where S is the scheduled outage time, and the reliability factor, $\text{RF} = H/[(H+F)]$, where F is the forced outage time and $Y = H + S + F$.

This decomposition of the capacity factor into the capacity utilization rate and the service factor (and the decomposition of the service factor into the scheduled availability rate and the reliability factor) allows a detailed analysis of NPP productivity. For example, Table II presents changes in CF, CU, SF, SA, and RF for NPPs in the United States (although the decomposition holds for the performance of each unit, the average of a product is not equal to the product of the averages):

- The mean and median CF increased 2 percentage points per year, or almost 3% per year.
- Median CU increased to 99.9% in 2000.
- Median SF increased to 91.8% in 2000, i.e., an average outage time of about 1 month.
- The forced outage rate ($1-\text{RF}$) declined to a median value of 0.8% ($= 1-99.2\%$) in 2000, i.e., less than 3 days per year.

Most forecasts of capacity factors for new nuclear technologies assume a value equal to the best average LWR capacity factor. With median capacity factors now above 90%, forecasts for new technologies assume a lifetime average capacity factor of 90%.

3.2.2 Average Fixed Costs

To calculate average fixed costs (AFCs), total capital costs are multiplied by the capital recovery factor (CRF) to determine equal annual payments and divided by average annual output, equal to the product of the (average) capacity factor, the size of the plant, and number of hours in a year. Total fixed (capital) costs include total plant costs (PCs) plus decommissioning costs (DCs). The CRF is equal to $[r \cdot (1+r)^T]/[(1+r)^T - 1]$, where r is the appropriate discount rate and T is the economic lifetime of the NPP. The appropriate discount rate is usually a weighted average cost of capital (debt and equity), consistent with the discount rate for calculating IDC. Also, the lifetime of the NPP is assumed to be the length of the operating license, which is 40 years in the United States. (However, NPPs can be closed early or their life can be extended.) If $r = 5\%$ and $T = 40$, $\text{CRF} = 5.8\%$. If $r = 10\%$ and $T = 40$, $\text{CRF} = 10.2\%$.

TABLE II
Changes in U.S. NPP Performance: 1990–2000^a

Year	Capacity factor		Capacity utilization		Service factor		Scheduled availability		Reliability factor	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1990	67.7	70.0	90.8	94.9	73.0	75.8	75.6	78.2	96.5	98.2
1992	71.7	74.9	93.1	96.3	75.3	79.5	78.1	81.4	96.5	98.3
1994	75.5	81.4	92.9	97.2	79.7	83.3	83.1	86.6	95.7	99.0
1996	77.5	82.7	95.1	98.6	78.9	84.3	81.9	87.2	96.4	98.8
1998	80.9	87.7	94.6	99.3	81.2	88.8	83.6	89.7	98.2	99.5
2000	89.3	91.9	99.2	99.9	89.5	91.8	93.5	93.6	95.9	99.2
Increase	21.6	21.9	8.4	5.0	16.5	15.9	17.9	15.4	−0.6	1.0

^aValues in percentages.

(Here, r is a real discount rate; the range of 5–10% is assumed to envelop most reasonable values for the average real cost of debt and equity.)

The average fixed costs for an ABWR built at a two-unit site in the United States can be calculated using the plant costs, decommissioning costs, and capital recovery costs as follows for the plant described in Table I [in millions (M) of U.S. dollars]:

$$\begin{aligned}
 \text{AFC} &= [(PC + DC) \cdot \text{CRF}] / \text{MWh} \\
 \text{AFC} &= [(2537\text{M} + 94\text{M}) \cdot 0.058] / 10,600,000 \\
 &= \$14.40 \quad \text{at } r = 5\% \\
 \text{AFC} &= [(2819\text{M} + 94\text{M}) \cdot 0.102] / 10,600,000 \\
 &= \$28.00 \quad \text{at } r = 10\%.
 \end{aligned} \tag{4}$$

Decommissioning costs are discounted at 5% to the start of commercial operation due to regulatory restrictions on NDT fund management; however, these costs are levelized at the average cost of capital, because this is the owner's opportunity cost. This example shows that the cost of capital has a great influence on the average fixed costs of NPPs.

3.3 Production (Variable) Costs

Three expenses that vary annually are fuel costs, operating and maintenance costs, and capital additions. Let average variable costs (AVCs) equal fuel costs per megawatt-hour (F) plus operating and maintenance (OM) costs per megawatt-hour plus capital additions per megawatt-hour (K):

$$\text{AVC} = F + \text{OM} + K. \tag{5}$$

Each of these costs is defined in the following section, with a discussion on how they have changed at U.S. NPPs during the 1990s.

3.3.1 Fuel Costs

Although fuel costs involve purchasing materials (e.g., uranium) and services (e.g., uranium enrichment and spent nuclear fuel management) over extended periods, fuel costs are treated as variable costs because they could be avoided if the NPP was not operating. As was seen in Fig. 1, there are three segments of the nuclear fuel cycle: the front end, in the reactor, and the back end. The front end of the nuclear fuel cycle involves uranium mining and milling, enrichment, and fuel fabrication. Uranium is primarily extracted through open-pit and underground mining. The uranium ore is crushed and leached with acids to produce a uranium oxide powder (U_3O_8), known as “yellowcake.” Since the creation of a commercial market in the early 1970s, uranium production has depended on the market price of uranium. The price of uranium increased in the late 1970s, but has declined continuously since 1980 in the United States and much of the world to \$20/kg (see Fig. 3).

To increase the percentage of fissile uranium (natural uranium contains 0.7% of uranium-235, a radioactive isotope, and 99.3% uranium-238), uranium oxide is converted to uranium hexafluoride, UF_6 (a gas above 56°C). The cost of conversion has been stable for years at \$8/kg of uranium. Then the UF_6 is enriched to between 3 and 5% uranium-235 for light water nuclear reactor fuel. Enrichment is done commercially with two methods, diffusion and centrifuge. With diffusion technology, the uranium gas is pumped through a series of slightly porous barriers that allow the smaller uranium-235 to penetrate more easily than the larger uranium-238. Diffusion technology requires large amounts of electricity for pumping the UF_6 . With centrifuge

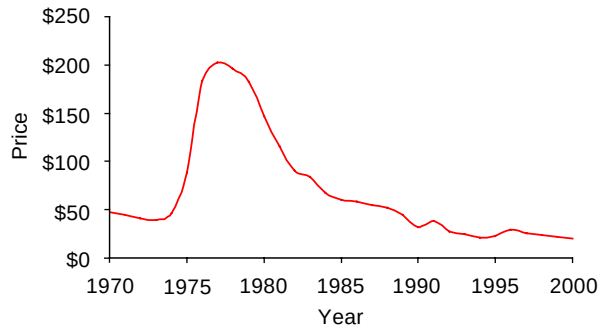


FIGURE 3 Uranium (U_3O_8) prices in kilograms/1996 U.S. dollars. Data for years 1970–1979 from NUEXCO (1979); for years 1981–2000, from Energy Information Administration (2000).

technology, the uranium gas is separated, with the lighter uranium-235 rising to the top when the UF_6 is spun. Enrichment is measured in “separative work units” (SWUs), i.e., the amount of effort required to increase the percentage of uranium-235. About 5 SWUs are needed to enrich 1 kg of uranium from 0.7 to 4% uranium-235. The SWU price has varied with changes in the price of electricity and changes in international competition. Between 1994 and 2003, SWU prices in the United States have varied between \$80 and \$110 (nominal dollars).

Enriched uranium is converted to uranium oxide and fabricated into nuclear reactor fuel. The cost of fuel fabrication varies between \$200/kg and \$400/kg of U_3O_8 . Approximately 20 metric tons of uranium is consumed each year in a 1000-MW NPP. Given the cost of uranium and conversion, fuel costs are dominated by enrichment and fabrication.

The back end of the nuclear fuel cycle includes spent nuclear fuel management, reprocessing, and high-level radioactive waste disposal. The cost of each operation varies from country to country and with the value of the uranium and plutonium recovered during reprocessing. The following approximate costs prevail in the United States: (1) for storage of spent nuclear fuel (SNF) at a NPP (on site) in nontransportable casks, \$250/kg; (2) for on-site storage, transportation, and storage at away-from-reactor facilities, \$500/kg; (3) for all storage, transportation, encapsulation, and disposal in a geologic repository, \$750/kg; and (4) for international commercial reprocessing (including transportation, recovery of uranium and plutonium, and disposal of radioactive waste), \$1000/kg. United States policy calls for storage of SNF at NPPs until the construction of a repository. The repository is to be financed through the collection of \$1/MWh, accumulated in the Nuclear Waste Trust Fund.

3.3.2 Operation and Maintenance Costs

Operating and maintenance expenses include labor and materials. Materials do not include equipment or other items with an economic life for more than 1 year. These items are capitalized and are known as “capital additions” (discussed later). Most of the OM expenses are fixed during the year, but they are treated as variable costs, because most could be avoided if the plant were to cease production. A majority of OM expenses pay NPP employees. There are four types of employees: administrative, operations, engineering, and maintenance, including security. These personnel work at the NPP or at corporate headquarters, for the NPP owner directly or for a subcontractor, and permanently or temporarily, e.g., during refueling outages.

3.3.3 Capital Additions

Capital additions, also known as refurbishments, include expenditures on equipment that lasts more than 1 year along with any labor that is involved in the installation or refurbishment of equipment. The largest capital additions at U.S. nuclear power plants have been the replacement of steam generators at PWRs. Capital additions are amortized over the remaining life of the plant, but because they could be avoided by retiring the plant, they are treated as a variable cost.

3.3.4 Production Costs: Empirical Evidence and Projections

Table III presents mean and median values for average variable costs at U.S. NPPs from 1990 to 2000: mean fuel expenses (F) declined from \$8.51 to \$4.59 and median fuel expenses declined from \$7.52 to \$4.40; mean OM costs declined from \$19.73 to \$11.47 and median OM costs declined from \$17.04 to \$10.47; and mean capital additions (K) declined from \$5.59 to \$3.00 and median capital additions declined from \$4.12 to \$1.27. These differences between means and medians show that unusually high costs at some NPPs have a large influence on mean variable costs.

To examine the influence of the changes in the capacity factor, CF, on AVCs, redefine the variables in Eq. (5), such that

$$\text{AVC} = \frac{\text{AVC}}{\text{CF}} = \frac{F}{\text{CF}} + \frac{\text{OM}}{\text{CF}} + \frac{K}{\text{CF}}, \quad (6)$$

where AVC is the average variable cost at a 100% capacity factor and all other costs are similarly defined. For example, if $\text{AVC} = \$20/\text{MWh}$ and $\text{CF} = 90\%$, $\frac{\text{AVC}}{\text{CF}} = \$18/\text{MWh}$. Table IV presents the means and medians of $\frac{\text{AVC}}{\text{CF}}$, $\frac{F}{\text{CF}}$, $\frac{\text{OM}}{\text{CF}}$, and $\frac{K}{\text{CF}}$ from

TABLE III

Mean and Median Production Costs for U.S. NPPs: 1990–2000^a

Year	Cost (\$)		Fuel (\$)		Operation/maintenance (\$)		Capital additions (\$)	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1990	32.24	29.28	8.51	7.52	19.73	17.04	5.59	4.12
1992	31.58	27.94	7.00	6.22	19.27	16.05	6.23	4.58
1994	26.11	23.37	6.41	5.74	16.12	13.95	3.82	3.11
1996	26.59	20.31	5.70	5.27	17.55	12.38	3.11	2.27
1998	23.85	19.24	5.34	5.02	15.86	11.70	2.68	2.27
2000	19.06	17.97	4.59	4.40	11.47	10.47	3.00	1.27
Decrease	13.19	11.32	3.92	3.12	8.26	6.57	2.59	2.85

^a All values in U.S. dollars.

TABLE IV

Mean and Median Production Costs at a 100% Capacity Factor for U.S. NPPs: 1990–2000^a

Year	Cost (\$)		Fuel (\$)		Operation/maintenance (\$)		Capital additions (\$)	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
1990	22.50	22.12	5.99	5.39	12.90	12.14	3.90	3.09
1992	21.59	21.05	5.13	4.79	12.79	12.37	4.05	3.41
1994	19.57	19.35	4.98	4.72	11.70	11.08	2.83	2.44
1996	18.11	17.24	4.51	4.45	11.26	10.58	2.32	1.85
1998	18.06	16.76	4.39	4.31	11.42	10.77	2.04	1.85
2000	18.20	17.03	4.22	4.06	11.08	10.16	2.97	1.14
Decrease	4.30	5.09	1.77	1.33	1.82	1.98	0.93	1.95

^a All values in U.S. dollars.

1990 to 2000. Holding the capacity factor constant, the decline in variable cost is much less dramatic: most of the decline was from an increase in the capacity factor (see Table II). For example, the decline in mean AVC was \$13.19, whereas the decline in mean AVC was only \$4.30, i.e., one-third of the decline in AVC was due to decreases in costs and two-thirds of the decline was due to increases in the capacity factor.

Although estimated AVC varies from country to country, costs per megawatt-hour are approximately \$5/MWh for fuel (including waste management and disposal costs), \$10/MWh for OM, and \$2/MWh for capital additions; i.e., near-term forecasts of AVC for both operating and new NPPs are approximately \$17/MWh. Therefore, the estimated average total costs (ATC = AFC + AVC) for new light water NPPs

(following the preceding assumptions) vary between \$31.40 with a cost of capital of 5% and \$45.00 with a cost of capital of 10%.

4. PRICE AND NET PRESENT VALUE

The empirical evidence and near-term projections for total NPP plant cost, plant construction time, average fixed costs, average variable costs, and total average generation cost were examined in Section 3. In this section, these costs are compared with the price of electricity. One method for making this comparison indirectly is to calculate a levelized electricity generation cost and compare it with electricity generation prices. This is appropriate under rate-of-return regulation. Under competitive

market structures, a net present value calculation is more appropriate.

4.1 Levelized Cost

The levelized electricity generation cost (ECG) is defined by the following formula:

$$\text{ECG} = \frac{\sum_t [C_t(1+r)^{-t}]}{\sum_t [\text{MWh}_t(1+r)^{-t}]}, \quad (7)$$

where the summation is over the period of construction, operation, and decommissioning, discounted to the start of commercial operation; C_t are all costs in the year t ; MWh_t is electricity generation in the year t ; and r is the discount rate, assumed to be equal to the cost of capital. During construction, C_t equals CX_t [see Eq. (1)]. These would be equal to base construction costs plus contingency (if an estimate) without IDC, given that discounting to commercial operation in Eq. (7) is equivalent to the IDC calculation. During operation, C_t equals AVC_t times MWh_t . During decommissioning, C_t equals DX_t , the decommissioning expenditures in year t . Finally, MWh_t is annual output during the years of operation.

Making the same assumptions as previously, for example, the following relationships apply:

- $\sum_t [C_t(1+r)^{-t}]$ during construction is equal to a total plant cost of \$2537 million for $r=5\%$ and \$2819 million for $r=10\%$, from Table I.
- $\sum_t [C_t(1+r)^{-t}]$ during operation is equal to \$17/MWh times 10,600,000 MWh, or \$180 million/year discounted over 40 years, or \$3100 million for $r=5\%$ and \$1760 million for $r=10\%$.
- $\sum_t [C_t(1+r)^{-t}]$ during decommissioning is equal to \$94 million, given restrictions on the discount rate for nuclear decommissioning trust funds.
- $\sum_t [\text{MWh}_t(1+r)^{-t}]$ during operation is 10,600,000 MWh/year, discounted over 40 years, or 182,000,000 MWh for $r=5\%$ and 104,000,000 MWh for $r=10\%$.

Here, $\text{ECG} = \$31.50$ for $r=5\%$ and $\$44.90$ for $r=10\%$. These values are very close to those for the ATC (calculated previously), showing that the two techniques of levelization are equivalent.

4.2 Price Regulation and Pricing under Deregulation

Under traditional rate-of-return regulation, the price charged by an integrated electric utility would be determined by the quantity of electricity sold and the

revenues required to generate electricity from its power plants and purchases made from other electricity generators. Under these conditions, the required revenues to operate the NPP would be equal to the levelized electricity generating cost.

In a competitive electricity market, the price in each period is usually determined by the bid of the highest cost producer. For example, the future price of electricity in Europe, Japan, and the United States will likely depend on the costs of generating electricity with natural gas. To approximate the price of electricity from natural gas, assume that the base construction costs of combined-cycle gas turbines (CCGTs) are \$500/kW throughout the world. With a construction time of 2 years, the IDC is approximately \$25/kW with $r=5\%$ and \$50/kW with $r=10\%$. With a 20-year capital recovery period, the CFR is 8% with $r=5\%$ and with 11.75% with $r=10\%$. The annual fixed cost is about \$42/kW for $r=5\%$ and \$65/kW for $r=10\%$. If the CCGT has a capacity factor of 90%, the average fixed cost is about \$5.30/MWh for $r=5\%$ and \$8.17/MWh for $r=10\%$. Second, the cost of natural gas per megawatt-hour depends on the heat rate. Assuming a heat rate of 7000 Btu/kWh, a natural gas price of \$4/MBtu yields an average fuel cost of \$28/MWh. Third, assuming OM plus capital additions are \$3/MWh yields a cost of generating electricity of \$36.30/MWh for $r=5\%$ and \$39.17/MWh for $r=10\%$. Therefore, with a price of natural gas at \$4/MBtu and with a cost of capital of 5%, nuclear power is cheaper than natural gas (\$31.50 < \$36.30), but with a cost of capital of 10%, nuclear power is more expensive than natural gas (\$44.90 > \$39.17). Of course, a complete analysis would consider all alternatives, including other nuclear technologies, coal, and renewables.

4.3 Net Present Value

Assuming natural gas sets the marginal price of electricity, a net present value analysis can determine the profitability of new nuclear power as a function of natural gas prices. Net present value (NPV) is defined as follows:

$$\text{NPV} = \sum_t [(\text{MWh}_t \cdot P_t) - C_t](1+r)^{-t} \quad (8a)$$

or

$$\begin{aligned} \text{NPV} = & \sum_t [\text{MWh}_t \cdot P_t](1+r)^{-t} \\ & - \sum_t C_t(1+r)^{-t}, \end{aligned} \quad (8b)$$

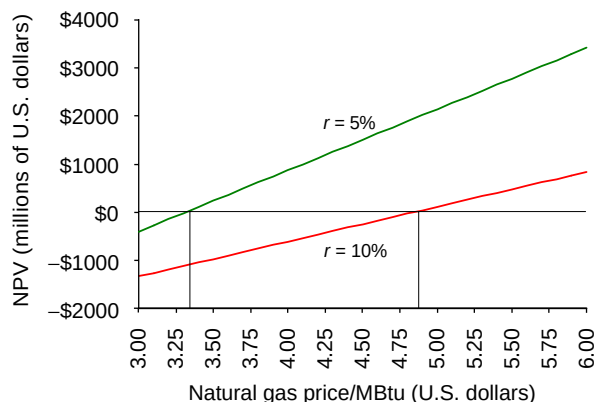


FIGURE 4 Net present value (NPV) of nuclear power as a function of the price of natural gas.

where P_t is the price of electricity and all other terms are the same as in Eq. (7). Notice that the last part of Eq. (8b) is identical to the numerator in Eq. (7). The first part of Eq. (8b) is the sum of discounted revenues. If P_t depends on the price of natural gas, NPV becomes a function of natural gas prices, holding all other variables constant (see Fig. 4). Nuclear power (under the assumptions in Section 3) is competitive for natural gas prices above \$3.30/MBtu with a cost of capital of 5% and for prices above \$4.85/MBtu with a cost of capital of 10%.

5. NUCLEAR POWER ECONOMIC UNCERTAINTIES

If electricity generators were concerned only with cost and price estimates, more nuclear power plants would likely be built in regions of the world with high fuel costs. But electric utilities are also interested in the uncertainties associated with nuclear power economics. For example, compare the total plant cost of a NPP with a CCGT. The NPP requires an investment of \$2500 million to \$2800 million for one 1400-MW unit (under the preceding sample assumptions). If two units are required to achieve optimal scale economies at a single site, the utility faces an investment of at least \$5 billion. The CCGT can be built in units of 300 MW, requiring an investment of \$150 million. Risk-averse utilities might prefer a smaller investment even if the average total costs of the CCGT are higher. Further, the CCGT can be built in 2 years, whereas the NPP requires 5 years. During construction, interest payments to lenders accumulate exponentially. Uncer-

tainties in the demand for electricity can also lead to larger risk premiums on capital invested in projects with long construction durations. Both effects favor technologies that can be built quickly.

Average fixed costs depend on the capacity factor. Though CCGTs are unlikely to operate at 90% capacity factors, their fixed costs are low compared to their variable costs. If NPPs are unable to achieve high capacity factors, average fixed costs increase. For example, if capacity factors at new NPPs were 70% (as they were in 1990 in the United States), average fixed costs would be \$18.40 with $r = 5\%$ and \$35.90 with $r = 10\%$. Nuclear power would not be competitive at capacity factors below 70%.

Average variable costs include fuel, OM, and capital additions. Although there are great uncertainties associated with long-term use of carbon-based fossil fuels, there are still uncertainties associated with the back end of the nuclear fuel cycle. Though most of these are political (and not economic) uncertainties, it is difficult to predict the cost of long-term geologic disposal. Also, most OM expenses are associated with labor costs. In the United States, regulation drove the increase in OM costs after the accident at Three Mile Island by requiring more employees at NPPs. Therefore, regulatory uncertainties, outside the control of the NPP owner, increase uncertainty in OM costs.

All of these uncertainties influence perception of total and average NPP generation costs by electric utilities, policymakers, and the public. The growth of commercial nuclear power depends on reducing NPP capital costs and resolving uncertainties associated with nuclear power economics. Simultaneously, reducing capital costs and uncertainties depends on new orders for NPPs. Given this dynamic interaction, market forces could lead to either the collapse or the self-sustaining expansion of the commercial nuclear power industry. This industry must change. It will be unable to do business as it has been doing under electricity deregulation and liberalization.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis •

*Nuclear Proliferation and Diversion • Nuclear Waste
• Public Reaction to Nuclear Power Siting and
Disposal*

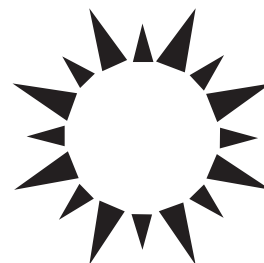
Further Reading

- Electric Power Research Institute (EPRI). (1997). "Technology Assessment Guide." EPRI, Palo Alto, California.
- Energy Information Administration (EIA). (2000). Annual Energy Outlook. EIA, Washington, D.C.
- Energy Information Administration (EIA). (2001). "The Electricity Market Module of the National Energy Modeling System: Model Documentation Report." EIA, Washington, D.C.
- International Atomic Energy Agency (IAEA). (1999). "Evaluating and Improving Nuclear Power Plant Operating Performance." IAEA, Vienna.
- International Atomic Energy Agency (IAEA). (2002). "Nuclear Power Reactors in the World." IAEA, Vienna.
- International Atomic Energy Agency (IAEA). (2002). "Power Reactor Information System (PRIS) Database." IAEA, Vienna.
- Nuclear Energy Agency. (1994). "The Economics of the Nuclear Fuel Cycle." Organization for Economic Cooperation and Development, Paris.
- Nuclear Energy Agency. (1998). "Projected Costs of Generating Electricity: Update 1998." Organization for Economic Cooperation and Development, Paris.
- Nuclear Energy Agency. (2000). "Uranium 1999: Resources, Production, and Demand." Organization for Economic Cooperation and Development, Paris.
- Nuclear Energy Agency. (2000). "Reduction of Capital Costs of Nuclear Power Plants." Organization for Economic Cooperation and Development, Paris.
- Nuclear Exchange Corporation (NUEXCO). (1979). "Monthly Report to the Nuclear Industry No. 129." NUEXCO, Menlo Park, California.
- Oak Ridge National Laboratory (ORNL). (1993). "Cost Estimate Guidelines for Advanced Nuclear Power Technologies." ORNL, Oak Ridge, Tennessee.
- Pasqualetti, M., Rothwell, G. (eds.). (1991). Special issue: Nuclear decommissioning economics. *Energy J.* 12.
- Rothwell, G.(ed.). (1998). Special issue on nuclear energy in Asia. *Pacific Asian J. Energy* 8(1).
- Rothwell, G., and Gomez, T. (2003). "Electricity Economics: Regulation and Deregulation." IEEE Press with John Wiley, Piscataway, New Jersey.

Nuclear Power, History of

ROBERT J. DUFFY

Colorado State University
Fort Collins, Colorado, United States



-
1. Dawning of the Nuclear Age in the United States, 1946–1954
 2. Creating a Commercial Atom, 1954–1965
 3. Nuclear Power Comes under Attack
 4. Regulatory Failure and the Changing Economics of Nuclear Power
 5. Nuclear Power in the 1980s
 6. Nuclear Power Today

Glossary

Atomic Energy Commission The powerful executive branch agency created in 1946 to oversee the U.S. atomic energy program; abolished in 1975.

atoms for peace The U.S. program launched in 1954 to share atomic information with European allies and to develop peaceful applications of atomic energy.

Joint Committee on Atomic Energy An unusually powerful congressional committee that dominated nuclear energy policy in the United States from 1946 to 1977.

Nuclear Regulatory Commission The agency created in 1975 to assume the regulatory duties of the Atomic Energy Commission.

Price–Anderson Act Legislation enacted in the United States in 1957 to encourage private sector interest in nuclear power by providing indemnification to the nuclear power industry in the event of a reactor accident. The law has been renewed several times over the program's history.

subgovernment Also known as policy monopolies or subsystems, these are small, stable groups of actors, both public and private, that dominate policy in specific issue areas.

The 50-year history of commercial nuclear power has been punctuated by dramatic policy changes. The first 20 years, marked by limited public participation, tight government control, and promises of clean, abundant energy, were followed by a period of intense social and political conflict over the technol-

ogy's environmental and safety implications. Nuclear policy in the United States and most European nations shifted from all-out support to a more ambivalent posture, which led to a dramatic slowdown in the construction of new plants. Although nuclear plants continue to be built, public opposition and high costs are obstacles to a large-scale comeback.

1. DAWNING OF THE NUCLEAR AGE IN THE UNITED STATES, 1946–1954

1.1 The Military Roots of Commercial Nuclear Power

From its birth in the highly secretive Manhattan Project during World War II, atomic energy was defined and perceived in military terms, which meant that information regarding the atom was subject to elaborate security precautions. This concern with maintaining the “secret” of the atom set the tone for the atomic program far into the future. From the beginning, both public involvement in the development of atomic energy programs and access to information about atomic energy would be severely restricted.

When the war ended, the sense of national emergency that had driven the atomic program dissipated, and many of the top scientists who had worked on the bomb project returned to their positions in universities and the private sector. With virtually no stockpile of atomic weapons at the war's end, government officials worried that if the atomic program ground to a halt, the nation's security would be at risk. A vigorous weapons program, it was believed, was essential to the nation's defense and security, especially because it was widely recognized that the American monopoly on atomic weapons would not last forever.

1.2 The Atomic Energy Act of 1946

In pursuit of these general goals, Congress in 1946 passed the Atomic Energy Act, which established the institutional framework within which atomic energy decisions would be made for approximately the next 30 years. The legislation created two bodies, the Atomic Energy Commission (AEC), which inherited ownership of all atomic materials, facilities, and information from the Manhattan Project, and the Joint Committee on Atomic Energy, which was to oversee AEC operations.

The AEC was to be the primary actor in formulating the nation's atomic energy policies. Its principal duties, as defined by Congress, were the production of atomic weapons and the fissionable materials required for their manufacture. In addition to establishing the AEC, the Atomic Energy Act mandated a government monopoly in atomic energy. Congress stipulated that responsibility for all atomic programs and information be transferred from the army to the AEC. Furthermore, the AEC was granted ownership of all fissionable materials and all facilities that used or produced them. The size and scope of the AEC's operations, together with the sweeping powers granted by Congress, clearly indicated that it was intended to be a powerful agency.

Composed of nine members from each chamber, the Joint Committee on Atomic Energy (JCAE) was clearly intended to be a powerful committee. It was, for example, the only joint committee created by statute rather than by the rules of both houses. The JCAE was also the only permanent joint committee granted the full legislative and investigative powers of a regular standing committee. Its formal status alone suggests that Congress intended to exercise control over atomic energy in general and over the AEC in particular. The Atomic Energy Act required the AEC to keep the JCAE "fully and currently informed with respect to the Commission's actions," mandated that all bills concerning the atom were to be referred to the JCAE, and it set no limits on the number of staff members or consultants the committee could hire. These sweeping and unusual grants of power, along with the JCAE's status as a special joint committee with exclusive jurisdiction over a glamorous and highly technical issue, made the committee one of the most powerful congressional bodies in the nation's history.

The normal tendency of Congress to defer to the advice of its specialized committees and subcommittees was reinforced in the case of atomic energy by several additional factors. First, atomic energy was largely perceived and justified as a defense program,

with primary emphasis on its military applications. As late as June 1961, approximately two-thirds of all research and development and construction activities of the AEC's reactor program continued to serve weapons and military applications. The definition and perception of the atomic program as primarily military in character lent credibility to both the program and the Joint Committee and legitimized its exclusive policymaking role.

1.3 Secrecy and Security

Because the United States was the only nation with the atomic "secret," security was considered to be an essential part of the effort to retain the atomic monopoly. The JCAE's inclination, like that of other security and defense committees, was to keep most information classified. With its exclusive control over atomic energy matters in Congress, the committee was able to dominate the legislative aspect of policymaking in these early years. The security restrictions meant that the atomic program was obscured by a blanket of secrecy that effectively limited the number of participants in the atomic program. Nuclear power would be relatively immune to challenge so long as access to information concerning the program was so closely held.

1.4 The Creation of an Atomic Subgovernment

In many respects, the atomic energy program in the 20 years following World War II was the archetypal subgovernment. A small, cohesive, and stable group of actors consisting of the Atomic Energy Commission, the Joint Committee on Atomic Energy, the nuclear power industry, and the scientists and engineers working on the program in universities and national laboratories exercised considerable autonomy in policymaking. Each of the participants involved in nuclear policymaking during this period had either an economic, political, or organizational stake in the development and use of atomic energy. There is, however, one highly unusual aspect of this policy community. Unlike the familiar example of subsystems in which government actors had been "captured" by the very business interests they were supposed to monitor, with nuclear power there was no industry "client" group until the government created one. It was only after policymakers in the mid-1950s took steps to overcome the serious technological, economic, and political obstacles confronting the generation of electricity from fission

that a commercial nuclear power industry emerged. This unusual government–industry partnership is crucial to understanding the history of nuclear power in the United States, as well as in most other nations.

The atomic energy subgovernment was endowed with additional prestige and power because of the program's identification with national security issues. The actors in this tightly knit monopoly were united by the conviction that the development of atomic energy, first as a weapon but later as a means of generating electricity, was both necessary and desirable for the nation's welfare. Indeed, this commitment was enthusiastically embraced by those involved in formulating and implementing policy and drove the program for over 20 years. For the most part, this consensus meant that only supporters of nuclear power would be mobilized for political action. Presidents rarely became involved in the program, Congress readily deferred to the advice of the experts on the Joint Committee on Atomic Energy, and, with the exception of the nuclear power industry, interest group activity was virtually nonexistent. This one-sided mobilization ensured that generous subsidies and the absence of political conflict marked government policy during this long period.

1.5 Nuclear Power and the National Interest

In the years following the World War II, nuclear power was equated with the national interest, and it was often said that politics and atomic energy did not mix. The cold war, the prevailing faith in science, and the recognition that the atomic monopoly could not last forever all contributed to a consensus that atomic energy was instrumental in securing the nation's future. To some, the term "national interest" could be used to promote almost anything pertaining to nuclear power. Furthermore, it was often claimed that nuclear energy would revolutionize the nature of industrial society by providing a relatively cheap and abundant source of power. Nuclear energy could lead to the discovery of new production techniques, it was said, and provide cheap electricity to areas where supplies were scarce. Others claimed that nuclear power could drastically reduce the costs of production, leading to increased economic growth and an improved standard of living. In short, a vigorous atomic energy program was thought to be necessary for a nation seeking to exert military, economic, and technological leadership in the postwar era. The national interest seemed to demand that the United States explore the atom's many possibilities.

2. CREATING A COMMERCIAL ATOM, 1954–1965

The dawn of the 1950s brought a subtle but significant shift in the nation's atomic energy plans. The principal focus remained on the atom's military aspects, but government officials recognized that the so-called peaceful uses of the atom could also serve a variety of important objectives in both domestic and foreign policy. Although the cold war dictated that the program's primary purpose would continue to be weapons production, the Joint Committee began to press the AEC to put greater emphasis on developing and promoting the peaceful uses of atomic energy. At the time, the most likely prospect for peaceful applications was the production of electricity from atomic fission.

2.1 The Peaceful Atom and Postwar Politics

Despite this subtle shift in the program's avowed goals, the program was still largely justified by national security arguments. Aside from enhancing the nation's scientific reputation and providing an abundant energy source, finding a constructive use for the atom would also further the nation's postwar geopolitical aims. The explosion of the Soviet atomic bomb in 1949 had punctuated the end of the American nuclear monopoly. American officials worried that the Soviets would try to develop the peaceful atom before the United States in an attempt to reap propaganda and political benefits. This prospect was not a pleasant one for those responsible for managing the atomic program.

Although American officials were preoccupied with the Soviets, they were also concerned that other nations would seek to develop their own atomic capabilities. The events of 1949 had shown that the United States no longer had a monopoly on the atom, and that the knowledge and ability to build atomic weapons would inevitably spread to other nations. It was commonly accepted that it was only a matter of time before Great Britain and France would join the atomic club. Other nations would surely seek to do the same and launch their own atomic programs, whether the United States wanted them to or not. Faced with this unpleasant reality, the United States wanted to be in a position to control nuclear proliferation to ensure that it was compatible with American perceptions of world order. Government decision makers recognized that in order to avoid

losing influence around the globe, the United States would have to assert its leadership and share some atomic information, and secrets, with other nations. With these objectives in mind, in a speech to the United Nations in December 1953, President Eisenhower launched an initiative that came to be known as “Atoms for Peace.” The program was clearly an effort to distinguish between the peaceful and military applications of the atom and to identify the United States with the peaceful applications. The Atoms for Peace plan also served a number of American foreign policy goals. The Atoms for Peace proposal, according to two researchers, was “in fact nothing less than a coherent global strategy for protecting Western Europe from Soviet domination.” If the United States could develop the atom as a source of electric power, it was argued, it would be a boon to Western European industries and economies while helping to contain the Soviets.

The Atoms for Peace proposal had the added attraction of assisting American industry by providing access to potentially lucrative markets. According to the terms of the bilateral agreements signed pursuant to the program, European nations contracted with American firms for all atomic technologies and materials. The attraction of these agreements, as one official from the American contractor Babcock and Wilcox wrote to the joint committee, was that they were a “means of securing a good and early foothold” in the potentially lucrative European market for American reactor vendors. One 1955 study estimated that Europe represented a \$30 billion market.

2.2 The Atomic Energy Act of 1954

In order to make more progress in developing the peaceful atom, Congress revised the Atomic Energy Act to permit greater access to nuclear information and materials. Although American firms were expressing an interest in atomic power, the calls for revising the 1946 law came primarily from the U.S. government and not from American industry, which was unprepared to assume the full responsibility for nuclear development because of the financial risks involved. In contrast to coal plants, the initial costs for nuclear plants would undoubtedly be very high, profits uncertain, and the technology new and potentially hazardous. In addition, projected demand for electricity in the early 1950s provided utilities with little incentive to build additional generating capacity. Furthermore, there were no reliable estimates of the possible damages stemming from a

serious reactor accident. Nevertheless, development of the peaceful atom was assumed to be vitally important to addressing a number of problems confronting the United States at the time. If the private sector would not take the necessary steps to solve these problems, the U.S. government would. The solution to the government’s problems was the Atomic Energy Act of 1954, which effectively created both commercial nuclear power and a commercial nuclear power industry. The development of commercial nuclear power in the United States, in short, emerged from an effort initiated by government rather than by private industry.

To open the door to greater industrial and international participation, Congress in 1954 loosened the restrictions and access to previously classified information, permitted the private ownership of nuclear facilities and allowed the private use of fissionable materials through a system of AEC licensing, liberalized patent laws in the atomic energy field, and offered a number of economic incentives to private industry.

2.3 The AEC’s Conflict of Interest

The law also made the AEC responsible for developing, promoting, and regulating nuclear power. From the outset, this statutory conflict of interest would be problematic; it became clear that the AEC would consistently emphasize its promotional and developmental responsibilities at the expense of its regulatory duties. Congress, through the Atomic Energy Act of 1954, offered little guidance on how to resolve this dilemma. The law, for example, directed the AEC to “protect the health and safety of the public,” but the statute offered little insight into how to achieve that broadly defined goal. Although such an expansive grant of authority was not unique, it would later become a problem, as the atomic program moved into its commercial phase.

2.4 Subsidies

Although the Atomic Energy Act of 1954 provided the AEC with little guidance on establishing regulations or defining safety goals, it did make clear that Congress did not want the AEC to impose burdensome regulations during the early stages of the development process. The section of the law under which all of the reactors in the AEC’s Power Reactor Demonstration Program were licensed instructed the AEC to “impose the minimum amount of such regulations and terms of license as will permit the

Commission to fulfill its obligations under this Act to promote the common defense and security and to protect the health and safety of the public.” In addition, members of the Joint Committee consistently pressed the AEC to move ahead with reactor development. Because reactors were not economically competitive with fossil fuel plants, however, the nation’s electric utilities did not have the normal incentives to build them. As a result, the AEC had to offer a wide array of incentives, including ample financial assistance, to the nuclear industry. The AEC thus embarked on several programs designed to encourage industry participation in the commercial power program. The AEC’s Power Reactor Demonstration Program, for example, assisted firms by waiving for 7 years the charges for the loan of source and special materials; by performing in its laboratories, free of charge, certain research and development work; and by agreeing to purchase technical and economic data from the participants under fixed-sum contracts.

Industry’s initial response to these inducements was tepid, so the government resorted to a mix of threats and incentives to induce interest. For example, even though the AEC retained ownership of all fissionable materials, through a system of AEC licensing the Atomic Energy Act of 1954 made these materials available to firms, including both the fuel for nuclear reactors and the fissionable materials produced by nuclear reactors. Because there was still a shortage of fissionable materials for atomic weapons, the AEC agreed to purchase all of the plutonium produced in private reactors at a guaranteed price that provided firms with handsome returns.

All told, by 1962, the federal government had spent over \$1.2 billion to develop reactor technology; this figure was double the amount spent by private business. One Department of Energy study conducted in the 1980s concluded that without federal subsidies nuclear electricity would have been 50% more expensive. These calculations did not include federal tax write-offs for utilities nor did they include money spent on programs with possible military implications.

2.5 The Price–Anderson Act

Even with these subsidies, the private sector was reluctant to assume responsibility for the risks involved with nuclear power. The underlying reason for this reluctance was the fear that they would be held liable for any and all damages resulting from a reactor accident. Although the AEC had an impres-

sive safety record with its reactors, the technology was new and largely untested, and because the potential damages from an accident were unknown, the reactor vendors and electric utilities were unwilling to make a full-scale commitment to nuclear power. The industry sought, and soon received, government protection from liability claims before proceeding with any plans to build reactors. The problem was that the potential liability from a reactor accident was so large that it was unlikely that either nuclear vendors or utilities would be able to obtain the necessary insurance from private insurers. The fear of financial disaster led the industry to seek legislation limiting their liability in the event of an accident. Because the potential damages from an accident could be so high, the industry argued that indemnification, in which the government would give general protection to the atomic industry against uninsured liability to the public, was needed to allow development to proceed. Private insurers agreed to provide \$60 million of protection for each nuclear facility, which was far more than had been made available to any other industry. Because it was entirely possible that damages from a reactor accident would far exceed that amount, Congress agreed to provide an additional \$500 million of protection for the industry.

2.6 The First Generation of Nuclear Reactors

After adoption of the Price–Anderson Act, private sector interest in nuclear power slowly picked up. Although the liability limit eased industry fears, industry participation remained at low levels for the next few years because nuclear electricity was still much more expensive to generate than electricity produced by conventional fuels like oil and coal. However, the words and actions of government officials convinced the nuclear industry that in the future there would be nuclear power in America, with or without private sector involvement. A number of utilities decided to buy reactors in order to gain an advantage for the future; others purchased reactors because they wanted to keep the government out of the power business. The big fear of many utilities was that the federal government would build reactors on its own and give the power to public corporations. Consequently, the fear of public power led utilities to invest in reactors before they were economically competitive.

A critical moment came in December 1963, when Jersey Central Power and Light signed a contract

with General Electric to purchase a 515-MW boiling water reactor for a plant in Oyster Creek, New Jersey. This agreement seemed to herald both the arrival of nuclear power as a viable alternative for generating electricity and the beginning of widespread commercial acceptance of nuclear power in the United States. The Oyster Creek reactor was to be the first built without any direct government subsidy, and the utility claimed the plant would be more economical to operate within 5 years, compared to conventional power sources. Although utilities began to order reactors in increasing numbers, in reality nuclear fission was not yet economically competitive with fossil fuels, primarily because of the higher capital costs for nuclear plants. In fact, the Oyster Creek plant was a “loss leader” designed to persuade utilities of the viability of reactor technology. Although it has been estimated that reactor vendors lost considerable sums of money on the earliest plants, the strategy worked—utilities began purchasing increasing amounts of nuclear generating capacity.

The sudden interest of the utilities in reactors can be attributed to several factors. First, demand for electricity in the 1960s was growing at an annual rate of 7%; at that rate of growth, demand would double every 10 years. Utilities thus needed to order new plants to meet projected demand. At the same time, many public utility commissions were lowering prices, which further increased demand for electricity. Next, because profits were determined by the size of the utility rate base, utilities had an incentive to build additional generating capacity to expand their rate base. Nuclear was an especially attractive option because it was by far the most capital-intensive source of electricity. Utilities thus had a number of financial incentives to build nuclear plants. Consequently, there was a rapid increase in both the number and the size of plants ordered after Oyster Creek. For example, American utilities ordered 70 reactors between 1963 and 1967, with approximately 80% of the orders being placed in 1966 and 1967. During that same period, the average capacity per reactor jumped from 550 to 850 MW. In 1967, nuclear power accounted for slightly more than half the total generating capacity ordered that year. The market for nuclear reactors was so good that some observers refer to this period as the “Great Bandwagon Market.”

But the Bandwagon Market was based on exceedingly optimistic projections and expectations of construction costs and operating performance. Because commercial generation of electricity from

fission was in its infancy and still largely untested, there was little in the way of hard data concerning actual costs and performance. In the early years of the commercial program, the utilities, the Commission, and the Joint Committee lacked the technical resources to assess the projections of reactor costs and efficiency offered by the reactor vendors. Not only did they lack the resources, but the AEC and JCAE also lacked the desire to evaluate the claims carefully. In their rush to commercialize reactor technology, government officials uncritically accepted the cost and efficiency estimates.

Electric utilities in the United States and abroad, caught up in the flood of good news about nuclear power, purchased more reactors and fewer fossil plants. The rush to nuclear power on the basis of the reactor vendors’ early and exceedingly optimistic price estimated eventually came back to haunt utilities that had purchased reactors. Many utilities possessed neither the technical staff nor the financial resources to build and operate nuclear reactors. It would also become clear that nuclear plants were not as efficient or reliable as vendors had claimed. The optimistic predictions for nuclear power thus succeeded in establishing a market for the technology, but they also set the stage for the future economic problems that would cripple the industry.

3. NUCLEAR POWER COMES UNDER ATTACK

In 1965, the politics of commercial nuclear power were of little concern to anyone not having a direct economic or political stake in the programs’ success. By the middle of the next decade, however, the policymaking arena was crowded and complex, and few issues in American politics were more visible or controversial. What changed is that, as nuclear reactors began to be built, more people came to believe that the issue affected them. In the process, nuclear power came to be understood in new, less positive ways. Long perceived as a national security issue, in the late 1960s, nuclear power came to be seen in the context of debates over environmental protection, public health and safety, and energy supplies. Eventually, it became part of even larger debates about government regulation of business, citizen participation, and democratic governance. As a result of these changing perceptions, subgovernment members lost the power to define the issue and control the parameters of debate. The subgovernment and nuclear policy were soon radically transformed.

As understandings and perceptions shifted and became increasingly negative, new actors were drawn to the nuclear issue. The new participants included other federal agencies, officials from state and local governments, and a number of prominent citizen groups. Many of the new participants would be critical of nuclear power. The influx of new participants, who brought their own opinions, shattered the consensus that had existed within the small atomic subgovernment, disrupted established patterns of policymaking, and created a decidedly less supportive political environment for the nuclear program.

3.1 The Emergence of an Antinuclear Movement

The boom in the market for nuclear plants coincided with the rise of the environment as an issue of national importance. Subsystem members initially perceived the growing concern with environmental values as a boon to the prospects of nuclear power. The AEC and the nuclear industry actively promoted the notion that nuclear power was cleaner than coal and other sources of electricity. This makes it all the more ironic that the first consistent opposition to nuclear power emerged in response to the potential environmental consequences of reactors. For the most part, the early opposition was essentially local in nature, being directed at particular reactors and not nuclear power. More specifically, concern during this period tended to focus on the thermal pollution caused by nuclear plants' discharge of heated water into nearby lakes and rivers. The AEC initially refused to consider the environmental effects of thermal pollution, arguing that it lacked statutory authority over nonradiological environmental matters. After a 3-year battle, and much bad publicity, Congress in 1970 adopted a law requiring federal agencies, including the AEC, to consider thermal pollution issue in the course of their licensing reviews. This controversy fundamentally altered the course of nuclear politics. It was the first critical step in the redefinition of the nuclear power issue; as concerns emerged about possible radioactive emissions from reactors and the issue of nuclear waste disposal, nuclear power was increasingly understood as an environmental issue, not a national security matter.

The most significant factor in the expansion of the debate and in the demise of the atomic subgovernment was the emergence of reactor safety as a prominent issue. The controversy over safety energized the burgeoning opposition to nuclear power and thrust the issue into the media spotlight, making

it impossible for public officials to ignore. The safety issue emerged and developed in much the same way the environmental issue had—as a response by local citizens to particular nuclear reactors—before rapidly escalating to include questions about the safety of nuclear plants in general. But concerns over the safety of nuclear reactors attracted more attention and generated more controversy compared to other environmental issues, largely because nuclear technology was still relatively new and unfamiliar to the American public. When questions of reactor safety emerged, the dramatic nature of the potential consequences of reactor accidents lent the nuclear issue a sense of drama and urgency that earlier concerns about thermal pollution could not. Furthermore, because the issue was so dramatic, it captured the attention of the national media. Shortly thereafter, the conflict over the safety of nuclear reactors occupied a prominent position in the public agenda. The AEC's credibility suffered a fatal blow when experts, including its own, began to disagree in public. Once some insiders began raising questions about reactor safety and the AEC's commitment to protecting it, the conflict rapidly expanded.

Antinuclear activists in the late 1960s and early 1970s were also extremely critical of the AEC's licensing procedures and rules. Critics charged, for example, that AEC procedures precluded meaningful public participation in agency proceedings while granting the nuclear industry privileged access to AEC personnel, especially in the critical early stages of the licensing process. Critics also claimed that AEC regulations were primarily responsive to the needs and desires of the industry, reflecting the commission's enthusiasm for building and licensing reactors. Perhaps most important, critics charged that the decision to approve reactors had usually been made before citizens were given an opportunity to register their opinions. They also charged that the AEC's public hearings, the principal forum in which citizens were allowed to participate, were nothing more than stacked-deck proceedings designed to foster the illusion of citizen input.

By the middle of the 1970s, the subgovernment had lost its ability to define the nuclear issue; the debate over nuclear power thus widened to include a variety of new issues. Some opponents of nuclear power were suggesting that rather than being merely a technical matter, the nuclear issue involved important social, political, and moral questions that scientists and engineers had no special ability to resolve. Antinuclear forces argued that citizens could learn enough about nuclear power to participate in a

responsible manner if given the opportunity. In this way, antinuclear forces were able to link their concerns about nuclear safety to more general concerns about accountability and responsiveness in government.

By the end of the 1970s, citizen groups were challenging, in court and in the streets, the licensing of almost every reactor. Indeed, some of the largest post-Vietnam protests and demonstrations concerned nuclear power. Several rallies drew crowds estimated to be in the hundreds of thousands. Nuclear power was plagued by similar developments in Europe, especially Germany, Great Britain, France, and Sweden. In most of these nations, as in the United States, the government became less supportive of nuclear power. France, which had fewer energy options, was a notable exception to this trend, and moved ahead with an aggressive nuclear program despite widespread public opposition.

3.2 Nuclear Power and the Courts

For most of the late 1960s and early 1970s, both the AEC and the Joint Committee consistently rebuffed persons or groups seeking to participate in the formulation of nuclear policy. In a textbook case of venue shopping, critics of nuclear power turned to the courts for assistance. One of the most significant consequences of increased oversight by the courts was that the commission, and its staff, devoted greater attention to procedural rights in an attempt to ensure that its procedures were seen as fair and capable of generating a record that could withstand judicial scrutiny. Rather than run the risk of being overturned by a reviewing court, the commission would try to show that it had solicited and considered many points of view.

The consequence of increased judicial oversight was a longer and more detailed review process that led, in turn, to more stringent environmental and safety standards. In an effort to satisfy its overseers, who were demanding more comprehensive and rigorous reviews, the commission upgraded the size and quality of its technical review staff. The larger, better, and more experienced regulatory staff began demanding more detailed information from utility applicants. Moreover, the commission was cognizant of the fact that the courts were more likely to overturn standards if the commission's decision-making process was procedurally deficient. As a result, the commission standardized its licensing review process in 1972. Specifically, the commission developed standard review plans for both safety and

environmental analyses. The standard review plans identified the information needed by the staff in performing their technical review of applications and suggested a format for its presentation.

Another factor contributing to the commission's increasingly tough regulatory standards was a dramatic increase in the size of nuclear reactors. At the end of 1966, the largest reactor in operation was under 300 MW, but by the end of the decade, utilities were placing orders for plants two to three times larger. The problem was that larger, more powerful reactors posed more troublesome questions from a safety perspective. The fuel in larger reactors would overheat more quickly in the event of a failure in the plant's cooling system, and larger reactors contained higher levels of radioactivity. In short, larger reactors seemed to pose a greater risk to the public health and safety, especially because utilities were pressuring the commission to approve reactors closer to population centers.

3.3 The Rise of Energy Issues

The Arab oil embargo in October 1973 and the ensuing energy crisis profoundly altered the perception of the nation's energy policies in the eyes of the public and government officials and catapulted the energy issue to the top of the nation's political agenda. Although this turn of events might have been expected to bolster the prospects of nuclear power, it actually undermined them by subjecting the various energy subgovernments to intense scrutiny. With this increased scrutiny, energy interests were placed in direct competition for federal research and development funds; nuclear power, which had received the bulk of such monies over the years, would now face stiff competition. At the same time, government officials, seeking to establish a comprehensive and coordinated national energy plan, responded by reorganizing both the administrative and legislative branches. These organizational changes, in turn, would prove to be incompatible with continued domination of energy politics by subsystems.

3.4 The Fall of the AEC and JCAE

Largely as a result of the controversies over thermal pollution and radiation, by the early 1970s the AEC was on the defensive, suffering from a deteriorating public image and an almost complete lack of credibility. Recognizing the inevitability of the separation of the AEC's regulatory and developmental functions, members of the AEC, the JCAE, and

the nuclear industry supported the Energy Reorganization Act of 1974 because they viewed it as an opportunity to emphasize the contribution nuclear power could make to solving the nation's energy problems. Hoping to enhance the prospects of nuclear power, subsystem members favored the proposal even though it meant sacrificing the AEC, whose responsibilities would be transferred to two new agencies—the Energy Research and Development Administration (ERDA) and the Nuclear Regulatory Commission (NRC).

The abolition of the AEC deprived the nuclear industry of a powerful patron and created a more complicated institutional framework for nuclear matters. Where there was once a small group of insiders restricting participation to program supporters, by the middle of the decade the range of governmental participants would be decidedly broader and less exclusive. In addition to the Nuclear Regulatory Commission and ERDA, the courts, the Environmental Protection Agency (EPA), the Fish and Wildlife Service (FWS), and state and local governments claimed jurisdiction. The AEC, for example, had not been very concerned with the environmental impacts of reactors, but the EPA and FWS were. More and more, licensing decisions were challenged in court, and state and local governments became involved through their siting and rate-making powers. For opponents of nuclear power, these multiple venues translated into new opportunities to shape nuclear decision making and to appeal unfavorable decisions to other, more receptive governmental actors. For these reasons, institutional shifts were a critical factor in the expansion of the conflict over nuclear power and in eventual shifts in government policy. Like the AEC, the Joint Committee on Atomic Energy came under attack in the 1970s for being overly protective of the nuclear power industry. The growing wave of criticism over nuclear power, coupled with the demise of the AEC, prompted increased scrutiny of the Joint Committee and fueled calls for its abolition as well. Over the next few years, the once-invincible Joint Committee struggled to ward off challenges to its authority and to the nuclear program it had supported for so long. The joint committee would not win this battle, however, and it was eliminated by Congress in 1977.

With the demise of the JCAE, longtime supporters lost control of the strategic junctures of policy-making. Responsibility for nuclear policy was transferred to some two dozen subcommittees, some of which were sympathetic to the concerns voiced by antinuclear groups. For opponents of nuclear power,

this decentralization reinforced previous venue changes and created additional access points to nuclear decision making. The new structural arrangements in Congress rendered the institution less exclusive and more permeable to outsiders, but also made it more complicated. Although nuclear policy once had been shaped exclusively by the Joint Committee, it now seemed that everyone had some claim to nuclear oversight. With responsibility for nuclear policy dispersed so widely, it became increasingly difficult for Congress to resolve major policy issues.

Congress in the 1970s and 1980s was inhospitable to decisive initiatives from either side, preferring to approach nuclear policy in an ad hoc manner. Still fearful of energy shortages, members of Congress were unwilling to close the door on the nuclear option, deeming it an essential short-term energy source, but at the same time they were unwilling to proceed with a full-scale nuclear program because of concerns about reactor safety. For the next few years, Congress would increase funding for both reactor safety and licensing programs and also would insist on greater oversight of the NRC and the nuclear power industry. In this way, legislators could claim that they were doing everything possible to provide the nation with safe and secure energy supplies.

4. REGULATORY FAILURE AND THE CHANGING ECONOMICS OF NUCLEAR POWER

In the late 1960s, the costs of producing electricity from coal and nuclear plants were comparable; nuclear plants were more expensive to build but had lower fuel and operating costs. By the early 1970s, utilities were ordering roughly equal amounts of coal and nuclear capacity. Within a few years, though, the bottom fell out of the market, sparing no sector of the industry. Significantly, nuclear power began its demise well before the accident at Three Mile Island in March 1979. Orders for new nuclear plants actually began to decline in 1974, and no orders would be placed after 1978. Furthermore, between 1974 and 1984, utilities canceled plans for over 100 reactors, many of which were already under construction. Some of the abandoned reactors were more than 50% complete.

The industry's economic woes were attributable to internal and external forces that caused construction and maintenance costs to skyrocket and utilities to scale back their plans to build additional nuclear

generating capacity. Among the exogenous forces, none was more important than the Organization of Petroleum Exporting Countries (OPEC) oil embargo, which sparked sudden price hikes in fossil fuels and fundamentally transformed energy consumption patterns. Higher oil prices and the desire for secure energy sources might have been expected to make nuclear-generated electricity more attractive. Instead, fears of energy shortages and rising apprehension about American dependence on imported oil prompted Americans to use less energy, thus undercutting the need for new sources of electricity.

The nuclear industry was also hurt by rapidly rising construction costs. Construction costs for all types of power plants increased between 1968 and 1977, but those for nuclear plants rose at twice the rate of coal plants and continued to increase into the next decade. By almost any measure, nuclear plants became prohibitively expensive to build. Reactors coming on-line during the 1980s, for example, cost an average of \$3500/kilowatt-hour in 1983 dollars, compared with an average of \$600/kWh for the 57 reactors completed before 1981. A significant portion of the increased construction costs was attributable to new regulatory requirements that increased the amount of materials, labor, and time needed to build nuclear plants. Although the nuclear industry argued that most of their problems, especially the rising construction costs, were the result of overly zealous NRC regulators, in reality, the claim of excessive regulation is simply inaccurate and ignores other more compelling explanations.

It seems clear that earlier regulatory failures and industry mismanagement were responsible for some of the cost increases. The AEC neglected its regulatory duties for most of its history and devoted the bulk of its attention and resources to working with the private sector to develop and promote the new technology. Although this process did facilitate reactor licensing in the short term, the case-by-case approach ultimately undermined the industry's long-term prospects. During this period of rapid scale-up, no two reactor designs were alike, so every application was unique, and every application review was unique. In addition, the larger plants were more complex and posed greater potential risks. This increase in the number and complexity of applications, along with the failure to standardize reactor designs, posed tremendous difficulties for the AEC staff, which lacked the resources and experience to determine if the designs were adequate and if existing regulatory standards would assure public health and safety. The staff could not keep up, and licensing

review times subsequently increased. The AEC responded to the growing licensing workload by nearly doubling its staff and by developing standard review plans to formalize the review process. As the licensing and regulatory staffs became larger and more sophisticated, they adopted new environmental and safety reviews and demanded ever more detailed information from applicants. Furthermore, as the quality and quantity of reviews increased, the staff discovered that existing standards were often inadequate and began to impose more stringent safety and environmental requirements on both new and existing plants.

Other regulatory changes were a product of increased operating experience. Once a sufficient number of reactors actually came on-line in the 1970s, the commission and the industry had a chance to evaluate the sufficiency of existing standards, designs, and operating procedures. Over time, previously unanticipated equipment and design problems were discovered, and remedial steps were taken to correct them.

What these examples show is that many of the industry's economic wounds were either self-inflicted or were the result of regulatory failures by the AEC. If the industry and the AEC had not rushed to commercialize nuclear power before the technology was mature, much of the political controversy, licensing delays, and economic ills might have been avoided. A strong regulatory program might have detected potential problems early on, before reactors were built, and even more important, before they started producing electricity. If such a program had been in place, costly design changes and construction stoppages would not have been unnecessary. But that is not what happened. Instead, the discovery of unanticipated problems and reluctant admissions that some existing standards were insufficient undermined confidence in the agency and the industry and fueled the growing safety controversy.

5. NUCLEAR POWER IN THE 1980s

5.1 The NRC after Three Mile Island

At approximately 4:00 a.m. on 28 March 1979, there was a serious accident at the Three Mile Island (TMI) nuclear facility near Harrisburg, Pennsylvania. Although it would take months to reconstruct the events and circumstances that caused the accident, it was much easier to calculate the political damage to the nuclear power industry and the NRC.

Put simply, the TMI accident renewed and deepened concerns about the safety of reactors. A blue-ribbon panel appointed by President Carter to investigate the accident at Three Mile Island was highly critical of the NRC and the nuclear industry and recommended major changes in the practices and attitudes of both. After concluding that there was “no well-thought-out, integrated system for the assurance of nuclear safety within the NRC,” the Kemeny Commission recommended a “total restructuring” of the agency. That did not happen, and the NRC never became an effective regulator.

In part, this was because the NRC was not really a new agency—it was merely a spin-off of the regulatory branch of the AEC. Although the Energy Reorganization Act succeeded in abolishing the Atomic Energy Commission, it never entirely displaced the commission’s deeply entrenched belief in the value of nuclear power. Part of the explanation stems from the fact that the NRC was essentially a carryover from the AEC in terms of personnel, regulations, and attitudes. Indeed, the NRC’s first official action was to adopt all of the AEC’s rules, regulations, and standards. Throughout the 1980s, the NRC behaved much like its predecessor. The agency consistently tried to reduce licensing delays, blocked the imposition of new safety requirements, and worked to curtail public involvement in agency proceedings. To a large extent, the NRC was responding to signals from Presidents Reagan and Bush, both of whom supported nuclear power and pushed a deregulatory agenda. Reagan, in particular, tried to control the NRC by appointing commissioners who shared his aversion to regulation.

When necessary, the NRC waived and changed rules that threatened to close existing plants or to delay the licensing of new ones. In several cases when plants did not meet the regulations, the NRC simply changed the regulations to facilitate licensing. The most notorious rule change involved the Shoreham, Long Island and Seabrook, New Hampshire reactors. After the accident at Three Mile Island revealed serious flaws in emergency and evacuation planning, the NRC issued new rules that required an emergency-planning zone within a 10-mile radius of nuclear plants. The rules also mandated state and local government approval of evacuation plans for all people living within the zone. When state and local officials adjacent to the two plants refused to prepare emergency plans, the NRC issued a new rule waiving the requirement that state and local officials participate in evacuation planning.

5.2 The Problem of Nuclear Waste Disposal

The question of what to do with nuclear waste did not receive much attention in the early years of the atomic program. The AEC and others in the subgovernment were more concerned with the scientific challenges of harnessing the atom. Nuclear waste, in contrast, was generally assumed to be a less complex, essentially technical problem that would be easy to solve. Reflecting its low priority within the AEC, action on nuclear waste disposal was deferred indefinitely. This decision, like so many others in the program’s early years, only made the problem worse, because the lack of an effective nuclear waste policy became one of the antinuclear movement’s most potent arguments in the 1980s.

The problem for utilities was that spent reactor fuel was piling up in cooling ponds at reactor sites around the nation. The cooling ponds, however, were not designed to store waste permanently, and most could hold no more than 3 years of spent fuel. Moreover, utilities could not expand their on-site storage capacity without state regulatory approval, which was by no means assured. By the late 1970s, with many utilities worried about running out of space and facing mounting on-site storage costs, the nuclear industry sought legislative relief, arguing that a federal commitment on high-level waste was essential to its future.

After several years of deliberation, Congress acted in 1982, adopting the Nuclear Waste Policy Act (NWPA). The NWPA established detailed procedures and a schedule for the selection, construction, and operation of two permanent high-level waste repositories. Under the law, the Department of Energy (DOE) would conduct an intensive nationwide search for potential sites, which would be evaluated with a demanding set of technical criteria and guidelines, including environmental assessments. The DOE was then to compile a list of potential sites from which the President would select two for further review and site characterization, one in the eastern United States and one in the west. By all accounts, implementation of the Nuclear Waste Policy Act was a failure. Although the DOE met the act’s 1983 deadline for issuing siting guidelines, virtually every other stage of the selection process was bitterly contested. All of the states considered as potential repository sites challenged the designation. With the states in open rebellion, Congress tried to repair the damage by amending the NWPA. In December 1987, Congress rejected the multiple-site

search process established in the legislation and instead designated Yucca Mountain, Nevada, as the repository site. The search for a second site in the east was abandoned.

The case of nuclear waste disposal demonstrates the increasingly decentralized nature of nuclear politics in the 1980s. The widespread search for suitable nuclear waste disposal sites touched many states and communities and brought multiple new actors, in the form of state and local governments, to the policymaking arena. Whatever their motivations, the new actors brought different perspectives and goals to the debate; indeed, most actively resisted federal efforts to site nuclear waste facilities. With states and local governments assuming a critical role in policymaking, it is clear that the conflict over nuclear waste had expanded far beyond the confines of subgovernment politics. At the same time, however, the nuclear waste issue illustrates that support for nuclear power within the federal government remained strong in the 1980s. The Department of Energy, the NRC, and Presidents Reagan and Bush supported nuclear power, and worked to reestablish a favorable political climate. With states impeding the siting of a high-level repository, the industry sought to shift the financial and legal risks of nuclear waste management to the federal government, which willingly obliged.

5.3 The Accident at Chernobyl

In 1986, a fire at a reactor in Chernobyl, in the Soviet Union, released a large radioactive cloud across much of the Ukraine and northern Europe. The world learned of the accident when monitors in Europe detected an unusual spike in radiation levels. Soviet officials initially denied that anything had happened, but eventually conceded that something had gone horribly wrong at Chernobyl. Hundreds died, thousands were evacuated, and large areas were rendered uninhabitable. The story dominated the news throughout the world and reinforced questions about the safety of nuclear reactors. Although nuclear officials in the United States and elsewhere correctly noted that the Soviet reactor had fewer redundant safety systems than reactors in their own nations, the industry suffered a severe crisis in public confidence. In the United States and many European nations, the Chernobyl accident further undermined the prospects for a nuclear resurgence. Outside of France, which remained committed to a nuclear future, few European nations built new reactors.

6. NUCLEAR POWER TODAY

6.1 Nuclear Power in the United States

American policymakers continue to be ambivalent about commercial nuclear power. Reflecting this ambivalence, nuclear policymaking in the past 20 years has been marked by incrementalism, with policymakers unwilling, or unable, to stray far from the status quo. Barring another energy or environmental crisis, it is extremely unlikely that American utilities will soon begin building any additional nuclear plants. And in the absence of a severe accident, policymakers are equally unlikely to require the shutdown of existing reactors or to rule out the possibility of future contributions from nuclear power. In the interim, the range of acceptable policy options will continue to lie somewhere in between the two extremes.

Commercial nuclear power peaked in the United States in 1990, when 112 reactors were in operation. Today, there are 103 operable reactors, in 31 states, generating approximately 20% of the nation's electricity. All told, U.S. utilities have placed orders for 259 nuclear reactors, but none has been ordered since 1978. Of the total order, 132 reactors received full power licenses; the rest were canceled. No plant has commenced operations since 1996, and 28 reactors have been permanently closed down.

Despite these downward trends, nuclear capacity in the United States is projected to increase slightly over the next 25 years, mostly due to improved performance from existing plants and fewer nuclear retirements due to licensing extensions. Owners of 10 reactors have applied for and received 20-year license extensions, and another 20 have applications pending. Owners of an additional 20 reactors are expected to apply for license extensions in the next 6 years. The George W. Bush administration is currently working to increase nuclear power's contribution to the nation's energy needs. In 2001, as part of the administration's national energy strategy, the Department of Energy issued a "roadmap" to deploy new reactors in the United States by 2010. In another important move long sought by the nuclear industry, President Bush approved the siting of a high-level waste repository at Yucca Mountain, Nevada. Although the decision has been challenged by the state, the facility is scheduled to begin operation in 2010. If the decision stands, nuclear advocates will have removed, at least temporarily, one of the major political roadblocks to the technology's ultimate fate.

6.2 Nuclear Power after September 11

The terrorist attacks of September 11, 2001 thrust nuclear power into the spotlight once again, although not in the way that industry supporters hoped. One of the planes that brought down New York's Twin Towers flew by the Indian Point nuclear facility located just 35 miles north of the city. The NRC and industry leaders were forced to confront the question of whether either group was prepared to deal with such attacks, and their responses were not encouraging. Specifically, there are concerns that both reactors and spent fuel storage pools, which are often located adjacent to reactors, are vulnerable to terrorist attack. Critics charge, moreover, that the NRC and reactor owners have consistently underestimated the risk of such attack, and that the agency's security rules are too lax and are often ignored. In designing its rules, for example, the NRC did not consider attacks by plane to be a credible threat. And although earthquakes and other natural disasters were considered in the design of spent storage fuels, they were not designed to withstand terrorist attack. According to some reports, attacks on reactors and storage facilities could have consequences worse than the 1986 accident at Chernobyl.

6.3 Nuclear Power around the World

There are now 442 nuclear plants in operation worldwide, with an additional 35 under construction. Nuclear power provides about 17% of the world's electricity. Thirty nations now have nuclear reactors, with the vast majority located in Organization for Economic Cooperation and Development (OECD) nations, where nuclear power accounts for almost 24% of all electricity. According to the Nuclear Energy Institute, the top 10 nuclear-generating nations are, in order, the United States, France, Japan, Germany, Russia, South Korea, United Kingdom, Canada, Ukraine, and Sweden. Nuclear power provides more than 75% of total electricity production in Lithuania and France, more than 50% in Belgium and Slovakia, more than 40% in Ukraine, Sweden, and Bulgaria, and about 30% in 10 other nations. The United States, with 104 operating reactors, has the most; France has 59 and Japan has 54. Japan also has plans to build up to 10 additional plants in the next 10 years.

Current projections are that global energy demand will more than double in the next 50 years, leading more nations to consider nuclear power as an answer

to their energy needs. The roster of nations that now have reactors includes a growing number that have been marked by economic, political, or social instability, or that are located in volatile regions of the globe. Among the nations that have added commercial reactors in the past 5 years are Pakistan, India, South Korea, and Brazil. Some of these nations and others, including Iran and North Korea, have reactors that have been linked to nuclear weapons programs. The spread of reactor technology to such nations raises considerable concerns about nuclear proliferation and, after September 11, about reactor vulnerability to terrorism. It is an open question whether these nations have the ability to secure their reactors and the nuclear materials they use and produce. As a result, the future of nuclear power remains clouded.

SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Electricity Use, History of • Hydrogen, History of • Hydropower, History and Technology of • Manufactured Gas, History of • Natural Gas, History of • Nuclear Engineering • Nuclear Power Economics • Nuclear Power Plants, Decommissioning of • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

- Balogh, B. (1991). "Chain Reaction: Expert Debate and Public Participation in American Commercial Nuclear Power, 1945–1975." Cambridge University Press, Cambridge.
- Bupp, I. C., and Derian, J.-C. (1978). "Light Water: How the Nuclear Dream Dissolved." Basic Books, New York.
- Campbell, J. L. (1988). "Collapse of an Industry: Nuclear Power and the Contradictions of U.S. Policy." Cornell University Press, Ithaca, New York.
- Duffy, R. J. (1997). "Nuclear Politics in America: A History and Theory of Government Regulation." University Press of Kansas, Lawrence.
- Falk, J. (1982). "Global Fission: The Battle over Nuclear Power." Oxford University Press, New York.
- Green, H. P., and Rosenthal, A. P. (1963). "Government of the Atom: The Integration of Powers." Atherton Press, New York.
- Hewlett, R. G., and Anderson, O. E. (1962). "A History of the Atomic Energy Commission, Vol. 1, The New World, 1939–1946." Pennsylvania State University Press, University Park.

- Hewlett, R. G., and Duncan, F. (1969). "A History of the Atomic Energy Commission, Vol. 2, Atomic Shield, 1947–1952." Pennsylvania State University Press, University Park, PA.
- Jasper, J. M. (1990). "Nuclear Politics: Energy and the State in the United States, Sweden, and France." Princeton University Press, Princeton, New Jersey.
- Nelkin, D., and Pollak, M. (1981). "The Atom Besieged: Extraparliamentary Dissent in France and Germany." MIT Press, Cambridge, Massachusetts.
- Union of Concerned Scientists. (1987). "Safety Second: The NRC and America's Nuclear Power Plants" (M. Adato, principal author, with J. MacKenzie, R. Pollard, and E. Weiss). Indiana University Press, Bloomington, Indiana.

Nuclear Power Plants, Decommissioning of

REBEKAH HARTY KRIEG

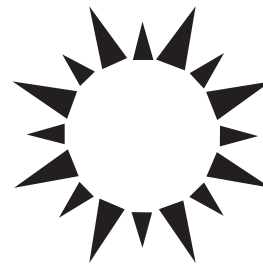
Rebekah Krieg Consulting
West Richland, Washington, United States

EVA E. HICKEY and JAMES R. WEBER

Pacific Northwest National Laboratory
Richland, Washington, United States

MICHAEL T. MASNIK

U.S. Nuclear Regulatory Commission
Washington, D.C., United States



-
1. Definition of Decommissioning
 2. Decommissioning Options and Activities
 3. Decommissioning Process
 4. Environmental Impacts
 5. Summary of Decommissioning Status of U.S. Nuclear Power Facilities
 6. Conclusion

Glossary

boiling water reactor A nuclear power reactor in which water, used as both coolant and moderator, is allowed to boil in the reactor core. The resulting steam can be used directly to drive a turbine and electrical generator, thereby producing electricity.

contamination Any undesired radioactive material or residual radioactivity that is deposited on or in structures, areas, objects, or people in excess of acceptable levels.

decontamination The reduction or removal of contaminating radioactive material from equipment, structures, areas, objects, or people.

greenfield One possible end state of decommissioning in which above-ground structures have been removed and efforts made to revegetate the site. Buildings may have been removed to below grade and then covered with soil.

high-level waste Material that consists of (1) irradiated (spent) reactor fuel, (2) liquid waste from the operation of the first cycle solvent extraction system and concentrated wastes from subsequent extraction cycles, in a facility for reprocessing irradiated reactor fuel, or

(3) solids into which such liquid wastes have been converted. Primarily in the form of spent fuel discharged from commercial nuclear power reactors, high-level waste also includes some reprocessed high-level waste from defense activities and a small quantity of reprocessed commercial high-level waste.

independent spent fuel storage installation (ISFSI) A complex designed and constructed for the interim storage of spent nuclear fuel and other radioactive materials associated with spent fuel storage. The most common design for an ISFSI at this time is a concrete pad with dry casks containing spent fuel bundles.

license termination Owners of nuclear power facilities in the United States hold a license that is granted by the U.S. Nuclear Regulatory Commission. This license is not terminated until the licensee can demonstrate compliance with specific criteria that detail the amount of radioactive material that can remain at the site following decommissioning. After the termination of the license, facility owners are free to use the facility or site for any nonnuclear uses they wish to pursue.

low-level waste A general term for a wide range of radioactive wastes. Industries, hospitals, research institutions, private or government laboratories, and nuclear fuel cycle facilities (e.g., nuclear power reactors and fuel fabrication plants) using radioactive materials generate low-level waste as part of their normal operations. These wastes have many physical and chemical forms and different levels of contamination. Low-level waste usually comprises radionuclide-contaminated materials (rags, papers, filters, solidified liquids, ion-exchange resins, tools, equipment, discarded protective clothing, dirt, construction rubble, concrete, or piping).

partial site release The release of a portion of an operating or decommissioning nuclear power reactor facility site for unrestricted use.

pressurized water reactor A power reactor in which heat is transferred from the core to an exchanger by high-temperature water kept under high pressure in the primary system. Steam is generated in a secondary circuit. Many reactors producing electric power are pressurized water reactors.

spent fuel The depleted nuclear fuel that has been removed from a nuclear reactor, for economic or other reasons, or because it can no longer sustain power production (cannot effectively sustain a chain reaction).

Decommissioning is a technical and administrative process by which a nuclear facility or site is safely removed from service and residual radioactivity is reduced to a level that permits release of the property for either partial or unrestricted public uses. Decommissioning is a planned, systematic activity that is both remediative and administrative. Environmental impacts are generally less during decommissioning than during construction and, in the context of water use, may be less than during operations. Nonetheless, decommissioning does impact land and water use, water and air quality, and local aquatic and terrestrial ecologies.

1. DEFINITION OF DECOMMISSIONING

During nuclear power plant reactor operation, a large inventory of radioactive fission products builds up within the fuel. Virtually all of the fission products are contained within the fuel pellets, which are encased in hollow metal rods, hermetically sealed to prevent further release. Occasionally, however, fuel rods develop small leaks, allowing a small fraction of the fission products to contaminate the reactor coolant. The radioactive contamination in the reactor coolant is the source of gaseous, liquid, and solid radioactive wastes generated in light water reactors during operation. Despite precautions to prevent the movement of contaminated material and to clean up any contaminated materials, such material is unintentionally moved through facilities by workers, equipment, liquids, and, to some degree, air motion. Contamination is most likely to occur in the reactor building, around the spent fuel pool, and around structures, systems, and components in areas near the reactor. During decommissioning,

remediative measures are utilized to eliminate fission product contaminants.

The second source of radioactive material encountered in reactors during decommissioning is from activation. Activation products are created when stable substances are bombarded by neutrons. Concrete and steel surrounding the core of the reactor are the most common locations of activated material.

The spent fuel contains the largest amount of radioactive material at a facility that is being decommissioned, followed by the reactor vessel, internals, and bioshield. Systems containing smaller amounts of radioactive material include the steam generator, pressurizer, piping of the primary system and other systems, as well as the radioactive waste systems. Minor contamination is found in the secondary systems and miscellaneous piping.

The definition of decommissioning provides for a definite starting point and ending point for the decommissioning process and for the evaluation of environmental impacts from decommissioning. Decommissioning starts after a facility has been removed from service. Thus, the decision to permanently cease operation of a facility and the activities and resulting environmental impacts from permanent cessation of operation (such as the change in water temperature of the cooling water body) are not considered a part of the decommissioning process. The decommissioning process starts with a planning and preparation process and then moves through a plant transition and deactivation phase, in which fuel is transferred to temporary spent fuel pool storage, the systems are drained and may be flushed, and support systems may be installed or moved. A third phase, decontamination and dismantlement precedes the fourth phase, compliance with relevant regulations for final disposition of the site. Some activities that may occur during the decommissioning process could also occur during normal operations (such as construction, maintenance, and decommissioning of an independent spent fuel storage installation; transportation and disposal of the spent fuel, away from the reactor location; or disposal of low-level waste). These activities should not be considered exclusive to decommissioning.

2. DECOMMISSIONING OPTIONS AND ACTIVITIES

Three approaches are generally the options used for decommissioning nuclear power facilities:

1. DECON. The equipment, structures, and portions of the facility and site that contain radioactive contaminants are promptly removed or decontaminated to a level that permits cessation of operations and release of the site for public use (requires approximately 5 to 10 years).

2. SAFSTOR. The facility is made safe and stable and is maintained in that state (safe storage) until it is subsequently decontaminated and dismantled. During SAFSTOR, a facility is left intact, but the fuel has been removed from the reactor vessel and radioactive liquids have been drained from systems and components and then processed. Radioactive decay occurs during the SAFSTOR period, thus reducing the quantity of contaminated and radioactive material that must be disposed of during decontamination and dismantlement. The definition of SAFSTOR also includes decontamination and dismantlement of the facility at the end of the storage period.

3. ENTOMB. Radioactive structures, systems, and components are encased in a structurally long-lived substance, such as concrete. The entombed structure is appropriately maintained, and continued surveillance is carried out until the radioactivity decays to a safe level.

These options can be combined. For example, after power operations cease at a facility, a short storage period could be used for planning purposes, followed by removal of large components (such as the steam generators, pressurizer, and reactor vessel internals). At this point, the facility could be placed in storage for 30 years, and eventually the decontamination and dismantlement process would be completed.

2.1 DECON

Numerous activities may occur during DECON. Draining (and potentially flushing) any contaminated systems and removing resins from ion exchangers are early steps. Setup activities include establishing monitoring stations or designing and fabricating special shielding and contamination control envelopes to facilitate decommissioning activities. Reduction of the site security area (setup of new security monitoring stations), modification of the control room or establishment of an alternate control room may also occur. Site surveys are an important step. Decontamination of radioactive components, using chemical decontamination techniques also occurs. This often precedes the removal of reactor vessel and internals, removal of other large components (including major radioactive components), and

removal of the balance of the primary system (charging system, boron control system, etc.). Activities related to removing other significant radioactive components are also undertaken. Decontamination and/or dismantlement of structures or buildings, temporary on-site storage of components, shipment and processing of low-level radioactive waste (including compaction or incineration of the waste), removal of the spent fuel and other non-low-level wastes to a storage facility at, or away from, the reactor site, and removal of hazardous radioactive (mixed) wastes are activities that may occur at any time during the process. Changes in management and staffing are also usually implemented.

There are several advantages to using the DECON option of decommissioning. By beginning the decontamination and dismantlement process soon after permanent cessation of operation, the available workforce can be maintained. This is advantageous because the facility staff is knowledgeable about the facility and will not require the retraining that would be necessary should the facility be placed in storage for 10 to 40 years. A second advantage may be the availability of facilities willing to accept low-level waste. Low-level waste disposal sites available to reactor facilities may be somewhat variable from year to year, and there is not always a certainty that a site will be available in future years. A third advantage is that costs for DECON may be lower than for other options because the price for decommissioning at a later date is greater, factoring in the cost of storage and inflation. Fourth, DECON also eliminates the need for long-term security, maintenance, and surveillance of the facility (excluding the on-site storage of spent fuel), which are required for the other decommissioning options. Fifth, completing the decommissioning process and being in compliance with relevant regulations for final disposition of the site allow the facility and site to become available for other purposes. The major disadvantages of DECON are higher levels of worker exposure to radiation and significant initial expenditures. Also, compared to SAFSTOR, DECON requires a larger potential commitment of disposal site space.

2.2 SAFSTOR

The SAFSTOR decommissioning option involves ensuring that the facility is in a safe, stable condition and maintaining that state for a period of time, followed by subsequent decontamination and dismantlement to levels that permit license termination. During the storage period of SAFSTOR, the facility is

left intact. The fuel has been removed from the reactor vessel and radioactive liquids have been drained from systems and components and processed. Radioactive decay occurs during the storage period, reducing the quantity of contaminated and radioactive material that must later be disposed of during decontamination and dismantlement. A number of activities typically occur during the preparation and storage stages of the SAFSTOR process. During preparation, there must be draining (and potential flushing) of some systems and removal of resins from ion exchangers, reconfiguration of the spent fuel pool cooling system, decontamination of highly contaminated and high-dose areas as necessary, performance of a radiological assessment as a baseline before storage, and removal of low-level waste that is ready to be shipped. These activities will occur at the same time or be followed by shipment and processing or storage of the fuel and other non-low-level wastes, deenergizing or deactivating systems and equipment, and reconfiguration of ventilation systems, fire protection systems, and the spent fuel pool cooling system for use during storage. Establishment of inspection and monitoring plans for use during storage, maintenance of any systems critical to final dismantlement during storage, are the final requirements of the preparatory stage. Changes in management or staffing could occur at any time.

During storage, there must be performance of preventative and corrective maintenance on plant systems that will be operating and/or functional during the storage period. There will also have to be maintenance to preserve structural integrity, maintenance of security systems, and maintenance of radiation effluent and environmental monitoring programs. Processing of any radioactive waste generated (usually small amounts) must also be done during the storage period. Following the storage period, the facility is decontaminated and dismantled to radiological levels that allow termination of the license. Activities during this period of time will be the same activities that occur for DECON.

There are several advantages to using the SAFSTOR option of decommissioning. A substantial reduction in radioactive material as a result of radioactive decay during the storage period reduces worker and public exposure to doses below those of the DECON alternative. Because there is potentially less radioactive waste, less waste disposal space is required. Moreover, the costs immediately following permanent cessation of operations are lower than costs during the first years of DECON because of

reduced amounts of activity and a smaller workforce. However, because of the time gap between cessation of operations and decommissioning activities, SAFSTOR can result in a shortage of personnel familiar with the facility at the time of dismantlement and decontamination and a necessity to train new personnel. In some cases, this creates a knowledge gap because the people most familiar with the facility have long since retired by the time the storage period is completed. Other disadvantages are that during the prolonged period of storage, the plant requires continued maintenance, security, and surveillance and there are uncertainties regarding the availability and cost of future low-level waste sites, which could mean higher future costs for decontamination and dismantlement.

2.3 ENTOMB

The ENTOMB decommissioning option involves encasement of long-lived radioactive contaminants in a structurally long-lived material, such as concrete. The entombed structure is appropriately maintained and surveillance is continued until the radioactivity decays to a level permitting release of the property. The purpose of the entombment process is to isolate the entombed radioactive waste so that the facility can be released to partial or unrestricted use. Therefore, prior to entombment, an accurate characterization of the radioactive materials that are to remain is needed, and the adequacy of the entombment configuration to isolate the entombed radioactive waste must be determined.

To characterize the environmental impacts from decommissioning via the ENTOMB option adequately, two different scenarios were developed to encompass a wide range of potential options by describing two possible extreme cases of entombment. The two scenarios, "ENTOMB 1" and "ENTOMB 2," differ primarily in the amount of decontamination and dismantlement that is done prior to the actual entombment. ENTOMB 1 assumes significant decontamination and dismantlement and removal of all contamination and activation involving long-lived radioactive isotopes prior to entombment. ENTOMB 2 assumes significantly less decontamination and dismantlement, significantly more engineered barriers, and the retention on site of long-lived radioactive isotopes. Both options assume that the spent fuel would be removed from the facility and either transported to a permanent high-level waste repository or placed in on-site storage.

ENTOMB 1 is envisioned to begin the decommissioning process in a manner similar to the

DECON option. The reactor would be defueled and the fuel initially either placed into the spent fuel pool for some period prior to disposal at a licensed high-level waste repository or placed in an on-site storage. Any decommissioning activity would be preceded by an accurate radiological characterization of structures, systems, and components (SSCs) throughout the facility. Active decommissioning would begin with draining and decontaminating of SSCs throughout the facility, with the goal of isolating and fixing contamination. The SSCs would either be decontaminated or removed and either shipped to a low-level waste burial site or placed inside the reactor containment building. Off-site disposal of resins and considerable amounts of contaminated material would occur. There would likely be a chemical decontamination of the primary system. The reactor pressure vessel and reactor internals would be removed, either intact or after sectioning, and disposed of off site.

Any other SSCs that have long-lived activation products would be removed. Interim dry storage of the vessel, vessel internals, and any other SSCs containing long-lived activation products could occur on site until a final disposal site for this waste is identified. Depending on whether the components are contaminated with long-lived radioisotopes, steam generators and the pressurizer would either be removed and disposed of off site or be retained inside the reactor containment. The spent fuel pool would be drained and decontaminated. The reactor building or containment would then be filled with SSCs contaminated with relatively short-lived isotopes from the balance of the facility. Material would be placed in the building in a manner that would minimize the spread of any contamination (i.e., dry, contamination fixed, or isolated). Engineered barriers would be put in place to deny access and to eliminate the possibility of release of any contamination to the environment. The reactor building or containment would be sealed and made weatherproof.

A partial site release could be completed for almost all of the site and the balance of the plant, thus allowing the remainder of the site to be used for other purposes. A monitoring period as long as 20 to 30 years would probably be necessary to demonstrate that the contamination was isolated and the structure was permanent. The general activities that would occur during ENTOMB 1 are planning and preparation activities, draining (and possibly flushing) contaminated systems and removing resins from ion exchangers, reduction of the site security area (optional), deactivation of support systems, and

decontamination of radioactive components, including use of chemical decontamination techniques. There would also be removal of the reactor vessel and internals, removal of other large components (including major radioactive components), removal of fuel from the spent fuel pool to a storage facility at or away from the reactor site, dismantlement of remaining radioactively contaminated structures and placement of the dismantled structures in the reactor building, installation of engineered barriers and other controls to prevent inadvertent intrusion and dispersion of contamination outside of the entombed structure, and filling the void spaces in the previous reactor building structure with concrete.

ENTOMB 2 was also envisioned to begin the decommissioning process in a manner similar to the DECON option. The reactor would be defueled and the fuel initially placed into the spent fuel pool for some period prior to disposal at a permanent high-level waste repository or placed in on-site storage. Any decommissioning activity would be preceded by an accurate radiological characterization of SSCs throughout the facility. Active decommissioning would begin with the draining and decontamination of SSCs throughout the facility, with the goal of isolating and fixing contamination. The spent fuel pool would be drained and decontaminated. SSCs would either be decontaminated or be removed and either shipped to a low-level waste burial site or placed inside the reactor containment building or the reactor building. Disposal off site of resins would occur. The primary system would be drained, the reactor pressure vessel filled with contaminated material, all penetrations sealed, the reactor pressure vessel head reinstalled, and the reactor vessel filled with low-density concrete. Reactor internals would remain in place. Emphasis would be placed on draining and drying all systems and components and fixing contamination to prevent movement, either by air or liquid means. The steam generators and pressurizer would be laid up dry and remain in place. The reactor building or containment would then be filled with contaminated SSCs from the balance of the facility. Material would be placed in the building in a manner that would minimize the spread of any contamination (i.e., dry, contamination fixed, or isolated).

Engineered barriers would be put in place to deny access and eliminate the possibility of the release of any contamination to the environment. The ceiling of the containment or reactor building, in the case of boiling water reactors, may be lowered to near the refueling floor and to the top of the pressurizer for

pressurized water reactors. The cavity of the remaining structure would be filled with a low-density concrete. The resulting structure would be sealed and made weatherproof and covered with an engineered cap designed to deny access, and prevent the intrusion of water or the release of radioactive contamination to the environment.

A partial site release would be completed for most of the site and the balance of the plant, thus allowing the remainder of the site to be used for other purposes if appropriate. A site-monitoring program would likely be initiated to demonstrate the isolation of the contamination and the permanence of the structure. Monitoring could be as long as 100 years. The general activities that would occur during ENTOMB 2 are planning and preparation activities, draining (and potentially flushing) of contaminated systems and removal of resins from ion exchangers, deactivation of support systems, and removal of fuel from the spent fuel pool to a storage facility at or away from the reactor site. It would be necessary to dismantle all radioactively contaminated structures (other than the reactor building), with placement of the dismantled structures in the reactor building. The ceiling of the reactor building would be lowered to near the refueling floor (in boiling water reactors) or near the top of the pressurizer (in pressurized water reactors), and engineered barriers installed and other controls put in place to prevent inadvertent intrusion and dispersion of contamination outside of the entombed structure, filling of the cavity of the reactor building structure with low-density concrete, and placement of an engineered cap over the entombed structure to further isolate the structure from the environment.

The advantages of both ENTOMB options are reduced public exposure to radiation due to significantly less transportation of radioactive waste to low-level waste disposal sites and corresponding reduced cost of low-level waste disposal. An additional advantage of ENTOMB 2 is related to the significant reduction in the amount of work activity, and thus a significant reduction in occupational exposures, as compared to the DECON or SAFSTOR decommissioning options.

3. DECOMMISSIONING PROCESS

The decommissioning process, as shown in Fig. 1, can be divided into four stages, beginning with planning, proceeding through deactivation activities, decontamination and dismantlement and finally to administrative

clearance for final disposition of the site. The first and last stages are primarily administrative.

Planning typically takes 1.5 to 2.5 years, regardless of the decommissioning option chosen. The main activities during this stage are determining the decommissioning option, making changes to the organizational structure (such as layoffs, strategic hiring, and redeploying personnel) to prepare for decommissioning activities, and starting the necessary regulatory activities. One preparation would be conducting postshutdown surveys of the facility for contamination or activation material. Planning activities vary, depending on the decommissioning option chosen and on whether the shutdown was planned or unplanned. If planned, some preparations may have already been made.

During stage 2, the plant moves from cessation of reactor operation to decommissioning activities. This stage will last from about 0.5 to 1.5 years, regardless of the decommissioning option chosen. Fuel will have to be transferred out of the reactor and into the spent fuel pool at all plants. Isolation, stabilization, or removal of all unnecessary systems is also conducted during this stage and new support systems needed for decontamination may be added. Changes to electrical systems, for instance, are common at this stage. Chemical decontamination of the primary system at this stage has resulted in reduction of total person-rem exposure during subsequent decommissioning activities. Chemically reducing levels of contaminants from piping, for instance, may significantly reduce doses to which workers are exposed and will keep costs lower.

Stage 3 involves decontamination and dismantlement of the plant for the DECON, SAFSTOR, and ENTOMB 1 options. For ENTOMB 2, stage 3 involves dismantlement of all radioactively contaminated

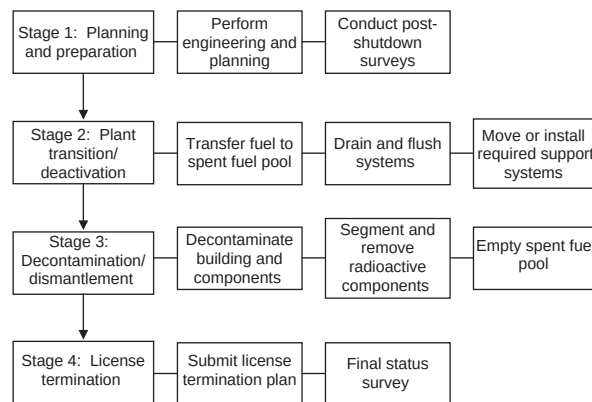


FIGURE 1 Reactor decommissioning process.

SSCs external to the reactor building and placement of these SSCs in the reactor building, followed by lowering the ceiling to the D-rings (pressurized water reactors) or the refueling floor (boiling water reactors). The greatest variability within a decommissioning stage is found in stage 3 because of the varieties of ways that systems can be dismantled. For instance, the secondary components have been dismantled at some plants in the United States, removing the moisture separators, diesel generators, and steam piping. Another plant has had its steam generators and pressurizer removed intact because transport and disposal facilities were easily available; the reactor vessel and internals were then removed whole (instead of having to be dismantled), filled with grout, welded closed, and shipped.

For the SAFSTOR option, stage 3 includes the time in storage. Decontamination and dismantlement activities can occur before, during, and after storage in SAFSTOR, with the time for stage 3 varying from a few years to more than 50 years. In both ENTOMB 1 and ENTOMB 2, grout and engineered barriers are installed. A monitoring program is also developed for the ENTOMB options. For ENTOMB 2, stage 3 involves dismantling all radioactively contaminated systems external to the reactor building and placing them in the reactor building. In pressurized water reactors, the ceiling is lowered to the D-rings; in boiling water reactors, the ceiling is lowered to the refueling floor.

Stage 4 is final administrative termination. Activities for this stage, which are similar for all options, include final site characterization, the final radiation survey, submission of a final termination plan, and a final site survey. The ENTOMB options would include both a partial site release and a detailed site monitoring program.

4. ENVIRONMENTAL IMPACTS

In order to adequately describe the environmental impacts of decommissioning, it is helpful to separate the impacts from those that occurred during site construction or operation and the immediate impacts that occur as a result of the decision to cease power operations. In general, environmental impacts occurring as a result of decommissioning a nuclear power facility are smaller than those impacts resulting from construction or operation of the facility. Immediate impacts as a result of the decision to cease power operation include the cessation of warm water flowing into the body of water that receives the

plant's thermal discharge, which may result in the loss of aquatic or marine organisms that have grown to be dependent on the additional heat from the facility. Potential environmental impacts during the decommissioning process include impacts on land and water use, water quality, air quality, aquatic ecology, and terrestrial ecology, in addition to radiological, socioeconomic, and aesthetic impacts.

4.1 Land Use

Land use during decommissioning is affected by the site layout. Most sites have sufficient area existing within the previously disturbed area (whether during construction or operation of the site) and, therefore, no additional land needs to be disturbed. However, some activities projected for decommissioning are expected to require land temporarily. Temporary changes may include expanded staging and laydown areas for removal of large facility components. In addition, the large number of temporary workers needed to accomplish the major decommissioning activities may require that temporary facilities be installed for on-site parking, training, site security access, office space, change areas, fabrication shops, and other needs. Widening and rebuilding access roads or creating or expanding gravel pits for building roads may occur off site, but plants that have been decommissioned thus far have not needed such additional land.

4.2 Water Use

Nuclear reactor facilities are usually located near or adjacent to significant water bodies (aquifers, rivers, lakes, etc.) that are important to the region. Operating nuclear reactor facilities use water from multiple sources. For example, water from an adjacent lake might provide cooling water, whereas potable water may come from groundwater wells located on site. Reactor cooling is the greatest use of water at an operating reactor; other uses include waste treatment, potable water, process water, and site maintenance. In general, the impact of nuclear reactor facilities on water resources dramatically decreases after plants cease operation. The operational demand for cooling and makeup water is largely eliminated after the facility permanently ceases operation.

Most of the impacts on water resources during decommissioning of a nuclear facility are typical of those during decommissioning or construction of any large industrial facility. For example, providing water

for dust abatement is a concern for any large construction project, as is potable water usage. However, the quantities of water required for decommissioning are trivial compared to those used during operations. Water use will be greater in facilities that are undergoing decontamination and dismantlement than in those that are in the storage phase.

4.3 Water Quality

The major decommissioning activities that could affect water quality include fuel removal, stabilization, decontamination and system dismantlement, and structure dismantlement. Surface waters are most likely to be impacted either by storm water or by releases of substances during decommissioning activities. However, unintentional releases of hazardous substances have been an infrequent occurrence at decommissioning facilities. Because the focus of decommissioning is the ultimate cleanup of the facility, considerable attention is placed on minimizing spills.

Certain decommissioning activities or options may result in changes in local water chemistry. For example, dismantling structures by demolition and disposing of the concrete rubble on site increase the possibility that hydration of the concrete could locally increase the alkalinity of the groundwater.

4.4 Air Quality

Decommissioning activities that have the potential to have a nonradiological impact on air quality include (1) worker transportation to and from the site, (2) dismantling of systems and removing of equipment, (3) movement and open storage of material onsite, (4) demolition of buildings and structures, (5) shipment of material and debris to off-site locations, and (6) operation of concrete batch plants. These activities typically take place over a period of years from the time the facility ceases operation until the decommissioning is complete. The magnitude and the timing of the potential impacts of each activity will vary from plant to plant, depending on the decommissioning option selected and the status of facilities and structures at the completion of decommissioning. The most likely impact of decommissioning on air quality is fugitive dust, although it should be less than during plant construction because the sizes of the disturbed areas are smaller, the duration of activity shorter, and paved roadways available. Best management practices, such as seeding or wetting, can be used to minimize fugitive dust.

4.5 Aquatic Ecology

Aquatic ecological resources may be impacted during the decommissioning process by either direct or indirect disturbance of plant or animal communities near the facility site. Direct impacts can result from activities such as removal of structures (e.g., the intake or discharge facilities) at the shoreline or in the water; the active dredging of a stream, river, or ocean bottom; or filling a stream or bay. Indirect impacts may result from effects such as runoff. If decommissioning does not include removal of shoreline or in-water structures, very little aquatic habitat will probably be disturbed. Thus, practically all aquatic habitat that was used during regular plant operations or, at a minimum, was not previously disturbed during construction of the site will not be impacted.

4.6 Terrestrial Ecology

Similar to aquatic resources, terrestrial resources may also be affected during the decommissioning process. Direct impacts can result from activities such as the clearing of native vegetation or filling of a wetland. Indirect impacts may result from effects such as erosional runoff, dust, or noise. During decommissioning, land at the site and the surrounding terrestrial ecology may be disturbed for the construction of laydown yards, stockpiles, and support facilities. Additionally, land away from the plant site may be disturbed by upgrading or installing new transportation or utility systems. In most cases, land disturbances will result in relatively short-term impacts and ecological systems will either recover naturally or will be landscaped appropriately for an alternative use after completion of decommissioning.

4.7 Radiological

Radioactive materials are present in the reactor and in the support facilities after operations cease and the fuel has been removed from the reactor core. Exposure to these radioactive materials during decommissioning may have consequences for workers. Members of the public may also potentially be exposed to radioactive materials that are released to the environment during the decommissioning process. To date, doses from occupational exposures to individual workers during decommissioning activities at nuclear power facilities in the United States are similar to, or lower than, the doses experienced by workers in operating facilities. In addition, these

doses have been below the regulatory limits. Off-site exposures, which could affect members of the public, during periods of major decommissioning have not differed substantially from those experienced during normal operations, which are also below the regulatory limits.

4.8 Socioeconomics

There are two primary pathways through which nuclear power plant activities create socioeconomic impacts on the area surrounding the facility. The first is through expenditures in the local community by the plant workforce and direct purchases of goods and services required for plant activities. The second pathway for socioeconomic impact is through the effects on local government tax revenues and services. When a nuclear power plant is closed and decommissioned, most of the important socioeconomic impacts will be associated with plant closure, rather than with the decommissioning process. The impacts occur either through changing employment levels and local demands for housing and infrastructure, or through decline of the local tax base and the ability of local government entities to provide public services. The effect on the local tax base and public services related to closure depends on the size of the plant-related tax base relative to the overall tax base of local government, as well as on the rate at which the tax base is lost. Experience has shown that publicly owned tax-exempt plants do not have an impact through this mechanism. In addition, fully depreciated plants or a plant that is located in an urban or urbanizing area with a large or rapidly growing tax base will also not have an impact by this mechanism. A large, newer, relatively undepreciated plant, located in a small, isolated community, is much more likely to have an impact. If the plant tax base is phased out slowly after closure, the impact is more likely to be mitigated.

4.9 Aesthetic Issues

The aesthetic impacts of decommissioning fall into two categories: (1) impacts, such as noise, associated with decommissioning activities that are temporary and cease when decommissioning is complete and (2) the changed appearance of the site when decommissioning is complete. During decommissioning, the impact of activities on aesthetic resources would be temporary. The impacts would be limited both in terms of land disturbance and the duration of activity and would have characteristics similar to those

encountered during industrial construction: dust and mud around the construction site, traffic and noise of trucks, and construction disarray on the site. In most cases, these impacts would not easily be visible off site. Aesthetic impacts could improve fairly rapidly in the case of an immediate DECON if the facility were dismantled, the structures removed, and the site regraded or revegetated. Impacts could also remain the same or would be similar in the case in which the structures are maintained during the decommissioning period and left standing, or throughout a long SAFSTOR or ENTOMB period. In the latter cases, the aesthetic impacts of the plant would be similar to those that occurred during the operational period.

5. SUMMARY OF DECOMMISSIONING STATUS OF U.S. NUCLEAR POWER FACILITIES

Twenty-two of the commercial nuclear reactors in the U.S. that are licensed by the U.S. Nuclear Regulatory Commission have permanently shut down and have either had their licenses terminated or are currently being decommissioned. Each facility has its own characteristics, depending on the decommissioning options being used for each facility and each facility's

TABLE I
Summary of Shutdown Plant Information

Parameter	Number
Reactor type	
Boiling water reactor	8
Pressurized water reactor	11
High-temperature, gas-cooled reactor	2
Fast breeder reactor	1
Decommissioning option	
SAFSTOR	14
DECON	7
Accident cleanup followed by storage	1
Fuel location	
Fuel on site in pool	13
No fuel on site ^a	8
Fuel on site in ISFSI ^b	1
Plan to move fuel to an ISFSI between 2000 and 2005	9

^aIncludes Three Mile Island, Unit 2, which has approximately 900 kg of fuel remaining on site due to the accident.

^bISFSI, Independent spent fuel storage installation.

decommissioning activities. An overview of the shutdown plants can be found in Table I, which includes 22 units shut down between 1963 and 1997. Table II summarizes important characteristics of the

shutdown plants. The thermal power capabilities of the reactors ranged from 23 to 3411 MWt. The reactors operated from just a few days (Shoreham) to 33 years (Big Rock Point).

TABLE II

Permanently Shutdown Plants in the United States (2003)

Nuclear plant	Reactor type ^a	Thermal power	Shutdown date ^b	Decommissioning option ^c	Location	Fuel status and license termination date
Plants currently in decommissioning process						
Big Rock Point	BWR	240 MW	08/30/97	DECON	Michigan	Fuel in ISFSI
Dresden, Unit 1	BWR	700 MW	10/31/78	SAFSTOR	Illinois	Fuel in ISFSI
Fermi, Unit 1	FBR	200 MW	09/22/72	SAFSTOR ^d	Michigan	No fuel on site
GE-VBWR	BWR	50 MW	12/09/63	SAFSTOR	California	No fuel on site
Haddam Neck	PWR	1825 MW	07/22/96	DECON	Connecticut	Fuel in pool
Humboldt Bay, Unit 3	BWR	200 MW	07/02/76	SAFSTOR ^d	California	Fuel in pool
Indian Point, Unit 1	PWR	615 MW	10/31/74	SAFSTOR	New York	Fuel in pool
La Crosse	BWR	165 MW	04/30/87	SAFSTOR	Wisconsin	Fuel in pool
Maine Yankee	PWR	2700 MW	12/06/96	DECON	Maine	Fuel in ISFSI ^e
Millstone, Unit 1	BWR	2011 MW	11/04/95	SAFSTOR	Connecticut	Fuel in pool
Peach Bottom, Unit 1	HTGR	115 MW	10/31/74	SAFSTOR	Pennsylvania	No fuel on site
Rancho Seco	PWR	2772 MW	06/07/89	SAFSTOR ^d	California	Fuel in ISFSI/partial DECON proposed in 1997
San Onofre, Unit 1	PWR	1347 MW	11/30/92	SAFSTOR ^d	California	Fuel in pool
Saxton	PWR	28 MW	05/01/72	SAFSTOR ^d	Pennsylvania	No fuel on site/ currently in DECON
Three Mile Island, Unit 2	PWR	2772 MW	03/28/79	Accident cleanup followed by storage	Pennsylvania	Approx 900 kg fuel on site/postdefueling monitored storage
Trojan	PWR	3411 MW	11/09/92	DECON	Oregon	Fuel in ISFSI
Yankee Rowe	PWR	600 MW	10/01/91	DECON	Massachusetts	Fuel in ISFSI ^e
Zion, Unit 1	PWR	3250 MW	02/21/97	SAFSTOR	Illinois	Fuel in pool
Zion, Unit 2	PWR	3250 MW	09/19/96	SAFSTOR	Illinois	Fuel in pool
Terminated licenses						
Fort St. Vrain	HTGR	842 MW	08/18/89	DECON	Colorado	Fuel in ISFSI/license terminated in 1997
Pathfinder	BWR	190 MW	09/16/67	SAFSTOR	South Dakota	No fuel on site/license terminated in 1992
Shoreham	BWR	2436 MW	06/28/89	DECON	New York	No fuel on site/license terminated in 1995

^a BWR, Boiling water reactor; FBR, fast breeder reactor; PWR, pressurized water reactor; HTGR, high-temperature, gas-cooled reactor.

^b The shutdown date corresponds to the date of the last criticality.

^c The option shown for each plant is the option that has been officially provided to the U.S. NRC. Plants in DECON may have had a short (1 to 4 years) SAFSTOR period. Likewise, plants in SAFSTOR may have performed some DECON activities or may have transitioned from the storage phase into the decontamination and dismantlement phase of SAFSTOR.

^d These plants have recently performed or are currently performing the decontamination and dismantlement phase of SAFSTOR.

^e Fuel to dry storage in an on-site independent spent fuel storage installation (ISFSI).

Three of the 22 plants (Fort St. Vrain, Shoreham, and Pathfinder) have completed decommissioning. Two of these three (Fort St. Vrain and Shoreham) used the DECON process for decommissioning. One facility, Shoreham, operated less than three full-power days before being shut down and decommissioned so there was relatively little contamination. One other facility, Pathfinder, was placed in SAFSTOR and subsequently decommissioned. Eleven of the plants shut down prematurely. Three Mile Island, Unit 2, ceased power operations as a result of a severe accident. Three Mile Island, Unit 2, has been placed in a monitored storage mode until Unit 1 permanently ceases operation, at which time both units are to be decommissioned.

Eleven of the permanently shutdown plants were part of the U.S. Atomic Energy Commission (AEC) Demonstrations Program, including Big Rock Point; Dresden, Unit 1; Fermi, Unit 1; GE-VBWR; Humboldt Bay, Unit 3; Indian Point, Unit 1; La Crosse; Pathfinder; Peach Bottom, Unit 1; Yankee Rowe; and Saxton. These plants were prototype designs that were jointly funded by the AEC and commercial utilities. The most recent of the Demonstration Program reactors to shut down was Big Rock Point, which operated for 33 years; it was permanently shut down in 1997.

Seven decommissioning facilities are located on multiunit sites in which the remaining units continue to operate, and one multiunit site shut down both units permanently. All eight of these licensees chose SAFSTOR as the decommissioning option. In most cases, SAFSTOR was chosen so that all units on a site could be decommissioned simultaneously. For various reasons, however, most shutdown units have done some decontamination and dismantlement. The reasons cited for choosing DECON have included the availability of low-level waste capacity, availability of staff familiar with the plant, available funding, the licensee's intent to use the land for other purposes, influence by state or local government to complete decommissioning, or a combination of other reasons.

A number of the plants have combined the DECON and SAFSTOR process either by entering shorter SAFSTOR periods or by doing an incremental DECON, allowing the plant to use resources and "decommission as they go." Sites have combined the options, usually to achieve economic advantages. For example, one site decided to shorten the SAFSTOR period and to begin incremental dismantlement out of concern over future availability of a waste site and future costs of disposal. One site that

prematurely shut down had a short SAFSTOR period to allow short-lived radioactive materials to decay and to conduct more detailed planning. Safety is another reason for combining the two options. Because of seismic safety concerns, one site undertook a major dismantling project to remove a 76-m (250-foot) concrete vent stack after it had been in SAFSTOR for 10 years.

The owner of the facility determines the physical condition of the site after the decommissioning process. Some intend to restore the site to "green field" status at the end of decommissioning, whereas others may install a nonnuclear facility. Some will leave structures standing at the time of license termination, and others will not. While undergoing the decommissioning process, some have opted for partial site release to decrease the size of the site area.

6. CONCLUSION

There are environmental impacts of decommissioning a nuclear power facility. However, of the different options and processes for decommissioning a nuclear power facility, all of the options used in the United States to date have resulted in environmental impacts that are smaller than those that occur during construction and operation of the facility.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fission Reactors: Boiling Water and Pressurized Water Reactors • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power, History of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

Further Reading

- Nordhaus, W. D. (1997). "The Swedish Nuclear Dilemma: Energy and the Environment." Resources for the Future Press, Washington, D.C.
- Slobodien, M. J. (ed.). (1999). "Decommissioning and Restoration of Nuclear Facilities: 1999 Health Physics Society Summer School Proceedings." Medical Physics Publ. Corp, Madison, Wisconsin.

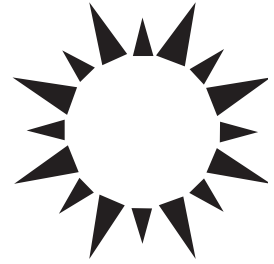
- U.S. Nuclear Regulatory Commission (NRC). (2002). "Generic Environmental Impact Statement on Decommissioning of Nuclear Facilities." NUREG-0586, Supplement 1. NRC, Washington, D.C.
- U.S. Nuclear Regulatory Commission (NRC). (1997). "Final Generic Environmental Impact Statement in Support of Rulemaking on Radiological Criteria for License Termination of NRC-Licensed Nuclear Facilities." NUREG-1496, Vol. 1. NRC, Washington, D.C.
- White, M. G., Norman, N. A., and Thomas, R. G. (eds.). (1995). "Decommissioning, Decontamination, and Environmental Restoration at Contaminated Nuclear Sites. American Nuclear Society, Winter Meeting 1994." American Nuclear Society, Washington, D.C.

Nuclear Power: Risk Analysis

B. JOHN GARRICK

Independent Consultant

Laguna Beach, California, United States



-
1. Why Risk Assessment?
 2. Nuclear Safety Theme
 3. Historical Development of Nuclear Plant Safety
 4. Nuclear Power Accident Experience
 5. Risk Assessment Methodology
 6. Important Applications and Benefits
 7. Risk Assessment and Regulatory Practice
 8. Future Direction of Risk Assessment

Glossary

- core** The nuclear fuel and fission reaction region in a nuclear reactor.
- core damage (melt)** An accident in which the nuclear fuel is damaged beyond recovery.
- defense-in-depth** A combination of multiple barriers and design basis accident analysis that reasonably assures public health and safety and protection of the environment.
- design basis accident** An accident forming part of the basis for the design of a nuclear power plant, for which there will be reasonable assurance that there will be no damage to the nuclear fuel for that accident.
- emergency core cooling system** A separate and independent cooling system for the reactor core in the event of a “loss-of-coolant accident.”
- event tree** A cause-and-effect representation of logic involving inductive reasoning.
- fault tree** An effect-and-cause representation of logic involving deductive reasoning.
- nuclear power plant** A commercial electric power-generating station that employs a nuclear reactor as the basic source of energy.
- nuclear reactor** An energy-generating system based on nuclear fission and a controlled self-supporting chain reaction.
- pinch points** The interfacing input and output states of the three models (plant, containment, site) that make up a full-scope risk assessment for a nuclear power plant.
- plant damage states** The output of the plant risk model representing the end states of accident scenarios that are a threat to the containment system.

quantitative risk assessment A process of systems analysis that calculates the likelihood and consequences of undesirable events and their uncertainties.

risk Answers the three questions about what can go wrong, how likely it is, and what the consequences are.

safety train A separate and independent set of engineered safety features for mitigating accidents.

scenario A sequence of events that describes the course of a nuclear plant accident.

scram The sudden shutdown of a nuclear reactor, usually by the rapid insertion of safety rods.

secondary containment An enclosure around a nuclear reactor to provide added protection from the release of radiation in the event of an accident that fails the primary containment system.

single-failure criteria A design criteria whereby the failure of a single system or piece of equipment will not result in any health and safety consequences.

trip An automatic or manual action that shuts down a system such as the reactor core or major piece of equipment, e.g., a turbine that provides the shaft power to the electric generators.

Nuclear power plant safety has been the principal driver for contemporary methods of quantitative risk assessment. Important contributions include a general definition of risk, methods for embracing and quantifying uncertainties, and the importance ranking of contributors to risk—an essential input for quantitative risk management. Most importantly, the adoption of risk assessment practices by the nuclear power industry coincides with an era of unprecedented safety in the performance of nuclear power plants.

1. WHY RISK ASSESSMENT?

The simple answer to “why risk assessment?” for nuclear power plants is that nations and the world have to make decisions about the best energy mix for

the future of planet Earth. Risk to people and the environment is a fundamental attribute of societal decision making. But there is more to it than just decision making. Early in the development of nuclear power, it became clear that large inventories of radiation required a level of safety analysis beyond standard practices. Nuclear reactors were being contemplated for use in generating electricity, and safety was a concern, especially in light of the stigma of the dangers of the fission process carried over from nuclear weapons development. The nuclear power industry was forced to seek new methods of safety analysis of nuclear power plants to overcome the “fear anything nuclear” syndrome that prevailed in the minds of some members of the public. The new methods needed to provide answers to three questions: what can go wrong with a nuclear power plant, how likely is it, and what are the consequences? The traditional methods of safety analysis, although somewhat effective in answering questions about what can go wrong and the consequences, profoundly failed to answer the question having to do with the likelihood of accidents. The “likelihood” question held the key for being able to quantify nuclear power plant risk. In short, for society to have access to nuclear energy systems that have the potential to end anxieties about energy resources, the industry was forced to come up with a more convincing safety case than was possible with past methods of analysis.

The nuclear electric power industry has been the leader in the development and widespread use of quantitative risk assessment (QRA). The U.S. nuclear electric power industry gave birth to the term “probabilistic risk assessment” (PRA); the international nuclear community sometimes uses the equivalent term “probabilistic safety assessment” (PSA). The concept that appears to be best received across different industries is that of quantitative risk assessment. In this discussion, quantitative risk assessment, probabilistic risk assessment, and just plain “risk assessment” are used interchangeably.

Risk assessment has survived and flourished in the U.S. nuclear power industry because it is an exceptional tool to make better decisions. QRA was able to satisfy the desire of nuclear plant owners to have a decision tool that quantitatively allows the evaluation of various options that have multiple input variables. The most important variables to the nuclear plant owners are cost, generation, and risk (public health, worker health, and economic). Although QRA started out as a tool to address the public health risk, it facilitated evaluating an entire

spectrum of variables. The industry’s recovery from the Three Mile Island Unit 2 accident in 1979 was greatly aided by the use of quantitative risk assessment because of the ability to better focus on the real safety issues. In fact, the industry has had an impeccable safety record since embracing contemporary methods of quantitative risk assessment, and safety is not the only benefit that has resulted from the widespread use of risk assessment in the nuclear power industry. Risk assessment provides the ability for plant personnel to balance cost, generation, and risk. Although there is no U.S. Nuclear Regulatory Commission (NRC) requirement for an existing nuclear power plant to maintain a risk assessment, the plants do so, following general industry guidelines. The NRC does require a prospective licensee to submit a PRA with the application for any proposed new nuclear electric power unit in the United States.

2. NUCLEAR SAFETY THEME

Today, nuclear power plant safety analysis employs the most advanced methods available for assessing the health and safety of the public. Many of the methods used for nuclear power plants have been adopted by high-technology industries such as those involved in space flight, defense systems, chemical plants, refineries, offshore platforms, and transportation systems. The probabilistic concepts currently spearhead the level of sophistication of the analyses, but there are basic tenets and themes that have guided the safety management of nuclear electric power plants from the beginning. The most fundamental of these basic tenets is the concept of multiple barriers.

Multiple barriers are a concept of providing enough barriers between radiation and the environment to provide assurance that the likelihood of simultaneous breach of all barriers is remote. Examples of barriers in a nuclear power plant are high-containment-capacity fuel with cladding, an isolated reactor coolant system, primary reactor building containment, secondary building containment, and exclusion distance. Other defense mechanisms include automatic control systems, single-failure criteria (no single failure threatens fuel integrity), and recovery capabilities from equipment malfunctions. QRA provides the ability to determine what risk levels are achieved by each barrier and at what cost. The value of each barrier is placed in the context of the overall risk. A principal factor in implementing multiple barriers is choosing design basis accidents to evaluate the multiple barriers. The principle behind

the design basis accident is the requirement that the plant design incorporate the capability to withstand specific hypothetical initiating events and failures without causing damage to the nuclear fuel. This process of multiple barriers and design basis accidents is referred to by the NRC as defense-in-depth.

The defense-in-depth concept has generally been implemented through the promulgation of very specific deterministic regulations. As the nuclear regulations come under review for risk-informing the regulatory process, it is not expected that the multiple-barriers tenet of regulatory practice will change in kind, but it will change in degree with much better knowledge of the real plant risk. In particular, with the availability of much more advanced methods for calculating the value of protective barriers, the ability exists to optimize the barriers in terms of risk and cost. Furthermore, the design basis concept can be assessed for its usefulness. This is considered important, as past application of defense-in-depth in the absence of a risk-informed approach has resulted in a definite increase in the complexity and cost of nuclear power plants without commensurate improvement to the overall public health risk. Quantifying defense-in-depth is one of the most significant benefits of QRA.

3. HISTORICAL DEVELOPMENT OF NUCLEAR PLANT SAFETY

Nuclear plant safety has two major fronts—the physical system and the analysis of the physical system. On the physical system front, improvements in safety design have included the advent of secondary containment systems (~1953), the inclusion of backup safety systems known as engineered safety features, especially with respect to emergency core cooling systems and electric power (~late 1950s and early 1960s), and the introduction of separate and independent safety trains (~1970s). In the 1980s and 1990s, nuclear power plants initiated programs for scram (sudden reactor shutdown) reduction based on a complete review and analysis of operating transients. As scrams were reduced, public health risk was reduced because there were fewer departures from normal steady-state operation. Also in the 1980s and 1990s, each nuclear power plant implemented the concept of “symptom-based procedures” for accident control and installed improved simulators for operator training.

On the analysis front, many events took place leading to a greatly improved understanding of the

safety of nuclear power plants. It was demonstrated that the consequences of accidents had little meaning without a better understanding of their likelihood. It became clear that it was not enough to do worst-case and maximum-credible accident analysis. Everyday transients followed by multiple failures of equipment and mistakes by operators were more likely than design basis accidents to result in reactor core damage.

The need for probabilistic analysis was recognized as early as the mid-1950s. However, detailed investigations of the probability of reactor accidents did not begin until about 1965. The first major reactor safety study to highlight the need for PRA of reactor accidents was the 1957 U.S. Atomic Energy Commission report WASH-740, “Theoretical Possibilities and Consequences of Major Accidents in Large Nuclear Power Plants.” Speculative estimates were made in WASH-740 that a major reactor accident could occur with a frequency of about one chance in a million during the life of a reactor. The report went on to observe that the complexity of the problem of establishing such a probability, in the absence of operating experience, made these estimates subjective and open to considerable error and criticism. Although it did not offer many specifics, this study did create interest in probabilistic approaches, and many studies were soon to follow. These included British and Canadian efforts, probabilistic analyses of military reactors, and several studies sponsored by the U.S. Atomic Energy Commission. At about the same time, Garrick wrote a Ph.D. thesis on unified systems safety analysis of nuclear power plants based on a total systems and probabilistic approach.

The breakthrough in the probabilistic risk assessment of nuclear power plants came in 1975 with the publication of the Reactor Safety Study by the U.S. Nuclear Regulatory Commission under the direction of Professor N. C. Rasmussen of the Massachusetts Institute of Technology. This project marked a turning point in the way people analyzed the safety of complex facilities and systems. The Reactor Safety Study was followed by several major industry studies, such as the risk assessments performed on the Zion and Indian Point nuclear power plants; the new methods that were introduced in these assessments have become standards of many QRA applications. These studies provided a major breakthrough in calibrating the worth of safety features and safeguards, and therefore, the safety margins of designs.

By the 1980s, the question was no longer “why,” but how soon a QRA could be developed for every nuclear power plant in the United States. That goal

has essentially been reached. The benefits of QRA for U.S. nuclear power plants have been demonstrated in terms of a reduction in frequency of core damage events (one reactor core lost in approximately the first 450 reactor years of experience versus zero reactor cores lost in over 2000 actual reactor years of experience since the Three Mile Island accident) and improved generation with a reduction in the cost of electricity. The most important benefit is nuclear power plants with reduced public health risk. QRA has been effective not only in calibrating the risk of nuclear power, but has provided better knowledge of the worth of safety systems and allowed the allocation of safety engineering resources to the most important contributors. Effective risk management of nuclear electric power plants in the United States has become a reality, not just a goal.

4. NUCLEAR POWER ACCIDENT EXPERIENCE

There have only been two accidents worldwide that have resulted in severe core damage of a nuclear power plant designed to generate electricity. The accidents involved the Three Mile Island Unit 2 plant near Harrisburg, Pennsylvania, in the United States, and the Chernobyl Nuclear Power Station in the Ukraine of the former Soviet Union. Both accidents permanently damaged the nuclear reactors involved, but only the Chernobyl accident resulted in known fatalities and injuries. The on-site consequences of the Chernobyl accident were very serious; an estimated 30 people are believed to have died from acute doses of radiation and some 300 people required hospital treatment for radiation and burn injuries. No off-site fatalities or injuries have yet been attributed to the Chernobyl accident, although the latent effects are yet to be quantified. It is important to put these two very serious accidents in context with the safety experience of the nuclear power industry. There are approximately 440 nuclear power plants in the world. Nuclear energy is just over 5% of the world primary energy production and about 17% of its electrical production. In the United States, there are 103 nuclear power plants operating, providing approximately 20% of the nation's electricity. The worldwide experience base is approaching 10,000 in-service reactor-years, of which about 3000 reactor-years is U.S. experience. The experience base is likely beyond 10,000 reactor-years if all types of reactors are included, such as research, test, weapons, and

propulsion reactors. Some 70% of the nuclear power plant experience worldwide involves light water reactors, for which only one accident has occurred, Three Mile Island. This safety record is most impressive. The challenge is to keep it that way.

4.1 The Three Mile Island Unit 2 Accident

The Three Mile Island Unit 2 (TMI-2) nuclear power plant, located near Harrisburg, Pennsylvania, went into commercial operation in December 1978. The plant was designed to generate approximately 800 MW of electricity and used a pressurized water reactor supplied by the Babcock and Wilcox Company. The accident occurred on March 28, 1979. Routine mechanical malfunctions with the plant resulted in an automatic shutdown ("feedwater trip") of the main feedwater pumps, followed by a trip of the steam turbine and the dumping of steam to the condenser. The loss of heat removal from the primary system resulted in a rise of reactor system pressure and the opening of the power-operated relief valve. This action did not provide sufficient immediate pressure relief, and the control rods were automatically driven into the core to stop the fission process.

These events would have been manageable had it not been for some later problems, such as with the emergency feedwater systems. Perhaps the turning point of the accident was that the opened pressure relief valve failed to close and the operators failed to recognize it. The result was the initiation of the well-studied small loss-of-coolant accident, known as the small LOCA. The improperly open valve, together with some other valve closures that had not been corrected from previous maintenance activities, created a shortage of places to put the heat loads of the plant. The response of the plant was the initiation of high-pressure emergency cooling. High pump vibration and concern for pump seal failure resulted in the operators eventually shutting down all of the main reactor coolant pumps. It was during the time that the coolant pumps were off, for 1 to 3 hours, that the severe damage to the core took place. At about 2 hours and 20 minutes into the accident, the backup valve (known as a block valve) to the stuck-open relief valve was closed. This action terminated the small LOCA effect of the stuck-open relief valve. Although the accident was then under some level of control, it was almost 1 month before complete control was established over the reactor fuel temperature, when adequate cooling was provided by

natural circulation. The consequences of the accident were minimal in terms of the threat to public health and safety, but the damage to the reactor was too severe to recover the plant.

4.2 Chernobyl Nuclear Power Station Accident

The Chernobyl nuclear power plant in the former Soviet Union involved a 1000-MW (electrical) boiling water, graphite-moderated, direct-cycle reactor. The Chernobyl accident occurred on April 26, 1986, and was initiated during a test of reactor coolant pump operability from the reactor's own turbine generators. The purpose of the test was to determine how long the reactor coolant pumps could be operated, using electric power from the reactor's own turbine generator under the condition of turbine coast down and no steam supply from the reactor. However, the experimenters wanted a continuous steam supply, so they decided to conduct the experiment with the reactor running—a serious mistake. The test resulted in a coolant flow reduction in the core and extensive boiling. Because of the inherent properties of this reactor design, the chain reaction increases on boiling, rather than decreases as in U.S. plants, and a nuclear transient occurred that could not be counteracted by the control system. The result was a power excursion that caused the fuel to overheat, melt, and disintegrate. Fuel fragments were ejected into the coolant, causing steam explosions and rupturing fuel channels with such force that the cover of the reactor was blown off. This accident resulted in approximately 30 fatalities from acute doses of radiation and the treatment of some 300 people for radiation burn injuries. The off-site consequences are still under investigation. Latent effects are expected, but they have not been quantified.

In summary, nuclear power suffered a severe setback from both of these accidents, although public support for nuclear power was already beginning to decline. Nuclear plants under construction were canceled and no new U.S. nuclear plants have been ordered since 1979. The fact that the TMI-2 accident did not result in any radiation injuries or fatalities and the Chernobyl reactor type is no longer in the mix of viable power reactors has not removed the fear that some segments of the public have of nuclear power. However, the superior performance and safety record in the United States since these two accidents has allowed the NRC to approve power upgrades and license extensions for several U.S. nuclear power plants.

5. RISK ASSESSMENT METHODOLOGY

Quantitative risk assessments of nuclear power plants are generally based on the following principles:

- The quantitative expression of risk should be in the form of a structured set of scenarios, each having a corresponding likelihood and consequence.
- Each risk scenario should take the form of a sequence of events, starting with an event that upsets an otherwise successful operation or system, and proceeding through a series of subsequent events to the end state that terminates the scenario (i.e., the consequences of the scenario).
- The set of scenarios must be complete in the sense that all of the important contributors to risk are included.
- The end states of the scenarios should reflect initial, cascading, and collateral consequences or levels of damage where appropriate.
- The scenarios must be quantified in terms of clearly defined risk measures, be realistic, incorporate uncertainties, and be based on the supporting evidence.
- The results should rank the contributors to risk in order of importance and must be presented in a way that supports decision making.

The overarching principle on which risk assessment methodology for nuclear power plants is founded is that when we ask the question “what is the risk?” we are really asking the following three questions: (1) what can go wrong, (2) how likely is it, and (3) what are the consequences? In the notation of the practitioners of risk assessment, this “triplet” definition of risk by Kaplan and Garrick is represented as follows:

$$R = \{(S_i, L_i, X_i)\}_c,$$

where S_i denotes risk scenario i , L_i denotes the likelihood of that scenario, and X_i denotes the consequences or damage level of that scenario. The angle brackets enclose the triplets, the curly brackets mean “a set of,” and the subscript c denotes complete, meaning that all of the important scenarios are included in the set.

An overview of the basic structure of a nuclear power plant quantitative risk assessment is shown in Fig. 1. There are basically three models within a full-scope quantitative risk assessment—the plant model, the containment model, and the site model. The plant model begins with the consideration of different

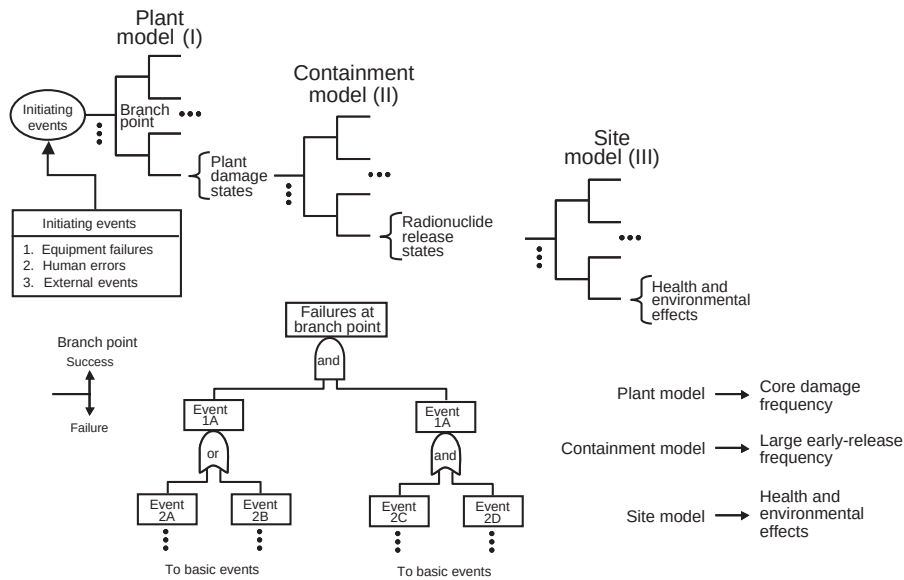


FIGURE 1 Event tree/fault tree structure of a nuclear plant quantitative risk assessment.

initiating event categories and has as output different plant damage states. The plant damage states are input to the containment model, and its outputs are different release states that become the input to the site model. The output of the site model is the calculated health and environmental risks. The interfaces between the models are “pinch-points” that allow the models to be developed independently, which greatly facilitates the organization and transparency of the analysis. Figure 1 also illustrates what is described in the next section as the “event tree,” “fault tree” format for structuring scenarios. A consulting firm, Pickard, Lowe and Garrick, Inc., developed the modular event tree structure for quantitative risk assessment, a variation of the Reactor Safety Study methodology. The U.S. Nuclear Regulatory Commission and the American Nuclear Society have adopted the same three-model structure in their procedures guide on quantitative risk assessment. They identified the plant, containment, and site models cumulatively as Levels I, II, and III, respectively.

5.1 Structuring the Scenarios (S_i)

Scenario structuring encompasses the methods, algorithms, and insights needed to identify and portray the risk scenarios (S_i). Two common methods (Approach 1 and Approach 2) are used for scenario development and are sometimes referred to as the bottom-up and top-down approaches. They have the following characteristics:

1. Given a set of initiating events, the structuring of scenarios is done so the end state of each scenario is the condition that terminates the scenario. That is, the scenario determines the end state.
2. Given an end state, project backward to determine the potential scenarios that could occur to arrive at that end state.

The most common logic diagrams for representing the two methods are event trees and fault trees. An event tree starts with an initiating event and proceeds to identify the succeeding events, including branches, that eventually terminate into possible undesirable consequences. An event tree, therefore, is a cause-and-effect representation of logic. Event trees are the logic diagrams of preference for Approach 1. A fault tree starts with the end state or undesired consequence of interest and attempts to determine all of the contributing system states. Therefore, fault trees are effect-and-cause representations of logic and are the logic diagram preference for Approach 2. That is, an event tree is developed by inductive reasoning whereas a fault tree is developed by deductive reasoning. A key difference in the two representations is that a fault tree is only in “failure space” and the event tree includes both “failure and success space.” The choice between the two is a matter of circumstances and preference. Often the two are used in combination such that the event tree provides the basic scenario space of events and branch points, and the fault trees are used to quantify

the “split fractions” at the branch points as illustrated in Fig. 1.

5.2 Defining Consequences (X_i)

Simply put, consequences are the end states of the scenarios. There are two perspectives of consequences and both are used in analyzing the risk of nuclear power plants. One perspective is to take each risk scenario to whatever point that is a reasonable termination of the scenario. The scenarios are then assembled by different consequences, or damage states. The other perspective is to define the damage state of interest in advance, such as “core melt,” “radionuclide releases,” “radiation dose to humans,” “injuries,” “fatalities,” and “property damage,” and consider only the scenarios that have as their end state the undesired damage state(s). Although examples of both approaches exist, current practice is to focus on “core melt” as the primary basis for measuring risk because core melt is a precursor to large radiation releases. One other damage state has been defined to serve as a surrogate for consequences beyond core melt, and that is “a large early-release frequency” of radiation. The choice is dependent on the requirements provided to the analyst as to how to structure the scenarios and end states.

5.3 Quantifying the Scenarios (L_i)

To quantify the likelihood, L_i , of different accident scenarios, it is first necessary to define the concept of likelihood. Most often the methodology adopts the “probability-of-frequency” principle to define likelihood. The frequency parameter is presented as a probability distribution to communicate frequency uncertainty. The actual quantification of the risk scenarios is done with the aid of the event tree (see Fig. 2). The event tree branch points are determined by actions, activities, and equipment (top events) that can alter or truncate the path of a scenario or sequence of events. An example of an action that

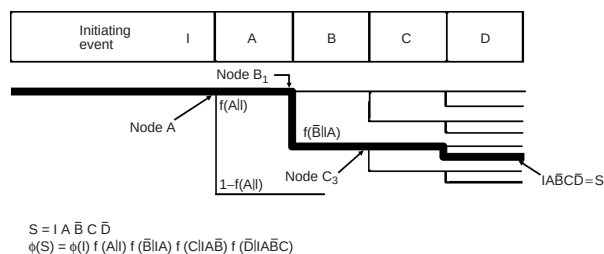


FIGURE 2 Quantification of a scenario using an event tree.

could alter the path of a scenario is a decision by a reactor operator to shut down a cooling system. An example of an activity might be the activation of a mitigating system, such as an emergency coolant system, and an example of an equipment event might be the failure of a source of electric power. Top events are placed in the boxes across the top of the diagram (Fig. 2) and are denoted A, B, C, and D. The event tree is a powerful tool because it makes visible all of the actions, equipment, processes, events, and features that affect an event. The diagram shown in Fig. 2 has only two outcomes emerging from a branch point (e.g., success or failure). However, an event tree can have multiple outcomes from a branch point to account for different degrees of degradation of a system. An individual scenario is a single path through the tree as illustrated by the highlighted lines.

Each scenario or path through the event tree can be described by an algebraic expression (shown in Fig. 2). Using input data for the initiating event and evidence-based split fractions at the branch points, the algebraic expression can be converted to an equation for calculating the frequency of individual scenarios. The remaining step is to embed the frequencies into appropriate probability distributions to communicate their uncertainties. There are various techniques for carrying out this operation, but the one often preferred is based on Bayes' theorem. Bayes' theorem is the fundamental, logical principle governing the process of inferential reasoning. It answers the question: “How does the probability of a given hypothesis change when we obtain a new piece of evidence?”

Once the scenarios have been quantified, the results take the form of the graph in Fig. 3. Each scenario has a probability-of-frequency curve quantifying its likelihood of occurrence. Figure 3 shows the curve for a single scenario or a set of scenarios

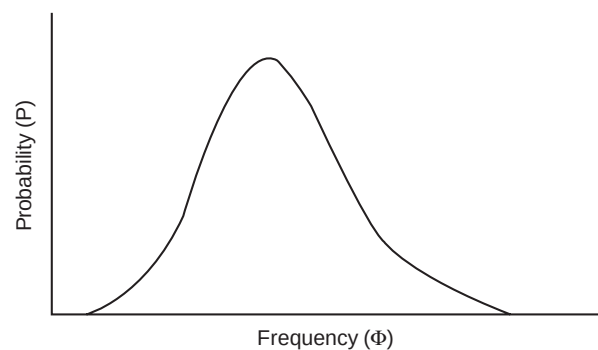


FIGURE 3 Probability-of-frequency curve for a specific consequence.

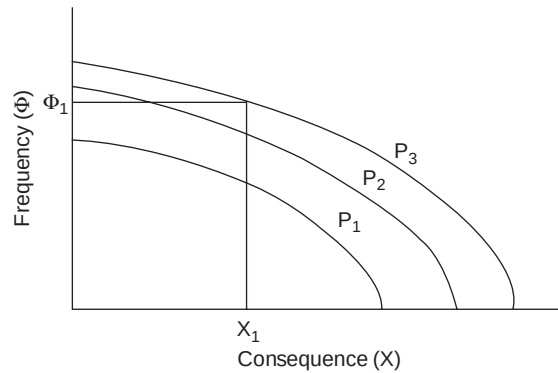


FIGURE 4 Risk curve for varying consequences, where the consequence is a variable.

leading to a single consequence. Showing different levels of damage, such as the risk of varying injuries or fatalities, requires a different type of presentation. The most common form is the classical “risk curve,” also known as the “frequency-of-exceedance” curve, or the even more esoteric label, the “complementary-cumulative-distribution function.” This curve is constructed by ordering the scenarios by “increasing levels of damage” and cumulating the probabilities from the bottom up in the ordered set against the different damage levels. Plotting the results on log-log paper generates curves, as shown in Fig. 4.

Although risk assessment results such as those illustrated in Figs. 3 and 4 can be beneficial in providing a perspective on the actual risks and in prioritizing contributors to risk, they are not the most important output of the risk assessment. The most important output is the revelation of the dominant contributors to the risk, which are necessary for effective risk management. The contributors are buried in the results assembled to generate such curves as those in Figs. 3 and 4. Numerous techniques can be used to extract and rank contributors. Most advanced risk assessment software packages contain algorithms for ranking the importance of contributors to a risk measure.

6. IMPORTANT APPLICATIONS AND BENEFITS

Since the Reactor Safety Study in 1975, major government and industry studies have provided the signature of what is meant by probabilistic risk assessment as practiced in the nuclear power field. These studies and applications include the joint effort of the U.S. Nuclear Regulatory Commission and the American Nuclear Society in the development of the

industry’s PRA procedures guide, the collaborative industry effort on the Oconee nuclear plant risk assessment, and the NRC’s evaluation of five light water reactor designs culminating in the 1990 report, NUREG-1150. But the studies most responsible for specializing the Reactor Safety Study methodology to plant-specific risk assessments were the industry-sponsored Zion and Indian Point studies. The two efforts, the Reactor Safety Study and the Zion and Indian Point studies, are now discussed further.

It was the U.S. Atomic Energy Commission that undertook the Reactor Safety Study under the direction of Professor Norman C. Rasmussen of the Massachusetts Institute of Technology. The study took 3 years to complete and was a turning point in the way to think about the safety of nuclear power plants, or, for that matter, the safety of any natural or engineered system. It should be noted that the Reactor Safety Study was initiated before the Energy Reorganization Act of 1974 abolished the U.S. Atomic Energy Commission. This Act transferred to the Nuclear Regulatory Commission all the licensing and related regulatory functions assigned to the Atomic Energy Commission by the Atomic Energy Act of 1954.

The Reactor Safety Study, using the Surry nuclear power plant (pressurized water reactor) and the Peach Bottom nuclear power plant (boiling water reactor) as reference designs, calculated the risk from the operation of 100 current-design light water reactors located in the United States. The methodology was founded on the principle of a structured set of accident sequences, or scenarios. The sequences were developed using fault tree and event tree logic diagrams, quantified on the bases of the supporting evidence, and then assembled into different consequences such as core damage frequencies, radiation release fractions, and off-site radiation effects to people and property. The principal findings of the study were that the risk associated with the operation of selected nuclear power plants is extremely small and that the dominant contributor to risk is not the large loss-of-coolant accident, previously emphasized as the design basis accident. Transients and small loss-of-coolant accidents often are the major contributors to risk. The Reactor Safety Study also highlighted the important role of the reactor operators in maintaining the safety of nuclear power plants.

The public and scientific community had mixed reactions to the Reactor Safety Study. Their primary criticism of the study was that the “uncertainty analysis” was weak and the report lacked “scrutability.” The initial reaction of the NRC to the

criticism was to withdraw their endorsement of the study. In spite of the criticisms of the study, there was strong support for the use of the methodology, especially from the plant owners. This favorable response, together with the fact that the Three Mile Island accident was among the event sequences addressed in the study, led the NRC to change their position and once again support the study.

The nuclear power industry was responsible for carrying the lessons learned from the Reactor Safety Study to the practical level of better understanding the safety of individual nuclear plants. Several studies followed on the heels of the Reactor Safety Study and the most comprehensive of these were the risk assessments performed on the Zion nuclear plant near Chicago and the Indian Point 2 and 3 plants near New York City. Under legal challenges at high-population-density sites, industry chose probabilistic risk assessment as a way to develop the necessary evidence of the safety of their plants. The Zion and Indian Point plants became, more or less, the test cases. The Zion and Indian Point assessments were full-scope studies, meaning that they analyzed the plant, its containment, and the off-site consequences. Challenged in the courts, the owners and operators of the plants defended the safety of their plants using the risk assessments as their primary evidence. They were successful. The conclusion was reached that the nuclear power plants at the high-population-density sites presented public health risk profiles similar to those at less populated sites. As a result, the nuclear power plants at the high-population-density sites were not shut down—a major achievement at the time.

The Zion and Indian Point studies contained many firsts, including the “triplet definition” of risk; they were the first comprehensive studies of core melt phenomena and containment response in a probabilistic format and were the first to employ a modularized event tree “pinch-point” format to represent beginning-to-end accident sequences, the first to perform uncertainty analysis at the component and basic event levels and to propagate the uncertainties through the scenarios, the first to explicitly include external events (such as earthquakes and fires) in the basic risk model, and the first to employ an atmospheric dispersion model for dose calculations that allowed for changes in plume direction. Other analytical concepts introduced by the Zion/Indian Point studies included the “probability-of-frequency” format for measuring risk and the “master logic diagram” method for determining initiating events. Many of these methods have become standards for contemporary risk assessments.

7. RISK ASSESSMENT AND REGULATORY PRACTICE

As a result of the Energy Reorganization Act of 1974, the U.S. Nuclear Regulatory Commission is responsible for licensing and regulating nuclear facilities, including nuclear power plants, and materials, and for conducting research in support of the licensing and regulatory process. The NRC’s primary regulatory responsibility is to provide reasonable assurance of adequate protection of public health and safety and protection of the environment from operations and accidents involving nuclear facilities and materials. The legacy of the NRC in the development and use of risk assessment is unique for regulatory bodies. The vision of the NRC of the need for probabilistic methods was reflected in their decision to sponsor the Reactor Safety Study when such an evaluation was not required by the regulations.

During the time when the Reactor Safety Study was being reviewed and the NRC had temporarily rejected it due to early criticisms, the nuclear power industry was taking the initiative to use the methodology for better assessing the risk of its plants. The impressive results coming out of the industry studies together with other events favorable toward the Reactor Safety Study caused the NRC to again embrace the technology and to seek ways of effectively applying it. Some of the actions taken were the publishing of frequency-based safety goals (1986), requiring limited-scope individual plant examinations based on probabilistic methods to determine if any nuclear power plant was an “outlier” with respect to public health risk (1988), and, in about this same time frame, issuing two new rules having to do with the treatment of loss of all electrical power and requirements to reduce the risk of transients. Perhaps the most significant action taken by the NRC toward embracing the concept of quantitative risk assessment was the 1995 publishing of a policy statement on the use of probabilistic risk assessment methods in nuclear regulatory activity. Quoting from the PRA policy statement, “The use of PRA technology should be increased in all regulatory matters to the extent supported by the state of the art in PRA methods and data, and in a manner that complements the NRC’s deterministic approach and supports the NRC’s traditional defense-in-depth philosophy.”

Even with what appeared to be an aggressive move on the part of the NRC toward encouraging the use of probabilistic risk assessment, there were only small changes in the regulations with respect to

the licensing process. The NRC recognized this problem and beginning in 1997 sought ways to make changes to the regulations to begin a more formal transition into a “risk-informed” approach to regulation. Several initiatives were put in place to stimulate risk-informing the regulatory process. Two examples are new rules having to do with the treatment of loss of all electrical power and requirements to reduce the risk of transients. Meanwhile, it was important for the NRC to make it clear what they mean by “risk-informed” regulation. The best answer to that came from a white paper prepared by the Commission in 1998 on “Risk-Informed, Performance-Based Regulation.” Quoting from the white paper, “A risk-informed approach to regulatory decision-making represents a philosophy whereby risk insights are considered together with other factors to establish requirements that better focus licensee and regulatory attention on design and operational issues commensurate with their importance to health and safety.” The current position of the NRC is a “risk-informed” approach to regulation. It is made very clear in the white paper “that the Commission does not endorse an approach that is risk-based,” if what is meant by risk based is that safety decisions are solely based on the numerical results of a risk assessment.

What all this means is that the licensing process is in a transitional phase, from deterministic and prescriptive regulations, to regulations that are less prescriptive and increasingly risk oriented, but not risk based. In the meantime, license applications and amendments must be accompanied with analyses that provide risk insights, but are in compliance with the deterministic requirements still in place.

8. FUTURE DIRECTION OF RISK ASSESSMENT

The future direction of quantitative risk assessment in the risk management of nuclear power plants is dependent on several factors: advancements in understanding accident phenomena, upgrading of risk assessment methodologies, regulatory activities, and security requirements.

8.1 Advancements in Understanding Accident Phenomena

The current practice for nuclear power plant risk assessments is to make two basic calculations, core

damage frequency and large-early-release frequency. In the future, calculations will need to include public health effects (immediate fatalities and latent cancer fatalities). These calculations of health effects will need to make use of the recent work with respect to realistic source terms for radiation releases. There is a lack of a dose-response model that reflects actual health effects for all levels and rates of radiation. Clearly, more accurate models of the real health effects are needed.

There is a need for realistic thermal hydraulic calculations that can form the basis of improved success criteria for the QRA models. In many cases, these success criteria are based on design basis accident calculations that do not effectively represent the performance capability of plant equipment. Design basis accident calculations also do not allow for the proper quantification of operator actions to activate equipment or to recover failed equipment. A realistic evaluation is needed of the thermal hydraulic interaction of the reactor coolant system and other fluid systems during accidents such as steam generator tube ruptures and intersystem loss of coolant accidents.

8.2 Upgrading of Risk Assessment Methodologies

QRA has greatly advanced since the breakthrough effort of the Reactor Safety Study. Improvements in the models include uncertainty analysis; the treatment of human reliability; consideration of external threats such as earthquakes, severe storms, and fires; importance ranking of contributors; and the fine-tuning of the models to better represent plant-specific details. There are still many areas for improvement. One very attractive direction would be for the risk assessments to be cast into different forms for use by different groups. The specialization could be not only for the risk assessment teams, but also for risk managers, those responsible for accident management and emergency response, and the public (risk communication). Each group has a different need for the information coming out of a risk assessment, and specializing the information by need could have a major impact on the acceptance and use of the results.

8.3 Regulatory Activities

The responsibility for the safety of the public in the United States with respect to radiation from nuclear reactors lies with the owners of the nuclear electric power plants. The owners of the nuclear units have

used their existing quantitative risk assessments to institute effective and efficient risk management programs at each nuclear power plant in the United States. Each nuclear power plant has a risk assessment that is maintained to industry guidelines. Engineers, maintenance persons, and operators at each nuclear power unit are aware of the results of the risk assessment and use them to actively manage risk. The superior safety record of nuclear electric power in the United States coincides with this new approach to risk management.

The NRC has a legal responsibility to “provide reasonable assurance of adequate protection of public health and safety.” The QRAs have demonstrated that most of the deterministic-based existing regulations do not efficiently address the dominant contributors to public health risk. Some of them do not address the dominant contributors at all. The regulations must be changed such that the regulations aid both the regulator and the licensee to manage public health risk in an effective and efficient manner. The NRC is slowly moving toward regulations that address public health risk using the insights gained from the licensee QRAs. This is evidenced in the recent changes to the rules governing the monitoring of maintenance and the implementation of the Reactor Oversight Process. But the pace of change to “risk inform the regulations” has been slow. The best course is for industry to play an active role in the development of future regulations that are “risk informed” and for the NRC to allocate the necessary resources to quickly change the regulations to be effective and efficient.

8.4 Security Requirements

An important new area for receiving benefits from risk assessment has to do with combating terrorism. Nuclear plants are often mentioned as a possible target for a terrorist attack and it is important that such a threat is linked to the vulnerability of the plants. Of course, there has already been a lot of work done in this area for nuclear power plants and events such as aircraft impact have always been a consideration in the safety assessment of nuclear plants. It is just that the threat now seems more real than ever before, and the question is what risk assessment can do to help protect plants from such threats. Although a great deal is known about the vulnerability of nuclear plants, there has not yet been a systematic process of connecting such vulnerabilities to specific threats. This will require cooperation between the experts on the threat of terrorism, the

intelligence community, and experts on risk assessment. Fortunately, some progress is being made, but it, too, is slower than it should be.

SEE ALSO THE FOLLOWING ARTICLES

Ecological Risk Assessment Applied to Energy Development • Nuclear Engineering • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power Economics • Nuclear Power Plants, Decommissioning of • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal

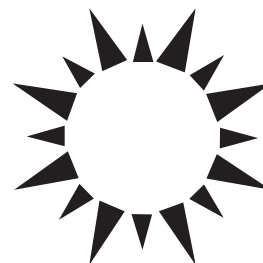
Further Reading

- Garrick, B. J. (1968). “Unified Systems Safety Analysis for Nuclear Power Plants.” Ph.D. Thesis, University of California, Los Angeles.
- Kaplan, S., and Garrick, B. J. (1981). On the quantitative definition of risk. *Risk Anal.* 1(1), 11–27.
- Pickard, Lowe and Garrick, Inc., Westinghouse Electric Corporation, and Fauske and Associates, Inc. (1981). “Zion Probabilistic Safety Study.” Prepared for Commonwealth Edison Company, Chicago, Illinois.
- Pickard, Lowe and Garrick, Inc., Westinghouse Electric Corporation, and Fauske and Associates, Inc. (1982). “Indian Point Probabilistic Safety Study.” Prepared for Consolidated Edison Company of New York, Inc., and the New York Power Authority, New York.
- U.S. Atomic Energy Commission (AEC). (1957). “Theoretical Possibilities and Consequences of Major Accidents in Large Nuclear Power Plants.” Report WASH-740 (March, 1957). AEC, Washington, D.C.
- U.S. Nuclear Regulatory Commission (NRC). (1975). “Reactor Safety Study: An Assessment of Accident Risks in U.S. Commercial Nuclear Power Plants.” Report WASH-1400 (NUREG-75/014). NRC, Washington, D.C.
- U.S. Nuclear Regulatory Commission (NRC). (1978). “Risk Assessment Review Group Report to the U.S. Nuclear Regulatory Commission” (H. W. Lewis, chairman). Report NUREG/CR-0400 (September, 1975). NRC, Washington, D.C.
- U.S. Nuclear Regulatory Commission (NRC). (1983). “PRA Procedures Guide—A Guide to the Performance of PRAs for Nuclear Power Plants.” Report NUREG/CR-2300 (January 1983). NRC, Washington, D.C.
- U.S. Nuclear Regulatory Commission (NRC). (1990). “Severe Accident Risks: An Assessment for Five U.S. Nuclear Power Plants.” Report NUREG-1150. NRC, Washington, D.C.
- U.S. Nuclear Regulatory Commission (NRC). (1995). “The Probabilistic Risk Assessment (PRA) Policy Statement. 60 FR 42622 (August 16, 1995).” NRC, Washington, D.C.
- U.S. Nuclear Regulatory Commission (NRC). (1999). “White Paper on Risk-Informed and Performance-Based Regulation. SECY-98-144 (February 24, 1999).” NRC, Washington, D.C.

Nuclear Proliferation and Diversion

H. A. FEIVESON

Princeton University
Princeton, New Jersey, United States



-
1. Introduction: There Is No Such Thing as a Peaceful Atom
 2. Nuclear Weapons
 3. Fissile Material Production
 4. Routes to Fissile Materials for Weapons
 5. Nuclear Reactors and Fuel Cycles
 6. Technical Barriers to Diversion
 7. Institutional Barriers to Proliferation: Safeguards
 8. Proliferation Resistance: An Initial Assessment
 9. The Long-Term Challenge for Nuclear Power: Enhanced Proliferation Resistance
 10. Conclusion

Glossary

fissile material Material that can sustain an explosive chain reaction, notably plutonium of almost any isotopic composition and highly enriched uranium.

highly enriched uranium Uranium containing greater than 20% uranium-235.

nuclear proliferation The spread of nuclear weapons to states that do not now possess them or to subnational groups.

once-through fuel cycles A civilian nuclear fuel cycle in which the spent fuel from the reactor is not reprocessed and the contained plutonium separated.

proliferation resistance The array of measures that weaken the links between the civilian nuclear fuel cycle and the capabilities of a nation or subnational group to acquire nuclear weapons.

plutonium recycling The use of plutonium recovered from spent fuel in fresh fuel for reactors.

reprocessing The chemical separation of spent fuel into constituent parts, typically the contained plutonium, fission products, and uranium.

spent fuel The fuel discharged from a nuclear reactor.

This article describes the current system of safeguards on peaceful nuclear activities to prevent

proliferation, gauges how well this system has succeeded and is likely to do so in the future, and examines various reactor and fuel cycle concepts that have been advanced to make nuclear power more “proliferation resistant” in the future.

1. INTRODUCTION: THERE IS NO SUCH THING AS A PEACEFUL ATOM

As early as 1904, in a talk to the Corps of Royal Engineers in England, the young scientist Frederick Soddy stated the fundamental dilemma of nuclear power in a now much-quoted passage:

It is probable that all heavy matter possesses—latent and bound up with the structure of the atom—a similar quantity of energy to that possessed by radium. If it could be tapped and controlled what an agent it would be in shaping the world's destiny! The man who put his hand on the lever by which a parsimonious nature regulates so jealously the output of this store of energy would possess a weapon by which he could destroy the earth if he chose.

This dilemma—that the atom can be exploited for almost boundless energy or for almost boundless destruction—was given a more practical policy-directed twist by Enrico Fermi in 1944, soon after the Manhattan Project laid the foundations for both nuclear reactors and nuclear weapons:

“It is not clear that the public will accept an energy source that produces this much radioactivity and that can be subject to diversion of material for bombs.”

It is now well understood that the fission of 3 kg of U-235 or plutonium could provide 1 day's worth of electricity for a moderate-size city or, alternatively, could generate an explosive yield of 50,000 tons of

TNT, three times the destructive power of the bomb that destroyed Hiroshima. The question regarding nuclear power is whether it can continue to provide a bountiful source of energy for peaceful purposes without creating unacceptable risks of proliferation of nuclear weapons.

In this article, proliferation resistance means the panoply of measures to weaken the links between the civilian nuclear fuel cycle and the capability of a nation or a subnational group to acquire nuclear weapons. The links are several. Nuclear power programs require cadres of nuclear scientists and technicians, a network of research facilities, research reactors, and laboratories—some of which are indispensable to a nuclear weapons program and which could also serve as a cover to a parallel, dedicated weapons program. The most direct connection between nuclear power and nuclear weapons, however, is through the production and use of fissile (weapons-usable) materials, especially plutonium and highly enriched uranium (HEU).

In summary, two critical diversion and proliferation risks confront civilian nuclear power:

That countries or terrorist groups could divert fissile or radioactive materials directly from the civilian nuclear fuel cycle into nuclear explosives or (for terrorists) so-called “dirty bombs,” which are conventional explosives surrounded by radioactive materials.

That countries aspiring to obtain nuclear weapons could use civilian nuclear facilities (power reactors, research reactors, reprocessing plants, uranium-enrichment plants, etc.) and trained cadres of nuclear scientists, engineers, and technicians as a cover and/or training ground for the dedicated acquisition of fissile material for nuclear weapons.

A third category of security risk posed by nuclear power is that terrorists could release substantial amounts of radioactivity through attacks on reactors, spent fuel pools, or certain other nuclear facilities. The consequences of such attacks could be devastating. However, the general issue of security of reactors and other nuclear sites raises questions outside the scope of this article and is not considered further here.

The proliferation-resistant barriers to confront these risks are both technical—(the quality of the plutonium and HEU used and produced in the fuel cycle, the degree to which the fissile material is or is not protected by a radiation shield, etc.) and institutional (e.g., safeguards, treaties, and agreements among suppliers not to export sensitive items). This article discusses both.

2. NUCLEAR WEAPONS

The critical fissile materials that can sustain an explosive chain reaction are plutonium (of almost any isotopic composition) and HEU (i.e., uranium enriched in the isotope U-235). Natural uranium contains approximately 0.7% U-235 and 99.3% U-238. Practically, uranium enriched to less than 20% U-235 cannot be used for weapons, and so-called weapon-grade uranium is generally taken to be more than 90% U-235. Although uranium with concentrations between 20 and 90% U-235 could sustain an explosive chain reaction, the critical masses needed, which go inversely as the square of the U-235 density, would be correspondingly very high. Although HEU includes all uranium above 20% U-235, here it generally refers to uranium at or above 90% U-235. Other isotopes could in principle be used in a weapon—notably U-233, americium-241, and neptunium-239—but so far these are not of practical significance. However, U-233, which is produced in a reactor containing thorium, could play a role in the future of civilian nuclear energy and is mentioned later.

Once it gathered the fissile material, a country or group could then assemble a fission weapon employing one of two fundamental designs: gun type and implosion. Fission weapons, such as were used in Hiroshima and Nagasaki, could also be used to ignite thermonuclear reactions in so-called two-stage devices, in which the fission component is termed a fission primary or “pit.” Thus, a potential proliferant, even if ultimately seeking a thermonuclear weapon, will first have to develop a fission device.

In a gun device, two subcritical masses of a fissile material are explosively brought together, as along the barrel of a gun. In an implosion weapon, a subcritical sphere or shell of fissile material is imploded, increasing the density of the material to a supercritical state. For reasons discussed later, a gun-type device cannot use plutonium. The rudiments of both designs are widely known, and it must be assumed that many countries would be able to put together a team of scientists and engineers to manufacture a weapon once the fissile material was obtained. The gun-type design is particularly simple (it was used in Hiroshima and was never tested prior to that), and a subnational group may be able to use this design to produce a device, once it obtains the necessary fissile material.

Thus, the critical barrier to proliferation is acquisition of fissile material, and a central question

for civilian nuclear power is whether it provides a ready route to such acquisition.

Generally speaking, the critical masses required are 10 kg for plutonium metal and 50 kg for uranium metal (90% U-235). These are the so-called bare crits. The use of a reflector—several inches of a material that will reflect neutrons back into the weapon assembly—could cut the critical mass by as much as one-half, for example, to 4–8 kg for plutonium and 25 kg for uranium. The significant quantities specified by the International Atomic Energy Agency (IAEA) are 8 kg for plutonium and 25 kg for 90% U-235. The actual critical masses depend on the precise designs employed and composition of the plutonium and uranium, but these approximations provide a good benchmark to compare with the flows of materials in the civilian nuclear fuel cycle.

3. FISSILE MATERIAL PRODUCTION

3.1 Plutonium

Plutonium is not naturally occurring and must be produced in a reactor. Reactors dedicated to the production of plutonium for weapons have used natural uranium as fuel with low burn-up, which is the measure of how long an element of fuel is left in the reactor and therefore how much energy the element produces before it is discharged. In a reactor dedicated to plutonium production, the burn-up is typically less than 1000 thermal MW-days per ton of fuel. In such reactors, approximately 0.9 g of plutonium is produced per thermal megawatt-day. For example, the Indian CIRUS research reactor, which has a thermal power of 40 MW, at a 50% capacity factor could produce approximately 6.6 kg of plutonium per year.

Plutonium is also produced in civilian power reactors. In light water reactors (LWRs), the net plutonium production is approximately 0.3 g of plutonium per thermal megawatt-day, and in natural uranium, heavy-water reactors of the CANDU type, the production is approximately 0.5 g of plutonium per thermal megawatt-day. Thus, a 1000-MW electric (3000 thermal MW) LWR at 70% capacity factor produces approximately 230 kg of plutonium per year. In the United States, which has the equivalent of 100 1000-MW electric commercial reactors, approximately 23,000 kg of plutonium is generated in the reactors each year.

The spent fuel of all reactors contains substantial amounts of plutonium. However, as long as the

plutonium remains locked in the spent fuel along with the highly radioactive fission products, it cannot be used for weapons. Furthermore, as long as the plutonium remains unseparated, it is protected from diversion by a very intense radiation field. For example, consider the dose rate from a pressurized water reactor (PWR) fuel assembly for a person at varying distances from the assembly. A typical PWR assembly weighs approximately 0.5 metric tons, has a burn-up of 35,000 MW-days per ton, and contains approximately 5 kg of plutonium. For such an assembly even for fuel out of the reactor for 15 years, a person 1 m from the assembly would receive a lethal dose in a few minutes. Moving 5 m away from such fuel would increase the time for a lethal dose to a couple of hours. After the first year of discharge, the exposure decays approximately by a factor of two every 30 years.

Separation of the plutonium has to be done in a reprocessing operation. With current technologies, the spent fuel needs to be decladded, chopped into small pieces, and dissolved in hot nitric acid to prepare for the plutonium extraction. The chemical extraction of the plutonium can then proceed by the use of blenders, pulse columns, centrifuge extractors, and the reduction and oxidation of the plutonium. All this has to be done behind heavy shielding and with remote handling. Therefore, reprocessing is not a trivial enterprise and requires substantial technical know-how.

3.2 Question of Reactor-Grade Plutonium

For a time, many in the nuclear industry believed that plutonium generated in commercial reactors could not be used for weapons. The reason for this belief (or hope) was that the so-called reactor-grade plutonium has a relatively high fraction of the isotope Pu-240, approximately 25% for plutonium from LWR spent fuel, compared to less than 6% for weapon-grade plutonium. Pu-240 fissions spontaneously, emitting large amounts of neutrons, leading to the possibility that one of the neutrons could initiate a chain reaction before the bomb assembly reaches its maximum supercritical state and thus creating a fizzle yield. Indeed, the prospect of such predetonation rules out the use of gun-type designs employing even weapon-grade plutonium, as noted earlier. This is because the assembly speed is low enough to virtually ensure a fizzle yield. Unfortunately, however, reactor-grade plutonium can be used for implosion weapons, for which the assembly speed is much greater. The issue was addressed in a 1994

National Academy of Sciences study and later described in a January 1997 U.S. Department of Energy release:

Virtually any combination of plutonium isotopes ... can be used to make a nuclear weapon. In short, reactor-grade plutonium is weapons-usable, whether by unsophisticated proliferators or by advanced nuclear weapon states. Theft of separated plutonium, whether weapons-grade or reactor-grade, would pose a grave security risk.

3.3 Highly Enriched Uranium

As noted previously, natural uranium contains approximately 0.7% U-235 and cannot sustain an explosive chain reaction. To obtain weapon-grade uranium, the uranium has to be enriched in isotope-separation facilities to approximately 90% U-235. This was, or is being, done in the nuclear weapon states mostly through the use of gaseous diffusion and centrifuge plants. The United States, in the Manhattan Project, and Iraq, in its nuclear program before the Gulf War, also used electromagnetic separation. All these methods require highly specialized equipment, considerable amounts of electricity (although the electricity intensity differs substantially among the different isotope-separation technologies), and a substantial effort to acquire usable quantities of weapon-usable material.

The isotope-separation work required to produce an enriched product depends on the tails assay of the depleted uranium streams but is independent of the method of separation used. It is usually measured in kilogram separative work units (SWUs). Thus, for example, to produce 1 kg of 90% uranium starting from natural uranium and with a tails assay of 0.2% requires 225 SWU and 180 kg of natural uranium (Table I). To put this in perspective, the kind of centrifuges that the Iraqis were assembling in 1990 each had a capacity of approximately 1 SWU per year. Thus, to produce one critical mass per year, for example, 25 kg of 90% uranium, would require more than 4000 centrifuges.

LWRs, the predominant commercial reactor type in the world today, and the high-temperature gas reactors currently under intensive study use low-enriched uranium (LEU) fuel, typically between 3 and 5% U-235 for LWRs and perhaps 8% for gas reactors. The production of such fuel requires isotope separation in which the uranium is enriched in the isotope U-235, with some typical data noted Table I. The isotope separation is accomplished in facilities in which in each “stage” the uranium is enriched slightly, with multiple stages connected in a “cascade”

TABLE I
Uranium and Enrichment Requirements^a

Feed enrichment	Product (1 kg)	Uranium feed (kg)	SWUs
Natural U	Weapon-grade U (90%)	180	225
Natural U	4% U	7.6	6.6
Natural U	8% U	15.6	17
4% U	Weapon-grade U	23.6	72
8% U	Weapon-grade U	11.3	38

^aThis is for 0.2% tails assay.

to produce the desired product. As already noted, the LEU cannot be used for weapons but the use of even LEU raises two proliferation concerns: The enrichment facilities used to produce the low-enriched fuel can be used to produce HEU, and if the low-enriched fuel is used as input to an enrichment cascade, the quantity of work required to produce HEU would be much reduced.

4. ROUTES TO FISSILE MATERIALS FOR WEAPONS

Although there are hybrid variants, the main routes to fissile material for weapons for an aspiring nuclear state are the following:

- Construct a dedicated uranium enrichment facility and produce HEU. All the declared nuclear weapons states—the United States, Russia, Great Britain, France, and China—did this. It is a route that was attempted by Iraq in the years leading up to the Gulf War. It was the route also followed by South Africa (before it gave up its nuclear weapon program) and, in a more ambiguous way, by Pakistan. North Korea also revealed in late 2002 that it had a program under way to produce HEU for weapons.
- Construct a dedicated production reactor and reprocessing plant, and produce and then separate plutonium. Again, all the declared nuclear weapon states did this. This route has also been tried by North Korea.
- Use a research reactor ostensibly built for peaceful purposes to produce the plutonium and then separate the plutonium in a reprocessing plant that may or may not be declared. This is the route taken by India with its CIRUS reactor and Israel at the Dimona site.

- Produce HEU at an enrichment facility nominally constructed for civilian uses. This is arguably what Pakistan did, although in the initial stages of the Pakistan program its facility was probably more accurately characterized as one dedicated to the production of HEU.
- Divert plutonium or HEU from the commercial civilian fuel cycle. As far as is known, this has not yet happened.

One critical implication of the previous discussion is that countries could, if they wished, choose a dedicated route to nuclear weapons material. They do not have to depend on the civilian fuel cycle (although such a cycle could provide a convenient cover). As a consequence, it is not essential that safeguards provide a foolproof barrier to diversion from the civilian fuel cycle. Rather, the principal task of safeguards must be to make diversion from civilian nuclear power substantially more difficult, time-consuming, and transparent than a dedicated route.

On the other hand, it should also be understood that civilian nuclear power could provide a country with the basis for eventually a dedicated weapons program by allowing the country to train scientists and engineers, to build research facilities, and to learn techniques of reprocessing and enrichment that could later be turned to weapons uses. In this manner, a civilian program could impel a country along a path of latent proliferation, in which the country moves closer to nuclear weapons without having to make an explicit decision to actually take the final step to weapons, or at least to make transparent its intention to take such a step. Latent proliferation is particularly germane to the consideration of the spread of civilian nuclear power to countries that do not currently have any and that therefore would not have a ready infrastructure to support a dedicated route to nuclear weapons.

The routes to weapons summarized previously are for states. At the present and foreseeable state of nuclear technology, it does not appear feasible for a subnational group to produce its own fissile material. Even if it started with LEU or spent fuel, a subnational group could not realistically produce HEU or separate plutonium. The principal prospect for diversion of weapons materials from civilian nuclear programs for subnational groups is if they could steal or buy on a black market already separated fissile material—HEU or plutonium. These do not exist in the once-through fuel cycle, predominant in the world today; but HEU is used in some civilian research reactors and plutonium is

separated in programs that reprocess and recycle plutonium. These potential sources of fissile material are discussed later.

Although a subnational group cannot realistically obtain fissile materials, it could conceivably acquire civilian spent fuel or high-level wastes for a dirty bomb or radiological weapon. However, the prospect that a terrorist group could use spent fuel or high-level wastes in a dirty bomb, although troubling, does not compare in severity to actual diversion of fissile material. A dirty bomb would not kill many people, and although if detonated in an urban concentration it would cause much economic dislocation, so could many other potential terrorist actions. With care, overseers of the nuclear fuel cycle could make it very difficult for a terrorist group to divert radioactive materials from nuclear facilities, especially compared to other routes terrorists could potentially take to acquire radioactive materials.

Naturally, countries or subnational groups could seek to obtain fissile materials not from peaceful nuclear activities but rather from nuclear weapons programs directly, with the security of materials in the sprawling nuclear weapons complex in Russia of particular concern. However, this possibility lies outside the scope of this article. It is critical that diversion be prevented in both the civilian and the weapons realms.

5. NUCLEAR REACTORS AND FUEL CYCLES

To understand the safeguards challenge, it is useful to understand the scope of the nuclear enterprise worldwide.

5.1 Nuclear Power Reactors

At the end of 2002, there were 442 reactors operating worldwide, with a total installed capacity of approximately 350 gigawatts electric (GW). The predominant reactor type operating today is the LWR, which is both cooled and moderated by light (ordinary) water. The LWR comprises two types of reactor: the PWR and the boiling water reactor (BWR). All LWRs use uranium fuel enriched to approximately 3–5% U-235 or mixed-oxide fuel (MOX) discussed later. Most LWRs have a capacity of approximately 1000 MW electric.

Canada and a few other countries operate heavy water reactors (HWRs), cooled and moderated by heavy water. HWRs typically use natural uranium fuel and thus do not require enrichment. This is a

proliferation-resistance plus in that a country would not have the excuse of producing enriched uranium for civilian uses. At the same time, the use of natural uranium has a drawback in that it allows a country to have independent control over its fuel cycle, not having to import enriched uranium. It would have to import or produce indigenously heavy water, however. Also, as already noted, a HWR produces more plutonium per year than a comparably sized LWR. The HWRs have capacities between 100 to several hundred megawatts.

In addition, a few other reactor types are currently deployed. Most important, the United Kingdom operates several reactors that are cooled by carbon dioxide and moderated by graphite. Some of these gas-cooled reactors use slightly enriched uranium and some use natural uranium metal. Russia operates several RBMK reactors (the Chernobyl type), which are water cooled and graphite moderated. RBMK reactors use enriched uranium.

Other reactor types are being investigated, with the belief that they might be more economic than current types and possibly also more proliferation resistant. Of special interest is the high-temperature gas-cooled reactor, including the pebble-bed modular reactor (discussed later), which is under intensive study. Also of interest is the fast breeder reactor, which is under development in Russia, Japan, and France. The type of fast reactor most fully developed is cooled by liquid sodium and uses plutonium or HEU as fuel. In such reactors, the chain reaction is carried by “fast” neutrons, which are not slowed to thermal speeds by a moderator as in the reactors described previously.

5.2 Research Reactors

In addition to the power reactors, most countries with nuclear power also operate one or more research reactors of several different types. These reactors are of concern for two reasons. First, they can be (and indeed have been) used to produce plutonium for weapons. The Indian CIRUS reactor produced the plutonium for India’s first bomb, and the Dimona reactor in Israel has by all accounts been used to produce plutonium for Israeli nuclear weapons. Second, some research reactors use HEU as fuel, which could be used for nuclear weapons. This proliferation risk is discussed further later.

5.3 Uranium Enrichment Facilities

Uranium enrichment plants producing LEU are limited to only a few plants, most located in the

nuclear weapons states. Minatom in Russia operates four plants. The U.S. Energy Corporation operates the Paducah plant in Kentucky. URENCO, a joint venture of three countries, operates three plants in the United Kingdom, Germany, and The Netherlands. Eurodif operates one plant in France. The total installed production capacity for enrichment is approximately 40 million SWU per year, with a current world demand of 37 million SWU.

5.4 Once-Through Fuel Cycles

Most nuclear power operates on a once-through fuel cycle in which the spent fuel is not reprocessed, and the fresh fuel for the reactors is either LEU or natural uranium, not usable for nuclear weapons. The spent fuel from the reactors for the most part remains at the reactor sites, either in spent fuel pools of water or in dry-cask storage, with the expectation that the fuel will eventually be put into deep geologic repositories (such as Yucca Mountain in the United States) for permanent disposal. In such a fuel cycle, weapon-usable material (plutonium or HEU) is never isolated.

5.5 Reprocessing and Recycling

Not all nuclear power is operated on once-through fuel cycles. The United Kingdom, France, Russia, and, to a lesser degree, Japan and India are reprocessing spent fuel, with commercial-size reprocessing plants located in Sellafield in the United Kingdom, La Hague in France, and Mayak (Chelyabinsk) in Russia. The U.K. and French reprocessing plants are reprocessing both their own spent fuel and spent fuel from other countries, notably Japan and Germany. In 1998, the total amount of plutonium separated from civilian spent fuel was approximately 195,000 kg, with most of this plutonium stored at the reprocessing plants: 70,000 kg in the United Kingdom, 76,000 kg in France, and 30,000 kg in Russia.

By 2002, approximately one-third of the spent fuel discharged from reactors each year worldwide was being reprocessed. Currently, the rate of plutonium separation is approximately 20,000–24,000 kg per year, although this may decrease during the next few years. By 2002, the total civilian inventory of separated plutonium exceeded the inventory of military plutonium, which has been estimated to be approximately 250,000 kg.

However, some of the separated plutonium is being fabricated into MOX at four plants in Europe. In 2000, these plants produced slightly less than 200,000 kg of MOX, incorporating

10,000–12,000 kg of plutonium, with MOX production capacity expected to double in the next few years. MOX is being burned in approximately 32 LWRs in France, Germany, Belgium, and Switzerland. Another 18 have been licensed to use MOX. Japan is planning to use MOX in one-third of its reactors by 2010. In almost all these cases, MOX is being used or is planned on being used in one-third of cores.

6. TECHNICAL BARRIERS TO DIVERSION

The principal technical barriers to diversion are evident from the previous discussion. They include materials barriers, such as the isotopic composition of the uranium fuel, the radiological protection provided by spent fuel, and the mass and bulk of materials that would have to be diverted. As an illustration of the last point, consider the effort that would be required to divert PWR spent fuel. Not only does the spent fuel present a strong radiation barrier to diversion but also each PWR assembly (containing approximately one critical mass of plutonium) weighs half a ton. It could not easily be handled or smuggled out of a facility. In transport, several assemblies would be put into a transport cask, perhaps weighing on the order of 100 tons—again representing a formidable barrier to diversion, certainly by a subnational group.

The technical barriers also include what might be called technology barriers, such as the complexity of facility modifications that would be required to acquire fissile material. One important illustration is to compare the relative proliferation resistance of two technologies for uranium enrichment—gaseous diffusion and centrifuges. A gaseous diffusion plant set up to produce LEU would be very difficult to convert so that the plant could instead produce HEU for weapons. In contrast, a centrifuge plant ostensibly designed to produce LEU could readily be converted to produce HEU. The reason for this is that for gaseous diffusion, it is very difficult and highly inefficient to reconfigure the cascade from one optimized to produce LEU to one designed to produce HEU, whereas it is relatively straightforward to do so for a centrifuge cascade.

7. INSTITUTIONAL BARRIERS TO PROLIFERATION: SAFEGUARDS

The Non-Proliferation Treaty (NPT), which entered into force in 1970, provides the principal institu-

tional barrier to proliferation. The treaty defines two classes of states: the declared nuclear weapon states—the United States, Russia, the United Kingdom, France, and China—which had tested nuclear weapons before the treaty came into force, and the nonnuclear weapon states. It imposes obligations on both.

With respect to the nonnuclear weapon states, the treaty requires that each party pledge not to acquire nuclear weapons and to subject all its nuclear activities to safeguards applied by the IAEA, an autonomous organization established in 1957 under the United Nations (UN). Under the IAEA safeguards agreements signed by the parties, the parties are obliged to report all nuclear activities and to allow IAEA inspections at all declared facilities. The requirement that all of a country's peaceful activities be put under safeguards is referred to as full-scope safeguards. The purpose of the inspections is to enable the IAEA to detect any significant diversion of materials from the civilian nuclear fuel cycle, with any such detection then reported to the UN Security Council. With the exception of India, Pakistan, and Israel, all countries with any kind of nuclear program have joined the treaty, including the five declared nuclear weapon states. South Africa initially did not join the NPT and then actually manufactured a small number of nuclear weapons. However, in 1990 and 1991, it dismantled the weapons and its secret nuclear program, and it joined the NPT as a nonnuclear state in 1991. North Korea withdrew from the NPT in early 2003.

Iraq and North Korea, despite their past formal adherence to the NPT, embarked on nuclear weapon programs while still parties to the Treaty. The United States has claimed that Iran also has a clandestine weapons program, but this has been denied by Iran, and all of Iran's declared nuclear facilities remain under IAEA safeguards.

During the years leading to the First Gulf War, Iraq maintained all its declared nuclear facilities under IAEA safeguards while it was undertaking an extensive clandestine nuclear weapons program hidden from IAEA inspectors. Since 1991, of course, Iraq has been subject (with a 4-year gap between 1998 and 2002) to extraordinary inspections by the IAEA for nuclear weapons programs and a special UN inspectorate for chemical and biological weapons, with the war launched against Iraq in March 2003 in part justified by the United States, the United Kingdom, and others on grounds that Iraq was still harboring weapons of mass destruction and not cooperating with the inspection agencies.

Aside from Iraq, North Korea is the other non-nuclear party to the NPT that has most clearly sought to evade safeguards. It delayed for several years signing a safeguards agreement with the IAEA, evidently separated some plutonium in the meantime, and in 1994 threatened to withdraw from the NPT when the plutonium production was discovered. The looming crisis in 1994 was averted when, under an agreement negotiated bilaterally with the United States (the so-called Agreed Framework), North Korea agreed to place all its declared facilities under safeguards by the IAEA and to freeze (although not dismantle) its nuclear program. However, the IAEA noted in its 2001 annual report that "the Agency is still unable to verify the correctness and completeness of the initial report of nuclear material made by the DPRK and is, therefore, unable to conclude that all nuclear material subject to safeguards has been declared." In October 2002, North Korea suggested that it had a hitherto undeclared program to enrich uranium; and in December 2002 it withdrew its declared stockpile of spent fuel and its reprocessing plant from IAEA safeguards. At the time of this writing, the situation in North Korea remains unsettled.

In addition to the NPT, many nonnuclear weapon countries have joined into nuclear-free zones, in which the countries of a specific region reaffirm their decision not to acquire nuclear weapons. These zones include Latin America (Treaty of Tlatelolco, 1967), the South Pacific (Treaty of Rarotonga, 1985), Southeast Asia (Treaty of Bangkok, 1995), and Africa (Treaty of Pelindaba, 1996).

Also supplementing the NPT, countries exporting nuclear technology have entered so-called supplier agreements not to export certain specified materials or technologies to nonnuclear countries and to ensure that whatever nuclear material or equipment that is exported is under safeguards. The critical supplier agreement in place is the Nuclear Supplier Group (NSG) instituted in 1977 (when it was initially called the London Club), including all the major nuclear suppliers. The NSG has formulated a set of guidelines for nuclear transfers. The guidelines define a "trigger list" of items that suppliers agree to export to nonnuclear weapon states only when the receiving state has brought into force an agreement with the IAEA requiring the application of full-scope safeguards on all its current and future peaceful activities. Suppliers also agree to exercise restraint on the transfer of sensitive technologies, such as reprocessing and enrichment facilities. These technologies could be exported only with explicit further commitments by the recipient states, for example, not to

produce HEU. Paralleling the NSG is another intergovernmental group, the Nuclear Exporters Committee, known as the Zangger Committee, set up in 1971, which also has established conditions and procedures to govern nuclear exports.

These multifaceted international arrangements have been effective. Today, essentially all nuclear facilities in nonnuclear countries (a category that does not include Israel, India, Pakistan, and North Korea) are under IAEA safeguards, which is a remarkable achievement.

Furthermore, if a state party to the NPT withdrew from the treaty, in many cases backup safeguards applied by the United States, Canada, France, or other nuclear supplier would come into force. That is, a country such as Japan could not simply shed all its safeguards obligations by withdrawing from the NPT. These residual safeguards obligations generally derive by virtue of the country importing reactor technology or nuclear fuel from supplier countries. It may be noted in passing that such backup to NPT safeguards may not persist in the future if nuclear power grows substantially and an increasing number of countries produce nuclear technologies and materials indigenously.

After the Gulf War of 1991 revealed that Iraq, although a party to the NPT and with all its declared nuclear facilities under safeguards, was pursuing a very extensive nuclear weapons program, the international community realized that safeguards would have to be extended to allow the IAEA to search for undeclared activities as well as monitoring declared sites. In response, the IAEA developed a strengthened safeguards system called INFCIRC 540, which authorizes the agency to employ environmental monitoring and other means to search for undeclared activities. This strengthened safeguard system has entered into force in several countries, including Japan and Canada. European countries have generally indicated their intention to accede to the new safeguards agreement. By the end of 2002, the United States had signed but not yet ratified. As of the summer of 2003, several countries with nuclear power ambitions had not signed, including Iran, (which has now indicated that it will sign and ratify), Egypt, Turkey, South Korea, Argentina, and Brazil.

8. PROLIFERATION RESISTANCE: AN INITIAL ASSESSMENT

How should one gauge the proliferation resistance of the nuclear fuel cycle? One starting point is an

October 2000 report by a task force under the auspices of the Nuclear Energy Research Advisory Committee (NERAC), set up by the U.S. Department of Energy, that was tasked with addressing the question of proliferation resistance. The task force considered various attributes of proliferation resistance, including the materials, technology, and institutional barriers described previously. The task force cautioned that the proliferation resistance provided by any particular barrier depends on the threat or scenario one has in mind, and so its report should be considered preliminary. Nevertheless, the summary table that the task force put forward captures the effectiveness of most of the proliferation-resistant ideas that have been put forward. Table II is a simplified form of the summary table. The table suggests two compelling conclusions: (i) that it will be extraordinarily difficult to contrive an effective proliferation-resistant nuclear fuel cycle for sophisticated states, and even difficult to do so for unsophisticated states, and (ii) that effective barriers could be erected against subnational diversions of nuclear material.

The first conclusion is not surprising. It is evident that any country with a serious nuclear energy program could convert it to weapons uses if it wished, although it might not be able to do this clandestinely. For example, even if a country operated only on a once-through fuel cycle, it could build a quick and dirty reprocessing plant able to extract plutonium reasonably rapidly within a year or less. Furthermore, it is unclear what sanctions could be brought against a country breaking out of its safeguards obligations. The IAEA is charged with bringing any violation of safeguards to the UN Security Council and conceivably the council or individual countries could take action against the

violating country, but whether and how this might be done are highly uncertain. Nevertheless, a country will be loathe to be caught cheating on safeguards or to directly break a safeguards agreement, and this provides some deterrent to such actions. However, it certainly does not present an ironclad one.

In the current world environment, the potential of a state to divert weapons materials from its civilian programs is not necessarily dismaying. First, nuclear power is overwhelmingly located in a relatively few industrialized democracies, a few countries in Eastern Europe, Ukraine, Lithuania, and Russia. Of the 350-GW installed capacity worldwide, less than 10 GW is in developing countries, generally thought to be the most likely arena for further proliferation. This includes 2.3 GW in India, 2.1 GW in China, 0.9 GW in North Korea, 0.4 GW in Pakistan, and 2.7 GW in South America. For the most part, the countries with nuclear power programs are either already nuclear weapon states or countries that for whatever reason do not aspire to become nuclear weapon states. Japan, South Korea, and Taiwan have at one time or another considered nuclear weapons, but they are deeply constrained from seeking weapons. Second, as already mentioned, the civilian program, even if it could be converted to weapons purposes, might not provide a quicker or easier path to nuclear material than a dedicated route.

This much is not in great dispute. More controversial is whether and to what degree reprocessing and recycling significantly weaken the proliferation resistance of the nuclear fuel cycle.

8.1 Reprocessing and Recycling

Opponents of recycling (this author included) emphasize the proliferation risks. First, the reprocessing

TABLE II
Relative Importance of Various Barriers to a Selected Threat

	Sophisticated state, overt	Sophisticated state, covert	Unsophisticated state, covert	Subnational group
Materials/barriers				
Isotopic	Low	Low	High	High
Chemical	Very low	Very low	High	High
Radiological	Very low	Low	Moderate	High
Mass and bulk	Very low	Low	Moderate	High
Detectability	Very low	Low	Moderate	High
Technical barriers	Very low-moderate	Very low-high	Low-moderate	Moderate-high
Institutional barriers	Very low-moderate	Very low-high	Low-high	Low-moderate

Source: NERAC Task Force on Technology Opportunities for Increasing the Proliferation Resistance of Global Civilian Nuclear Power, October 2000.

generates a tremendous quantity of separated plutonium that has to be very carefully accounted for and guarded. As already noted, at the end of 2002, well over 200 metric tons of separated plutonium was being stored in the form of plutonium oxide at several sites throughout the world, with most in France, the United Kingdom, and Russia. The material in France and the United Kingdom is reasonably secure. However, a large quantity is in Russia under less certain security, and there are appreciable quantities in Japan.

Second, the use of MOX in reactors means that there will be supplies of fresh plutonium fuel at MOX fabrication plants, reactor sites, and in transport from reprocessing plants to fabrication plants to reactors—all potential sources of diversion for subnational groups. With the recycling activities restricted to Europe, the standards of security and safeguards applied to the MOX are probably high but may not always remain so. Third, whatever the risks today, they will be multiplied if a real market develops for MOX, with middlemen and agents arranging for the purchase and sale of MOX.

Finally, safeguards at bulk-handling facilities, such as reprocessing plants and MOX fabrication plants, will be much more difficult to apply than at reactors, with far more diversion paths potentially available to a country.

These risks seem all the more exacerbating to the critics of recycling because the enterprise appears so dubious on economic grounds alone. Even if the plutonium were free (i.e., even if the MOX fabrication plants received the plutonium free, with all the costs of separating the plutonium in reprocessing plants written off), the cost of a ton of MOX fuel exceeds that of a ton of ordinary LEU at current uranium prices. Unless the cost of uranium increased very sharply, the small savings in uranium and in enrichment due to plutonium recycling cannot make up for the far greater cost of fabrication of MOX, and the economic penalty of recycling is all the more apparent if the cost of reprocessing is taken into account.

Supporters of reprocessing and recycling downplay the proliferation risks, especially since these activities are limited to Europe and Japan. They also point out that a country currently relying only on a once-through fuel cycle could build a quick and dirty reprocessing plant to separate plutonium.

Furthermore, recycling plutonium and burning it in reactors sharply reduces the buildup of plutonium in spent fuel compared to the once-through fuel cycle. This could be important because one potential risk of the once-through fuel cycle is that the

plutonium contained in the spent fuel could eventually become accessible to potential proliferators as the radioactive barrier provided by the fission products decays, by a factor of approximately 10 every 100 years. Thus, potentially the geologic repositories to which the spent fuel is destined could become plutonium mines. How seriously one should take this consideration is unclear. As long as a country has an operating nuclear power program, diversion of plutonium appears far easier than it would be in a mining operation.

8.2 Research Reactors Using HEU

By far the greatest inventories of HEU in the world are in nuclear weapons and in materials recovered from dismantled nuclear weapons—on the order of 1000–2000 tons. The greatest security challenge is to secure these materials until they can be denatured (i.e., rendered not usable for weapons) by mixing the HEU with natural or depleted uranium. This denaturing of weapon-grade uranium in fact is being done in Russia, with the United States purchasing the denatured product for LEU for reactors.

The quantity of HEU at research reactors is far smaller than the weapons-derived HEU—approximately 20 tons. However, this 20 tons is widespread and also must be carefully secured. In 2001, there were approximately 40 operating HEU reactors with thermal power greater than 1 MW. Many of these are in the United States and Russia. However, several are in nonnuclear weapon states. Reactors supplied by the United States are in Australia, Belgium, Germany, Israel, The Netherlands, Portugal, and South Africa. Reactors supplied by Russia are in the Czech Republic, Hungary, Kazakhstan, North Korea, Libya, Poland, Ukraine, and Uzbekistan. There are also a small number of shutdown HEU reactors supplied by Russia that still contain some HEU.

Despite the number of HEU reactors still in operation, the situation has greatly improved during the past two decades. In 1978, the United States launched the Program for Reduced Enrichment for Research and Test Reactors, which has successfully converted many HEU reactors with thermal power greater than 1 MW in the United States and abroad to low-enriched fuel. The Soviet Union and now Russia have made similar efforts since 1978. The United States and Russia had been the principal suppliers of HEU for research reactors. Indeed, most of the reactors still using HEU were constructed before 1980. Since then, only four HEU reactors with thermal power more than 1 MW have begun

operation or are planned: a 5-MW reactor in China, a 10-MW reactor in Libya, a 10-MW reactor in Russia, and a 20-MW reactor in Germany.

Still, one of the main unfinished nonproliferation undertakings is shutting down or converting the remaining HEU reactors and removing any fresh and spent fuel inventories at the reactor sites to their countries of origin.

9. THE LONG-TERM CHALLENGE FOR NUCLEAR POWER: ENHANCED PROLIFERATION RESISTANCE

9.1 The Challenge

In the long term, if nuclear power worldwide becomes a robust contributor to a strategy to address global warming, the challenge to ensure that nuclear power is adequately proliferation resistant will be truly formidable. It is useful for illustrative purposes to take as a benchmark a global nuclear capacity of 3000 GW—an eightfold increase from today's worldwide capacity of 350 GW.

An increase of at least this magnitude will be necessary for nuclear power to make a dent in global warming. For example, under the central business-as-usual projection of the Intergovernmental Program on Climate Change (IPCC), if nuclear power increased to 3000 GW by 2075 (50% of world electricity then projected) and 6500 GW by 2100 (75% of world electricity), the total carbon emissions avoided cumulatively would be approximately 290 billion tons through 2100—only approximately one-fourth the projected cumulative carbon emissions through 2100 projected by the IPCC.

The management of a nuclear system of 3000 GW would be deeply challenging. First, whereas today the countries with nuclear power programs are either already nuclear weapon states or industrialized democracies with no current intentions to acquire nuclear weapons, an exuberant nuclear future will present a different picture. It is widely recognized that significant nuclear growth during the next half century will have to be largely in the developing countries. This is where by far the greatest increase in electricity production is projected. If one considers the top 25 countries by population projected for 2050 by the UN, among them are several that today have essentially no or a negligible amount of nuclear power: Indonesia, Nigeria, Pakistan, Bangladesh, Ethiopia, Congo, the Philippines, Vietnam, Egypt, Iran, Saudi Arabia, Tanzania, Turkey, Sudan, Uganda,

Yemen, and Thailand. No doubt, several of these countries, and others less populous, will not actually develop nuclear power on a large scale. However, if nuclear power truly comes to play a substantial role in the world energy economy, it will have to be located in many of these countries and on a substantial scale. Therefore, it will be a very substantial undertaking to extent safeguards to all these countries, some of which may in time aspire to nuclear weapons.

Second, the flows of material will be immense. If based on a once-through fuel cycle using LWRs, a 3000-GW system would generate annually approximately 600 tons of plutonium and also require on the order of one-half million tons of natural uranium. If based on liquid-metal plutonium breeder reactors, it would involve the fabrication into fresh fuel annually of more than 4000 tons of plutonium (although the cumulative inventory of plutonium would be much less than that for a system based on LWRs).

9.2 Proliferation Resistance Concepts

Confronted by this challenge, nuclear engineers and scientists in the United States and abroad have been exploring advanced nuclear technologies and novel concepts to improve the proliferation resistance of nuclear power (among other goals), including reactor types and/or new fuels that allow very high burn-up and produce less plutonium than do current reactors (such as the pebble-bed high-temperature gas-cooled reactor) and that may be operated on once-through fuel cycles; breeder or particle-accelerator-driven reactors that, to the extent possible, colocate sensitive processes (such as reprocessing) with the reactor and do not separate the plutonium from other actinides; and schemes that restrict nuclear power to large, international energy parks that would then export to individual countries electricity, hydrogen, or small, sealed reactors. The reactors envisioned in this last scheme would be 40 or 50 MW and would be fueled at some central nuclear park and then sealed and sent to client countries. The reactors would have lifetime cores, not requiring refueling, and at the end of the core life (e.g., 15–20 years) would be returned to the central facility unopened. This can be called a hub-spoke configuration.

9.3 Once-Through Fuel Cycles: Thorium Fuels and the Pebble-Bed Gas-Cooled Reactor

Although the once-through concepts under consideration differ sharply in terms of reactor type,

economics, and inherent safety, the proliferation resistance attractions of the different concepts are generally similar. Two technologies, the Radkowsky thorium fuel (RTF), currently under development by the Radkowsky Thorium Power Corporation, and the pebble-bed high-temperature gas-cooled reactor illustrate the enhanced proliferation resistance promised by the new concepts.

In the short term, the RTF fuel, employing a seed-blanket concept, allows a technical fix for current LWRs aimed at making the spent fuel difficult to use for weapons purposes. The fuel is designed to operate on a denatured uranium–thorium once-through fuel cycle in current LWRs and to achieve very high burn-up. In such a fuel cycle, the reactor would generate approximately one-fifth the plutonium generated in today's LWRs per kilowatt-hour of electricity produced. Moreover, the Pu-238 concentration in the discharged fuel would be 6.5% in the seed and 12% in the blanket compared to approximately 1% in today's PWR uranium fuel. This is significant because Pu-238 has a very high specific decay heat (560 compared to 1.9 W/kg for Pu-239 and 6.8 W/kg for Pu-240). The decay heat emission for the RTF seed is approximately three times that of normal PWR reactor-grade fuel and for the RTF blanket fuel approximately six times that of PWR fuel. The higher heat loads are likely to require special heat-removal measures in the design of a weapon. The spontaneous fission rate of the RTF plutonium isotopic mixture will also be far greater than that for PWR uranium fuel, making the design of a weapon more difficult and increasing the probability of a fizzle yield for a crude weapon. The U-233 (which, like U-235, is a fissile material) generated by thorium capture of a neutron would always be denatured with a sufficient quantity of U-238 so that the uranium cannot be used for weapons, and it is contaminated with the gamma-emitting U-232 decay chain, which would make any attempt to use the material for weapons more difficult and also any diverted material more detectable.

The pebble bed modular reactor derives its proliferation resistance from the fact that the spent fuel would be high burn-up material in thousands of tiny carbon-coated spheres, making it a comparatively unattractive source from which to recover weapons-usable materials. At any given time, the core of the reactor (nominally 100–110 MW) would consist of 360,000 pebbles (60 mm in diameter), each containing approximately 7 g of uranium (at approximately 8% U-235) in the fresh fuel in 11,000 microspheres (0.9 mm in diameter). In the spent fuel, there will be plutonium. At low burn-up, however,

when the weapon-grade quality of the plutonium is relatively high the plutonium content will be very low. At full burn-up, there will be more plutonium but far from weapon-grade quality, with a high fraction of Pu-238 (approximately 6%). The total content of the plutonium will be approximately 5 kg per ton of uranium fuel so that perhaps 200,000 pebbles would have to be diverted to obtain a critical mass.

There are some disadvantages. First, over time the decay of the approximately 30-year half-life fission products will lower the radiation barrier of spent fuel, whereas the decay of the 88-year half-life Pu-238 will make it easier to use the extracted plutonium for weapons. Perhaps more important, both fuel cycles use uranium that is more highly enriched than is typical today: the RTF fuel, 20% U-235 for the seed, and the pebble bed reactor, 8% U-235. Uranium enriched to 8–20% cannot be used for weapons, but the routes to weapons-grade uranium from such feed materials are easier than if one started with natural or 4% low-enriched uranium. The fabrication of the pebble-bed fuel and the fuel-handling operations will require special attention when safeguards are developed. Despite these cautions, these concepts, if they can truly be developed, promise enhanced proliferation resistance, especially during the next few decades.

In the long term, however, any nuclear system relying primarily on once-through fuel cycles will face formidable difficulties, even if sufficient uranium can be obtained from extraction from seawater or terrestrial sources to sustain a once-through cycle. (The natural uranium required will be on the order of one-half million tons annually.) For specificity, assume a 3000-GW nuclear capacity comprising half pebble-bed high-temperature gas reactors, each 100 MW, and half LWRs, each 1 GW. In such a world, there would be 15,000 pebble-bed reactors and 1500 LWRs and an enrichment requirement worldwide of approximately 400 million SWU per year. If one takes 2 million SWU per year as a nominal capacity of one enrichment plant (approximately the size of a URENCO plant), 200 such plants would be required. A 2-million SWU plant could make approximately 600 bombs per year starting with natural uranium. It could make 3500 bombs per year starting with 8% uranium, the fuel enrichment of the gas reactor fuel. Overall, the system will involve massive flows of natural and LEU and powerful economic incentive for innovation to make isotope separation cheaper and quicker. This is especially of concern because terrorist groups could

far more readily make a nuclear weapon from HEU than from separated plutonium.

Plutonium will also be a matter of concern. The spent fuel would be on the order of 50,000–70,000 tons of heavy metal per year, approximately the capacity that has been planned for Yucca Mountain (70,000 tons). Thus, nominally we can imagine one Yucca Mountain being constructed every year worldwide. Each one will have to be secured and safeguarded indefinitely since after several decades, the radioactivity surrounding the plutonium will decay substantially, making the spent fuel repositories prospective plutonium mines, as discussed earlier. Each repository (using the Yucca Mountain scale) would contain on the order of 1000 tons of plutonium-239.

9.4 Closed Fuel Cycles

Because a once-through fuel cycle will involve such prodigious use of uranium and enrichment services, it is likely that a 3000-GW nuclear power system will drive the nuclear industry toward breeder reactors. Such a development would minimize flows of uranium; but it would increase material flows of plutonium. For this reason, breeder concepts in which the plutonium is never fully separated from the fission products and/or other actinides have appeared worthy of study on proliferation-resistance grounds. Among these are integral fast reactors (IFR), which employ pyroprocessing technologies that do not separate the plutonium from the minor actinides and that integrate the reprocessing and reactor operation at a single site, and particle-accelerator driven reactors, which also use pyroprocessing, one version of which has been developed by Carlo Rubbia.

In general, scenarios to divert plutonium from these fuel cycles will require off-normal operation and specialized equipment, which in principle could be detectable by an inspection agency. On the other hand, even where pyroprocessing is used, fuel recycling equipment could be operated in a manner that would allow the extraction of large amounts of plutonium fairly quickly by adding a cleanup stage. In addition, safeguards that relied on materials accountability would be difficult to apply and therefore more than customary reliance would consequently have to be placed on containment and surveillance. Partly for these reasons, a study by the Martin-Marietta Corporation commissioned by the Departments of State and Energy in 1992 concluded that “in the absence of compelling circumstances to

the contrary, access to IFR technology should be preferentially limited at the outset to countries that have good nonproliferation credentials.”

This last point is the most troubling. Which countries are considered “safe” or to have “good nonproliferation credentials” will depend on who is making the judgment. It is illusory to think that the world can build a secure nuclear future that depends on two classes of countries—one where certain activities and fuel cycles are barred and one where anything, or almost anything, goes.

9.5 Hub-Spoke Configurations: International Energy Parks

One manifestation of future nuclear power is not open to many of the objections noted previously. This is to cluster all sensitive nuclear facilities in centralized, heavily guarded nuclear parks, preferably under international control. Then, electricity directly or hydrogen produced through electrolysis could be sent out from the parks. Also, the possibility of exporting small reactors from the park has been examined.

In this last case, long-life reactor cores would be assembled at the central facility. The reactors would be sealed and then exported to users in other countries, where they could be “plugged in” to the electric generation system. After some years (15–20), the core/spent fuel would be returned to the central facility or to some international spent fuel repository. During the 15–20 years of operation, there would be no refueling. In such a system, a country would need relatively few research facilities, operators, and other trained nuclear technicians and engineers.

This reactor concept has impressive proliferation-resistance credentials. These may be summarized as follows, adopting the analysis presented by engineering teams at the Berkeley Department of Nuclear Engineering, the Lawrence Livermore National Laboratory, the Argonne National Laboratory, and Westinghouse Electric Company for one version of a small reactor, the encapsulated nuclear heat source (ENHS) reactor. The following description and analysis are adapted from a paper by Ehud Greenspan *et al.* presented at the Global 2001 Conference in Paris in September 2001.

First, appropriation by a subnational group of the reactor, although it is transportable, would be a daunting challenge. The ENHS reactor is approximately 20 m long, with a 3-m diameter, and weighs approximately 200 tons. The fuel, which could either be 13% enriched uranium or a uranium-plutonium

fuel having 11 or 12% plutonium, is embedded in a mass of lead-bismuth (solid during transport and liquid during operation) throughout the core life. It would further be possible to seed the reactor with gamma-emitting cesium-137 before shipment, thus surrounding even the fresh fuel in a radiation shield. Furthermore, the ENHS does not give a country a useful source of neutrons: It is not possible to insert fertile material for irradiation. As noted previously, the core life of the reactor would be 15–20 years and during this period there would be no refueling. If operated on the hub-spoke concept, the client country would need no fuel fabrication facility and no fuel management capability. Because the reactor operates “almost autonomously,” the client country would need few operators of the nuclear system. Overall, the hub-spoke concept could diminish the rationale and opportunities for a country to develop research facilities and trained cadres of scientists and technicians that could later be diverted to weapons activities.

Presumably, the client country would be able to break into the sealed reactor, but it should be possible to ensure that such an attempt to obtain nuclear fuel could not be done undetected. Moreover, the acquisition of the fuel after a break-in would probably take days to weeks.

There are some matters of concern. First, the spent fuel of the reactor (using a nominal 40-MW capacity) will contain approximately 600 kg of plutonium if the uranium fuel is used and 1.4 tons of plutonium if the uranium-plutonium fuel is used. Also, if the fuel is removed from the reactor before its complete lifetime, the plutonium could be weapon grade or close to it. The uranium fuel, if obtained by a would-be proliferant, would not be weapons usable but would require less separative work for production of weapons-grade uranium than ordinary light water uranium fuel.

These problems notwithstanding, a nuclear system based on international energy parks, if it could be developed, does present an arguably proliferation-resistant strategy for nuclear power in the long term.

Are international energy parks realistic alternatives on political and economic grounds? International energy parks run against the strong wish of many countries to become energy independent. Also, they will require either that client countries accept discriminatory restrictions on their nuclear activities not accepted by the countries hosting the nuclear parks or that all countries, including the industrialized countries, accept a high degree of international control over their nuclear energy programs. Beyond these considerations, countries will also be wary of concentrating

too much of their energy future in a few places, with their attendant risks of common-mode failures, disruption of transmission lines or shipping, etc.

10. CONCLUSION

With the growth of the international terrorist threat, the diversion and proliferation risks of nuclear power are sure to gain in importance compared to the many other problems confronting nuclear power, such as economics, waste disposal, and safety. It may be that with nuclear power remaining at current levels and with almost all reactors and enrichment and reprocessing plants located in a relatively small number of highly industrialized countries, the international community will view the proliferation resistance of the current system of technologies and safeguards as tolerable, especially if next-generation reactors can allow markedly increased burn-up of fuel and if the IAEA strengthened safeguards system is widely implemented.

In the long term, if nuclear power grows substantially, the development of new proliferation-resistant technologies and fuel cycle configurations will gain new urgency. The concept under study that holds the most promise is the development of a hub-spoke arrangement in which all sensitive activities are performed at a central, perhaps international, energy park. However, such a strategy faces enormous political and practical obstacles. All the more so does the extreme of this strategy—to place all nuclear power under international control. This indeed was the central recommendation of the first comprehensive study of how to control nuclear energy, the “Report on the International Control of Atomic Energy,” prepared for the Secretary of State’s Committee on Atomic Energy (the so-called Acheson-Lilienthal Report) that formed the basis for the Baruch Plan for international control of nuclear weapons (submitted to the United Nations by the United States in 1946). It is worth remembering the sober view of this report:

There is no prospect of security against atomic warfare in a system of international agreements to outlaw such weapons controlled only by a system which relies on inspection and similar police-like methods. The reasons supporting this conclusion are not merely technical but primarily the inseparable political, social, and organizational problems involved in enforcing agreements between nations, each free to develop atomic energy but only pledged not to use bombs. So long as intrinsically dangerous activities may be carried out by nations, rivalries are inevitable and fears are engendered that place so great a pressure on a system of enforcement by police methods that no degree of ingenuity or technical competence could possibly cope with them.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fuel: Design and Fabrication • Nuclear Fusion Reactors • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Waste • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal • Radiation, Risks and Health Impacts of • Uranium Mining: Environmental Impact

Further Reading

- Albright, D., Berkhout, F., and Walker, W. (1997). "Plutonium and Highly Enriched Uranium 1996." Stockholm International Peace Research Institute/Oxford Univ. Press, Stockholm/Oxford.
- Bennett, R., and Koteck, J. (2002). Next generation nuclear power. *Sci. Am.* 286(1), 70–79.
- Cirincione, J. (2002). "Deadly Arsenals: Tracking Weapons of Mass Destruction." Carnegie Endowment for International Peace, Washington, DC.
- Feiveson, H. A. (2001.). The search for proliferation-resistant nuclear power. *Public Interest Rep.* 54(5).
- Garwin, R., and Charpak, G. (2001). "Megawatts and Megatons." Knopf, New York.
- Goldblat, J. (2002). "Arms Control: The New Guide to Negotiations and Agreements." Stockholm International Peace Research Institute, Stockholm.
- Krass, A., *et al.* (1983). "Uranium Enrichment and Nuclear Weapons Proliferation," pp. 1–40. Stockholm International Peace Research Institute/Taylor & Francis, Stockholm/London.
- Lloyd, W. R., Sheaffer, M. K., and Sutcliffe, W. G. (1993, October). Dose rate estimates from an irradiated pressurized-water-reactor fuel assembly in air, Report No. UCRL-ID-115199. Lawrence Livermore National Laboratory, Livermore, CA.
- Rhodes, R., and Beller, D. (2000). The need for nuclear power. *Foreign Affairs* 79, 30–44.

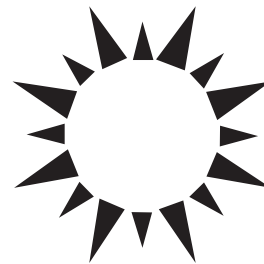
Nuclear Waste

PAUL K. ANDERSEN and ABBAS GHASSEMI

New Mexico State University
Las Cruces, New Mexico, United States

MAJID GHASSEMI

New Mexico Institute of Mining and Technology
Socorro, New Mexico, United States



-
1. Nuclear Fission
 2. Nuclear Fuel Cycle
 3. Radionuclides in Nuclear Waste
 4. Classification of Radioactive Wastes
 5. Disposal of Low-Level Wastes
 6. Management of Spent Nuclear Fuel
 7. Disposal of High-Level Wastes
 8. Radioactive Waste Management in the United States
 9. Radioactive Waste Management in France
 10. Radioactive Waste Management in Japan

Glossary

actinides The 14 elements having atomic numbers in the range from $Z = 90$ (thorium) to $Z = 103$ (lawrencium); the element actinium ($Z = 89$) is sometimes included as well.

activation Creation of a radioactive element from a stable one by the absorption of neutrons or protons.

alpha radiation (α) Nuclear radiation consisting of helium nuclei, each comprising two neutrons and two protons.

Becquerel (Bq) The SI unit of radioactivity, equal to one disintegration per second.

bentonite A soft plastic clay formed by the chemical alteration of volcanic ash, which expands on the absorption of water.

beta radiation Energetic electrons (denoted β^-) or positrons (denoted β^+) emitted from unstable nuclei.

calcining (or calcination) The process of heating a material to a high temperature below its melting point to drive off volatile species or promote chemical changes.

centrifugal separation An enrichment process that uses a series of high-speed centrifuges to separate $(^{238}\text{U})\text{F}_6$ and $(^{235}\text{U})\text{F}_6$ according to the difference in their molecular masses.

clay Unconsolidated sediments consisting of particles less than 1/256 mm in diameter.

daughter A nuclide that results from the radioactive decay of another nuclide (the “parent” nuclide).

depleted uranium Uranium containing less than the natural fraction of U-235 (0.71%).

electronvolt (eV) A unit of energy, defined as the kinetic energy acquired by an electron that is accelerated through a potential difference of 1 volt; an electronvolt is equal to 1.6022×10^{-19} J.

enrichment The process of increasing the proportion of a fissile isotope in a nuclear fuel above its natural fraction.

fertile element A nonfissile element that can be made fissile by the absorption of neutrons.

fissile element A heavy element that readily undergoes fission by the absorption of thermal (i.e., slow) neutrons.

fission The splitting of a heavy nucleus into two roughly equal fragments, with the accompanying release of energetic neutrons and (sometimes) a light nucleus such as tritium.

gamma radiation (γ) High-energy photons emitted from unstable nuclei.

gaseous diffusion A separation process that depends on differences in diffusion rates of gaseous compounds through a membrane; gaseous diffusion is used to separate $(^{238}\text{U})\text{F}_6$ and $(^{235}\text{U})\text{F}_6$.

half-life The time required for half of the nuclei in a radioactive sample to decay.

hex Abbreviation for uranium hexafluoride, UF_6 .

igneous rock Rock that solidified from molten or partly molten material.

ion exchange A separation process in which ions are removed from a liquid solution to an insoluble solid (typically a polymer resin) and are replaced by ions from the solid.

isotope Any one of multiple forms of the same element having the same atomic number Z but different mass numbers A .

leaching The removal of soluble constituents from a solid by the action of a liquid solvent.

nuclide A species of atom existing for a measurable length of time and having an atomic number Z , a mass

number A , and nuclear energy distinct from any other species of atom.

radionuclide A radioactive nuclide.

radiotoxicity A measure of the potential of a radioactive substance to cause harm to living tissue by radiation after introduction into the body.

reprocessing The separation of spent nuclear fuel into useful materials (especially uranium and plutonium) and wastes.

solvent extraction A process for removing chemical solutes from a liquid solution using a second immiscible liquid phase.

spent nuclear fuel Fuel after it has been used in a nuclear reactor, containing fission products, plutonium, and unreacted uranium.

transuranic element An element having an atomic number Z greater than 92.

tuff Rock formed from consolidated volcanic ash; welded tuff results when ash particles are fused together by heat and pressure.

vitrification The process of converting or incorporating nuclear wastes into a glass for permanent disposal.

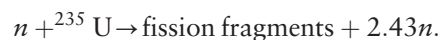
yellowcake The solid product of uranium milling, nominally U_3O_8 but actually consisting of mixed uranium oxides, hydrides, and impurities.

Radioactive waste may be defined as any material that contains a higher concentration of radionuclides than is considered safe and for which no further use is contemplated. Nuclear waste is a radioactive by-product of nuclear power generation, nuclear weapons production, and the medical and industrial applications of radioactive substances. This article focuses on radioactive waste from civilian nuclear power. In 2002, there were 438 nuclear power plants operating in 31 nations, producing more than 16% of the electricity generated worldwide. In some countries, nuclear power plants supply more than half of the electricity generated. Worldwide, nuclear power generation produces more than 200,000 m³ of nuclear waste annually, of which less than 5% is highly radioactive. The total volume of nuclear wastes is nearly 20 million m³.

1. NUCLEAR FISSION

All commercial nuclear power plants rely on the fission of uranium and, to a lesser extent, plutonium. (An alternative cycle, based on thorium, has been explored but never commercialized.) Naturally occurring uranium consists of two major isotopes:

U-235 (0.71% by weight) and U-238 (99.28% by weight). Only U-235 is fissile, meaning that it readily undergoes fission by the absorption of slow neutrons. The fission of U-235 can be represented by



A fission reaction produces two (or more) fission fragments or products and additional neutrons (2.43 neutrons on average); the latter can initiate further fission reactions. Each fission reaction releases approximately 200 MeV (million electronvolts) of energy, predominantly as the kinetic energy of the fission fragments. In contrast, the chemical combustion of a hydrocarbon fuel typically releases less than 7 eV per molecule of CO₂ formed.

The more abundant uranium isotope U-238 is not fissile. However, U-238 is fertile, meaning that it can capture neutrons and be converted to plutonium-239:



Plutonium-239 is fissile:



2. NUCLEAR FUEL CYCLE

The nuclear fuel cycle based on uranium is depicted in Fig. 1. The complete cycle consists of nine major steps: (1) mining and milling, (2) conversion, (3) enrichment, (4) fuel fabrication, (5) reactor operation, (6) interim storage, (7) reprocessing, (8) stabilization and immobilization, and (9) final disposal.

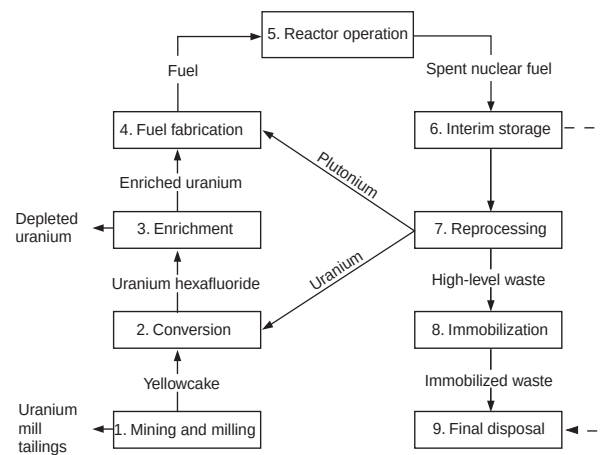


FIGURE 1 Nuclear fuel cycle based on uranium.

2.1 Mining and Milling

Uranium is excavated from open-pit or deep-shaft mines using conventional mining techniques (step 1 in the fuel cycle). The ore is crushed and the uranium is leached out using sulfuric acid or sodium carbonate and bicarbonate. The uranium is recovered from the liquid using processes such as ion exchange, solvent extraction, precipitation, and drying. The product is yellowcake, a solid consisting of U_3O_8 and various other oxides, hydroxides, and impurities.

2.2 Conversion

In the second step of the fuel cycle, yellowcake is reacted with hydrogen fluoride and fluorine to produce uranium hexafluoride (UF_6) or “hex.” Depending on the temperature and pressure, hex may be handled as a solid, a liquid, or a gas.

2.3 Enrichment

Most commercial reactors require enriched fuel in which the concentration of U-235 has been increased from its natural value of 0.71% to as much as 5% by weight. (A few power reactors, most notably the early British gas-cooled reactors and the Canadian CANDU reactors, do not require enriched fuel.) Weapons-grade uranium must be enriched even further, to approximately 90% U-235. The most common enrichment processes are gaseous diffusion and centrifugation, both of which take advantage of the minute difference in mass between U-235 and U-238. The by-product from enrichment is depleted uranium, which contains less than 0.71% U-235 (typically 0.25–0.30%).

2.4 Fuel Fabrication

Fuel for commercial power reactors may contain uranium oxide (UO_2), metallic uranium, or mixed uranium and plutonium oxides ($\text{UO}_2 + \text{PuO}_2$ or “MOX”). The majority of commercial power reactors use uranium oxide, produced by chemically converting enriched UF_6 to UO_2 powder. The powder is pressed, sintered, and machined to form cylindrical pellets, which are then stacked inside metal tubes called fuel pins or fuel rods. A number of fuel pins are gathered to form fuel bundles or assemblies.

2.5 Reactor Operation

The core of a nuclear power plant that produces 1000 MW of electricity typically contains approxi-

mately 75 metric tons (165,000 pounds) of enriched uranium oxide fuel. Approximately 25 metric tons (55,000 pounds) of fresh fuel is required each year. Thus, each fuel assembly spends about 3 years in the reactor core. The heat from the fission reactions is used to generate steam for running power turbines. Most of the power output comes from the fission of U-235. However, some U-238 is transformed into Pu-239, the fission of which provides about a third of the reactor’s energy output.

2.6 Interim Storage

When first withdrawn from the reactor, reactor fuel is both thermally hot and highly radioactive because of fission products. Spent nuclear fuel (SNF) assemblies are kept in cooling pools at the reactor site until their radioactivity and heat generation have declined to safer levels. This generally requires several months. (However, in the United States, spent fuel has remained in cooling ponds for decades awaiting the development of a permanent national disposal site.)

2.7 Reprocessing

Irradiated fuel from reactors may be reprocessed to remove fission products and to recover the uranium and plutonium. The fuel rods are chopped up and dissolved in acid, and the various components are separated chemically. In a true cycle, the uranium and plutonium are recycled for use as fuel. However, some countries (including the United States) employ the so-called once-through or open fuel cycle, which omits this step.

2.8 Stabilization and Immobilization

The high-level wastes from the reprocessing of SNF are mostly in liquid form. The liquid waste is calcined to produce a dry powder, which is then converted to a glass (a process called vitrification) or ceramic. (In a once-through cycle, this step is omitted because spent uranium oxide fuel is already in ceramic form.) The glass or ceramic is then sealed in corrosion-resistant canisters for long-term disposal.

2.9 Final Disposal

High-level wastes resulting from reprocessing (and SNF from a once-through cycle) must be deposited for hundreds or thousands of years to allow the radionuclides to decay to safe levels. Most proposals for final deposition envision some kind of deep

geological repository in stable crystalline rock, salt deposits, or clay.

3. RADIONUCLIDES IN NUCLEAR WASTE

More than 2500 nuclides have been discovered or created. Most of these are too short-lived to be of practical interest. The Index of Radioisotopes, published by the U.S. Nuclear Regulatory Commission (NRC), lists some 770 isotopes. Those of concern in radioactive wastes generally fall into one of three categories:

- *Activation products.* Substances used as coolants, moderators, and reactor structures can become radioactive by absorbing neutrons. These are usually light- or medium-mass nuclides.
- *Medium-mass fission products.* The fission of uranium produces more than 200 different radionuclides, covering about a third of the periodic table from copper to the lanthanides. Fission fragments

tend to have neutron-rich nuclei with mass numbers from approximately 70 to 160, with peak production occurring near mass numbers 95 and 135. Most fission products are beta or beta/gamma emitters.

- *Heavy nuclei.* Nuclei having mass numbers between 232 and 246 are used as fuels for nuclear fission or are produced by neutron capture in fission reactions. Most of these are long-lived alpha emitters.

Table I lists some radionuclides of concern in nuclear wastes. These radionuclides are also discussed in detail in what follows.

3.1 Americium

Americium (Am, atomic number 95) is an artificial actinide. Its most important isotope is americium-241 (half-life 432.7 years), which is produced by the beta-decay of plutonium-241. Americium-241 is used commercially in applications such as smoke detectors, thickness gauges, and fluid level gauges. Although americium-241 has very high radiotoxicity, it is not very mobile in the environment. It decays by alpha emission to neptunium-237.

TABLE I

Some Radionuclides in Nuclear Wastes

Nuclide	Z	Half-life	Decay mode	Decay energy (MeV)	Relative radiotoxicity
Americium-241	95	432.7 years	α	5.4857	Very high
Carbon-14	6	5,730 years	β^-	0.157	Moderate
Cesium-137	55	30.17 years	β^- , γ	1.17	High
Cobalt-60	27	5.272 years	β^- , γ	2.824	High
Hydrogen-3	1	12.26 years	β^-	0.0186	Low
Iodine-129	53	16×10^6 years	β^-	0.193	Low
Iodine-131	53	8040 days	β^-	0.971	High
Iodine-133	53	20.8 hours	β^- , γ	1.76	Moderate
Krypton-85	36	10.72 years	β^- , γ	0.687	Low
Neptunium-237	93	2.14×10^6 years	α	4.957	Very high
Plutonium-238	94	87.7 years	α	5.593	Very high
Plutonium-239	94	24,110 years	α	5.244	Very high
Plutonium-242	94	376,000 years	α	4.983	Very high
Radium-226	88	1,600 years	α , γ	4.870	Very high
Radium-228	88	5.75 years	β^-	0.045	Very high
Radon-222	86	3.82 days	α	5.590	Moderate
Ruthenium-106	44	372.2 days	β^-	0.039	High
Strontium-90	38	29 years	β^-	0.546	High
Technitium-99	43	213,000 years	β^-	0.293	Low
Thorium-232	90	1.40×10^{10} years	α	4.081	High
Uranium-234	92	2.46×10^5 years	α	4.776	Very high
Uranium-235	92	7.04×10^8 years	α , γ	4.679	Low
Uranium-238	92	4.46×10^9 years	α	4.039	Low

3.2 Carbon

Carbon (C, atomic number 6) occurs in nature predominantly as the stable isotopes carbon-12 (98.89%) and carbon-13 (1.1%). Its most important radioactive isotope is carbon-14, a weak beta-emitter having a half-life of 5730 years. Carbon-14 is formed naturally in the upper atmosphere by the action of cosmic rays on nitrogen. It is also produced by neutron capture in nuclear explosions and fission reactors if nitrogen-containing materials are present. Carbon-14 can replace carbon-12 in biological systems. The radiotoxicity of carbon-14 is moderate.

3.3 Cesium

Cesium (Cs, atomic number 55) is an alkali metal having 40 isotopes, ranging in mass from 112 to 151. Of these, the fission product cesium-137 (half-life 30.17 years) is of greatest concern. Cesium-137 is one of the nuclides most responsible for the radioactivity of used reactor fuel and high-level waste (with the other being strontium-90). Cesium-137 is highly radiotoxic. It is widely used in moisture, liquid level, and thickness gauges as well as to treat cancer.

3.4 Cobalt

The transition metal cobalt (Co, atomic number 27) is brittle, hard, and magnetic. Naturally occurring cobalt, consisting of the isotope cobalt-59, is used to make very hard and corrosion-resistant alloys. The absorption of a neutron converts cobalt-59 to radioactive cobalt-60 (half-life 5.272 years), which decays by the emission of a beta particle and a high-energy gamma ray to nickel-60.

3.5 Hydrogen

Hydrogen (H, atomic number 1) has two stable isotopes: hydrogen-1 and hydrogen-2 (deuterium). Hydrogen-3 (tritium) is a low-energy beta-emitter having a half-life of 12.26 years, produced naturally by the action of cosmic rays in the upper atmosphere. Tritium also occurs as a product in a small fraction of fission reactions and is produced commercially by the irradiation of lithium-6. Tritium is most likely to be found bound to oxygen to form tritiated water, which has chemical properties that are virtually identical to those of ordinary water. Tritium is usually taken into the body by inhalation and skin absorption, although ingestion is also possible. Blood

flow distributes the tritiated water evenly throughout the soft tissues of the body. Its radiotoxicity is low.

3.6 Iodine

Iodine (I, atomic number 53) is a halogen occurring below bromine in the periodic table. Two radioactive isotopes of iodine are of importance in nuclear wastes. Iodine-129 (half-life 1.6×10^7 years) and iodine-131 (half-life 8.040 days) occur as fission products. Both are beta-emitters that tend to concentrate in the thyroid. The radiotoxicity of iodine-129 is low; the radiotoxicity of iodine-131 is high. Other isotopes (e.g., iodine-123, iodine-124) are commonly used in medical imaging and treatment.

3.7 Krypton

Krypton (Kr, atomic number 36) is one of the noble gases, which as a group have little tendency to react chemically with other elements. A total of 32 isotopes of krypton have been identified, having atomic masses ranging from 69 to 100. Six of these are stable. The fission product krypton-85 is a beta-emitter having a half-life of 10.72 years. Because it is chemically inert, krypton is difficult to separate from process streams. Although krypton is not absorbed appreciably in body tissues, it may be inhaled and affect the lungs. Its radiotoxicity is low.

3.8 Neptunium

Neptunium (Np, atomic number 93) is the first transuranic element. Neptunium-239 (half-life 2.36 days) occurs as an intermediate in the formation of plutonium from uranium but is too short-lived to be of concern in most wastes. In contrast, neptunium-237 (half-life 2.14×10^6 years) is a major source of alpha radiation in spent fuel and high-level wastes. It is formed by alpha decay of americium-241 or beta decay of uranium-238. The radiotoxicity of neptunium-237 is very high.

3.9 Plutonium

Isotopes of plutonium (Pu, atomic number 94) are generated by neutron capture in uranium-238 or thorium-232. Plutonium-239 (half-life 24,110 years) is of greatest concern because it is fissionable. Other important isotopes are plutonium-238 (half-life 87.7 years), plutonium-240 (half-life 6564 years), plutonium-241 (half-life 13 years), and plutonium-242

(half-life 3.76×10^5 years). All of these isotopes have very high radiotoxicity.

3.10 Radium

Radium (Ra, atomic number 88) is a naturally occurring radioactive element. Radium-228 (half-life 5.76 years) is a decay product of thorium; radium-226 (half-life 1600 years) is a decay product of uranium. Radium-226 and its daughters (including radon-222) are the major sources of radioactivity in uranium mine tailings. Chemically, radium is an alkaline earth metal like calcium and tends to concentrate in the bone. The radiotoxicity of radium is very high.

3.11 Radon

Radon (Rn, atomic number 86) is a radioactive noble gas produced by the decay of radium. The two most important isotopes are the alpha-emitters radon-220 (half-life 54.6 s) and radon-222 (half-life 3.82 days). Being a gas, radon is highly mobile; in some areas, radon is a concern because it can accumulate in buildings. Both radon-220 and radon-222 are moderately radiotoxic.

3.12 Ruthenium

Ruthenium (Ru, atomic number 44) belongs to the platinum family of metals. It has seven stable natural isotopes. The most important radioactive isotope is ruthenium-106 (half-life 372.2 days), a fission product that decays by low-energy beta emission to rhodium-106 (a high-energy beta/gamma emitter with half-life of 30 s). Ruthenium-106 has high radiotoxicity.

3.13 Strontium

Strontium (Sr, atomic number 38) is an alkaline earth metal lying below calcium in the periodic table. Four stable isotopes are found in nature, and 16 radioactive isotopes have been produced artificially. Most of the radioactive isotopes of strontium are short-lived and of little interest. The exception is strontium-90 (half-life 29.1 years), a beta-emitter that is produced as a fission fragment. Much of the radioactivity of SNF results from strontium-90. Because its chemical properties resemble those of calcium, strontium-90 tends to concentrate in the bones. It is highly radiotoxic, decaying into yttrium-90 (itself an energetic beta-emitter of moderate radiotoxicity).

3.14 Technetium

The first artificially produced element, technetium (Tc, atomic number 43) occurs in nature only in minute amounts. At least 34 isotopes have been produced, having atomic masses ranging from 85 to 118; of these, the beta-emitter technetium-99 (half-life 2.13×10^5 years) is of special concern because it is produced in fission reactors, both as a fission fragment and as an activation product. Like iodine, technetium is absorbed by the thyroid. The radiotoxicity of technetium is relatively low.

3.15 Thorium

Thorium (Th, atomic number 90) is a naturally occurring radioactive element, several times more abundant than uranium, which has been considered as the basis for an alternative fuel cycle. The most important natural isotope is thorium-232 (half-life 14.0×10^9 years), which is weakly radioactive, decaying by alpha and gamma emission to radium-228. The radiotoxicity of thorium-232 is relatively low. Another isotope, thorium-230 (half-life 75,400 years) is produced naturally by the decay of uranium-234. The radiotoxicity of thorium-230 is very high. A third isotope, thorium-228 (half-life 1.9 years) also has very high radiotoxicity.

3.16 Uranium

Uranium (U, atomic number 92) is the heaviest element found in appreciable amounts in nature, making up 2 to 4 grams per metric ton of the earth's crust. Natural uranium consists of 99.283% by weight uranium-238 (half-life 4.468×10^9 years), 0.711% uranium-235 (half-life 7.04×10^8 years), and 0.005% uranium-234 (half-life 245,000 years). All three isotopes are alpha-emitters. Natural uranium has relatively low radiotoxicity.

4. CLASSIFICATION OF RADIOACTIVE WASTE

There is no generally accepted classification scheme for radioactive wastes. The following categories are commonly defined in the United States:

- Uranium mill tailings (UMTs) are the residues from the mining and milling of uranium ores. Most uranium ores in the United States contain uranium at a concentration of approximately 0.2% (2000 ppm)

by weight. Consequently, the production of 1 kg of uranium metal requires the excavation of 500 kg (or more) of ore. Milling creates additional solid and liquid wastes. Tailings generally contain relatively low concentrations of radioactive materials; however, because of the high volumes of tailings, their overall radioactivity can be high. The principal radionuclides of concern in UMTs are radium-226 (half-life 1600 years) and its daughter radon-222 (half-life 3.8 days).

- Depleted uranium, the major by-product from the enrichment process, is uranium containing less than 0.71% uranium-235. Depleted uranium currently has few civilian uses but could eventually be converted to fissile plutonium in breeder reactors. The U.S. Department of Energy (DOE) stores approximately 730,000 metric tons of depleted uranium, most of it as uranium hexafluoride.

- Low-level waste (LLW) has relatively low levels of radioactivity and contains little or no transuranic radioisotopes (i.e., radionuclides beyond uranium in the periodic table). LLW may be further divided into classes A, B, and C, with A being the least hazardous and C being the most hazardous. (Some LLW is classified as “greater than C.”) Mixed low-level waste (MLLW) contains both low-level radioactive and chemically hazardous substances. Although not shown in Fig. 1, LLW is generated at every step of the nuclear fuel cycle and may include used protective clothing, rags, paper, tools, filter elements, and so on.

- Spent nuclear fuel (SNF) is the irradiated fuel discharged from nuclear reactors. It contains fission products, plutonium, and unreacted uranium and is both thermally hot and highly radioactive.

- Transuranic waste (TRUW) contains alpha-emitting radionuclides having atomic numbers greater than 92. To qualify as TRUW under U.S. law, the radionuclides must possess half-lives greater than 20 years and be present in concentrations greater than 100 nCi/g. Although their radioactivity is relatively low, the long lives of the radioisotopes require long-term storage. Virtually all of the TRUW in the United States is the by-product of nuclear weapons production.

- High-level waste (HLW) includes the residues resulting from the reprocessing of spent reactor fuel and the production of nuclear weapons. HLW contains highly radioactive fission products and small amounts of plutonium isotopes. SNF may be considered as HLW in countries that do not reprocess fuel.

Under the foregoing classification scheme, commonly followed in the United States, wastes are categorized according to their source as well as their radioactivity. However, other countries tend to categorize wastes according to their radioactivity alone. For example, in the United Kingdom, solid wastes are classified as high-, intermediate-, low-, or very low-level wastes as follows:

- Very low-level waste (VLLW) can safely be discarded as ordinary refuse. To qualify, bulk waste must contain less than 400 kBq of beta/gamma activity per 0.1 m³ of material; individual items must contain less than 40 kBq of beta/gamma activity.

- Low-level waste (LLW) has higher levels of radioactivity than can be safely handled as ordinary refuse but not exceeding 4 GBq per metric ton of alpha radiation or 12 GBq per metric ton of beta/gamma radiation.

- Intermediate-level waste (ILW) has higher radioactivity levels than does LLW but not so high that heating becomes a concern.

- High-level waste (HLW) has such high levels of radioactivity that appreciable thermal heating occurs. Consequently, storage and processing facilities have to be designed to contain the radiation and dissipate the generated heat.

5. DISPOSAL OF LOW-LEVEL WASTE

LLW accounts for approximately 90% of the total volume of all wastes produced in the nuclear fuel cycle but only 1% of the total radioactivity. Various methods have been adopted for disposing of LLW, depending on the type of waste and its levels of radioactivity. For example, the NRC requires that Class A and B wastes be contained for at least 100 years, and Class C waste must be contained for 500 years. Combinations of natural and engineered barriers are used to prevent radioactivity from escaping to the environment. A major goal is to prevent the intrusion of ground water and surface water.

5.1 Shallow Land Burial

The cheapest way to dispose of LLW is to bury it in shallow trenches dug in stable soil (Fig. 2). A typical disposal trench may be 6 to 10 m deep, 30 m wide, and 100 to 300 m long. In some cases, the trench may be lined with a layer of clay, polymer, or concrete.

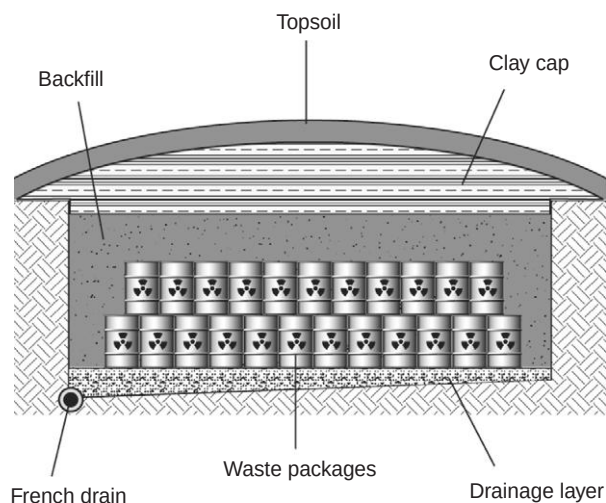


FIGURE 2 Shallow land burial for LLW.

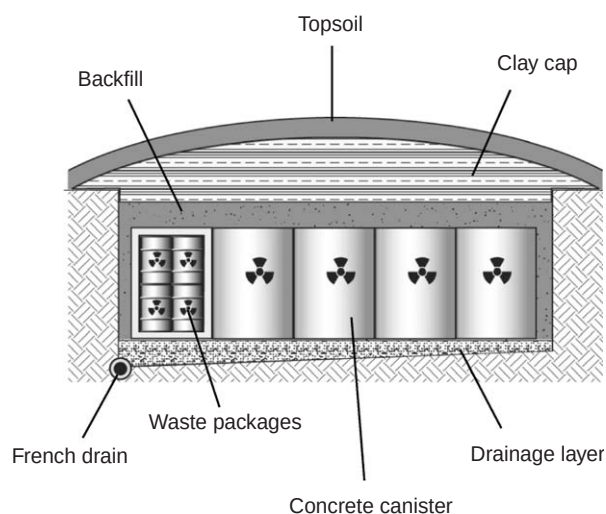


FIGURE 3 Modular concrete canister disposal.

The floor of the trench is sloped and covered with a layer of sand, gravel, or crushed stone. A drainage system keeps the waste containers dry. Once the LLW containers are in place, the trench is backfilled with sand or earth, which is then compacted and capped with a layer of clay to reduce water infiltration. The clay cap may be stabilized by a layer of topsoil planted with vegetation. Shallow land burial (SLB) has been used throughout the world.

5.2 Modular Concrete Canister

Modular concrete canister (MCC) disposal (Fig. 3) differs from SLB in that the waste containers are first placed inside concrete canisters, which are then

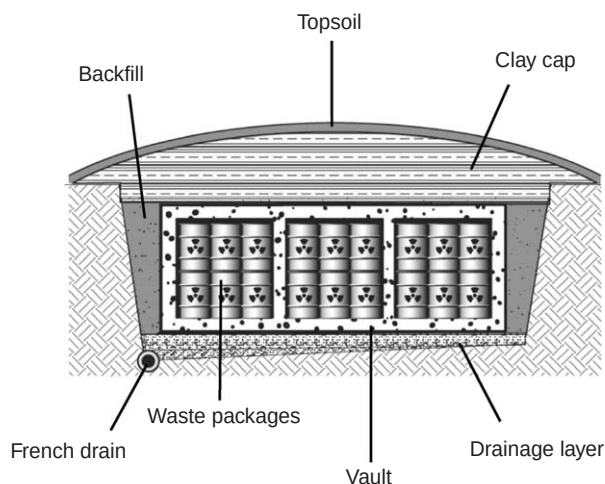


FIGURE 4 Belowground vault disposal.

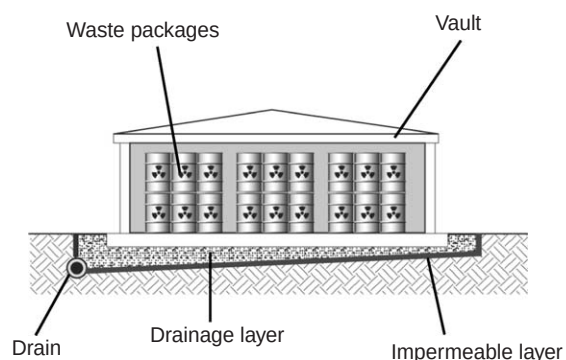


FIGURE 5 Aboveground vault disposal.

buried in trenches. The canisters serve as an additional barrier against the release of radioactivity.

5.3 Belowground Vault

In belowground vault (BGV) disposal (Fig. 4), a specially designed concrete, metal, or polymer vault is placed in a trench below ground level. The floor of the trench is typically sloped and equipped with drains. After the waste containers are placed in the vault, the trench is backfilled with permeable sand or gravel and capped with impermeable clay, topsoil, and vegetation. BGV disposal is used in Japan at the Low-Level Waste Disposal Centre at Rokkasho.

5.4 Aboveground Vault

An aboveground vault (AGV) is a concrete, metal, or masonry structure built above ground level (Fig. 5). A vault may be a simple reinforced concrete box with a poured concrete roof, or it may be a more elaborate

structure resembling a large warehouse. The principal advantage of an AGV is that it allows easy recovery of wastes for further processing or reuse. Its principal disadvantage is that the structure is not protected by earth from the effects of weather, intrusions, or accidents. Consequently, an AGV is often considered more for temporary storage than for permanent disposal. AGVs have been used at the West Valley Demonstration Project in New York and at the Point Lepreau and Bruce sites in Canada.

5.5 Earth-Mounded Concrete Bunker

An earth-mounded concrete bunker (EMCB) is a structure built partly below ground and partly above ground (Fig. 6). Wastes are segregated according to activity; higher activity (class B/C) wastes are embedded in concrete vaults or “monoliths” below ground level, and lower activity (class A) wastes are then placed atop the vaults. The wastes are covered with an impermeable layer of clay and a layer of topsoil to form a mound or tumulus. This method is used in France at the Centre de La Manche and the Centre de l’Aube as well as in Canada at the Whiteshell Nuclear Research Establishment.

5.6 Shaft Disposal

In a typical shaft disposal (SD) scheme, a hole some 3 m in diameter is sunk 40 or 50 m into the ground (but well above the water table). The walls of the shaft may be lined with clay, metal, or concrete, and the bottom of the shaft may be left unlined for drainage.

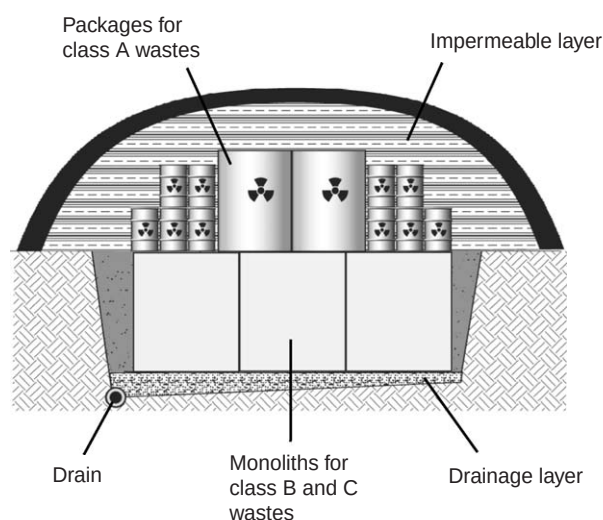


FIGURE 6 Earth-mounded concrete bunker (or tumulus) disposal of LLW.

After the waste containers have been stacked in the shaft, it is backfilled with sand or clay and capped with concrete. Shaft disposal of LLWs has been used at a number of sites in the United States and Canada.

5.7 Deep Geological Disposal

Some countries have chosen to dispose of LLW and ILW in deep geological repositories similar to those planned for HLW. These repositories are to be located in stable geological media such as igneous rock, salt, or clay.

6. MANAGEMENT OF SPENT NUCLEAR FUEL

Nuclear fuel rods are highly radioactive when first withdrawn from the reactor core because they contain fission products, principally relatively short-lived radionuclides. For this reason, the SNF is transferred to an interim storage facility adjacent to the reactor to allow the radioactivity to decline to safer levels. The radioactivity of the spent fuel decreases by approximately 99% during a year of storage.

Most interim storage facilities keep SNF in pools of water, commonly known as cooling ponds. The water serves both as a radiation shield and a heat transfer medium. Where cooling ponds have become crowded, some facilities put the SNF in dry storage instead of building new ponds. Dry storage facilities may keep SNF in concrete or metal casks, reinforced concrete vaults, or dry wells. Cooling is accomplished by natural convection of ambient air.

As shown in Table II, there is no uniform policy regarding the management of SNF. Some countries (e.g., Japan, France, Russia, the United Kingdom) currently reprocess their SNF. Other countries (e.g.,

TABLE II

National Policies on Reprocessing of Spent Nuclear Fuel

Reprocess (closed cycle)
Belgium, China, France, Germany, India, Japan, The Netherlands, Russia, Switzerland, United Kingdom
Do not reprocess (once-through cycle)
Canada, Finland, Lithuania, Slovenia, Spain, South Africa, Sweden, United States
Undecided on reprocessing
Argentina, Brazil, Czech Republic, Hungary, Mexico, Pakistan, Slovenia, South Korea, Taiwan

Canada, Finland, the United States) do not reprocess. Still other countries (e.g., Czech Republic, South Korea, Taiwan) have deferred the decision on whether to reprocess, relying on storage of SNF for the near term.

7. DISPOSAL OF HIGH-LEVEL WASTES

7.1 Objectives

Disposal of HLW (and SNF treated as waste) will require the isolation of the wastes from the biosphere for thousands of years. The principal concern is the contamination of ground water that may be transported to the biosphere. The design of a typical repository is intended to accomplish all or most of the following objectives:

- Delay the intrusion of water into the repository.
- Delay the release of radionuclides from the waste package.
- Retard the dissolution of radionuclides.
- Limit the migration of any radionuclides that escape containment.
- Dilute any escaped radionuclides to safe levels.
- Permit retrieval of the SNF for reuse.

A variety of methods for the disposal of HLW/SNF have been studied over the years. Most employ multiple barriers to provide “defense in depth.” A typical scheme for HLW might involve solidification and vitrification of the waste. The vitrified wastes would then be placed inside one or more specially engineered metal, plastic, or ceramic containers. The containers would then be placed in cavities in a natural medium such as rock, clay, or salt and perhaps backfilled with a material such as bentonite, which swells when wet and slows the passage of water.

7.2 Seabed Disposal

In the past, low-level nuclear wastes have been dumped somewhat haphazardly into the ocean. Although most nations ended ocean disposal in the 1970s, several proposals have been put forth for the long-term disposal of HLW in the seabed. One approach would place the waste in specially designed canisters that would then be buried in deep sediments far from inhabited coasts. The sediments would provide a natural barrier to the movement of radionuclides, and the seawater would act as a dilution medium for any radionuclides that managed

to escape containment. Seabed disposal is still considered a viable option for the future. However, an international agreement, the 1993 London Convention, banned the disposal of all radioactive wastes at sea for 25 years, after which such disposal methods may be reconsidered every 25 years.

7.3 Very Deep Hole Disposal

Nuclear waste containers could be stacked in holes as deep as 10,000 m (~6 miles) underground, putting them out of reach of flowing ground water. Despite uncertainties about the behavior of both the waste containers and the surrounding rock at such depths, recent advances in drilling technology have renewed interest in this approach.

7.4 Space Disposal

The National Aeronautical and Space Administration (NASA) and the DOE studied the possibility of disposing of nuclear waste in space. Various schemes were considered, including placing waste in orbit around the earth, the sun, or the moon; shooting waste into the sun; placing waste on the moon; and sending waste into deep space. Although space disposal would place the waste permanently outside the biosphere, it would be expensive. Also, there is the possibility of an accident during launch, which could scatter waste over a large area.

7.5 Ice Sheet Disposal

Radioactive waste could be buried in or under the ice sheets that cover much of the continent of Antarctica. Because the ice can be several kilometers thick and there is no permanent human population on the continent, the waste would be effectively isolated from human contact. However, ice disposal would be expensive because of the harsh climate and the need to transport the waste to a remote site. Climate changes also could cause melting of the polar ice, releasing the waste. Moreover, the Antarctic Treaty of 1959 prohibits the disposal of radioactive waste on the continent.

7.6 Deep Geological Disposal

Most current plans for the long-term disposal of nuclear waste involve the excavation of deep repositories in stable geological formations such as igneous rock, salt deposits, and clay. Each of these has advantages and disadvantages.

TABLE III
Plans for High-Level Waste Repositories

Country	Geological medium	Estimated opening	Status
Belgium	Clay	2035 or later	Searching for site
Canada	Granite	2035 or later	Reviewing repository concept
Finland	Crystalline bedrock	2020	Site selected (Olkiluoto)
France	Granite or clay	2020 or later	Developing repository concept
Germany	Salt	Unknown	Moratorium on development
Japan	Granite or sedimentary rock	2030 or later	Searching for site
Russia	Not selected	Unknown	Searching for site
Sweden	Crystalline rock	2020	Searching for site
Switzerland	Crystalline rock or clay	2020 or later	Searching for site
United Kingdom	Not selected	After 2040	Delaying decision until 2040
United States	Welded tuff	2010	Site selected (Yucca Mountain)

7.6.1 Igneous Rock

Intact igneous rock, such as granite or tuff, tolerates relatively high temperatures and resists the flow of water. On the other hand, hard rock may fracture, permitting the intrusion of water through the resulting fissures. Both the Yucca Mountain repository in the United States and the SFR repository in Sweden are sited in igneous rock.

7.6.2 Salt Deposits

The presence of a stable salt deposit indicates that little or no flowing water has been present at the site for geological ages. Left undisturbed, the salt should remain dry for thousands of years. Salt is easily mined. It deforms plastically under the influence of heat and pressure, eventually flowing around the waste and sealing it in a tight, resolidified mass. Salt has very low permeability. It also has a relatively high thermal conductivity, which would aid in the dissipation of heat from the waste. On the other hand, if water found its way into the repository, the resulting salt brine would tend to be highly corrosive. The Waste Isolation Pilot Plant (WIPP) in the United States is located in bedded rock salt. The Gorleben repository in Germany is located in a salt dome.

7.6.3 Clay

By definition, clays are unconsolidated sediments consisting of particles less than 1/256 mm in diameter. (Clay particles are too small to feel gritty between the fingers or teeth.) Clays are commonly composed of hydrous aluminum silicates having a sheet-silicate structure, producing grains that tend to be flat. Clay resists the intrusion of water and retards the migration of radionuclides. Like salt, clay tends

to be self-sealing. Clay layers have been under consideration for repositories in Belgium, France, and Switzerland.

Table III summarizes the HLW repository plans of major nuclear countries. The nuclear waste management policies of the three leading nuclear countries (the United States, France, and Japan) are examined in more detail in the following sections.

8. RADIOACTIVE WASTE MANAGEMENT IN THE UNITED STATES

The United States has by far the largest civilian nuclear program in the world, with a total of 104 reactors having a power capacity of 97.5 GW. In 2000, these plants produced approximately 20% of the country's electricity.

8.1 History

The world's first nuclear reactor, built under the direction of Enrico Fermi, was put into operation at the University of Chicago on December 2, 1942. This led to the rapid development of an immense industrial complex for the wartime production of weapons-grade uranium and plutonium under the aegis of the Manhattan Project. On July 16, 1945—less than 3 years after Fermi's reactor achieved the first chain reaction—the first atomic bomb (a plutonium device) was exploded near Alamogordo, New Mexico. In August 1945, the United States dropped atomic bombs on Hiroshima and Nagasaki, Japan.

The U.S. Atomic Energy Commission (AEC) was established in 1946 and given the responsibility for both civilian and military nuclear programs. Initially, waste management was given lower priority than the development of nuclear weapons and (later) a nuclear power industry. It was widely expected that plants would be constructed to reprocess SNF. Uranium and plutonium would be reused; reprocessing wastes would be solidified and transferred to a permanent repository for disposal.

The National Academy of Sciences recommended in 1957, and again in 1970, that a federal waste repository be built in a stable bedded salt deposit. Accordingly, in 1970, the AEC investigated salt beds near Lyons, Kansas, as a possible site for a repository. However, the Lyons site was rejected in 1972 because of concerns about its geology and hydrology as well as the large number of boreholes left by previous oil and gas drilling in the area. In 1974, the AEC began exploring salt beds in southeastern New Mexico. The next year, a site for the WIPP was selected.

In 1974, the AEC was abolished by Congress, and its responsibilities were transferred to two new agencies. The Energy Research and Development Administration (ERDA) was created to develop and promote nuclear power, and the NRC was created to regulate civilian nuclear programs. In 1976, the U.S. Environmental Protection Agency (EPA) was given the responsibility of setting radiation protection standards for nuclear wastes. The next year, the ERDA was replaced by a new cabinet-level agency, the DOE. The new department was also given responsibility for nuclear weapons programs.

Given these organizational changes, it is perhaps not surprising that the executive branch of the federal government made little progress in formulating a coherent policy for nuclear waste management during the 1970s. Congress has repeatedly stepped in and formulated policies for waste disposal. Some of the major pieces of legislation include the following:

- The Low-Level Radioactive Waste Policy Act of 1980 requires each state to be responsible for disposing of LLW generated within its borders, either by providing a suitable disposal site or by entering into compacts with other states to develop a shared site. The disposal sites were to be in operation by 1986.
- The Nuclear Waste Policy Act (NWPA) of 1982 created the Office of Civilian Radioactive Waste Management in the DOE and gave it the responsibility for developing a system to manage SNF and HLW. The act also directed the secretary of energy to

nominate five sites for study as possible SNF/HLW repositories. Based on these initial studies, the secretary was to recommend (by January 1, 1985) three of the five sites for intensive study. Finally, one site was to be recommended by March 31, 1987. The goal was to have the first repository in operation by 1998.

- The Low-Level Radioactive Waste Policy Amendments Act (LLRWPA) of 1985 was crafted to address difficulties arising out of the Low-Level Waste Policy Act of 1980. The formation of state compacts and the development of disposal facilities had taken longer than was expected. The LLRWPA kept open the three existing commercial disposal sites and set regulations for the disposal of LLW.

- The Nuclear Waste Policy Amendments Act of 1987 specified that Yucca Mountain be the sole site for study as an HLW repository instead of having three site studies proceeding in parallel. The amendment also canceled the requirement from the NWPA of 1982 that a second repository be created. The selection of Yucca Mountain was a political decision, not a scientific or technical one; local opposition to other sites made them less attractive politically.

- The Waste Isolation Pilot Plant Land Withdrawal Act of 1992 transferred control of the WIPP site from the Department of Interior to the DOE. The act also gave the EPA responsibility for ensuring compliance with environmental regulations.

- The Energy Policy Act of 1992 was a large and comprehensive piece of legislation covering a broad range of energy policy issues, including energy efficiency and conservation, integrated resource planning, alternative fuels, and competitive restructuring of wholesale electricity markets. The act required that the EPA set standards for the Yucca Mountain repository based on recommendations by the National Academy of Sciences. The NRC was to assume long-term responsibility for the repository.

8.2 Low-Level Waste Repositories

Six commercial LLW disposal sites have been created in the United States: Barnwell, North Carolina; Beatty, Nevada; Richland, Washington; Maxey Flats, Kentucky; Sheffield, Illinois; and West Valley, New York. All of these sites have used some form of SLB. Because of leaks, the latter three sites have been closed.

Low-level military wastes have been stored at a number of sites around the country, including Amarillo, Texas; Hanford, Washington; the Idaho National Engineering and Environmental Laboratory;

Los Alamos, New Mexico; Paducah, Kentucky; and Savannah River, South Carolina.

8.3 Waste Isolation Pilot Plant

The WIPP is an underground waste repository located in bedded salt formations in southeastern New Mexico, approximately 40 km (25 miles) east of Carlsbad. The WIPP repository was created for the permanent disposal of TRUW resulting from nuclear weapons production. It is the first deep geological repository designed specifically for long-term disposal of nuclear waste.

The waste destined for the WIPP site consists of a variety of materials contaminated with transuranic isotopes, including items such as protective clothing, laboratory equipment, used reagents, and solidified sludge. The principal radionuclides are plutonium, americium, and neptunium; small amounts of thorium, uranium, and other nuclides may also be present. Although the overall radioactivity of TRUW is much lower than that of SNF and other high-level waste, TRUW contains long-lived radionuclides that remain hazardous for thousands of years.

The WIPP repository is located in the 240 million-year-old salt beds of the Permian Salado Formation. The salt beds are the remains of a shallow saltwater sea that once covered what is now west Texas and southeastern New Mexico. The Salado Formation is 200 to 400 m thick, consisting mostly of halite (rock salt) containing thin layers (0.1–1.0 m thick) of nonhalite materials such as anhydrite (calcium sulfate) and potash.

The WIPP repository is situated 658 m (2160 feet) below the surface. It comprises a network of excavations connected to the surface by four vertical shafts: a salt-handling shaft, a TRUW shaft, an air intake shaft, and an air exhaust shaft. Underground roadways called drifts lead from the shafts to the waste disposal areas. TRUW is placed in rectangular rooms that are 4 m high, 10 m wide, and 91 m long. Adjacent rooms are separated by pillars of unexcavated salt that are 31.5 m wide. A group of seven rooms is called a panel. A total of eight panels are planned, but only one had been excavated by the time the first shipment of waste arrived in 1999. Other tunnels and rooms have been excavated for various scientific experiments.

8.4 Yucca Mountain

The Yucca Mountain repository is located in the Mojave Desert of southern Nevada, approximately

120 km (75 miles) northwest of Las Vegas. The repository is intended for the permanent disposal of SNF and HLW from both civilian and defense sources.

As planned, the Yucca Mountain repository will hold SNF that is currently kept in 131 temporary storage facilities in the United States. By 2002, these sites contained some 45,000 metric tons of SNF from commercial reactors and 2500 metric tons from defense reactors. Another 380,000 m³ of liquid HLW from nuclear weapons production was stored at sites operated by the DOE; this waste is to be stabilized in the form of borosilicate glass.

Yucca Mountain is composed of tuff, a material formed from volcanic ash and dust. Some of the tuff is welded, meaning that the fine glassy particles of the tuff have been fused together by natural heat and pressure. The repository would be sited at a depth of approximately 350 m in a layer of welded tuff known as the Topopah Spring Member. Above and below the repository are layers of nonwelded tuffs that contain few fractures.

Yucca Mountain is located in an arid region, where the mean precipitation is less than 20 cm per year and an estimated 95% of the water runs off or evaporates before it can seep into the ground. The water table lies approximately 300 m (1000 feet) below the level of the repository. Ground water tends to flow to the south toward the Amargosa Desert, which lies in an isolated hydrological basin having no outlet to the ocean.

In July 2002, the U.S. Congress voted in favor of a resolution recommending that Yucca Mountain be considered for development as a permanent repository. At that time, no final design for the repository had been adopted, but two cases were being considered.

The base or statutory case would provide an eventual storage capacity of 70,000 metric tons of heavy metal (MTHM). The base case repository would store waste containers in 58 tunnels called emplacement drifts, each 5.5 m (18 feet) in diameter. The emplacement drifts would be spaced 81 m (266 feet) apart, and their total length would be 56,222 m (184,455 feet). An additional 12,988 m (42,612 feet) of excavations would be created for access to the repository, with another 6542 m (21,463 feet) of excavations being created for the exhaust main.

The full inventory case would have a capacity of 97,000 MTHM. It would consist of 90 emplacement drifts having a total length of 74,214 m (243,484 feet). The design could be expanded to accommodate as much as 119,000 MTHM if such expansion ever becomes necessary.

9. RADIOACTIVE WASTE MANAGEMENT IN FRANCE

France has the second largest civilian nuclear program in the world, with 59 reactors in operation. In 2000, French nuclear plants had a total capacity of 63.2 GW and produced 76% of the country's power, a higher percentage than in any other country.

9.1 Low- and Intermediate-Level Wastes

Disposal of nuclear waste is the responsibility of ANDRA (Agence Nationale pour la Gestion des Déchets Radioactifs). Waste is classified according to half-life as well as activity. Waste having half-life less than 30 years is designated as "short-lived" (SL), and waste having half-life greater than 30 years is designated as "long-lived" (LL). Four activity levels are defined: very low-level waste (VLLW), low-level waste (LLW), intermediate-level waste (ILW), and high-level waste (HLW).

The Centre de la Manche Disposal Facility, located near Cherbourg in Normandy, operated from 1969 to 1994 as a disposal site for LLW and ILW. Initially, waste was placed in simple trenches. Over the years, the EMCB concept was developed. The la Manche facility holds nearly 530,000 m³ of waste and is no longer accepting waste shipments.

In 1992, the Centre de l'Aube Disposal Facility near Troyes began accepting low- and intermediate-level waste (LILW-SL). The Centre de l'Aube has a capacity of 1 million m³, expandable to 2 million m³.

9.2 Management of Spent Nuclear Fuel

SNF is kept in storage ponds for a year to allow its radioactivity to decline to safe levels, after which it is transported to a reprocessing plant.

France has reprocessed its own nuclear fuel since 1958 and has provided reprocessing services for Belgium, Germany, Japan, The Netherlands, and Switzerland. The French nuclear fuel company COGEMA (Compagnie Générale des Matières Nucléaires) currently operates two reprocessing plants at La Hague, west of Cherbourg in Normandy.

9.3 High-Level Wastes

HLW from reprocessing is vitrified at the reprocessing plant in La Hague. Vitrified HLW from French sources is currently stored at La Hague awaiting

disposal. Vitrified HLW from foreign sources is returned to the countries of origin.

Since 1991, ANDRA and CEA (Commissariat à l'Energie Atomique) have been studying several options for long-lived radioactive waste, including deep geological disposal in clay or granite. A decision to begin construction on a deep repository could be made as early as 2006.

10. RADIOACTIVE WASTE MANAGEMENT IN JAPAN

Japan operates 53 nuclear power plants, giving it the third-largest civilian nuclear program. In 2000, Japan's nuclear plants had a total capacity of 43.5 GW, producing about a third of the country's electrical power.

10.1 Low-Level Wastes

Since 1992, Japan has operated the Low-Level Waste Disposal Centre at Rokkasho at the northern end of Honshu Island. The BGV is the preferred method of disposal. Ultimately, the facility at Rokkasho will hold 600,000 m³ of waste.

10.2 Management of Spent Nuclear Fuel

SNF is stored in cooling ponds for 4 years or longer before being sent on for reprocessing.

Since 1977, Japan has operated a small reprocessing plant at Tokai Mura for SNF; however, most SNF is shipped to France or Great Britain for reprocessing. Japan is constructing its own reprocessing plant at the Rokkasho nuclear complex. It is expected to begin full-scale operations in 2005.

10.3 High-Level Wastes

Vitrified nuclear wastes are returned to Japan from the reprocessor. High-level vitrified wastes are sent to the High-Level Radioactive Waste Storage Facility at Rokkasho to be stored for 30 to 50 years.

In 2000, the Nuclear Waste Management Organization (NUMO) was created to develop a disposal repository in granite or sedimentary rock. Construction of the repository is expected to begin around 2030.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Fuel: Design and Fabrication • Nuclear Fuel Reprocessing • Nuclear Fusion Reactors • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Occupational Health Risks in Nuclear Power • Public Reaction to Nuclear Power Siting and Disposal • Radiation, Risks and Health Impacts of

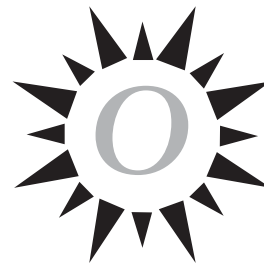
Further Reading

- Bodansky, D. (1996). "Nuclear Energy: Principles, Practices, and Prospects." American Institute of Physics, Woodbury, NY.
- Murray, R. L. (1994). "Understanding Radioactive Waste." 4th ed. Battelle Press, Columbus, OH.
- National Research Council. (1996). "The Waste Isolation Pilot Plant: A Potential Solution for the Disposal of Transuranic Waste." National Academy Press, Washington, DC.
- Wilson, P. D. (ed.). (1996). "The Nuclear Fuel Cycle: From Ore to Waste." Oxford University Press, Oxford, UK.

Obstacles to Energy Efficiency

MARILYN A. BROWN

Oak Ridge National Laboratory
Oak Ridge, Tennessee, United States



-
1. Background
 2. The Energy Efficiency Gap
 3. Obstacles to Energy Efficiency
 4. The Role of Government
 5. Conclusions

Glossary

efficiency gap The difference between the actual level of investment in energy efficiency and the higher level that would be cost-beneficial, from the points of view of consumers and society.

externalities The goods or services that people consume as by-products of other people's activities; such goods and services are "external" to market transactions and are therefore unpriced.

market barriers The obstacles that contribute to the slow diffusion and adoption of energy-efficient innovations.

market failures The conditions of a market that violate one or more of the neoclassical economic assumptions (e.g., rational behavior, costless transactions, and perfect information) that define an ideal market.

Homes, offices, factories, and vehicles rarely incorporate the full complement of available cost-effective, energy-efficient technologies, despite the sizable costs that inefficient designs impose on consumers and society. There is compelling evidence that vast cost-effective opportunities exist for improvements in energy efficiency, implementation of which could significantly curtail the rapid growth in energy consumption in the United States, resulting in a more competitive economy, cleaner air, lower greenhouse gas emissions, and enhanced national security. These opportunities exist because of large-scale market failures and barriers that prevent markets from operating efficiently, slowing the diffusion of innovations. Assessments of energy policies and programs suggest that public interventions can overcome many of these obstacles.

1. BACKGROUND

Examination of energy trends following the 1973/1974 oil embargo highlights the great strides in energy efficiency that have made the U.S. economy much less energy intensive today than it was in 1970. Despite this progress, an energy efficiency gap still exists. Many investments that appear to be cost-effective to the consumer are not being made in energy-efficient technologies and practices. This would not be surprising if a relatively small number of such investments were being foregone, or if only a small portion of future energy growth would be impacted by making these investments. However, numerous studies indicate the continued existence of a sizable untapped reservoir of highly cost-effective investments that could significantly cut U.S. energy use. If these investments were to occur, the result would be a more competitive economy, cleaner air, lower greenhouse gas emissions, and less dependence on foreign oil.

If an energy-efficient technology is cost-effective, why does more of it not just happen? If individuals or businesses can make money from energy efficiency, why are they all not just doing so? Why does the cost-effectiveness of an energy-efficient option not simply propel it to commercial success? These questions can be answered, in large part, by describing the range of obstacles that prevents the full exploitation of energy efficiency. Markets for energy-efficient technologies are not unique in their imperfections. Other products and services face obstacles that impede their adoption, even when their consumer economics appear to be favorable. Conditions hindering cost-effective investments in energy efficiency are particularly noteworthy because of their widespread environmental, national security, and macroeconomic repercussions.

This article begins with a series of individual technology case studies that present compelling evidence of an efficiency gap. Next, the market

failures and institutional barriers that explain the existence of this gap are described. The sector differences in market failures and barriers are then characterized. This lays the groundwork for discussing potential government roles and the pros and cons of different policies and programs that could tackle barriers to energy efficiency.

2. THE ENERGY EFFICIENCY GAP

The term “efficiency gap” refers to the difference between the actual level of investment in energy efficiency and the higher level that would be cost-beneficial from the points of view of consumers and society. It is difficult to quantify the exact magnitude of this gap. One approach to characterizing its size is through modeling and scenario development. This typically involves enumerating on a technology-by-technology basis the difference between current practice and best practice, where best practice is defined as the utilization of the energy technologies with the lowest life cycle costs. Keeping in mind the natural rate of turnover of new equipment and consumer purchases, the size of the lost opportunities or the gap that exists can then be estimated. Using this technology-based accounting approach, a study by five U.S. Department of Energy laboratories concluded that removing obstacles to energy efficiency through policy interventions initiated in the year 2000 could have reduced the forecasted U.S. energy consumption in 2010 by 10%. By 2020, the reduction could grow to nearly 20%.

Modeling exercises and policy scenarios are constructive methods for bounding the potential for improved energy efficiency, but their theoretical nature may not persuade skeptics. The “natural experiment” provided by the California electricity crisis provides a basis for characterizing the efficiency gap that is grounded in actual events. Beginning in June of 2000, the State of California experienced an unprecedented electricity supply crisis. Electricity prices escalated as electricity demand exceeded supply, and rolling blackouts occurred in June 2000 and in January, March, and May of 2001. Energy efficiency and conservation played a prominent role in avoiding further problems that were forecast for the summer of 2001. In response to a vigorous energy conservation campaign, Californians averaged a 10% reduction in their electricity consumption during peak hours in the summer of 2001, compared with the previous summer. Weather data show similar temperature patterns across the two

summers, thus the reduction in usage is attributable to investments in energy efficiency and altered lifestyle choices. This consumer response was prompted by a number of utility and state-run programs that promoted no-cost energy conservation actions, installation of low-cost energy efficiency measures, and investments in energy-efficient equipment, appliances, and building shell retrofits. Because a minority of households contributed the bulk of the conservation gains, the 10% reduction in consumption is probably an underestimate of the full efficiency gap.

Another way to characterize the efficiency gap is through case studies of individual technology choices. Many different case studies could be cited showing that consumers and businesses often choose not to purchase highly cost-effective energy technology. Skeptics might conclude that these technologies must carry hidden liabilities that are difficult to identify and quantify. The following case studies, however, describe efficiency gaps associated with energy-efficient options that are clearly superior to the technologies being replaced, and with no significant “hidden costs” to the consumer.

The most energy-efficient substitute for the Edison light bulb is fluorescent lighting. Fluorescents use one-quarter to one-third as much energy as incandescent bulbs, and they burn 10 to 15 times longer. A study of lighting investments using data from the Environmental Protection Agency (EPA) Green Lights Program has shown that there is a large potential for profitable energy-saving investments in fluorescent lighting that is untapped because of impediments based on organizational decision making. A key impediment is capital rationing—when a business unit is allocated a quota of capital to cover all improvements, including new product designs and cost-cutting production improvements such as energy efficiency. Another obstacle is the lack of organizational rewards for energy managers who reduce utility bills. Often the energy savings are not even measured. All fluorescents need a ballast to operate. The ballast is a device that alters the electric current flowing through the tube, which activates the gas inside, causing the tube to emit light. The most efficient ballasts today are electronic. Electronic ballasts superseded two prior technologies, the original “standard” ballast and the more efficient magnetic ballast. Efficient magnetic ballasts were commercially available as early as 1976. They were a well-tested technology, with performance characteristics equal to or better than standard ballasts by the early 1980s. By 1987, five states—including

California and New York—had prohibited the sale of standard ballasts. But the remaining three-quarters of the population chose standard ballasts over efficient ballasts by a ratio of 10:1, even though the efficient magnetic ballast typically paid back its investment in less than 2 years. The time required to establish retail distribution service networks and to gain consumer confidence were key obstacles to the adoption of the efficient magnetic ballast, as is the case with many innovations. (Since 1990, federal standards have prohibited the sale of the standard ballast.)

The purchase of home appliances illustrates additional impediments to energy efficiency. In one study by Lawrence Berkeley National Laboratory (LBNL) scientists in the 1980s, consumers were given a choice in stores throughout the United States of two refrigerators that were identical in all respects except two: energy efficiency and price. The energy-efficient model (which saved 410 kilowatt-hours (kWh) per year, more than 25% of energy usage) cost \$60 more than the standard model. The energy-efficient model was highly cost-effective in almost all locations of the country. In most regions, it provided an annual return on investment of about 50%. In spite of these favorable economics, which were easily observed by the purchaser, more than half of all purchasers chose the inefficient model. The higher purchase price of the efficient model was the principal barrier to its purchase. Yet the consumer would save more money by investing in the more expensive, efficient refrigerator than by investing a similar amount in any standard investment mechanisms such as a certificate of deposit, U.S. Treasury Bill, or stock portfolio.

To enable the use of remote controls for televisions in the early 1990s, it became necessary for televisions to consume some amount of power continuously. Typical televisions with remote controls at that time used 5 to 7 W of standby power for that purpose. The Energy Star television program was able to reduce these power losses by requiring that televisions qualifying for the Energy Star label must reduce standby power to 3 W or less, a savings of roughly 50%. The resulting price increase had a payback period of 1 to 2 years for consumers. Because this savings was no more than a few dollars a year per television, there was no public outcry for manufacturers to deliver the improvement. At the same time, the aggregate savings to the nation of widespread market penetration were significant. Through the labeling program, the lack of consumer interest could be overcome. Most major manufacturers now offer such televi-

sions, and several of them have reduced standby losses to 0.5 W.

Case studies of manufacturing facilities illustrate that barriers to energy efficiency also exist in the industrial sector. Motor systems represent the largest single end use of electricity in the American economy—23% of U.S. electricity consumption—and they present a very substantial energy efficiency potential. The results of a recent market assessment involving on-site surveys of 265 industrial facilities document that technologies offering a simple payback of 3 years or less can typically save businesses 11–18% of the energy used to drive motors. The Department of Energy (DOE) Best Practices program conducts audits, demonstrations, and technical assistance to encourage the use of proved, cost-effective technologies to improve industrial motor systems. Monitoring and validation of energy use data from these activities confirm the profitability of these investments, underscoring the large gap between current practice and economically smart investments. Limited information, expertise, and capital cause this gap.

3. OBSTACLES TO ENERGY EFFICIENCY

Numerous obstacles, including market failures and barriers, contribute to the efficiency gap (Table I). “Market failures” occur when there are flaws in the way markets operate. They are conditions of a market that violate one or more of the neoclassical economic assumptions (e.g., rational behavior, costless transactions, and perfect information) that define an ideal market. Market failures can be caused by (1) misplaced incentives, (2) distortionary fiscal and regulatory policies, (3) unpriced costs such as air

TABLE I
Market Failures and Barriers Inhibiting Energy Efficiency

Market failures (flaws in the way markets operate)	Market barriers (other factors preventing rapid adoption)
Misplaced incentives	Low priority of energy issues
Distortionary fiscal and regulatory policies	Capital market barriers
Unpriced costs	Incomplete markets for energy efficiency
Unpriced benefits	
Costly, imperfect and asymmetric information	

pollution, (4) unpriced benefits such as education, training, and technological advances, and (5) costly, imperfect, and asymmetric information. By failing to account for such market imperfections, assessments of energy policies based on neoclassical economic models often underestimate their potential benefits. It is widely argued by neoclassical economists that the existence of market failures is a prerequisite for market intervention. However, the existence of such failures is also seen as an insufficient justification for government involvement. Feasible, low-cost policies must be available that can eliminate or compensate for these market failures.

“Market barriers” refer to other obstacles that contribute to the slow diffusion and adoption of energy-efficient innovations. To the extent that it is in society’s best interest to use its energy more efficiently and to reduce emissions from fossil fuel combustion, it is important to understand the full range of obstacles to energy-efficient technologies. These barriers include (1) the low priority of energy issues among consumers, (2) capital market imperfections, and (3) incomplete markets for energy-efficient features and products.

3.1 Market Failures

3.1.1 *Misplaced Incentives*

Misplaced incentives inhibit energy-efficient investments in each sector of the economy. This problem occurs when an “intermediary” has the authority to act on behalf of a consumer, but does not fully reflect the consumer’s best interests. Examples of this failure are numerous:

- Architects, engineers, and builders select equipment, duct systems, windows, and lighting for the individuals who occupy buildings.
- Landlords purchase appliances and equipment for tenants who pay the energy bills.
- Industrial buyers choose technologies that manufacturers use in their factories.
- Specialists write product specifications for military purchases.
- Fleet managers select the vehicles to be used by drivers.
- New car buyers determine the pool of vehicles available to buyers of used cars.

The involvement of intermediaries in the purchase of energy technologies limits the ultimate consumer’s role in decision making and leads to an underemphasis on life cycle costs. This is exacerbated in the construction industry, where typical fee struc-

tures for architects and engineers cause incentives to be distorted in ways that penalize efficiency. Additional first costs are typically needed to enable the installation of superior heating, ventilation, and air-conditioning systems that reduce operating costs. These additional expenditures beyond the typical “rule-of-thumb” equipment sizing used by most engineers result in a net penalty for designers of efficient systems. Even though this type of fee structure has been strongly discouraged in the United States, both the designer and procurer of design services still generally base their fee negotiation on percentage-of-cost curves.

The landlord–tenant relationship is a classic example of misplaced incentives. If a landlord buys the energy-using equipment and the tenants pay the energy bills, the landlord is not incentivized to invest in efficient equipment unless the tenants are aware of and express their self-interest. Thus, the circumstance that favors the efficient use of equipment (when the tenants pay the utility bills) leads to a disincentive for the purchase of energy-efficient equipment. The case that favors the purchase of efficient equipment (when the landlord pays the utility bills) leads to a disincentive for the tenants to use energy efficiently. About 90% of all households in multifamily buildings are renters, which makes misplaced incentives a major obstacle to energy efficiency in this segment of the housing market.

Misplaced incentives can also involve significant time lag. For instance, new car purchasers have a dominant influence on the design decisions of automakers. Yet they are not representative of the entire driving public, many of whom purchase their vehicles secondhand. In particular, compared to the average driver, new car purchasers are substantially wealthier, which skews their purchase preferences away from fuel economy and toward ride quality, power, and safety. Similar obstacles to energy efficiency exist in the secondary markets for appliances and homes.

3.1.2 *Distortionary Fiscal and Regulatory Policies*

Distortionary fiscal and regulatory policies can also restrain the use of efficient energy technologies. A range of these distortionary policies was recently identified in an analysis of projects aimed at installing distributed generation. Distributed generation is modular electric power located close to the energy consumer, including photovoltaics, gas turbines, and fuels cells. Regulatory barriers to these new technologies include prohibitions against their

use and state-to-state variations in environmental permitting requirements that result in significant burdens to project developers. Tariff barriers include excessive backup and standby charges as well as utility buy-back rates that do not provide credit for on-peak electricity production. An example of a distortionary fiscal policy is the tax treatment of capital versus operating costs. United States tax rules require capital costs for commercial buildings and other investments to be depreciated over many years, whereas operating costs can be fully deducted from taxable income. Because efficient technologies typically cost more than standard equipment on a first-cost basis, this tax code penalizes efficiency. Similarly, many states are uneven in their sales tax policies. For instance, most states charge sales tax on residential energy-saving devices but not on residential fuels and electricity.

Electricity pricing policies of state legislatures and regulatory commissions also prevent markets from operating efficiently and subdue incentives for energy efficiency. The price of electricity in most retail markets today is not based on time of use. It therefore does not reflect the real-time costs of electricity production, which can vary by a factor of 10 within a single day. Because most customers buy electricity as they always have (under time-invariant prices that are set months or years ahead of actual use), consumers are not responsive to the price volatility of wholesale electricity. Time-of-use pricing would encourage customers to use energy more efficiently during high-price periods. Metering, communications, and computing technologies are needed to support such dynamic pricing and voluntary-load-reduction programs. The cost of designing and installing this infrastructure represents another potential barrier to real-time pricing. Although this might be cost prohibitive for some customers, the cost of this infrastructure would be affordable for large commercial and industrial customers.

3.1.3 Unpriced Costs

A strong case can be made that fossil fuels (and other energy resources) are underpriced because market prices do not take full account of a variety of social costs associated with their use. Fossil energy using today's conversion technologies produces a variety of unpriced environmental costs (i.e., negative externalities), including air, water, and land pollution and greenhouse gas emissions. For example, more than 90% of U.S. emissions of carbon dioxide, NO_x , and SO_2 results from the combustion of fossil fuels (see Fig. 1). Other types of energy-related environmental

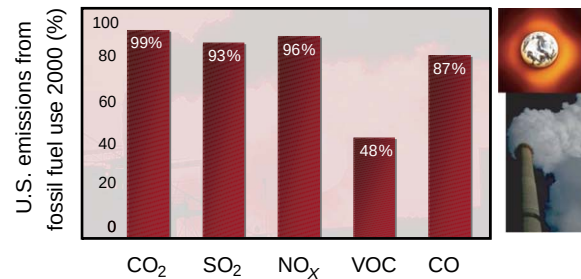


FIGURE 1 Energy-linked emissions as a percentage of total U.S. emissions. Sources: Environmental Protection Agency, *National Air Pollutant Emissions Trends Report*, 1998, EPA-454/R-00-0002-March 2000; and Energy Information Agency, *Emissions of Greenhouse Gases in the U.S., 2000*, DOE/EIA-0573(00), October 2000.

impacts include water and ground contamination from petroleum leaks from vessels and underground storage tanks. Nuclear power production generates fission by-products that require long-term storage. Sources of renewable energy present specific environmental concerns as well, such as fish kills from hydropower, avian mortality from wind power, and particulate air pollution from biomass combustion. In addition to imposing unpriced environmental costs, oil markets also inflict unpriced national security costs. Under relatively tight market conditions, the physical concentration of oil reserves in a relatively small number of countries generates the potential for physical and price-setting supply disruptions. These market conditions impose national security costs by reducing foreign policy flexibility and complicating military strategy, especially during periods of rising oil demand and tightening world markets.

Some of these environmental costs are being addressed through regulation. Absent regulation that effectively internalizes the full costs of environmental impacts and national security, externality costs exist. When the market does not value these costs, more energy is consumed than is socially optimal. Negative externalities associated with fossil energy combustion can be “internalized” through policy interventions. Domestic carbon trading is one example of such a policy. The idea of the carbon trading system is to create fossil fuel prices that better reflect the full cost of fossil fuel consumption, causing consumers to make decisions that take into account the complete cost of the resource. These higher prices should cause consumers to use less fossil fuel. At the same time, the government-collected carbon permit revenues can be recycled to consumers.

Existing environmental control costs are embedded in some energy costs. For instance, the U.S. Environmental Protection Agency regulations enforcing the Clean Air Act and other federal legislation impose control costs on the marginal emitter of criteria pollutants such as SO₂ and NO_x. However, not all existing fossil generators incur operating cost penalties. Furthermore, there are several emissions produced by fossil fuel combustion that are not capped today. These include carbon, mercury, and smaller (2.5 µm) particulates. No costs are currently included to account for damages from these pollutants. Because energy prices do not include the full cost of environmental externalities, they understate the societal cost of fossil energy use based on today's combustion technologies. This market imperfection can be addressed through public policies and programs that bring market choices more in line with full societal costs.

3.1.4 Unpriced Benefits

Unpriced benefits also dampen the energy productivity of the economy. A public good is a good or service that has two principal characteristics. First, one person's consumption of it does not reduce the amount of it available for other people to consume. This characteristic is called "inexhaustibility." Second, once such a good is provided, it is difficult to exclude other people from consuming it, a characteristic called "nonexcludability." Because public goods are unpriced, markets tend to underproduce them.

The public goods nature of education, training, and research is an important rationale for government support. Investments by employers in creating a well-educated, highly trained workforce, for instance, are dampened because of the firm's inability to ensure that the employee will work long enough for that firm so as to repay its costs. The difficulties of selecting and installing new energy-efficient equipment compared to the simplicity of buying energy may prohibit many cost-effective investments from being realized. This is a particularly strong barrier for small and medium-sized enterprises. In many firms (especially with the current trend toward lean firms), there is often a shortage of trained technical personnel who understand and can explain the ability of energy-efficient technologies to generate a stream of cost savings that more than pay for any up-front installation premium. Government programs that pay university engineering faculty and students to conduct energy audits of industrial plants can overcome this barrier by training the next generation of energy professionals while delivering

energy diagnostics and audit recommendations to plant managers.

Investing in research and development often results in benefits that cannot be fully captured by the sponsoring private entities. Although benefits might accrue to society at large, individual firms cannot realize the full economic benefits of their research and development investments. Further, companies that absorb the market risk of introducing new technologies are generally unable to reap the full benefits of their trailblazing. The payback from advances in energy-efficient technologies is not only experienced by the sponsoring company, but also flows to the public, to the company's competitors, and to other parts of the economy. The problem is especially pronounced when an industry is highly fragmented, as is the case with the construction and homebuilding industries, in which the largest 400 homebuilders serve just 40% of the residential market. Fragmentation increases the number of potential external beneficiaries of the research and development results. In highly fragmented markets, individual companies also may lack the scale of resources needed for advanced research and development investments.

The risk of innovation leakage and exploitation by competing firms puts pressure on firms to invest for quick returns. Technology innovation is typically a longer term investment fraught with risks to the investor. The result is an underinvestment in research and development from the standpoint of overall benefits to society. The problem is particularly difficult in the newly restructured electric sector, for which research and development funding has decreased dramatically. Companies will not fund the optimal societal level of basic research and development of new technologies because many of the benefits of such research will flow to their competitors and to other parts of the economy. This is true of many industries, and is one of the main rationales for government-funded long-term, precompetitive research in industries that have a vital role in the U.S. economy.

The Council of Economic Advisers estimates that the private returns from research and development are 20–30%, whereas social returns (including energy security and environmental benefits) are 50% or higher. This gap limits the extent to which the private sector can supplant a government role in maintaining nationally beneficial research and development. With the development of international markets, fragmentation is growing and industry's priorities are shifting further away from basic and

applied research and toward near-term product development and process enhancements. Over the past several decades, business spending on applied research has dropped to 15% of overall company research and development spending, whereas basic research has dropped to just 2%. In addition, corporate investments in energy research and development, in particular, are down significantly.

3.1.5 Costly, Imperfect, and Asymmetric Information

Insufficient and incorrect information can also be a major obstacle to energy efficiency. Reliable information about product price and quality allows firms to identify the least costly means of production, and gives consumers the option of selecting goods and services that best suit their needs. Yet information about energy-efficient options is often incomplete, unavailable, expensive, and difficult to obtain. With such information deficiencies, investments in energy efficiency are unlikely. Asymmetric information can also cause problems; when one party to a transaction has substantially more product information, market opportunities will tilt toward the entity with the information advantage.

It is difficult to learn about the performance and costs of energy-efficient technologies and practices because the benefits are often not directly observable. For example, residential consumers get a monthly electricity bill that provides no breakdown of individual end-uses, making it difficult to assess the benefits of efficient appliances, televisions, and other products. Appliance energy labels help, as does “Energy Star” branding, but studies have shown that many consumers do not understand them. Efficiency choices tend to be bundled with other purchases. For instance, the price paid for different levels of vehicle fuel economy is buried in base prices or in the price of complete subsystems such as engines. Further, efficiency differences are coupled with substantive differences in other critical consumer attributes such as acceleration performance, level of luxury, and vehicle handling. Reliable information on the marginal cost of fuel economy may be obtainable, but the effort required for an individual consumer to secure such information could be prohibitive.

Decision-making complexities are another source of imperfect information that can confound consumers and inhibit “rational” decision making. Even while recognizing the importance of life cycle calculations, consumers often fall back to simpler first-cost rules of thumb. Although some energy-efficient products can compete on a first-cost

basis, many of them cannot. Properly trading off energy savings versus higher purchase prices involves comparing the time-discounted value of the energy savings with the present cost of the equipment—a calculation that can be difficult for purchasers to understand and compute. This is one of the reasons builders generally minimize first costs, believing (probably correctly) that the higher cost of more efficient equipment will not be capitalized into a higher resale value for the building. The complexities of decision making are one form of transaction cost. Note, however, that if consumers were extremely concerned about life cycle energy savings and were determined to base their purchasing decisions on them, product manufacturers would have a strong incentive to provide consumers with better information about energy efficiency and with clearer trade-offs. It can be argued that the lack of such information and choices is simply the consequence of consumer disinterest in using energy efficiently—the first of several market barriers.

3.2 Market Barriers

3.2.1 Low Priority on Energy Issues

Energy efficiency is not a major concern for most consumers because energy costs are not high relative to the cost of many other goods and services. In addition, the negative externalities associated with the U.S. energy system are not well understood by the public. The result is that the public places a low priority on energy issues and energy efficiency opportunities. In turn, this reduces producer interest in providing energy-efficient products. In most cases, energy is a small part of the cost of owning and operating a building, a factory, or a car. Of course, there are exceptions. For low-income families, the cost of utilities to heat, cool, and provide other energy services in their homes can be a very significant part of their income—averaging 15% compared to 4% for the typical U.S. citizen. For energy-intensive industries such as aluminum and steel, energy can represent 10–25% of their production costs. Many companies in these more energy-intensive firms have decided to incorporate energy management as a key corporate strategy. These companies have made impressive gains in efficiency. A means of sharing successful strategies and best practices can help improve overall industry performance and extend benefits to smaller, less profitable firms.

Because energy costs are typically small on an individual basis, it is easy (and rational) for

consumers to ignore them in the face of information-gathering and transaction costs. However, the potential energy emissions savings can be important when summed across all consumers. This is one reason why government agencies such as the EPA and DOE work directly with manufacturers to improve product efficiency. A little work to influence the source of mass-produced products can pay off in significant efficiency improvements and emissions reductions that rapidly propagate through the economy due to falling production costs as market shares increase.

3.2.2 Capital Market Barriers

Capital market barriers can inhibit efficiency purchases. Although, in theory, firms might be expected to borrow capital any time a profitable investment opportunity presents, in practice, firms often ration capital—that is, they impose internal limits on capital investment. The result is that mandatory investments (e.g., required by environmental or health regulations) and those that are most central to the firms' product line often are made first.

Different energy producers and consumers have varying access to financial capital, and at different rates of interest. In general, energy suppliers can obtain capital at lower interest rates than can energy consumers, resulting in an "interest rate gap." Differences in these borrowing rates may reflect differences in the knowledge base of lenders about the likely performance of investments as well as the financial risk of the potential borrower. At one extreme, electric and gas utilities are able to borrow money at low interest rates. At the other extreme, low-income households may have essentially no ability to borrow funds, resulting in an essentially infinite discount rate for investments in energy efficiency. The broader market for energy efficiency (including residential, commercial, and industrial consumers) faces interest rates available for efficiency purchases that are also much higher than the utility cost of capital. In addition, it has been shown that firms typically establish internal hurdle rates for energy efficiency investments that are higher than the cost of capital to the firm. Information gaps, institutional barriers, short time horizons, and nonseparability of energy equipment all contribute to this gap, and each is amenable to policy interventions that could move the rates down toward auto loan, mortgage, and opportunity costs. Energy prices, as a component of the profitability of an investment, are also subject to large fluctuations. The uncertainty about future energy prices, especially in

the short term, seems to be an important barrier. Such uncertainties often lead to higher perceived risks, and therefore to more stringent investment criteria and a higher hurdle rate. An important reason for high hurdle rates is capital availability. Capital rationing is often used within firms as an allocation means for investments, leading to hurdle rates that are much higher than the cost of capital, especially for small projects.

3.2.3 Incomplete Markets for Energy Efficiency

Incomplete markets for energy efficiency are often a serious obstacle. Energy efficiency is generally purchased as an attribute of a product intended to provide some other service. Fuel economy in automobiles, for example, is one of a large number of features that come in a package for each make and model. If higher fuel economy were treated as an optional item, available at a higher price, then consumers would have a choice of efficiency levels. But such a separate option does not presently exist. Similar circumstances in the commercial buildings market constrain investor choices. The complexity of design, construction, and operation of commercial buildings makes it difficult to characterize the extent that a particular building is energy efficient. As a result of this host of market failures and barriers, the discount rate that consumers appear to use in making many energy efficiency decisions is higher than the interest rate at which consumers could borrow money. This discount rate gap has been widely observed in the literature and is reflected in some key energy models, such as the Energy Information Administration's National Energy Modeling System.

3.3 Sectoral Differences in Market Failures and Barriers

Each end-use sector functions differently in the U.S. energy marketplace. One of the reasons for this variation is the distinct supply chain and market structure for delivering new technologies and products in different sectors. Residential and commercial building technologies are shaped by thousands of building contractors and architectural and engineering firms, whereas the automotive industry is dominated by a few manufacturers. As a result, the principal causes of energy inefficiencies in manufacturing and transportation are not the same as the causes of inefficiencies in homes and office buildings, although there are some similarities. For example, in the manufacturing sector, investing in cost-effective

energy efficiency measures (which cut operating costs and therefore increase profits) is hampered by a common preference to invest resources to increase output and market share as a preferred route to expanding profits. Because energy uses are so diverse in the industrial sector, “one-size-fits-all” products rarely play a large role. Achieving widespread use of technological breakthroughs can require dissemination of technical documentation and technical assistance to help firms tailor the technology to their particular production processes. In the building sector, information gaps prevent the energy-efficient features of buildings from being capitalized into real estate prices. This is partly due to the lack of widely adopted building energy-rating systems. These information gaps are less characteristic of the transportation sector, in which fuel economy is well understood in terms of miles per gallon.

The end-use sectors also differ in terms of their ability to respond to changing energy prices. This is partly due to the varying longevity of the equipment that is used. For example, cars, lighting, and air conditioners turn over more quickly than do large industrial boilers, which can operate for the better part of a century. Stock turnover must be taken into account when assessing the potential market impacts of policies and programs. There are also differences in fuel flexibility. The U.S. transportation system today is relatively fuel inflexible, being primarily dependent on petroleum, whereas energy use in buildings and industry is more fuel flexible.

The vast differences in the research and development capability of the sectors also influence their ability to respond quickly to changing energy prices and market signals. The private sector as a whole spends more than \$180 billion per year on research and development, dwarfing the government expenditure on all nondefense technology research and development. Of the private-sector research and development expenditure, the automobile manufacturers stand out—Ford alone spends more than \$8 billion per year in research and development. Next comes the rest of the industrial sector. Here manufacturers account for a majority of research and development expenditures. In the buildings sector, the construction industry has virtually no indigenous research and development. The Council on Competitiveness estimates that the construction industry spends less than 0.2% of its sales on research and development, far less than the 3.5% that other industries spend on average.

Finally, each of the sectors is distinct in terms of the primary societal benefits from improved energy

efficiencies. Fuel economy in transportation is essential to improving air quality and protecting against oil price volatility. Energy productivity in industry is essential to economic competitiveness and pollution prevention. Energy efficiency in buildings makes housing more affordable on a life cycle basis; it is also important in ensuring electric grid stability and in reducing SO₂, NO_x, and particulate matter, because most of the energy consumed in buildings is fossil-generated electricity. As a result of these differences, public policies and programs need to be customized specifically to meet the needs of each sector.

4. THE ROLE OF GOVERNMENT

The existence of market failures and barriers that inhibit socially optimal levels of investment in energy efficiency is the primary reason for considering public policy interventions. In many instances, feasible, low-cost policies can be implemented that either eliminate or compensate for market imperfections and barriers, enabling markets to operate more efficiently to the benefit of society. In other instances, policies may not be feasible; they may not fully eliminate the targeted barrier or imperfection; or they may do so at costs that exceed the benefits.

To foster energy efficiency, transaction costs need to be reduced. Several of the problems discussed herein, particularly those related to information, can be viewed as transaction costs associated with energy decision making. Examples include the costs of gathering and processing information, making decisions, and designing and enforcing contracts relating to the purchase and installation of energy-using technology. These costs are real, in the sense that they must be borne by the consumer and should be included in the cost of the energy efficiency measure. A key question is whether there are institutional interventions that can reduce these costs for individual consumers. For example, the time and effort required to find a refrigerator with the maximum cost-effective level of energy efficiency could be significant.

Information programs (e.g., fuel economy Web sites) and technical assistance (e.g., industrial energy assessments) can help make up for incomplete information by reducing the consumer’s cost of acquiring and using needed information. They can also simplify decision making and can help consumers focus on energy issues that may seem small to an individual consumer but can be large from a national perspective.

The direct installation of home weatherization measures addresses the lack of access of low-income households to capital. Programs that support financing through energy services companies and utilities also address this barrier. More indirectly, but just as important, technology demonstrations provide financial markets with evidence of performance in the field, which is critical to reducing the cost of capital. For instance, electric utility companies in many regions have demonstrated the value of advanced lighting technologies through various incentive programs that have subsequently led to the widespread acceptance of these products and the increased availability of financing through mechanisms such as energy-saving performance contracts.

The public goods nature of research and development can be addressed through direct government funding. Great potential exists for public/private research and development partnerships to produce scientific breakthroughs and incremental technology enhancements that will produce new and improved products for the marketplace. If public policies reorient even a tiny fraction of the private sector's research expenditures and capability to address the nation's energy-related challenges, it could have an enormous impact. One way to reorient private-sector investments is through industry/government research and development alliances that involve joint technology roadmapping, collaborative development of advanced energy-efficient technologies, and cost sharing.

5. CONCLUSIONS

Homes, offices, factories, and vehicles rarely incorporate the full complement of cost-effective, energy-efficient technologies, despite the sizable costs that inefficient designs impose on consumers and nations. The evidence of this efficiency gap is compelling, and the reasons for it are numerous. Models of the U.S. energy system, statistical analyses, and case studies underscore the widespread existence of this gap and the array of market obstacles that cause it. By improving our understanding of these obstacles, it may be possible to fine-tune existing policies, invent entirely new types of policy interventions, and better explain their benefits to the public.

SEE ALSO THE FOLLOWING ARTICLES

Commercial Sector and Energy Use • Conservation Measures for Energy, History of • Consumption,

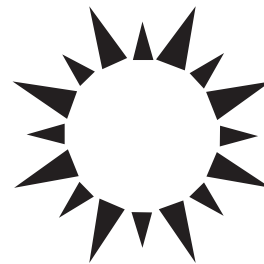
Energy, and the Environment • Discount Rates and Energy Efficiency Gap • Economics of Energy Efficiency • Energy Efficiency and Climate Change • Energy Efficiency, Taxonomic Overview • Industrial Energy Efficiency • Labels and Standards for Energy • Market Failures in Energy Markets • Material Efficiency and Energy Use • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

- Alderfer, R. B., Eldridge, M. M., and Starrs, T. J. (2000). "Making Connections: Case Studies of Interconnection Barriers and Their Impact on Distributed Power Project." NREL/SR-200-28053 (May 2000). National Renewable Energy Laboratory, Golden, Colorado.
- Brown, M. A., Levine, M. D., Short, W., and Koomey, J. G. (2001). Scenarios for a clean energy future. *Energy Policy* 29(14), 1179–1196.
- Council of Economic Advisers (CEA). (1995). "Supporting Research and Development to Promote Economic Growth: The Federal Government's Role." CEA, Washington, D.C.
- DeCanio, S. J. (1998). The efficiency paradox: Bureaucratic and organizational barriers to profitable energy-saving investments. *Energy Policy* 26(5), 441–454.
- Environmental Protection Agency (EPA). (1999). "Driving investment in Energy Efficiency: Energy Star and Other Voluntary Programs." EPA, Washington, D.C.
- Geller, H., and McGaraghan, S. (1996). "Successful Government–Industry Partnership: The U.S. Department of Energy's Role in Advancing Energy-Efficient Technologies." American Council for an Energy Efficient Economy, Washington, D.C.
- Greening, L. A., Sanstad, A. H., and McMahon, J. E. (1997). Effects of appliance standards on product price and attributes: An hedonic pricing model. *J. Regul. Econ.* 11(2), March
- Hassett, K. A., and Metcalf, G. E. (1993). Energy conservation investment, do consumers discount the future correctly? *Energy Policy* 21, 710–716.
- Hirst, E., and Brown, M. A. (1990). Closing the efficiency gap: Barriers to the efficient use of energy. *Resour. Conserv. Recycling* 3, 267–281.
- Horowitz, M. J., Lewis, K., and Coyle, A. (2000). Green Lights and Blue Sky: Market Transformation Revealed," 2000 ACEEE Summer Study on Energy Efficiency in Buildings, Vol. 6. American Council for an Energy-Efficient Economy, Washington, D.C.
- Jaffe, A. B., and Stavins, R. N. (1994). The Energy-efficiency gap. *Energy Policy* 22(10), 804–810.
- Koomey, J., Sanstad, A. H., and Shown, L. J. (1996). Energy-efficiency lighting: Market data, market imperfections, and policy success. *Contemp. Econ. Policy* 14(3), July.
- Levine, M. D., Koomey, J. G., McMahon, J. E., Sanstad, A., and Hirst, E. (1995). Energy efficiency policy and market failures. *Annu. Rev. Energy Environ.* 20, 535–555.
- Lovins, A. (1992). "Energy-Efficient Buildings Institutional Barriers and Opportunities." Strategic Issues Paper (December). E-Source. Boulder, Colorado.
- Lutzenhiser, L., Gossard, M. H., and Bender, S. (2002). "Crisis in Paradise: Understanding the Household Conservation Response to California's 2001 Energy Crisis." Proceedings of the Summer Study on Energy Efficiency in Buildings, pp. 8.153–8.166. American Council for an Energy-Efficient Economy, Washington, D.C.

- Martin, M., Tonn, B., Schmoyer, R., Overly, J., Schexnayder, S., and Johnson, D. (1999). "Industrial Assessment Center Program Impact Evaluation." ORNL/CON-473 (December). Oak Ridge National Laboratory, Oak Ridge, Tennessee.
- Meier, A., and Whittier, J. (1983). Consumer discount rates implied by purchases of energy-efficient refrigerators. *Energy* 8(12).
- Parfomak, P., and Lave, L. (1997). How many kilowatts are in a negawatt? Verifying ex post estimates of Utility conservation impacts at the regional scale. *Energy J.* 17(14).
- Romm, J. J. (1999). "Cool Companies: How the Best Companies Boost Profits and Productivity by Cutting Greenhouse Gas Emissions" Island press, Covelo, California.
- Ross, M. H. (1990). Capital budgeting practices of twelve large manufacturers. In "Advances in Business Financial Management" (P. Cooley, Ed.), pp. 157–170. Dryden Press, Chicago.
- Sanstad, A. H., Blumstein, C., and Stoft, S. E. (1995). How high are options values in energy efficiency investments? *Energy Policy* 23, 739–743.
- U.S. Department of Energy (DOE), Office of Policy and International Affairs. (1996). "Policies and Measures for Reducing Energy Related Greenhouse Gas Emissions." DOE/PO-0047 (July). DOE, Washington, D.C.
- U.S. Department of Energy (DOE), Office of Energy Efficiency and Renewable Energy. (2000). "Clean Energy Partnerships: A Decade of Success." DOE/EE-0213 (March). DOE, Washington, D.C.
- Xenergy, Inc. (1998). "United States Industrial Electric Motor Systems Market Opportunity Assessment." Xenergy, Inc., Burlington, Massachusetts. Available on the Internet at <http://www.motor.doe.gov/>.

Occupational Health Risks in Crude Oil and Natural Gas Extraction



JOEL M. HAIGHT

Pennsylvania State University

University Park, Pennsylvania, United States

-
1. Specific Health Risks
 2. General Health Risks during Specific Operations
 3. Summary

Glossary

blowout prevention (BOP) equipment Usually a hydraulically controlled rubber sealing device that is pressured to close around the drill string to seal the well bore and contain the contents and pressure of a well blowout.

cellar A pit dug during the preparation of the location to be used for drilling an oil or gas well. It is created to accommodate the stack of equipment making up the well head, including the blowout prevention equipment.

corrosion inhibitor Commercially available chemical mixtures designed to protect steel or other alloys against the effects of corrosion.

diverter A sealing device and collection of valves and pipes positioned near the top of the wellhead so that in the event of a sharp, rapid pressure increase (“kick”), workers can shut in the well and open the diverter valves to direct flow and pressure away from the wellhead.

heavy metals Any of a series of dense metallic elements naturally occurring in the earth’s crust, such as chromium, cadmium, lead, nickel, and iron.

hydrogen sulfide A toxic, colorless, heavier-than-air gas formed as a by-product of organic decomposition of sulfur compounds.

Raynaud’s syndrome A phenomenon in which there is an abnormal constriction of the blood vessels of the fingers characterized by numbness and white color and which is exacerbated by cold conditions.

risk A measure of the combination of the severity and likelihood of the occurrence of undesirable consequences.

well program A written description of the plans involved in drilling an oil or gas well. It contains information about the size and type of rig to be used, the depth and type of

well to be drilled, the expected formation pressures, drilling fluid design, casing design for the well, etc.

The extraction and production of crude oil and natural gas are complex operations that involve the potential for exposure to many health risks. This article discusses each of these individual health risks as well as the operational phases of drilling and extraction of these valuable natural resources. These risks include potential exposure to hydrogen sulfide, naturally occurring radioactive material, hydrocarbons, corrosion inhibitors, solvents, and ergonomic, biological, and physical hazards. This article also covers the drilling and completion of a well and surface production operations. This discussion includes treatment of the health risks associated with each phase of drilling and extraction operations, as well as the techniques used by industry to reduce the risk of exposure to these materials.

1. SPECIFIC HEALTH RISKS

1.1 Hydrogen Sulfide

One of the most common and toxic health risks associated with crude oil and natural gas extraction is exposure to hydrogen sulfide (H₂S). The oil and gas extraction industry has taken great pains over the years to minimize overexposure to H₂S. The industry in general has done much to design, build, and install equipment that safely contains the hydrogen sulfide; it has provided extensive education, training, and protective equipment for its employees and has installed alarm and detection systems in appropriate

locations. The effort appears to have accomplished its objective as the Bureau of Labor Statistics (BLS) reports that during the years 1993–2001 there was only one year (1999) in which there were more than 50 cases of respiratory illness reported (of which only a fraction would have been from hydrogen sulfide exposure). The BLS reports similar data for the mining industry.

H₂S is a colorless, heavier-than-air gas that possesses an odor of rotten eggs at low concentrations, but deadens the sense of smell at higher concentrations. It is highly corrosive to certain metals, is generally considered more toxic than carbon monoxide, and has a toxicity similar to that of hydrogen cyanide. When it is burned, it produces sulfur dioxide (SO₂), which is also considered toxic to humans. Once H₂S is released into the atmosphere, it remains there for approximately 18 h. Its decomposition and (or) transport in the atmosphere are thought to contribute to the formation of SO₂ and sulfuric acid, which contribute to the formation of acid rain.

The main exposure route for workers in the oil and gas industry is through inhalation. At lower concentrations, H₂S acts as a local irritant, producing irritation of the eyes and mucous membranes. At higher concentrations, it can be fatal (>600–800 ppm). According to a 1991 report by Amdur and co-workers, at these concentrations, H₂S stops respiration at the cellular level (cells do not obtain the oxygen they need). H₂S forms a complex bond with the iron in the mitochondrial cytochromes and stops aerobic metabolism. The transfer of oxygen from the blood to the cells cannot occur. This toxic action is similar to that of cyanide. H₂S toxicity is thought to be associated with carboxyhemoglobin or methemoglobinemia. A notable trait of H₂S is that it also produces olfactory fatigue by deadening the olfactory nerve. Thus, the strong rotten egg-like odor should not be used as a measure of the presence or absence of H₂S.

The effects from short-term, high-dose exposures are well documented and have, in some cases, led to death. Long-term, low-dose exposures are of concern as well. In 1981, Ludvich reported that symptoms such as respiratory irritation, sore throat, cough, shortness of breath, accumulation of fluid in the lungs, and even memory changes are possible. One positive aspect about exposure to H₂S is that it does not accumulate in the body. With chronic exposure at very low doses at or slightly above the odor threshold of 10 ppb, effects such as headaches, nausea, appetite loss, irritability, and fatigue are possible.

H₂S is a by-product of the organic decomposition of sulfur compounds and can be found in many other

settings in addition to the oil industry, such as mining, sewage treatment, and the paper industry. H₂S is found in solution in some crude oils and as a component of natural gas. It begins to flash out of solution as the oil and (or) natural gas are brought to the surface. In operations where the gas is intentionally separated from the oil as it is processed, the hydrogen sulfide will most often go with the gas fraction. This is accelerated if there is any additional heat input during extraction. Potential exposure to hydrogen sulfide can occur during drilling operations through exposure to recycled drilling muds, during exposure to the water fraction of sour crudes (sour is the name given to crude oils with a relatively higher sulfur and H₂S content), or during a blowout. Once the oil has been extracted and is being gathered, transported, and treated, there are many potential opportunities for exposure. Workers can be exposed when gauging tanks containing sour crude, when performing wellhead maintenance such as replacing valve packing, when pulling pumping rods, during entry into confined spaces such as valve pits and cellars, and during work in areas where leaks have occurred. Since H₂S is heavier than air, once a leak has occurred it tends to hug the ground and collect in low-lying areas, such as a cellar at a wellhead.

According to a 1981 study by Ludvich, since H₂S can react with some metals, corrosion and embrittlement are possible and leaks can result. H₂S causes stress cracking in certain metals used in vessels that contain or transport the crude oils containing it. It does so because the metals are subjected to high stress in corrosive environments. H₂S will also dissolve in water to form a weak acid that can cause pitting if it is also in the presence of oxygen or carbon dioxide. The most significant form of attack is the ability for H₂S to create hydrogen embrittlement. This is also known as sulfide stress cracking. In this case, brittle failure can occur rapidly. The significance of this phenomenon to human health is that H₂S creates an environment where it can cause significant leaks in containment vessels. Releasing H₂S to the work environment increases the risk of exposure. There are many alloys that have been developed and successfully used to contain hydrogen sulfide (e.g., Monel 400 and K-500, Incoloy 800 and 825, Inconel 600, 625, 718, and X-750, and Hastelloy C, G, and X). These and other appropriate alloys should be considered when designing a system to store, transport, or process sour crude oil or natural gas containing hydrogen sulfide.

Additional protection schemes used in minimizing the risk of exposure include the extensive use of

hydrogen sulfide detection equipment positioned around potential leak or collection areas. This detection equipment is connected to dedicated alarm systems that alert workers in the area that there has been a release or leak. On receiving this alarm, workers are able to don appropriate respiratory equipment and respond to stop the leak or safely evacuate the area before exposure occurs. The use of supplied or self-contained breathing-air respiratory protection is common in areas where exposure to hydrogen sulfide is possible. This can be anywhere in the oil field; however, the protection is generally offered near wellheads, in motor control centers, around oil treatment plants, and in any low-lying area that might require entry by workers.

1.2 Naturally Occurring Radioactive Material

The risk of exposure to naturally occurring radioactive material (NORM) must be considered in discussions of the oil and natural gas extraction industry. NORM is found to be associated with barium sulfate (BaSO_4 , also known as barite). This is sometimes a component found in drilling mud. In a 1994 report, Wilson explains that as drilling mud or fluids are circulated down hole, sometimes barite is lost in the rock formation as drilling progresses. It can later be brought to the surface in the production stream as the pressure profile in the well or in the formation changes over time. BaSO_4 will also absorb radium as it precipitates out on tubing, casing, and production equipment when the well is flowing. The barium sulfate is deposited on the piping and tubing along the production string. This way, it is exposed to any material contained in the production stream as it passes up the production string. Occasionally, the production stream will contain radium, which is then absorbed onto the BaSO_4 deposits. This can result in radioactive deposits that require protection for workers having to handle equipment or properly dispose of it on dismantling at the end of its useful life. NORM production also seems to increase as brine production increases in the well. In this case, the NORM may accumulate in the drill string.

Natural gas flows from the formation and to the surface at high flow rates under turbulent conditions. Wilson further explains that radon gas is often mixed with the natural gas as it flows into the production system. Radon's half-life is only ~ 3.8 days and it then decomposes to radioactive lead, Pb^{210} . The route the natural gas takes as it travels up the production tubing

and into the transportation and processing system is sometimes found to have a thin layer of Pb^{210} on the surfaces of the piping and tubing. This thin accumulation layer of ~ 0.004 in. can also affect the disposal of this equipment at the end of its useful life.

In the petroleum industry, it is estimated that the typical radium 226 concentration in oil sludge and scale is 155 pCi/g. It can be as high as 40,000 pCi/g. It is also estimated that the petroleum industry generates as much as 360,000 m^3 of NORM waste each year. Proper disposal is critical to reducing the risk of exposure.

According to Wilson, in oil and gas extraction, radiation exposure levels could be expected to be relatively low and to occur over a long period (career). The long-term effects of radiation exposure are difficult to determine. However, the delayed effects of radiation exposure are generally thought to be the precursor of cancer and genetic defects. In general, exposure to radiation at high enough levels is thought to damage or change cells in the body. Cell changes or damage can occur at almost any place in the body. Examples include the blood system (anemia due to loss of red cells and lower resistance from loss of white cells), the lymphatic system (signs of hemorrhaging in lymph nodes and weight loss in the spleen), the digestive tract (nausea and vomiting from damage to the lining of the organs in the digestive canal), and the lungs (lung cancer).

To reduce exposures, many considerations must be incorporated into well programs and drilling operations. The best approach to exposure reduction is through time, distance, and shielding. Exposure dose follows the inverse square law, so basically, doubling the distance to the source cuts the dose in half. Other interventions applied include the use of shielding, warning labels, and personal protective equipment, such as gloves, coveralls, and boots. Full-face air-purifying respirators with appropriate NORM cartridges are also worn. An active detection program is important to the protection of the workers. As radioactive material is detected, the workers should be removed from the exposure site as soon as possible to minimize the time of exposure. NORM-contaminated equipment as well as sludge, scale, and other deposits must be disposed of properly in designated disposal facilities.

1.3 Hydrocarbons

The hydrocarbons associated with oil and natural gas extraction are generally thought to be of lower health risk than many of the contaminants found in or

associated with oil and gas production. Methane, ethane, propane, and butane are found in a vapor state under atmospheric conditions. This makes them, in theory, the easiest to be exposed to through inhalation. According to the American Conference of Governmental Industrial Hygienists' 2001 Threshold Limit Values for Chemical Substances and Physical Agents Guide, methane and ethane are considered nothing more than simple asphyxiants and propane and butane have threshold limit values—time-weighted average (TLV–TWA) of 2500 and 800 ppm respectively. Thus, exposure to these fractions is considered a low health risk.

In general, as the carbon number of the crude oil fractions increases, the health risk also increases. The TLV–TWA for pentane, heptane, and octane ranges from 600 ppm for pentane (five carbons) to 300 ppm for octane (eight carbons). This is not true for the normal hexane isomer (*n*-hexane, six carbons). It is a higher health risk, with a TLV–TWA of 50 ppm. Exposure to these fractions would occur on contact exposure to liquid crude oil during spills or leaks; however, a portion of each of these components would likely volatilize depending on the atmospheric conditions. This would increase the potential for inhalation exposure to occur. It is important to point out that crude oil exposure would have to be carefully measured to determine whether a person was overexposed due to the combined exposure to several components or fractions at one time. Crude oil (all fractions still bound up in solution) does not necessarily behave in the body in same way as the individual fractions would.

Some of the potential health effects from hydrocarbon exposure include central nervous system depression, vomiting, and mucous membrane irritation. Since the TLV–TWAs are relatively high for most of the hydrocarbon fractions, the incidence of overexposure during oil and gas extraction is less common than exposure to hydrocarbon distillates and hydrocarbon contaminants present during oil and gas refining.

Although the crude oil is not refined, some of the health effects associated with these crude oil fractions and “reformates” can still generate health effects. Benzene is a carcinogenic component of petroleum and can represent a significant percentage of a barrel of crude oil. Benzene can attack the blood system if exposures are high enough. The inhalation route is considered the most critical and most common since benzene has a relatively high vapor pressure under ambient conditions. It can be absorbed through the skin, so excessive or long-term

skin contact exposure should be avoided. Fortunately, in crude oil, benzene is not present in its concentrated form during extraction operations. Therefore, this naturally reduces exposure potential.

Exposure control programs involve the application of leak minimization principles, working systems to manage confined space entry and hot work, detection equipment, exposure monitoring, and the use of respiratory protection as appropriate.

1.4 Heavy Metals (Chromium, Nickel, Iron, Lead, etc.)

Heavy metals are naturally occurring elements within the earth's crust. Those that would be of the greatest health risk during oil and gas extraction include lead, chromium, cadmium, nickel, and iron. There are many more. These metals are redistributed throughout the environment by geologic and biologic cycles. Heavy metals become potential exposure agents with an associated health risk to oil and gas workers during the extraction process. In 1991, Devereaux reported that the metals reside in the rock formations below the ground's surface. On drilling or extraction, heavy metals become entrained in or included with drilling fluids, oil, water, brine, and rock cuttings that are returned to the surface. Exposure to these metals in solid form would not be significant at this point, as it is not expected that workers would regularly ingest these materials. Exposure could occur if materials containing the metals became airborne either from an aerosolized mist in a spill or pressurized leak or from the creation of a dust cloud of dried material. This would make inhalation of the material possible. Although the risk of exposure to heavy metals during most extraction activities would generally be considered low, burning the oil, such as is sometimes done after a flow test at a production well, would increase the risk of exposure. Metal fumes would be created as the heat of the fire evaporates the metals and they then quickly condense to respirable-size particles. These particulates would become entrained in the hydrocarbon smoke and potentially be inhaled by workers in the area.

Metals are neither created nor destroyed in the body. Some are considered essential to the diet in trace amounts. When the “trace” amount is exceeded, not only are there immediate health risks to be concerned about, but some metals accumulate in the body unless intervention is implemented, such as administration of chelating agents. Chronic effects from exposure to heavy metals are usually of greater

concern than acute exposure. This concern is based on the accumulation of metals in the body over time. In the oil and gas industry, acute exposures are relatively easy to avoid, since in most circumstances the metals are in solid form. The potential health effects from chronic or acute exposure to heavy metals are many and varied, depending on the metal involved. Some of the effects include chronic pulmonary disease and chronic renal disease from cadmium; peripheral neuropathy and anemia induced by lead; dermatitis caused by nickel; and renal failure and hepatic cirrhosis due to iron (acute exposure). Chromium is a carcinogen and each of the metals listed above exhibits carcinogenic effects.

During drilling operations, the risk of exposure to heavy metals in drilling fluids, brine, and rock cuttings is higher than during extraction and production operations. This is true because the drilling system is much less closed than the production system. However, during drilling, the metals are in solid form, with the highest risk of exposure coming from exposure to metal-containing mist or dust as noted above. The exposure risk can be minimized here by wearing proper protective equipment, such as coveralls, gloves, and eye protection. It may also be appropriate to wear respiratory protection when exposure to metal-containing dust is possible. Proper maintenance of the drilling and circulating equipment will also help minimize the risk of exposure by reducing the likelihood for drilling fluid, brine, water, and oil leaks.

For production operations workers, the highest risk of exposure comes when burning metal-containing crude oil following a flow test. The risk of exposure to heavy metals can be minimized in this case by finding alternate means to dispose of the flow test oil or using appropriate respiratory protection such as an organic vapor air-purifying respirator. Alternate disposal for the oil may include discharge into a clean holding tank and redirecting the oil back into the production stream for processing.

1.5 Corrosion Inhibitors

The equipment used in oil and gas extraction is exposed to many different types of fluids and other materials. Some of these materials are not compatible with the steels and other alloys used in manufacturing oil and gas handling equipment. Corrosion can result and subsequent leakage of the production stream is then possible. The equipment is designed using materials that can withstand attack from many different contaminants and process streams under

many different conditions. To supplement this, many companies use corrosion inhibitors to help protect their equipment. These inhibitors are used as part of a regular, on-going treatment program that complements the equipment's preventive maintenance program. Many different corrosion inhibitors are safe and present little health risk. However, in years past, two materials that were used for this purpose contained either hydrazine or arsenic. These materials are no longer widely used, if at all, but residual material is or can still be left on equipment internals. Hydrazine and arsenic have been identified as carcinogens. The inorganic compound hydrazine has been shown to induce pulmonary tumors in sensitive strains of mice at high dose levels and arsenic has been associated with the development of lung and skin cancer and lymphomas in humans. This association has been shown for the trivalent, inorganic form of arsenic in drinking water.

Since these materials are no longer widely used in corrosion inhibitors, it is likely that the risk of exposure is becoming lower as time passes. However, it may still be safe to suggest that if there is evidence of buildup of any kind on the internals of drilling, extraction, or processing equipment, one would do well to ascertain whether the buildup contained arsenic or hydrazine. The literature suggests that the highest exposures to these materials occur through ingestion of drinking water. If effort is made to avoid ingestion (ensuring the drinking water system does not contain harmful corrosion inhibitors) and skin contact (protective gloves, coveralls, etc.), the risk of exposure is reduced. Dismantled equipment, left out in the open air, may contain arsenic- or hydrazine-contaminated scale or buildup. If this material dries out, it could become airborne if disturbed. Small particles could be picked up in the wind and dispersed into work areas. Removal of the equipment from work areas or removal of any potentially contaminated buildup on equipment would help to reduce the risk of exposure.

1.6 Organic Solvents (Diesel Fuel, Kerosene, etc.)

Nearly everyone is exposed to solvents in some form or fashion over his or her lifetime. In the crude oil and natural gas extraction industry, exposure to solvents can potentially be high. Since some of the contaminants and components of crudes, lubricants, drilling fluids, brine, and other materials can collect and build up on equipment, solvent use is often required to

remove deposits and clean down-hole and surface handling equipment. Many solvents have been made to be environmentally friendly and health-friendly. However, some solvents used in the oil and gas industry are not. Solvents such as gasoline, diesel fuel, and kerosene are often used. Exposure to solvents can occur through inhalation and skin contact. Concentration and time of exposure are two important considerations, as is the chronic or acute nature of the exposure. Since there are so many different solvents in use, with varying health effects, it is difficult to characterize the health effects of all solvents. In general, solvents can dissolve protective oils in the skin, drying it out to the point where cracking occurs. Cracking allows solvent to be more readily absorbed into the system on continued exposure. Inhalation exposures to solvents (gasoline, diesel, kerosene, etc.) at high enough concentrations can result in dizziness, nausea, headache, and other central nervous system effects.

Solvents are used in the oil and gas extraction industry to remove buildup of hydrocarbons and other contaminants from equipment (valve bodies, pump impellers, shafts, etc.). This is done by first placing equipment in vats, tubs, or other containers of the solvent and then, through rinsing, brushing, or other active means, the buildup or contaminant is removed. There are a number of types of solvent application devices and methods, such as high-pressure sprays as well as soaking.

The type of application is critical to assessing and protecting against exposure. For any application that aerosolizes the solvent, ventilation or respiratory protection or both should be used. For liquid contact, gloves and eye, face, and body protection (aprons, coveralls, etc.) should be worn at all times when solvents are being handled. When and where possible, it is effective to reduce exposure time and opportunities to a minimum. Many solvents are flammable (especially diesel and kerosene), so appropriate fire precautions should be taken. The ventilation system should be rated for flammable materials and the potential for static spark generation should be minimized. Proper disposal of spent solvents is important to keeping exposures low.

1.7 Ergonomic, Physical, and Biological Health Risks (Noise, Heavy Lifting, Mosquitoes, Vermin, Heat, and Cold)

The oil and gas drilling and extraction industry has, like most industries, become more automated over

the past 100 years. Even though there is not as much lifting and handling of heavy equipment and supplies as there once was, there are still ergonomic and physical health risks associated with drilling and extraction operations. With the need for compressors, pumps, turbines, steam generators, drilling motors, and rotary tables, etc., noise levels around drilling and extraction operations can be high. Noise levels around compressors, turbines, and oil-fired steam generators can be as high as 95 to 100 dBA (decibels measured on an A-weighted scale). Noise exposure around equipment during drilling and completion activities can be of greater concern than during production operations because workers are exposed for longer and more continuous periods. It is a 24 h operation and workers sometimes remain in high-noise-level areas for their full shift. Oil and gas production operations do involve the use of compressors, turbines, and steam generators. Even though the sound levels are similar to those around drilling operations, exposure periods are usually shorter and more intermittent. Many companies make an effort to use only equipment that is designed and built to generate lower noise levels, but this does not always completely reduce the noise levels to acceptable levels. The goal of many companies is to reduce noise levels and exposures to the point where they do not have to implement the Occupational Safety and Health Act's Hearing Conservation Standard. For time-weighted average exposures above 85 dBA, implementation of this standard is mandatory in the United States. Hearing protection, training, and audiometric testing can supplement sound level reduction interventions.

Although the drilling and extraction operations have become automated over the years, there is still a need for the transport, installation, application, and handling of heavy, bulky, and awkward-to-move equipment and supplies. Equipment and supplies, such as large wrenches, drums of chemicals, chains, rigging equipment, and valves, can be heavy or require awkward body positions to lift or handle. Opening and closing large valves can present an ergonomic risk. There is a risk of back and shoulder strains in this type of operation. At times, work on a drilling rig requires the use of supplied-air respiratory equipment. Although this respiratory equipment is designed and built to be worn and carried, when it is combined with the carrying and use of other tools over a long shift, the risk of muscle strain or, at least, fatigue is present. Although automation and lift-assist devices have been successfully implemented and used in the oil and gas extraction industry, many

tasks still require manual handling and brute strength (e.g., manipulating and attaching drill pipe from the monkey board in the drilling derrick). The search for improved means and use of automated and lift devices continues. In the meantime, education and training can go a long way to helping workers understand their physical limitations and how to accomplish their job without straining or stressing their muscles and bones.

The climate in oil and gas extraction areas can be extreme and present certain health risks. The hot humid climates of Africa, South America, and the South Pacific region present the risk of heat stress. Without proper acclimatization, hydration, and work/rest regimens, workers risk the onslaught of heat-related illness, including the potential for heat stroke. Care must be taken when tasks are to be performed under high heat and humidity conditions. The cold in areas such as Siberia, the north slope of Alaska, and western Kazakhstan also presents the risk of frostbite and freezing. In these climates, it is also possible to have difficulty operating equipment that requires manual dexterity. If gloves make it difficult to operate small equipment, such as drills, torquing guns, and wrenches, workers will be more likely to remove gloves meant to protect them from the cold. In extreme cold, blood flow to the extremities is decreased and the vibration of the equipment then contributes to the development of Raynaud's syndrome (an abnormal constriction in the blood vessels of the fingers when exposed to cold). This phenomenon results in a loss of feeling in the fingers. Adequate, layered clothing helps protect workers who must work in the cold environment. Providing heated shelters, as well as hand- and foot-warming devices to allow workers temporary and regular relief from the cold, can reduce the risk of frostbite, Raynaud's syndrome, and freezing.

Oil and gas drilling and extraction operations are found all over the world. Some of these locations present a health risk due to the occupation or infestation by insects and other vermin. The risk of contracting mosquito-borne malaria is present in some locations, such as sub-Saharan Africa. According to local government environmental agency representatives, the fleas on the rats and other rodents of western Kazakhstan reportedly carry various diseases. Mosquito control programs involving spraying and minimization of standing water can help reduce the risk of exposure to malaria. Workers can also reduce the risk of contracting malaria through prophylactic drug regimens. Efforts to reduce the encroachment of disease-carrying

rodents into living areas of drilling and production operations involve the use of poisons, traps, and continued food removal and habitat destruction.

2. GENERAL HEALTH RISKS DURING SPECIFIC OPERATIONS

2.1 Drilling Operations—Land Based

Once it has been determined that a particular formation has the necessary characteristics for oil or gas production, and the necessary permits have been approved, a well program is developed. This program contains various technical information about how the well is to be drilled, such as how the rig will be positioned, depth of the well, drilling fluid design, maintenance requirements, geological information, casing design, and kick tolerances. It also contains information about the hazards that are expected to be encountered. This is critical to the overall program as some of the hazards have the potential to result in loss of life.

2.1.1 Location Preparation and Driving the Conductor

The first step in drilling an oil or gas well is the preparation of the location for the drilling rig. During this phase, the site is cleared, the boundary is marked, and access roads are installed. Depending on the geographical area, consideration must be given to hazards such as mosquito-borne malaria, unexploded war ordinance, and hot and cold ambient conditions. During this phase, a number of pits are dug. These excavations include a rig pit to accommodate blowout prevention equipment. This pit is called the "cellar" and it is lined with concrete. Health-related concerns here include the potential for collapse of the sides of the pit as workers are preparing it for concrete. In the event of a leak in production equipment, the pit provides a location for the potential collection of the heavier-than-air, toxic hydrogen sulfide. After completion of drilling and commencement of production, work inside the cellar requires diligent application of a confined-space-entry permit program including air testing and use of appropriate respiratory protection.

A waste pit is dug for disposal of drill cuttings after they have been separated from the drilling fluids. In desert locations, a concrete-lined water pit is constructed and a water well is drilled to provide water for the drilling operations. Extensive quantities of water are required for mixing drilling mud and

cement. At this point, conductor pipe is driven into the ground with a pile driver. This provides the structure for the beginning of drilling operations. The process generates high-impact noise levels. This is common to pile driving operations in general. Since the exposure limit for impact noise is 140 dBA, there is probably less concern for overexposure to noise at this point in the operation. The health hazards associated with this phase of the operation are minimal; however, there is a health risk concern when oil begins to flow. The pits are low-lying areas that present a place for heavier-than-air gases and vapors such as H_2S to collect.

2.1.2 Moving the Rig on Location and Attaching the Diverter

The rig substructure is constructed of steel beams in a large frame, which supports the rig floor. On land, the rig floor is generally ~15–20 ft above the ground level. The derrick rises up from the substructure above the rig floor (sometimes as much as 100 ft or more). The drill floor supports the rotary table, which turns the drill string. The structure is secured with large steel pins. Although much of this work is done with cranes, winches, and other lift devices, there is much manual lifting and manipulating of heavy equipment during this phase. The risk of musculoskeletal sprains and strains is high during this operation. Once the traveling block and block line are in place, the rig is ready to start operating.

During drilling operations, there is always a danger of an unexpected “kick.” The formation releases pressure and flow as formation restrictions are cleared. If the pressure exceeds the gravity head and friction pressure of the drilling fluids, the pressure and flow can escape up the well. This potentially exposes the workers on the rig floor and the environment to high-pressure, potentially high-temperature fluids, hydrocarbons, and hydrogen sulfide. One of the tools drillers use to minimize the risk of this exposure is to install a flow diverter above the conductor pipe. When the rig is ready to start, a diverter is attached to the conductor pipe. This diverter contains a large rubber seal forced under hydraulic pressure in and around the drill string to seal it. Two large pipes are installed under this seal. They lead away from the rig in opposite directions. If, during drilling, there is a kick in the well below the conductor pipe, the flow is directed away from rig when the diverter is closed and the valve on the pipe leading from the rig is opened. The flow is usually directed to a containment system,

where the release of energy and flow can be controlled and contained.

After the conductor pipe has been driven into the ground with the pile driver, the inside of the pipe will be full of rocks, dirt, and other rubble. This is removed by drilling. During this process, the drill bit is rotated as it is lowered into the rubble. The rocks are removed as drilling mud is pumped through the drill string and the mud rises back up the annular space between the drill string and the conductor pipe, carrying the rocks with it. As the conductor pipe area is drilled, there is a potential to penetrate a shallow pocket of gas. This gas pocket can contain hydrocarbon or hydrogen sulfide and if it is penetrated, the pressure, hydrocarbons, and other contaminants can be released to the surface where the workers can be exposed. It is easier to recognize and control this release if the size of the drill hole is kept small. A smaller drill hole would help to keep the initial flow small. This would allow a few extra seconds for the drillers to close the diverter and to begin rapidly introducing drilling fluid down the drill string to control the pressure and flow. This gas release can be of great enough intensity that removal of large amounts of rock also introduces the potential danger of projectiles.

2.1.3 Spudding and Drilling the Well

When the drill bit reaches the bottom of the conductor pipe and begins the drilling process, this phase is called spudding the well. As the well bore is drilled during the spudding phase, rock cuttings are generated and must be removed from the hole. This is done by pumping drilling fluid down through the hollow drill string, out the holes in the bottom of the drill bit, and back up the annular space between the hole and the drill pipe. This phase continues until the borehole is deep enough for the surface casing.

When spudding is complete, drilling of the first hole section and installation of the intermediate casing sections begin. The drilling process is the same in that a rotating bit at the bottom of the drill string removes rock cuttings from the bottom of the hole while drilling fluid is pumped through the drill string. In this section of the well, the surface casing is installed and cemented in. The blowout prevention (BOP) equipment will also be installed and tested. This equipment helps to minimize the potential for exposure to the unexpected sudden release of pressure and well fluids during drilling operations. As the automatically controlled seal material is tightened around the drilling and production equipment to

form a seal, the blowout is contained in the well and kept away from workers.

When the BOP equipment is installed and tested, the casing is drilled out and the strength of the formation must be tested. Testing the strength of the formation is necessary to help ensure that it will support the well without collapsing. This test is done by first drilling out a cement valve called a float shoe at the bottom of the surface casing and then removing all cuttings and approximately 15 ft of new rock below the bottom of the surface casing. Then the drill bit is removed to a point above the BOP equipment. The BOP equipment is closed and the hole is incrementally pressurized by adding drilling fluid a little at a time until the pressure created by the addition of fluid to a closed system begins to reduce as more fluid is added. This tells the drillers that some “leak-off” to the formation is occurring at that pressure. The addition of more fluid at this point could fracture the formation and weaken it. This pressure is determined to be the formation strength. Although there is always a risk of release of pressure and fluids from the formation, the risk of exposure of workers to hydrogen sulfide, hydrocarbons, etc., at this point, is low due to the containment provided by the BOP equipment.

When the BOP equipment is operational and the formation has been tested, the program is ready for drilling of the intermediate hole sections. This is where the diameter of the well bore is reduced and the depth increased significantly. With the intermediate casing sections installed, the production hole section is next. This is the smallest diameter section and is generally the bottom of the well. This section coincides with the presence of both liquid and gaseous hydrocarbons. This is also the point where blowouts, the release of hydrogen sulfide and other hydrocarbons, and exposure to drilling fluids and formation contaminants become more likely. Hydrocarbon and other gases sometimes dissolve in the drilling fluids and then escape at the surface in circulation tanks. This introduces the risk of exposure to the workers as well as the risk of fire should flammable hydrocarbon gases find an ignition source. Throughout this phase, casing and liner pipe are being installed, cemented in place, and sealed.

2.2 Drilling Operations—Offshore

Offshore drilling often involves drilling several wells from one platform. This requires directional drilling to access several locations in the formation without the expense and risks of a production platform at

every well location. The health risks for offshore drilling are, for the most part, the same as they are for onshore drilling. The contaminants present in the underground formations will be the same. Hydrogen sulfide exposure will be of concern, as will hydrocarbon, heavy metals, and NORM exposure. The differences would be related to the isolated and confined location. If there are contaminant leaks or releases, there is no immediate escape. Exposure to a cold environment in locations such as the North Sea and the Maritime Provinces in Canada would be an additional risk brought on by exposure to the cold-water environment.

2.3 Surface Production Operations

Surface production operations include production of the oil and natural gas at the wellhead, collection of produced oil and gas at collection manifolds, and transport to the production treatment facility. Control stations at well sites are enclosed and often climate-controlled spaces where the instrumentation that controls the well operation is housed. In the event of a leak, the control stations can become a collection point for hydrogen sulfide or other hydrocarbons. These spaces are often fitted with hydrogen sulfide, oxygen, and flammable material detectors with alarms to indicate when concentrations of these contaminants may be an exposure concern for operators. As mentioned earlier, cellars or pits dug to support the drilling operation may be left at a wellhead. These also serve as a collection point for hydrogen sulfide and other heavier-than-air contaminants should a leak develop.

As production at a particular well progresses over time, there is a need to test its production capability. Flow tests are conducted to test this capacity. During this test, oil is allowed to flow into a pit or similar reservoir. The flow is measured over time to gauge this flow capacity and ultimately to ascertain the well's production rate to help determine production life. This oil is sometimes burned as a means of disposing of it. This practice potentially exposes workers to hydrogen sulfide, sulfur dioxide (converted from hydrogen sulfide in the combustion process), carbon smoke particulates, and unburned hydrocarbons. Under normal conditions, the oil is transported through pipelines to collection manifolds until it results in one stream entering the treatment facility.

As oil flows through the pipelines and collection manifolds, two situations that can have an impact on human health occur. In some climates, and with some

crude oils, hydrates can precipitate out of the oil and collect in the pipelines. This can occur under low-temperature conditions, or when there are changes in pressure, or at higher salt concentrations. These hydrates can plug the lines and trap pressure. If they free themselves as thermal cycling occurs throughout a day, they can be forced rapidly through the pipe by the pressure behind them. This, in essence, is an out of control projectile that can rupture pipelines. If a line ruptures, workers in the area could be struck by the hydrate material or piping fragment, injured by the moving pipeline, or exposed to the hydrocarbons or hydrogen sulfide that would be released in a rupture. Methanol is sometimes injected into the line to help dissolve the hydrate buildup before it becomes a plug in the line. Although done under procedure-driven control, this practice has the potential to expose the workers to alcohol or fire.

Lines also become plugged over time with sediment and other contaminants and occasionally must be cleared. The method of such clearing is called "pigging" the line. A barrel-shaped device of various designs and sizes, with various scraping mechanisms positioned around the device's circumference (called a pig), is pressured through the line to remove any blockage or restrictions in the line. The pig is slightly smaller in diameter than the internal diameter of the line. It is loaded at a closed pig-launching station, the loading door is closed, the production line is opened, and the pig is pressured through the line with natural gas, nitrogen, or some other nonreactive pressurizing material. The sludge or blockage ahead of the pig is either routed back into the production stream or removed and disposed of at the pig-receiving station. At the receiving station, the pig is removed after the production side is closed and the pressure is bled off through a vent. This operation clears the line and improves flow efficiency.

The health risks associated with this operation include potential exposure to high pressure, hydrogen sulfide, or hydrocarbons. If an attempt is made to open a pig trap while it is still pressured, the door would swing open rapidly; serious injury is likely. Because the system is opened during launching and receiving, the possibility of exposure to hydrogen sulfide and other hydrocarbons is increased.

The goal of the production facility is to separate the production stream into oil, gas, and water. This is achieved through a series of separators, which make use of pressure differentials to separate oil and gas and make use of viscosity and gravity differences to separate oil and water. Sometimes the gas contains hydrogen sulfide, which must be further processed.

Natural gas is sometimes distilled to separate methane, ethane, propane, and sometimes butane into separate streams. Prior to this step, depending on the hydrogen sulfide content in the production stream, which can be as high as 15–25% by volume, the gas must pass through an absorption process. This step strips the hydrogen sulfide from the gas by placing it in contact with an amine or similar absorptive chemical with an affinity for hydrogen sulfide. The amine or other absorptive chemical must also be treated to remove the H_2S so it can be reused. The chemical is passed through a "regeneration" process, such as distillation, to remove the H_2S . The gas containing the hydrogen sulfide and other acid gases, collectively referred to "sour gas," are processed in sulfur treatment units utilizing heat and reaction with catalysts to convert the hydrogen sulfide to elemental sulfur. This elemental sulfur can then be sold as a product. At several stages in the sulfur treatment process, small quantities of hydrogen sulfide and sulfur dioxide can still be present. Exposure is still a possibility; therefore, the use of appropriate respiratory protection during sulfur blocking or loading may be necessary depending on the concentration.

The crude oil fraction, which is usually considered the main product, is sometimes passed through desalting equipment, undergoes solids centrifuging, and undergoes distillation to remove salts, unwanted produced solids, and any residual gases. All of these steps are not necessary for all crude oils. Some crude oils need only undergo static retention time in a holding tank to allow the water to settle out and it is then ready for sale. This would be true for crude oils containing less gas and hydrogen sulfide. Although the health risks at this stage are relatively low compared to stages where the hydrogen sulfide concentration is high, an appropriate level of precaution is necessary. Proper respiratory protection should be nearby and calibrated and operational leak detection equipment should be functioning.

3. SUMMARY

The extraction and production of crude oil and natural gas are complex operations that involve the potential for exposure to many health risks. Each of these individual health risks has been well studied over the years. Sound-exposure reduction programs have been established and successfully implemented in many oil and gas operations around the world. The risk of exposure to hydrogen sulfide, NORM,

hydrocarbons, solvents, heavy metals, and other hazards has been successfully managed in the oil and gas industry. These risks have been managed by properly designing, installing, and using equipment that contains these contaminants. Oil and gas companies have reduced exposures by providing appropriate detection and alarm equipment that helps to shut down equipment that is malfunctioning or leaking or by alerting workers of a leak or other operational problem. Additional management of these risks has been achieved by implementing effective permit to work programs and by providing appropriate protective equipment and educating workers so that they are better able to recognize and avoid the risk of exposure to the contaminants found during crude oil and natural gas extraction.

The crude oil and natural gas industry has done well to minimize the overall health risks associated with drilling and extraction operations. Even though the severity of exposure to contaminants such as hydrogen sulfide or sulfur dioxide is quite high, the likelihood of exposure has been greatly reduced through the diligent efforts of this industry.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Crude Oil Spills, Environmental Impact of • Natural Gas Processing and Products • Natural Gas Transportation and

Storage • Occupational Health Risks in Nuclear Power • Radiation, Risks and Health Impacts of

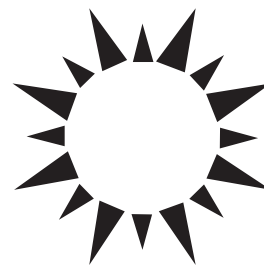
Further Reading

- Amdur, M. O., Doull, J., Klaassen, C. D. (Eds.). (1991). "Casarett and Doull's Toxicology: The Basic Science of Poisons, 4th ed. Pergamon, New York.
- American Conference of Governmental Hygienists. (2001). "Threshold Limit Values for Chemical Substances and Physical Agents—Biological Exposure Indices." ACGIH Worldwide, Cincinnati, OH.
- Arnold, K., and Stewart, M. (1998). "Surface Production Operations, Design of Oil-Handling Systems and Facilities." 2nd ed. Gulf Publishing, Houston, TX.
- Bureau of Labor Statistics, United States Government, Web site available at <http://www.bls.gov>.
- Devereux, S. (1999). "Drilling Technology in Non-technical Language." PennWell, Tulsa, OK.
- Ludvich, D. (1981). "The Hydrogen Sulfide Technical Manual." Safety Technology and Oilfield Protectors, Lafayette, LA.
- Mandavia, S. (2001). "Toxicity, Hydrogen Sulfide" (D. C. Lee, J. T. VanDeVoort, J. Benitez, J. Halamka, and R. J. Roberge, Eds.). eMedicine Web site available at <http://www.emedicine.com>.
- McGavran, P. (2001). "Literature Review of the Effects Associated with the Inhalation of Hydrogen." Idaho Department of Environmental Quality, Boise, ID.
- Perrin, D. (1999). "Oil and Gas Field Development Techniques, Well Completion and Servicing." Editions Technip. Paris and Institut Français du Pétrole, Rueil-Malmaison, France.
- Plog, B. A. (Ed.). (1988). "Fundamentals of Industrial Hygiene," 3rd ed. National Safety Council, Chicago, IL.
- Williams, P. L., James, R. C., Roberts, S. M. (Eds.). (2000). "Principles of Toxicology Environmental and Industrial Applications," 2nd ed. Wiley, New York.
- Wilson, W. F. (1994). "N. O. R. M. A Guide to Naturally Occurring Radioactive Material." PennWell, Tulsa, OK.

Occupational Health Risks in Nuclear Power

DAVID B. RICHARDSON

University of North Carolina, Chapel Hill
Chapel Hill, North Carolina, United States



-
1. Ionizing Radiation
 2. Radiation Exposure and Radiation Dose
 3. Characteristics of Radiation Exposures in the Nuclear Power Industry
 4. Biological Effects of Ionizing Radiation
 5. Health Risk Following Occupational Exposure to Ionizing Radiation
 6. Radiation Hazards in the Context of the Nuclear Fuel Cycle

Glossary

absorbed dose A measure of the joules of radiation energy per kilogram of matter.

alpha particle A positively charged nuclear particle consisting of two protons and two neutrons emitted during radioactive decay.

beta particle A high-speed electron emitted during radioactive decay.

effective dose A measure calculated by multiplying equivalent dose by a tissue-specific weighting factor that is intended to take into account differences in sensitivity of different organs and tissues of the human body to the stochastic effects of ionizing.

equivalent dose A measure calculated by multiplying absorbed dose by a radiation-specific weighting factor that is intended to take into account differences in the density of ionization produced by different types of ionizing radiation.

gamma ray High-energy photons of electromagnetic radiation.

ionizing radiation Photons of high-energy electromagnetic radiation and particle forms of radiation that have sufficient energy to produce ions by removing electrons from atoms or molecules.

This discussion focuses on the health risks faced by workers in the nuclear power industry who operate

and maintain nuclear reactors used for electrical power generation. These workers may encounter a range of hazards on the job, including exposures to solvents, welding fumes, and asbestos, as well as occupational risks for traumatic injury and electric shock. In addition, many workers in the nuclear power industry are exposed to radiological hazards arising from external sources of ionizing radiation and from internal deposition (via inhalation, ingestion, or absorption) of radionuclides. Understanding of the effects of these occupational exposures comes from a number of disciplines, including biology, toxicology, and epidemiology. Biology and toxicology provide evidence derived from laboratory research about the effects of ionizing radiation exposure on molecules, cells, and experimental animals. Epidemiology, in contrast, provides evidence about radiation health effects derived from studies of disease in human populations. The radiological hazards associated with nuclear power generation are notably different from the occupational hazards associated with other methods for power generation and consequently will be the focus of this review.

1. IONIZING RADIATION

The term ionizing radiation refers to photons of high-energy electromagnetic radiation as well as corpuscular or particle forms of radiation. Occupational exposure to ionizing radiation occurs in the nuclear power industry as a result of radioactive decay of unstable isotopes of elements. Certain isotopes of elements, such as uranium and plutonium, can emit alpha particles (two protons and two neutrons) during radioactive decay. Alpha emitters are primarily found among the elements that are heavier than

lead, although some isotopes of elements lighter than lead decay by alpha emission. Radioactive decay in most elements that are lighter than lead occurs via beta decay in which an electron, called a beta particle, is ejected from the nucleus. Workers employed at heavy water nuclear power reactors, for example, may have substantial potential for exposure to beta radiation from tritium (a radioactive isotope of hydrogen). Gamma radiation often accompanies alpha and beta particle emission. The emission of neutrons occurs as the result of nuclear reactions (for example, fission of uranium-235 nuclei).

As ionizing radiation is transmitted through human tissue, energy is transferred to the tissue. This can lead to ionization of the atoms or molecules in the surrounding tissue. Alpha particles, which are relatively heavy and have a double positive charge, readily transfer energy to human tissue, resulting in the slowing of the alpha particle along its path. The energy of the alpha particles will almost entirely be deposited in the outer (dead) layers of skin if a person is externally exposed to alpha radiation. In contrast, if a radionuclide that emits alpha particles is deposited within the human body (for example, through inhalation, ingestion, or injection), damage will occur to the living cells of organs and tissues. Beta particles, which are electrons traveling at high speeds, can penetrate to a greater depth in human tissue than alpha particles. The distance that a beta particle travels through human tissue depends on the particle's energy and can range from several millimeters (for beta particles of medium energy) up to a depth of 2 cm in human tissue (for high-energy beta particles). Neutrons, which are particles without an electrical charge, can also travel short distances through human tissue. Neutrons produce ionization indirectly, by capture (for low-energy neutrons) or collision with surrounding material (for higher-energy neutrons). Gamma rays are high-energy photons of electromagnetic radiation and can completely penetrate the human body, depositing energy and causing ionization of atoms and molecules along the photons' path.

2. RADIATION EXPOSURE AND RADIATION DOSE

Radiation dose may be expressed in terms of absorbed dose, equivalent dose, and effective dose. An absorbed dose is the amount of radiation energy absorbed per unit mass of an organ or tissue. In 1976 the gray was adopted as the International System of Units (SI) unit

of absorbed dose; one gray is equal to one joule of radiation energy per kilogram of matter. Absorbed dose is also sometimes reported in a unit known as the rad, which is an acronym for radiation absorbed dose. One rad is equal to 1/100th of one gray.

A measure of equivalent dose to an organ or tissue is a quantity that takes into account differences in ionization density for different types of radiation (e.g., alpha particles, beta particles, and gamma rays). Equivalent dose is calculated by weighting the absorbed dose to an organ or tissue by a radiation weighting factor that reflects the relative biological effectiveness of the photon or particle that produced the ionization within the tissue (it is assumed that the biological effect associated with a photon or particle is related to the ionization density). Values for this radiation weighting factor are assumed to be independent of the organ or tissue that is exposed as well as the biological end point under consideration. For outcomes that are assessed at the cellular or molecular level, such as DNA damage, it has been observed that a given absorbed dose of ionization radiation produces a greater amount of damage when the dose is due to alpha radiation than photons of electromagnetic radiation. For human health outcomes, such as cancer, however, there is uncertainty about the relative biological effectiveness of different forms of ionizing radiation because of the limitation of available empirical evidence about the magnitude of radiation-cancer associations for absorbed doses of different types of ionizing radiation. The International Commission on Radiological Protection currently recommends a radiation weighting factor of 1 for photons and electrons and a radiation weighting factor of 20 for alpha particles. While absorbed dose is a physical quantity, equivalent dose estimates may change as recommendations about weighting factors are revised. Equivalent dose is expressed in the unit sievert (Sv). An earlier unit used to express equivalent dose was the rem, which is an acronym for radiation equivalent dose; one rem is equal to 1/100th of one Sv.

A measure of effective dose is intended to express the biological effect associated with irradiation, taking into account differences in radiosensitivity of the body's various organs and tissues. Effective dose is calculated by weighting the equivalent dose in a tissue or organ by a tissue weighting factor that reflects the contribution of the tissue to the detriment of health when the body is uniformly irradiated. Doses to tissues and organs such as the lung are assigned a greater weight than doses to tissues and organs that are assumed to be relatively insensitive to adverse

effects of ionizing radiation. Effective dose is expressed in the unit Sv, and, like equivalent dose estimates, effective dose estimates may change as recommendations about weighting factors are revised.

3. CHARACTERISTICS OF RADIATION EXPOSURES IN THE NUCLEAR POWER INDUSTRY

Workers in commercial nuclear power plants are primarily exposed externally to gamma rays from activation products, such as Co-60, as well as from fission products, such as Zr-95 and Cs-137, when fuel failures occur. On average, workers in the commercial nuclear power industry accrue effective doses of 2-3 mSv per year. However, the average dose reported for nuclear power plant workers varies by reactor type as well as by job activity. For commercial boiling water reactors in the United States, the reported average annual measurable effective dose per monitored work has fallen between 1973 and 2001 from approximately 8.5 mSv to approximately 1.7 mSv; for pressurized water reactors over the same period, the average annual measurable effective dose per work has fallen from 10.0 mSv to 1.7 mSv. Maintenance work, particularly during outages (when refueling, maintenance, inspections, design modifications, and other activities take place), tends to be associated with relatively high accrued doses. It is estimated that most occupational radiation exposure are accrued during these special maintenance activities. Furthermore, the energy and geometry of radiation fields typically differs for exposures accrued during outage work when compared to routine operations. Doses tend to be accrued at higher dose rates and in an anterior/posterior geometry during outage work. In contrast, lower dose rate exposures accrued in a rotational geometry tend to be accrued during routine operations. The characteristics, in terms of energy and geometry, of radiation fields may be important for critical evaluation of reported occupational doses. For workers in areas where there is potential for occupational exposure to external sources of ionizing radiation, doses are typically monitored by a dosimeter that is worn on the chest. The response of a radiation dosimeter is influenced by the energy and geometry of the radiation field, and doses may be missed if the dose is exclusively delivered to an extremity of the body.

Other characteristics of the work environment can also affect radiation doses received on the job; these

include the design of the facility, the rate of fuel failure, and radiation protection practices (particularly during repair work). In many countries, including the United States, contract workers are hired to perform short-term maintenance tasks and to provide additional support during outages. Record keeping and retrospective evaluation of accrued doses for contract workers is often more difficult than for plant and utility personnel because contractors may work at more than one facility per monitoring year. Workers employed at multiple facilities in a year are often referred to as transient workers. Analyses of transient worker data in the U.S. nuclear power industry in 2001 indicate that their average measurable effective dose per worker (3.7 mSv) tend to be higher than the average for all workers in the industry. In 2001, transient workers' doses accounted for 53% of the industry's collective dose. A small proportion of nuclear industry workers are exposed to neutrons on the job, although neutron doses account for only a small proportion of the total collective dose accrued by nuclear power plant workers.

Internal deposition of radionuclides occurs among some workers employed in the commercial nuclear power industry. In the United States in 2001, for example, the licensed operators of commercial nuclear power reactors reported occupational intake by ingestion of radionuclides including CO-60, RU-106, PU-238, AM-241, and CM-243. Reported occupational exposure by inhalation at power reactors included CO-58, CO-60, AM-241, CM-242, H-3, and CU-64. At heavy water reactors, workers may accrue significant internal doses from tritium, a beta emitter produced by neutron activation of the heavy water (deuterium). On average, internal exposure contributes about 30% of the annual collective occupational radiation effective dose among workers employed at heavy water reactors.

The International Commission on Radiological Protection recommends that occupational doses be limited to 20 mSv per year. Currently U.S. regulations require that employers limit the total effective dose equivalent from internal and external sources to 50 mSv per year. Few workers in the commercial nuclear power industry accrue doses in excess of this annual regulatory limit, although relatively high annual doses may be received by workers at nuclear power facilities in situations of accidents and remedial activities.

While average annual radiation dose estimates for workers in the nuclear power industry are low, if employment in the industry continues for several decades some workers may accrue cumulative doses of a hundred mSv or more. In a study of workers

employed at a U.S. nuclear power facility that started operation in 1975, over an average of 5.4 years of follow-up, the average cumulative external radiation dose accrued by plant and utility workers was 2 mSv and the average dose accrued by contract workers was 30 mSv. In contrast, long-term workers (those employed in the industry for 15 years or more) accrued an average cumulative dose of 111 mSv.

4. BIOLOGICAL EFFECTS OF IONIZING RADIATION

As charged particles, such as alpha and beta particles and uncharged neutrons and gamma rays, pass through living tissue energy may be transferred to the tissue. Although usually involving only a small fraction of all cells in the body, this interaction can produce harmful health effects. It is believed that the primary mechanism by which biological damage occurs is via the creation of ionized atoms and molecules that become chemically reactive. This can occur directly via ionization of a critical molecule, such as DNA, or indirectly via ionization of nearby molecules, such as water.

4.1 Deterministic and Stochastic Effects

At relatively high doses, exposure to ionizing radiation produces adverse consequences such as nausea, radiation burns, cataracts, decreased fertility, spontaneous bleeding, and, at very high doses, death. An acute equivalent dose of 150 mSv or more to the testes is likely to produce temporary sterility; at high doses, permanent sterility may occur. An acute equivalent dose of 500 mSv or above to the lens of the eye will lead to detectable opacities; such a dose to the bone marrow will lead to depression of red blood cell production. Cell killing is crucial to the development of most of these effects. These are referred to as deterministic effects; the incidence and severity of the deterministic effects observed in an individual exposed to ionizing radiation tends to increase with increasing dose. The rate at which the radiation dose is accrued is also an important factor in the production of deterministic effects; delivery of the same magnitude of radiation dose over a longer period of time reduces the likelihood of producing these adverse effects.

At lower radiation doses, the primary concern is that ionizing radiation increases the probability of certain stochastic effects, most notably cancer and genetic effects. Unlike deterministic effects of ioniz-

ing radiation, the severity of these stochastic effects is not believed to be related to the radiation dose; rather, the radiation dose is believed to increase the probability that one of these stochastic effects will occur in an exposed population. Current regulatory guidelines are premised on the assumption that any dose of ionizing radiation will increase the risk of stochastic effects.

4.2 Genetic and Somatic Effects

The damage to chromosomal DNA produced by exposure to ionizing radiation may lead to adverse genetic effects as well as somatic effects. Genetic effects are those that are caused by mutations of any germ cell (sperm or oocytes), which led to permanent changes in genetic material that are subsequently inherited by offspring of the exposed individual. Since ionizing radiation may damage chromosomal material, preconceptional exposure to ionizing radiation may lead to mutations in genetic material that will be transmitted to the exposed person's children. Genetic effects may result in decreased fertility, birth defects, or a chromosomal defect impacting offspring and future generations. Such effects have been documented in laboratory experiments involving irradiation of nonhuman animals; however, heritable genetic effects of ionizing radiation exposure have thus far not been documented in human populations such as the (F1) cohort of A-bomb survivors.

Human exposure to ionizing radiation may also lead to adverse somatic effects. Somatic effects are those mutation events that occur in the cells of tissue and organ systems other than the germ cells and that led to health outcomes such as cancer. Although a full understanding of the molecular basis of cancer has not been developed, point mutations, which are single base nucleotide changes in DNA, are believed to be involved in the malignant transformation of cells as part of a multistep process of carcinogenesis. Since ionizing radiation can damage the genetic material of any cell in the body, it has been suggested that radiation-induced cancer can occur in most, if not all, sites or tissues of the body although the radiosensitivity of tissues may vary.

5. HEALTH RISK FOLLOWING OCCUPATIONAL EXPOSURE TO IONIZING RADIATION

The primary quantitative basis for regulations that are used for radiation protection efforts in the

nuclear power industry are risk estimates derived from the life span study (LSS) of atomic bomb survivors in the cities of Hiroshima and Nagasaki. When the atomic bombs were dropped in 1945, the estimated population of Hiroshima was about 330,000 and the estimated population of Nagasaki was about 250,000. It is not reliably known how many people survived these attacks; however, beginning in 1950 during U.S. military occupation of Japan, a systematic survey of atomic bomb survivors was undertaken. A nationwide survey of A-bomb survivors who were still alive in October 1950 was conducted as a supplement to Japanese decennial census. The survey indicated that 159,000 people who had been exposed in Hiroshima and 125,000 people who had been exposed in Nagasaki were alive at the time of the census. Subsequently, individuals were assigned radiation dose estimates based on information collected by questionnaire on survivors' location and position, as well as consideration about shielding and other factors influencing radiation dose. Recent analyses, using DS86 dose estimates, have included 75,991 A-bomb survivors. Among that group of survivors, 34,272 survivors had estimated doses equal to 0 Gy.

These data have been the basis for estimation of the long-term cancer risk following radiation exposure at a wide range of ages; risk estimates for noncancer causes of death have also been reported. Results are often reported as the estimated excess relative risk of disease or death per sievert radiation dose (ERR_{1Sv}). Estimates of excess relative risk per unit dose are derived using a model in which it is assumed that the ratio of the radiation-induced cancer risk to the baseline cancer risk (the risk for those unexposed to radiation) is constant over time. An $ERR_{1Sv} = 0.40$, for example, indicates that a group of people exposed to 1 Sv is estimated to have a 40% higher cancer mortality rate than a comparable group that was not exposed to radiation. In workers' compensation cases, a related measure is sometimes used, which is the probability of causation (or assigned share); this is the probability that a cancer that has occurred was caused by radiation and is equal to $ERR/(1 + ERR)$.

Based on follow-up of atomic bomb survivors through 1990, the estimated increase in all solid cancer mortality was 45% per Sv ($ERR_{1Sv} = 0.45$). The excess relative risk for the incidence of all solid tumors (as opposed to cancer mortality) diagnosed between 1958 and 1987 among members of the LSS was $ERR_{1Sv} = 0.63$. The larger estimate when evaluating cancer incidence than when evaluating

cancer mortality has been attributed to greater diagnostic accuracy of the incidence data and the lack of full representation of radiosensitive but relatively nonfatal cancers, such as breast and thyroid, in the mortality data. Analyses of leukemia incidence reported strong evidence of radiation-induced risks for leukemia, with ERR at 1 Sv for leukemia subtypes (ALL, AML and CML, respectively) were 9.1, 3.3, and 6.2. Lung ($ERR_{1Sv} = 0.95$), breast ($ERR_{1Sv} = 1.59$), and thyroid ($ERR_{1Sv} = 1.15$) were also sites with relatively large estimates of association.

There are limitations to the study A-bomb survivors, however, which led to uncertainty about the true magnitude of hazard posed by occupational radiation exposures. One source of uncertainty relates to the appropriateness of using information derived from studies of the effects of high dose-rate acute exposures received by the atomic bomb survivors to derive risk estimates for workers who are typically exposed to relatively low-dose rate chronic exposures. This requires extrapolation from high-dose effects to low-dose effects (and from the settings of an acute exposure to settings of protracted or chronic exposures). Applying risk estimates derived from the LSS also requires the assumption that the effects of ionizing radiation on the population of Japanese survivors exposed in 1945 apply to contemporary populations of workers. There are also uncertainties about the validity of extrapolating over time and between countries that have different background cancer incidence and mortality rates. In addition, questions have been raised about the validity of conclusions drawn from a study in which the study population is restricted to those people who survived not only the initial atomic blast(s) but also epidemics of infectious diseases, malnutrition, and the devastation of local infrastructure. It is unknown how large an impact this initial selection effect may have on estimates of cancer risk derived from these survivors. Finally, questions persist about uncertainty and bias in radiation dose estimates for A-bomb survivors; the accuracy of survivors' dose estimates depends on the accuracy of the information about location of survivors that was collected by questionnaire, as well as on assumptions about shielding and other factors that influenced survivors' doses.

Given these concerns, studies of workers who have been occupationally exposed to ionizing radiation are also of substantial interest for considerations about worker protection. Studies of workers can provide direct evidence about low dose chronic radiation exposures accrued in the settings of

regulatory interest. Furthermore, occupational studies of radiation health effects offer the opportunity to study humans who can be followed through employment records and occupational medical departments, who were often individually monitored for radiation exposures, and who can be traced through pension benefits and other record sources. Studies of workers in the nuclear power and nuclear weapons industries have focused primarily on the effects of external penetrating radiation, in part because external doses easier to quantify than are doses from internally deposited radionuclides. A notable limitation of studies of workers in the nuclear industry, however, is the relatively low statistical power of most studies when compared to the LSS. Most occupational cohorts include fewer people than the LSS, and the magnitude of radiation doses accrued by workers in the nuclear industry is far lower than the doses accrued by many people in the LSS. Consequently, in most studies, the confidence intervals of the radiation risk estimates are much wider than those derived from analyses of the LSS. However, continued follow-up of these occupational cohorts will continue to increase the weight of evidence derived from study populations that can provide direct evidence about the health effects associated with chronic low level radiation exposures. Ongoing studies of workers employed at nuclear power plants offer the possibility of contributing to this literature and may have greater statistical power than studies of nuclear weapons workers given the higher average annual exposures reported for workers in the nuclear power industry.

A number studies of workers in the nuclear weapons industry have investigated the relationship between recorded occupational radiation doses and mortality. Some of these studies have observed significant increases in all cancer mortality with increasing cumulative radiation doses, which were typically accrued at dose rates below current radiation protection guidelines, while other studies have reported no evidence of radiation-induced cancer. Among those studies that reported evidence of radiation-induced cancer among U.S. nuclear workers are studies of Oak Ridge National Laboratory (ORNL), Fernald, and Rocketdyne. Significant positive trends in cancer mortality with radiation dose have also been reported among workers in the United Kingdom's Atomic Weapons Establishment (AWE) and among workers included in the Canadian National Dose Registry. Significant association between leukemia and radiation dose were observed in

studies of workers employed at Rocketdyne and Mound in the United States, Sellafield in the United Kingdom, and AECL in Canada, as well as in the combined analyses of data from workers cohorts in the United Kingdom.

For example, in a study of 124,743 workers in the United Kingdom's nuclear power, fuel cycle, and weapons production industries, the association between external radiation dose and all cancer mortality was $ERR_{1Sv} = 0.09$ (90%CI, $-0.28, 0.52$), and for leukemia it was $ERR_{1Sv} = 2.55$ (90%CI, $-0.03, 7.2$). A study of radiation-mortality associations among 206,620 Canadian workers, which included people employed in a variety of (nonmining) occupations in which monitoring for radiation exposure occurs, reported that for males the ERR_{1Sv} for all cancers was 3.0 (90%CI, 1.1, 4.9). No statistically significant association between radiation dose and mortality from all cancers was seen in a combined analysis of radiation-mortality associations among 95,673 nuclear industry workers in the United States, the United Kingdom, and Canada; the ERR_{1Sv} for all cancers excluding leukemia was -0.07 (90%CI, $-0.4, 0.3$). Under a 2-year lag assumption the ERR_{1Sv} for leukemia excluding CLL was 2.2 (90%CI, 0.1, 5.7). Although these risks estimates are lower than the linear estimates obtained from studies of atomic bomb survivors, they are compatible with a range of possibilities, from a reduction of risk at low doses, to risks twice those on which current radiation protection recommendations are based. Variation in findings between cohorts may reflect complex patterns of bias related to selection into cohorts (as well as selection within cohorts in specific jobs or areas), exposure measurement error, and confounding exposures. Differences in radiation-mortality associations between cohorts could also reflect differences between study populations in the distribution of modifying factors.

Summary estimates of radiation-cancer associations are by definition a weighted average of the dose response relationships for subgroups with different susceptibilities. Differences in sensitivity to the carcinogenic effects of ionizing radiation may arise from genetic predispositions as well as from induced susceptibility from exposure to nonradiological occupational or environmental agents. Some markers of innate differences between people in radiosensitivity have been identified; for example, specific germline mutations are associated with increased vulnerability to the carcinogenic effects of radiation. More generally, there are observations that suggest

that the process of aging may itself influence susceptibility to radiation-induced cancer.

For radiation protection purposes, consideration of age at exposure is crucial. For example, it is important to limit the radiation doses accrued by pregnant females in the workplace because the developing human fetus which is believed to be extremely sensitive to the carcinogenic effects of ionizing radiation. Strong evidence of the radiosensitivity of the human fetus comes from the Oxford Survey of Childhood Cancers (OSCC), which was a national case-control study that included children under 16 years of age in Great Britain who died from any form of malignancy and a matched control (a living child of the same sex, living in the same administrative district in which the death occurred, and born on approximately the same day as the deceased child). The most recent analyses of the OSCC were based on data for 14,759 traced cases who died at ages 0 to 15 years during 1953 to 1979 and their matched controls. Findings from the OSCC have led to estimates of the excess relative risk for childhood cancer at 1 milligray that range from 0.038 to 0.051 (depending on fetal dose estimates and the birth cohorts included in analyses).

For occupational exposures accrued by adults who are not pregnant, age at exposure may also be an important consideration, although there is divergent evidence about age-related changes in sensitivity to the carcinogenic effects of ionizing radiation among adults. For all cancers, estimates of excess relative risk per unit dose in the LSS decline as age at exposure increases. In contrast, some studies of nuclear workers exposed to chronic low levels of ionizing radiation suggest that at older adult ages sensitivity to the carcinogenic effects of ionizing radiation increases. Such increases in radiosensitivity have been postulated to reflect declines in the efficiency of most biological systems, including the body's cellular repair and immune systems. Increases in radiation-cancer associations with age at exposure have been reported in studies of workers employed at Hanford, Oak Ridge National Laboratory, Fernald, and Rocketdyne, although evidence of variation in radiation effects with age at exposure was not found in the combined analyses of U.S., UK, and Canadian nuclear workers. For lung cancer, there is some evidence of increase in the magnitude of the radiation-cancer association with age at exposure in the LSS as well as in several occupational cohort studies. Again, such variation in findings between cohorts may reflect patterns of bias, exposure measurement error, and confounding exposures;

furthermore, estimates of the excess relative risk of cancer associated with doses accrued in very narrow time windows, such as those dose response estimates reported for occupational exposures accrued at very old ages, are expected to be particularly unstable.

Effects of internal exposure to ionizing radiation from internal deposition of radionuclides encountered in the nuclear power industry are less well studied in occupational epidemiology. It is known that alpha and beta radiation are highly effective at ionizing and that their effects on cellular molecules can be extremely damaging. There is clear evidence of mutations and structural changes to DNA resulting from alpha radiation; therefore, these exposures are presumed to be carcinogenic. Studies of workers in the former Soviet Union nuclear weapons complex provide evidence of increased cancer following relatively high doses from plutonium exposure; however, internal dosimetry is complex, particularly when reconstructing historical dose estimates using records for these workers.

While accidents can result in larger occupational exposures to ionizing radiation, epidemiological investigations of accidental exposures to ionizing radiation have typically been hampered by poor dosimetry data. The most notable of these was the nuclear accident at the Chernobyl power station on April 26, 1986, which resulted in large-scale occupational exposures to ionizing radiation. More than 600,000 workers participated in the clean-up after the accident. Only a small proportional of these workers, often referred to as "liquidators," received individual dosimeters and, among those who were monitored, problems arose in recording and maintaining dosimetric information. Cleanup workers have been reported to exhibit substantially increased incidence of chromosomal aberrations; while studies of cytogenetic damage can offer some information about an individual's exposure, such approaches are poorly suited to systematic investigation of populations. In a nested case control study that included 34 leukemia cases occurring more than 2 years after the accident and 136 controls, no difference was found in official dose estimates. In contrast, increases in the incidence of leukemia have been reported among cohorts of liquidators in Russia, Belarus, and Ukraine. However, the results of these studies must be interpreted with caution because of problems related to the accuracy of diagnoses and differences in surveillance and registration of cancer cases among liquidators compared to population referents.

6. RADIATION HAZARDS IN THE CONTEXT OF THE NUCLEAR FUEL CYCLE

The activities in the nuclear power industry are dependent on the work done in early and later stages of the nuclear fuel cycle. The nuclear fuel cycle begins with the work of uranium mining and milling, uranium enrichment, and either fabrication of nuclear fuel for reactor operations or weapons production. Workers are also employed in the transportation and disposal of wastes and, in some cases, the reprocessing of spent nuclear fuel. These activities take place in different geographic locations and involve workers employed in many different industries and also involve potential occupational exposures to ionizing radiation. The United Nations Scientific Committee on the Effects of Atomic Radiation has published a useful review of the radiation doses accrued by workers in each of stages of the nuclear fuel cycle. For most workers employed in the nuclear fuel cycle, radiation doses tend to be accrued at low dose rates and may include external as well as internal sources of radiation exposure.

Experimental evidence indicates that DNA damage can be produced by an individual radiation track, suggesting that even very low doses of radiation produce observable biological effects. However, there remains uncertainty about the magnitude of risk for adverse health effects associated with low levels of ionizing radiation and uncertainty about the factors that influence individual susceptibility to radiation-induced disease. A better understanding of the occupational risks associated with chronic low level occupational radiation exposures should come from ongoing

studies of workers in the nuclear industry, which, it is hoped, will help to inform worker protection and compensation programs.

SEE ALSO THE FOLLOWING ARTICLES

Nuclear Engineering • Nuclear Power, History of • Nuclear Power Plants, Decommissioning of • Nuclear Power: Risk Analysis • Nuclear Proliferation and Diversion • Nuclear Waste • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Public Reaction to Nuclear Power Siting and Disposal • Radiation, Risks and Health Impacts of

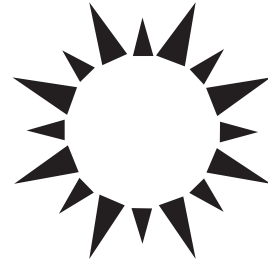
Further Reading

- Committee for the Compilation of Materials on Damage Caused by the Atomic Bombs in Hiroshima and Nagasaki. "Hiroshima, Nagasaki: The Physical, Medical, and Social Effects of the Atomic Bombings." (1981). Iwanami Shoten, Tokyo.
- International Agency for Research on Cancer. "Ionizing Radiation, Part 1: X- and Gamma-Radiation, and Neutrons." (2000). IARC Press, Lyon, France.
- National Research Council, Committee on the Biological Effects of Ionizing Radiation (BEIR V). (1990). "Health Effects of Exposure to Low Levels of Ionizing Radiation (BEIR V)." National Academy Press, Washington, DC.
- United Nations Scientific Committee on the Effects of Atomic Radiation. (2000). "Sources and Effects of Ionizing Radiation." United Nations, New York.
- U.S. Nuclear Regulatory Commission. (2001). "Occupational Radiation Exposure at Commercial Nuclear Power Reactors and Other Facilities, 2001: Thirty-Fourth Annual Report." Report No. NUREG-0713, Vol. 23. U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, Washington, DC.

Ocean, Energy Flows in

RUI XIN HUANG

Woods Hole Oceanographic Institution
Woods Hole, Massachusetts, United States



-
1. Basic Conservation Laws for Ocean Circulation
 2. Energetics of the Oceanic Circulation
 3. Heat Flux in the Oceans
 4. The Ocean Is Not a Heat Engine
 5. Mechanical Energy Flows in the Ocean
 6. Major Challenge Associated with Energetics, Oceanic Modeling, and Climate Change
 7. Conclusion

Glossary

Ekman layer A boundary layer within the top 20 to 30 m of the ocean where the wind stress is balanced by Coriolis and frictional forces.

mixed layer A layer of water with properties that are almost vertically homogenized by wind stirring and convective overturning due to the unstable stratification induced by cooling or salinification.

thermocline A thin layer of water at the depth range of 300 to 800 m in the tropical/subtropical basins, where the vertical temperature gradient is maximum.

thermohaline circulation An overturning flow in the ocean driven by mechanical stirring/mixing, which transports mass, heat, freshwater, and other properties. In addition, the surface heat and freshwater fluxes are necessary for setting up the flow.

wind-driven circulation A circulation in the upper kilometer of the ocean, which is primarily controlled by wind stress. In addition, this circulation also depends on buoyancy forcing and mixing.

Oceanic circulation plays an important role in the climate system because it transports heat and freshwater fluxes meridionally. There is a huge heat flux through the upper surface of the ocean, but the ocean is not a heat engine; instead, the ocean is driven by external mechanical energy, including wind stress and tides. Although mechanical energy flux is 1000 times smaller than the heat flux, it controls the

strength of the oceanic general circulation. In particular, energy input through wind stress is the most important source of mechanical energy in driving the oceanic circulation.

1. BASIC CONSERVATION LAWS FOR OCEAN CIRCULATION

Oceanic circulation is a mechanical system, so it obeys many conservation laws as other mechanical systems do, such as the conservation of mass, angular momentum, vorticity, and energy. It is well known that these laws should be obeyed when we study the oceanic circulation either observationally, theoretically, or numerically. However, this fundamental rule has not always been followed closely in previous studies of the oceanic circulation. As a result, our knowledge of the energetics of the ocean circulation remains preliminary.

1.1 Mass Conservation Law

This is one of the most important laws of nature. However, mass conservation has not been applied in many existing numerical models. In fact, most models are based on the Boussinesq approximations, which consist of three terms: mass conservation is replaced by a volume conservation, *in situ* density in the horizontal momentum equations is replaced by a constant reference density, and tracer prognostic equations are based on volume conservation instead of mass conservation. For models based on the Boussinesq approximations, heating (cooling) can induce an artificial loss (gain) of gravitational potential energy. Furthermore, precipitation (evaporation) is often simulated in terms of the equivalent virtual salt flux out of (into) the ocean; thus, an artificial sink of gravitational potential energy is generated in the models.

1.2 Angular Momentum Conservation Law

In a rotating system, the linear momentum is not conserved; instead the angular momentum conservation is more relevant (i.e., the rate of change of the angular momentum of the oceanic general circulation must be balanced by torques). The angular momentum conservation law has been used extensively in the study of the atmospheric circulation. The Antarctic Circumpolar Current is the corresponding part in the oceans where the angular momentum conservation law dictates the balance.

1.3 Potential Vorticity Conservation Law

The potential vorticity equation can be obtained by cross-differentiation of the horizontal momentum equations and subtracting the results; it has been extensively used in oceanographic studies. According to the theory, total potential vorticity is conserved in the interior of a stratified fluid, and the only source of potential vorticity must be on the outer surfaces of the stratified fluid.

1.4 Energy Conservation Law

Motions in the oceans must be associated with friction; thus, to maintain circulation in the ocean, mechanical energy source is required to balance the loss of mechanical energy. Energy source and dissipation are the most important means of controlling the oceanic circulation system. However,

over the past years the mechanical energy balance of the oceanic general circulation has been overlooked, and there has been very little discussion about such a fundamental issue. In fact, energy balance equations for the oceanic general circulation in most previously published books or papers are incomplete because one of the most important terms in the mechanical energy balance, energy required for vertical mixing in a stratified environment, is not included. Therefore, there is no complete statement about mechanical energy balance for the oceanic general circulation. In addition, most numerical models do not conserve total energy, especially in the finite difference scheme in time stepping. Thus, energetics diagnosed from these models are unreliable.

2. ENERGETICS OF THE OCEANIC CIRCULATION

Energy in the ocean can be classified into four major forms: the gravitational potential energy, kinetic energy, internal energy, and chemical potential, as shown in Fig. 1. Although the chemical potential can be considered as part of the internal energy, it is listed as a separate item in the diagram.

There are four sources (sinks) of gravitational potential energy generated by external forcing. A major source is the tidal force, which is due to the temporal variation of the gravitational forces of the moon and the sun. Although tidal motions have been studied separately from the oceanic general circulation in the past, it is now recognized that a strong

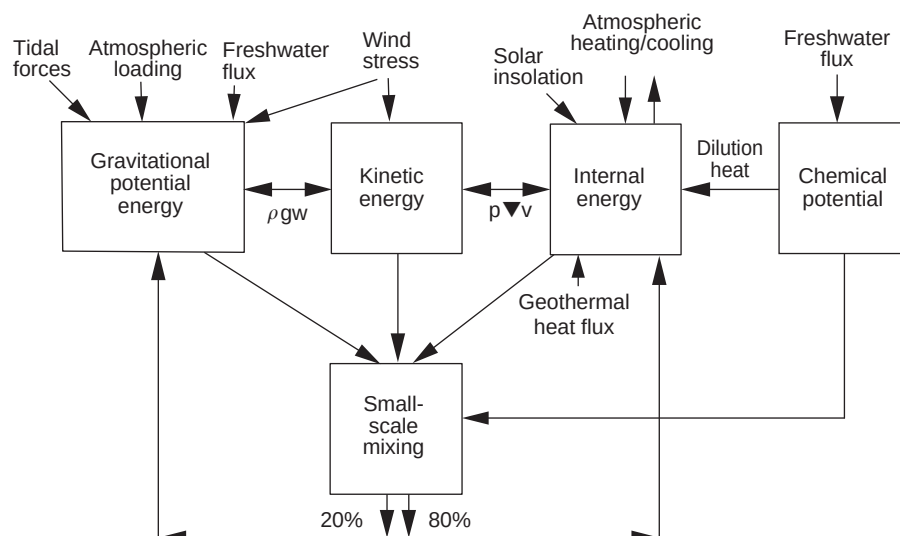


FIGURE 1 Energetics diagram for the ocean circulation.

connection exists between the tidal motions and the oceanic general circulation.

In addition, changes of the sea level atmospheric pressure (called the atmospheric loading) can generate gravitational potential energy. The total amount of this energy flux remains unclear. Furthermore, freshwater flux across the air-sea interface can generate gravitational potential energy. This turns out to be a small sink of energy, and the global integrated amount of energy flux is about $-0.007TW$ ($1TW = 10^{12}$ Watts). The reason for this loss of gravitational potential energy is that evaporation (precipitation) prevails at subtropics (high latitudes) where sea level is high (low) due to warm (cold) temperature. Wind stress is one of the most important sources of mechanical energy for the ocean. In fact, wind stress generates surface waves, thus providing a source of mechanical energy, which is equally partitioned into the gravitational potential energy and kinetic energy. The other source of kinetic energy is the energy input from the wind stress to the surface geostrophic currents and the ageostrophic currents or the Ekman drift.

The external sources of internal energy include three terms: the solar radiation, heat exchange to the atmosphere, and geothermal heat flux. Heat exchange to the atmosphere includes latent heat flux, long wave radiation, and sensible heat flux. Geothermal heat flux through the sea floor is much smaller than the air-sea heat fluxes, but it is one of the important factors controlling the deep circulation in the abyssal ocean.

The source of the chemical potential is associated with the salinity, which is generated by freshwater flux through the air-sea interface in forms of evaporation and precipitation. The general role of chemical potential in the oceanic circulation remains unclear. One of the major roles played by the chemical potential is driving the molecular mixing of salt.

2.1 Energy Conversion

Energy in any given form can be converted into other forms through specific processes, as highlighted in Fig. 1. For example, gravitational potential energy and kinetic energy can be transferred to each other through the vertical velocity conversion term. In addition, kinetic energy can be converted into gravitational potential energy through small scale mixing by internal waves and turbulence. On the other hand, gravitational potential energy can be converted into kinetic energy of meso-scale eddies through baroclinic instability. This may be one of the major pathways of energy flow in the ocean.

2.2 Conversion from Kinetic Energy to Gravitational Potential Energy

In a turbulent ocean, the conversion from kinetic to potential energy is represented by $\overline{\rho w g}$, where the overbar indicates the ensemble mean. This exchange can be separated into several terms:

$$\overline{\rho w g} = \bar{\rho} \bar{w} \bar{g} + \overline{\rho' w' g} + (\overline{\rho' w'} + \overline{\rho' \bar{w}} + \overline{\bar{\rho} w'}).$$

The first term on the right-hand side represents the kinetic to potential energy conversion associated with the ensemble mean density and velocity fields. The second term on the right-hand side represents the contribution due to turbulent mixing. In a stratified fluid, upward motion is usually correlated with positive density anomaly, so turbulent mixing in a stratified ocean increases the gravitational potential energy of the state, and thus requires a source of kinetic energy. The rest of the terms on the right-hand side represent the kinetic to potential energy conversion due to the time-dependent component of the gravitational force, such as the tidal forces.

In the stratified oceans, the vertical eddy mixing rate κ can be related to the energy dissipation rate through

$$\kappa N^2 = \alpha \varepsilon, N^2 = -\frac{g}{\rho_0} \frac{\partial \rho_\theta}{\partial z},$$

where N^2 is the buoyancy frequency, ε is energy dissipation rate, and $\alpha \approx 0.2$ is an empirical coefficient. Therefore, the decline in turbulent kinetic energy does not entirely become “wasted” heat. In fact, a small percentage of the turbulent kinetic energy “dissipation,” represented by the empirical coefficient α , is actually transformed to large-scale potential energy. The energy required to sustain mixing is one of the most important dynamic factors controlling the meridional overturning rate.

2.3 Conversion of Internal Energy to Mechanical Energy

Internal energy can be converted to kinetic energy through the compression term. Since sea water is nearly incompressible, this conversion rate is rather low. In addition, it is very important to note that energy transformation is not unconditional. Instead, the conversion between internal energy and mechanical energy is subject to the second law of thermodynamics—that is, mechanical energy can be converted entirely into internal energy; however, internal energy can be converted into mechanical energy only partially, with the upper bound set by the

Carnot efficiency. As will be discussed later, however, the ocean is not a heat engine; therefore, the equivalent thermal-mechanical energy conversion efficiency is negative.

3. HEAT FLUX IN THE OCEANS

Oceans play an important role in the earth's climate system in many ways. In fact, most of the solar radiation penetrates the atmosphere and is first absorbed by the ocean and the land and then sent back to the atmosphere in forms of latent heat, long wave radiation, and sensible heat. To a large degree, these processes are controlled by the oceanic circulation. In the quasi-steady state, all heat flux absorbed by the ocean must be released back into the atmosphere. The spatial distribution of air-sea heat flux is one of the critical components of the earth's climate system.

3.1 Heat Flux through the Upper Surface of the Ocean

The dominating form of energy flux in the oceans is the heat flux. Solar radiation on Earth can penetrate the atmosphere, with only a small fraction being absorbed by the greenhouse gases, such as water vapor and CO_2 . Since oceans cover about 70% of the earth, most of the solar radiation is directly absorbed by the oceans (total amount of 52.4PW).

At low latitudes, solar radiation absorbed by the ocean exceeds the heat flux from the ocean into the atmosphere, including the latent heat flux, long wave radiation, and the sensible heat flux. Thus, the ocean at low latitudes gains net heat through the upper surface, and this extra heat is transported poleward and released at middle and high latitudes. The most important areas of the ocean-to-atmosphere heat release include the Gulf Stream, Kuroshio, and the Northern North Atlantic.

3.2 Poleward Heat Fluxes in the Climate System

Heat fluxes in the current climate system consist of three components: the atmospheric sensible heat flux, the oceanic sensible heat flux, and the latent heat flux associated with the hydrological cycle in the atmosphere-ocean coupled system (Fig. 2). The sum of these three fluxes is the total poleward heat flux diagnosed from satellite measurements. The ocean plays the role of heat storage and heat redistributor. The most important part for climate study is the total

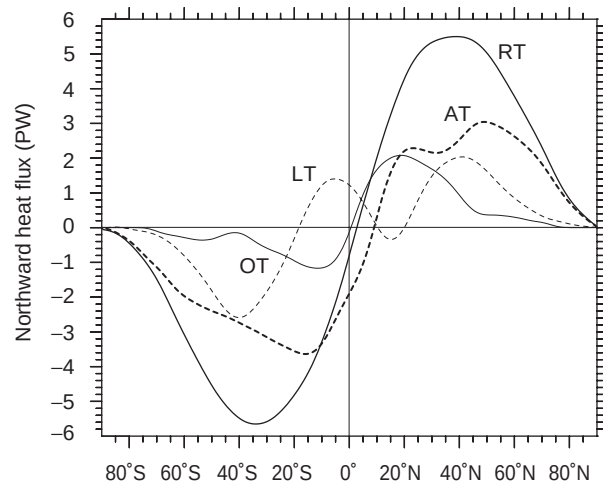


FIGURE 2 Northward heat flux in PW: RT indicates the heat flux required for radiation balance as diagnosed from satellite data; AT indicates the atmospheric sensible heat flux; OT indicates the oceanic sensible heat flux; and LT indicates the latent heat flux associated with the hydrological cycle.

amount of poleward heat flux associated with the oceanic current, which is about 2PW . Note that the latent heat flux, which has been traditionally associated with the moisture flux in the atmosphere, can be considered as an atmosphere-ocean coupled mode.

Although the atmospheric sensible heat flux is the largest component in both hemispheres, it is not much larger than the other components. Thus, the total oceanic contribution to the poleward heat flux may consist roughly of 50% of the total heat flux.

The poleward sensible heat flux in the ocean can be separated into two major components: the flux associated with the meridional overturning cell and the flux associated with the horizontal gyration. Each component can be further divided into the contribution by the mean Eulerian flow and eddies. For example, poleward heat flux in the North Atlantic reaches its maximum around 25°N , where both the mean meridional overturning and horizontal wind-driven gyre contribute. In comparison, however, the contribution by eddies is relatively small. In the North Pacific, the mean meridional overturning cell induces an equatorward heat flux; however, the sum of heat fluxes is still poleward because it is associated with the domination of the wind-driven gyre.

3.3 Total Heat Flux through the Oceans

The oceanic to atmospheric heat flux consists of three components: the latent heat flux, 31.1PW ; the long wave radiation, 17.8PW ; and the sensible heat flux, 3.5PW (Fig. 3). Although air-sea interface heat flux is

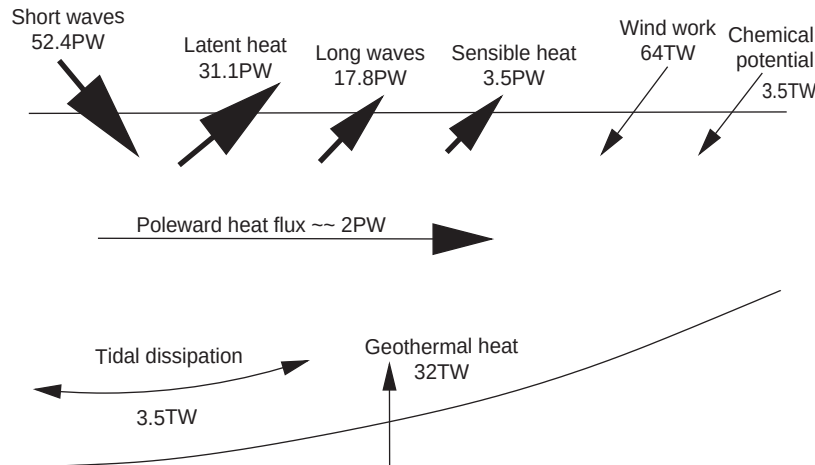


FIGURE 3 A diagram of external sources of energy for the ocean circulation.

of major importance to climate study, it does not directly drive the oceanic general circulation. Instead, air-sea heat flux is controlled by the oceanic circulation or, more accurately, controlled by the atmosphere-ocean-land coupled system. From the energetic point of view, the oceanic circulation is governed by many laws of conservation, in particular the energy conservation law. The thermohaline boundary conditions at the sea surface are preconditions for the thermohaline circulation only, while the ocean circulation is primarily controlled by the sources and sinks of mechanical energy, with the total air-sea heat flux as the by product of the circulation.

In Fig. 3, energy fluxes into the ocean in other forms are also included because this additional energy input eventually turns into internal energy and is released in to the atmosphere in forms of heat flux. The amount of geothermal heat flux is about 32TW , which is eventually lost in to the atmosphere through the air-sea heat exchange. Furthermore, there are three other fluxes: the tidal dissipation, 3.5TW ; the chemical potential generated through evaporation and precipitation; and the wind work, 64TW . Thus, the total outgoing heat flux is the sum of the incoming heat flux, 52.4PW , plus four additional energy sources: 64 (wind) + 32 (geothermal) + 3.5 (tides) + 3.5TW (chemical potential) = 103TW .

4. THE OCEAN IS NOT A HEAT ENGINE

4.1 Is the Ocean a Heat Engine?

The atmosphere and oceans work together as a heat engine, if we neglect the small contribution of tidal

energy to the circulation in the oceans. The atmosphere can be considered as a heat engine, which is driven by differential heating, with an efficiency of 0.8% (the corresponding Carnot efficiency is about 33%). In this sense although the oceans are subject to differential heating similar to the atmosphere, they are not at all a heat engine. As will be discussed here, the driving force for the oceanic circulation is the mechanical energy in the forms of wind stress and tides. In comparison, differential heating is only a precondition for the thermohaline circulation, and not the driving force of the oceanic general circulation. Thus, the ocean is not a heat engine; instead, it is a machine driven by external mechanical energy that transports thermal energy, freshwater, CO_2 , and other tracers.

4.2 Sandstrom's Theorem

Sandstrom theorem states: A closed steady circulation can be maintained in the ocean only if the heating source is situated at a level lower than the cooling source. Sandstrom's theorem can be demonstrated by laboratory experiments, as shown in Fig. 4. In the first experiment, the heating source is put at a level higher than the cooling source. There is no detectable circulation, and a stable stratification can be observed between the heating and cooling levels. Theoretically, there is still a circulation driven by the molecular mixing, but it is so sluggish that there is practically none. In the second experiment, the heating source is put at a level lower than the cooling source. A vigorous circulation can be observed between the levels of heating and cooling, just as we put a pot of water over the kitchen range.

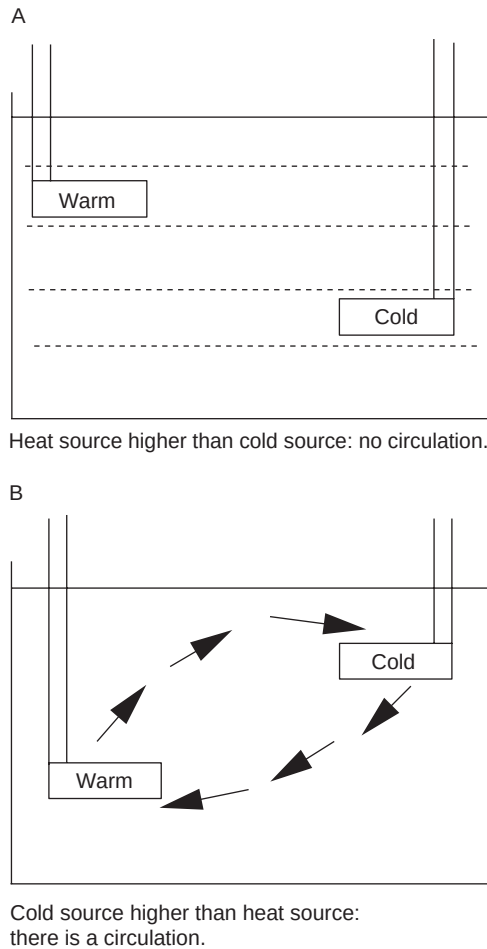


FIGURE 4 Laboratory experiments demonstrating the Sandstrom theorem.

Sandstrom's theorem applies to both the atmosphere and oceans. The atmosphere is heated from below and cooled from above; thus, it is a heat engine. On the other hand, the ocean is heated and cooled from the sea surface. Since heating and cooling take place at the same level, there is no circulation driven by the thermal forcing (Fig. 5A). Under such a condition, thermal forcing can create a very thin thermocline, on the order of meters on the surface of the oceans. Below such a thin thermocline, temperature is virtually constant, equal to the coldest temperature set up by the deepwater formation at high latitudes. Since there is no density difference at depth, there is only a very sluggish circulation. However, the effective depth of heating is moved downward by wind and tidal mixing. The large horizontal density gradient thus created drives a strong circulation in the world oceans (Fig. 5B).

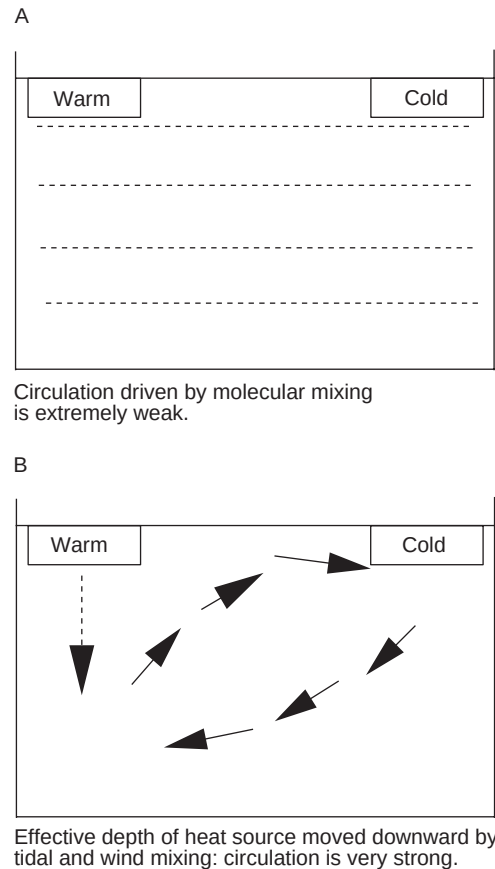


FIGURE 5 Application of the Sandstrom theorem to the world's oceans.

The essential point is that to maintain a quasi-steady circulation in the ocean, external sources of mechanical energy are required to balance the loss of mechanical energy caused by friction associated with motion. Although there is a tremendous amount of thermal energy going through the ocean, it cannot be converted into mechanical energy needed to sustain the oceanic circulation, as dictated by the fundamental law of thermodynamics. Therefore, instead of utilizing the huge amount of thermal energy going through the ocean, the maintenance of the oceanic circulation depends on the availability of external mechanical energy, with a flux rate of three to four orders of magnitude smaller than the thermal energy fluxes (Fig. 3).

5. MECHANICAL ENERGY FLOWS IN THE OCEAN

Mechanical energy sources and sinks for the oceans are illustrated in Fig. 6. Wind stress applies to the

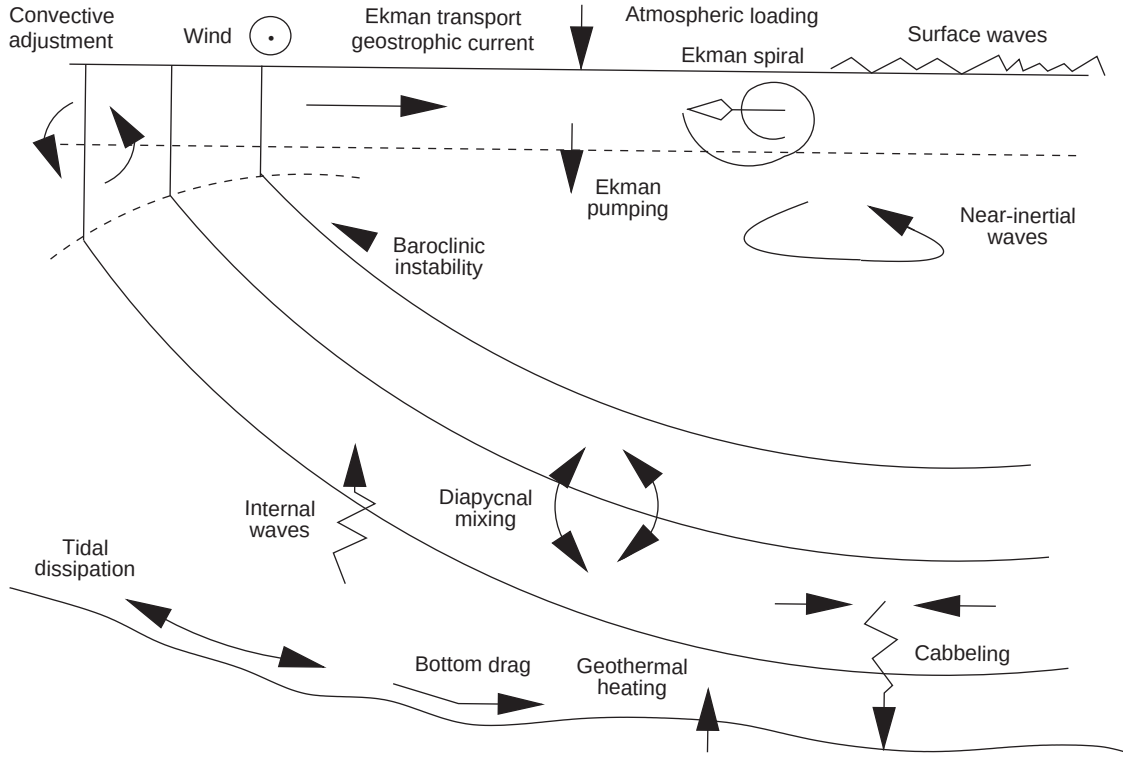


FIGURE 6 Mechanical energy diagram for the ocean circulation.

surface of the ocean and drives both surface currents and waves. The mechanical energy input from wind stress into the ocean W_{wind} is defined by

$$W_{wind} = \langle \sigma_{ij} \rangle \cdot \vec{u} = \vec{\tau} \cdot \vec{U}_0 + \overline{\vec{\tau}' \vec{u}'_0} + \overline{p' w'_0},$$

where $\vec{\tau}$ and $\vec{U}_0$ are the spatially averaged tangential stress and surface velocity respectively, $\vec{\tau}'$ and $\vec{u}'_0$ are the perturbations, and p' and w'_0 are perturbations of the surface pressure and velocity component normal to the surface. The perturbations are defined in terms of the scale of the surface wavelength.

The first term on the right-hand side is the wind stress work of the quasi-steady current on the surface, and quasi-steadiness is defined in comparison to the timescale of the typical surface waves. $\vec{U}_0$ has two components: the geostrophic current and the Ekman drift:

$$\vec{U}_0 = \vec{U}_{0,G} + \vec{U}_{0,Ekman}.$$

5.1 Geostrophic Current

Energy input through the surface geostrophic current can be calculated as the scalar product of wind stress and surface geostrophic velocity. The recent calculation by Carl Wunsch puts the global total as 1.3TW. This energy is directly fed into the large-

scale current, so it can be efficiently turned into gravitational potential energy through the vertical velocity conversion term.

5.2 Surface Drift

There is a frictional boundary layer (Ekman layer) on the top of the ocean, in which the wind stress is balanced by frictional force. The surface drift is called the Ekman drift. The exact amount of wind stress energy input through the Ekman spiral is unclear, and the preliminary estimate made by Wei Wang and Rui Xin Huang is about 2.4TW for the frequency higher than 1/(2 days). In addition, there is a large amount of energy input through the near-inertial waves, which is due to the resonance at a frequency of $\omega = -f$; that is, for inertial frequency of the clockwise (anticlockwise) wind stress in the Northern (Southern) Hemisphere. Recent calculations by Watanaba and Hibiya gave an estimate of 0.7TW.

5.3 Geopotential and Kinetic Energy of the Wind-Driven Circulation

The convergence of the Ekman flux gives rise to a pumping velocity at the base of the Ekman layer W_E ,

which is responsible for pushing the warm water into the subsurface ocean and thus forming the main thermocline in the subtropical ocean. The main thermocline is a relatively thin layer of water at the depth of 500 to 800 m in the upper ocean where the vertical temperature gradient is maximum. Since salinity contribution to the density is relatively small in most places of the oceans, the thermocline is also close to the pycnocline where the vertical gradient of density is maximum. Ekman pumping sets up the wind-driven circulation and the associated bowl-shaped main thermocline in the world's oceans. In this process, gravitational potential energy is increased, which is a very efficient way of converting kinetic energy into gravitational potential energy. Simple scaling analysis shows that the amount of available gravitational potential energy associated with the main thermocline is about 1000 times larger than the amount of kinetic energy associated with the mean flow of the wind-driven circulation.

5.4 Loss of Gravitational Potential Energy through Convective Adjustment

Another important process taking place in the mixed layer is the convective adjustment due to cooling and salinification. According to the Sandstrom theorem, if both heating and cooling apply to the same level (the sea surface), there is no gravitational potential energy being generated by the thermal forcing alone. However, heating and cooling in the mixed layer is not a linear process. During the cooling process, water at the sea surface becomes heavier than water at the depth, thus a gravitational unstable stratification appears. Due to this unstable stratification, a rapid convective adjustment process takes place and

density becomes nearly homogenized in the upper part of the water column. During this process, gravitational potential energy of the mean state is converted into kinetic energy for turbulence and internal waves (Fig. 7A).

As a result, the effective center of cooling is not at the sea surface; instead it is located halfway of the well-mixed layer's depth. Because of this asymmetry associated with cooling/salinification, surface buoyancy forcing gives rise to a sink of gravitational potential energy. The total amount of energy loss through this process remains unclear. A preliminary estimate based on the monthly mean climatology by Rui Xin Huang and Wei Wang for this sink term is about 0.24TW. However, calculation in which the diurnal cycle is resolved may give rise to a much larger value.

5.5 Baroclinic Instability

The steep isopycnal surface in the oceans, such as that along the meridional edges of the wind-driven gyre, is unstable due to baroclinic instability. Because of this, a large amount of gravitational potential energy of the mean state can be converted into the eddy kinetic/potential energy (Fig. 7B). It is estimated that the amount of eddy kinetic energy is about 100 times larger than the kinetic energy of the time-mean flow. Despite great effort in carrying out field observations and numerical simulation, there is no reliable estimate on the total amount of eddy kinetic energy in the world's oceans. However, satellite observations have provided global distribution of both the mean and eddy kinetic energy on the sea surface. The ratio of these two forms of kinetic energy is on the order of 100, consistent with

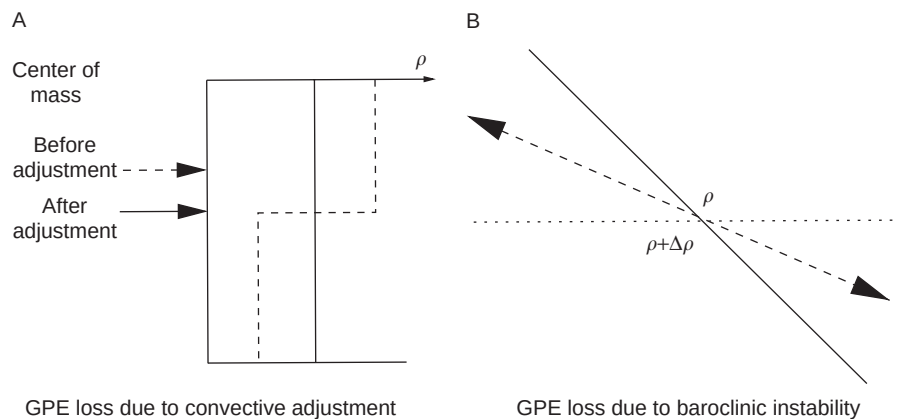


FIGURE 7 Gravitational potential energy loss due to convective adjustment and baroclinic instability.

theoretical prediction based on scaling. The rate of energy conversion over the world's oceans through the baroclinic instability remains unknown, and the recent estimate by Rui Xin Huang and Wei Wang is about 1.1TW. Since most eddy energy is dissipated through small-scale processes, the conversion from the mean state to eddies is considered as a sink of the gravitational potential energy of the mean state.

5.6 Surface Waves

Another major source of energy caused by wind stress is fed through the surface waves. Wind stress drives surface waves in the oceans, and this energy input can be treated as the form drag for the atmospheric boundary layer. The rate of energy input through surface waves remains unclear; however, the preliminary estimate by Wei Wang and Rui Xin Huang is about 60TW. Although this seems to be a large amount of energy, it is believed that most of it is transformed into long waves and propagates away from the source area. It is likely that only a small fraction of this energy is dissipated locally through wave breaking and white capping, but a major fraction of this incoming energy may be dissipated to remote places in forms of swell. These long waves have a rather low dissipation rate, their energy is likely to dissipate cascading into other forms of energy or through wave breaking along the beaches in the world's oceans; however, the details of this dissipation mechanism remain unclear.

5.7 Diapycnal and along Isopycnal Mixing

Tracers, including temperature and salinity, are mixed in the oceans through internal wave breaking and turbulence. Mixing can be classified into diapycnal mixing and along isopycnal mixing. Since along isopycnal mixing involves the least amount of gravitational potential energy, it is the dominating form of mixing on large scales.

On the other hand, in stratified fluid vertical mixing increases gravitational potential energy because light fluid is pushed downward and heavy fluid is pushed upward. Although molecular diffusion can play the role of mixing, the corresponding rate is too small and thus the result is irrelevant to the oceans. In the upper ocean, below the mixed layer, diapycnal mixing rate is on the order of $10^{-5} m^2 s^{-1}$, which is much larger than the mixing rate due to molecular mixing. In other places of the oceans, the mixing rate

can be much higher than this background rate. The strong diapycnal (or vertical) mixing in the oceans is driven by strong internal wave and turbulence.

The energy sources supporting diapycnal mixing in the oceans include wind stress input through the sea surface and tidal dissipation, which will be discussed in the next section. Turbulent mixing associated with fast current, especially that associated with flow over sill and the down-slope flow afterward, can provide a strong source of energy supporting mixing. There is much observational evidence indicating that mixing can be on the order of 10^{-3} – $10^{-1} m^2 s^{-1}$. The mixing rate can be as large as $0.1 m^2 s^{-1}$ within the Romanche Fracture Zone, where water drops more than 500 m within 100 km of a downward flow.

Internal lee waves generated by flow over topography, such as the Antarctic Circumpolar Current, are probably one of the most important contributors.

The total amount of energy supporting diapycnal mixing in the oceans remains unclear because mixing is highly nonuniform in space and time. According to the theory of thermohaline circulation, the meridional overturning rate is directly controlled by the strength of the external mechanical energy available for supporting mixing in the oceans, including mixing in the mixed layer and mixing in the subsurface layers. Therefore, understanding the physics related to the spatial and temporal distribution of mixing is one of the most important research frontiers in physical oceanography.

5.8 Tidal Dissipation

The primary source of mechanical energy supporting mixing likely comes from both the wind stress applied to the ocean's surface and tidal dissipation in the deep ocean. The total amount of tidal dissipation in the world's oceans is 3.5TW (Fig. 8), which is calculated from accurate tracking of the moon's orbit. The spatial distribution of tidal energy dissipation remains inaccurate. Based on satellite altimeter data assimilation, Walter Munk and Carl Wunsch estimated that 2.6TW is dissipated within the shallow seas of the world's oceans, and the remainder 0.9TW is believed to be distributed in the deep ocean (Fig. 6). Tidal dissipation in the deep ocean takes place primarily over rough topography, where barotropic tidal energy is converted into energy for internal tides and internal waves that sustain the bottom-intensified diapycnal mixing on the order of $10^{-3} m^2 s^{-1}$. The conversion of kinetic energy to gravitational potential energy through

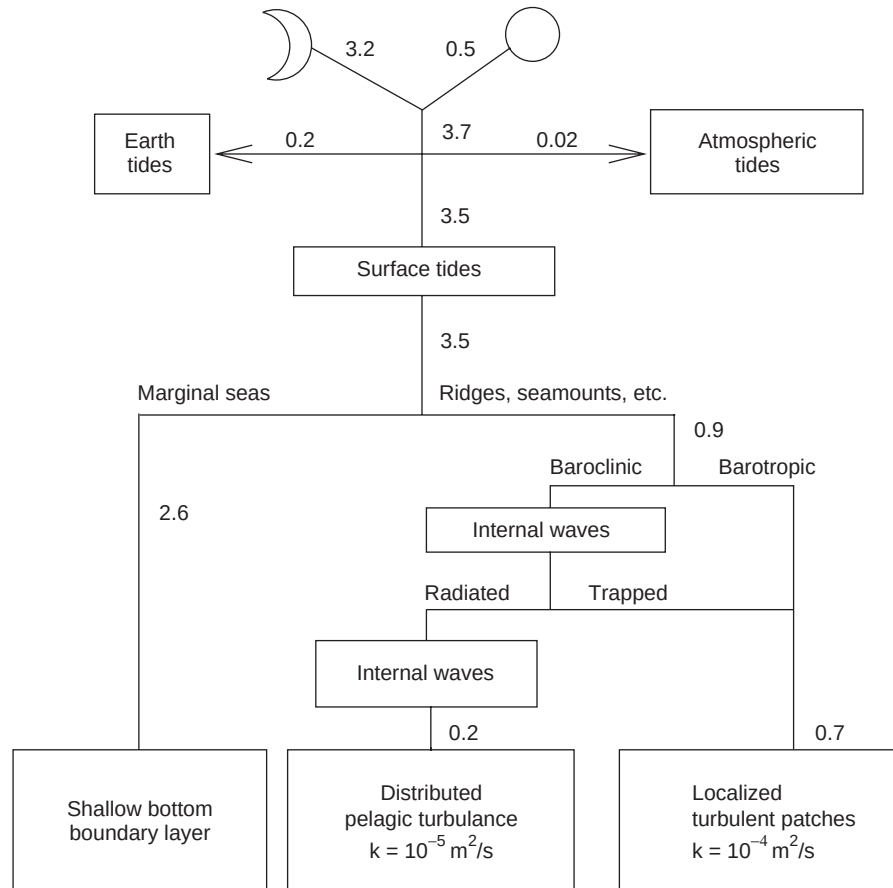


FIGURE 8 Tidal dissipation diagram, unit in TW. Modified from Munk and Wunsch (1998).

internal waves and turbulence is of rather low efficiency, typically on the range of 20%; thus, the amount of gravitational potential energy generated by tidal dissipation in the deep ocean is about 0.18TW.

Tidal dissipation has varied greatly over the geological past, thus the energy supporting diapycnal mixing may vary as well. The reader is referred to the review by Kagan and Sundermann.

5.9 Geothermal Heat

In addition, geothermal heat or hot plumes provide a total heat flux of 32TW. Although this is much smaller than the heat flux across the air-sea interface, it may be a significant component of the driving force for the abyssal circulation. Since the geothermal heating is applied at great depth, and the corresponding cooling takes place at the sea surface, geothermal heat can be more efficiently converted into gravitational potential energy. The rate of this conversion is about 0.05TW, which is a small term compared to other major terms; nevertheless it is not negligible,

especially for the abyssal circulation and temperature distribution in the abyss.

5.10 Bottom Drag

Oceanic currents moving over rough bottom topography must overcome bottom or form drag, as shown in Fig. 6. The total amount of bottom drag for the world's oceans circulation remains unclear; however, the preliminary estimate for the open oceans is about 0.4TW.

5.11 Atmospheric Loading

Finally, sea level atmospheric pressure varies with time, and the sea surface moves in the vertical direction in response. As a result, changes in sea level atmospheric pressure can input mechanical energy into the oceans. The total amount of energy due to this source for the world's ocean circulation remains unclear; however, the preliminary estimate for the open oceans is about 0.04TW.

5.12 Miscellaneous

Many other mechanisms contribute to the mechanical energy balance in the oceans.

Cabbeling. Due to the nonlinearity of the equation of state, water density is increased during both diapycnal and isopycnal mixing. The newly formed water with higher density sinks is called cabbeling, as shown in Fig. 6. As a result, gravitational potential energy is converted into energy for internal waves and turbulence. The total amount of gravitational potential energy loss associated with cabbeling in the oceans remains unclear.

Double diffusion: Seawater contains salt, thus it is a two-component chemical mixture. On the level of molecular mixing, the heat diffusion is 100 times faster than the salt diffusion. The difference in heat and salt diffusion for laminar fluid and turbulent fluid environment is one of the most important aspects of the thermohaline circulation in the oceans.

Double diffusion in the oceans primarily manifests itself by two types: salt fingers and diffusive convection. Salt fingers appear when warm and salty water lie over cold and freshwater. The mechanism that drives the instability is illustrated in Fig. 9A, assuming a vertical pipe connects the cold and freshwater in the lower part of the water column with the warm and salty water in the upper part of the water column. The wall of the pipe is very thin, so it allows a rather efficient heat flux into the pipe thus warming up the cold water. At the same time, the low salt diffusivity preserves the freshness of the ascending water parcel. Thus, the buoyancy difference drives the upward motion of water inside the pipe, and a self-propelled fountain can be built in the ocean.

The subtropical gyre interior is a salt-finger's favorite place, where strong solar insolation and excessive evaporation leads to warm and salty water above the main thermocline. In this salt-finger favorite regime, warm and salty fingers move downward and cold and fresh plumes move upward. Since heat is 100 times more easily diffused between the salt fingers and the environment, salt fingers lose their buoyancy and continue their movement downward. In this way gravitational potential energy released is used to drive the mixing. According to the new insite observations in the upper thermocline of the subtropical North Atlantic, the mixing rate for the temperature (salinity) is on the order of $4 \times 10^{-5} m^2 s^{-1}$ ($8 \times 10^{-5} m^2 s^{-1}$), and the equivalent density mixing rate is negative. The total amount of gravitational energy loss due to salt fingering in the world's oceans remain unknown.

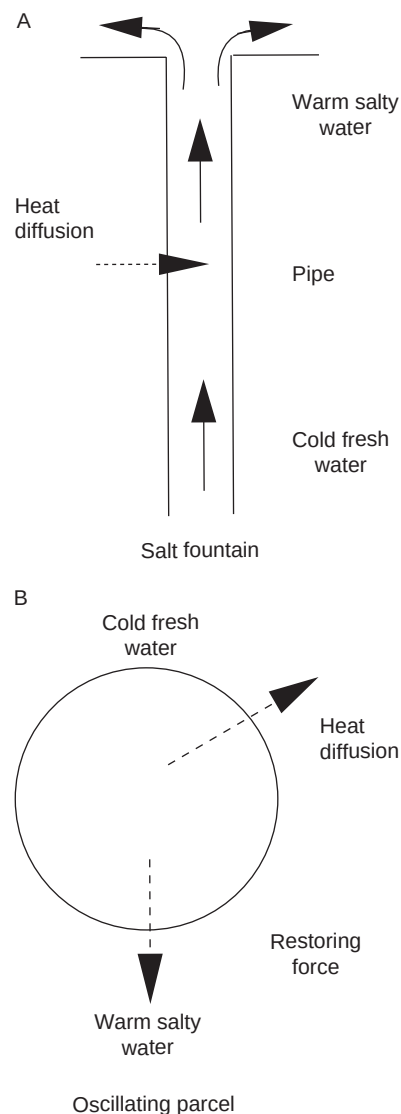


FIGURE 9 Two possible cases for double diffusion in the oceans.

The other possible double diffusive process is the diffusive convection which takes place if cold and freshwater overlay warm saltwater. The system is relatively stable and allows an oscillatory instability (Fig. 9b).

5.13 Attempt at Balancing the Mechanical Energy in the Ocean

The balance of mechanical energy, including both the kinetic energy and gravitational potential energy, is shown in Fig. 10. It is clear that at this time we do not know even the lowest-order balance of mechanical energy. There is much kinetic energy input into

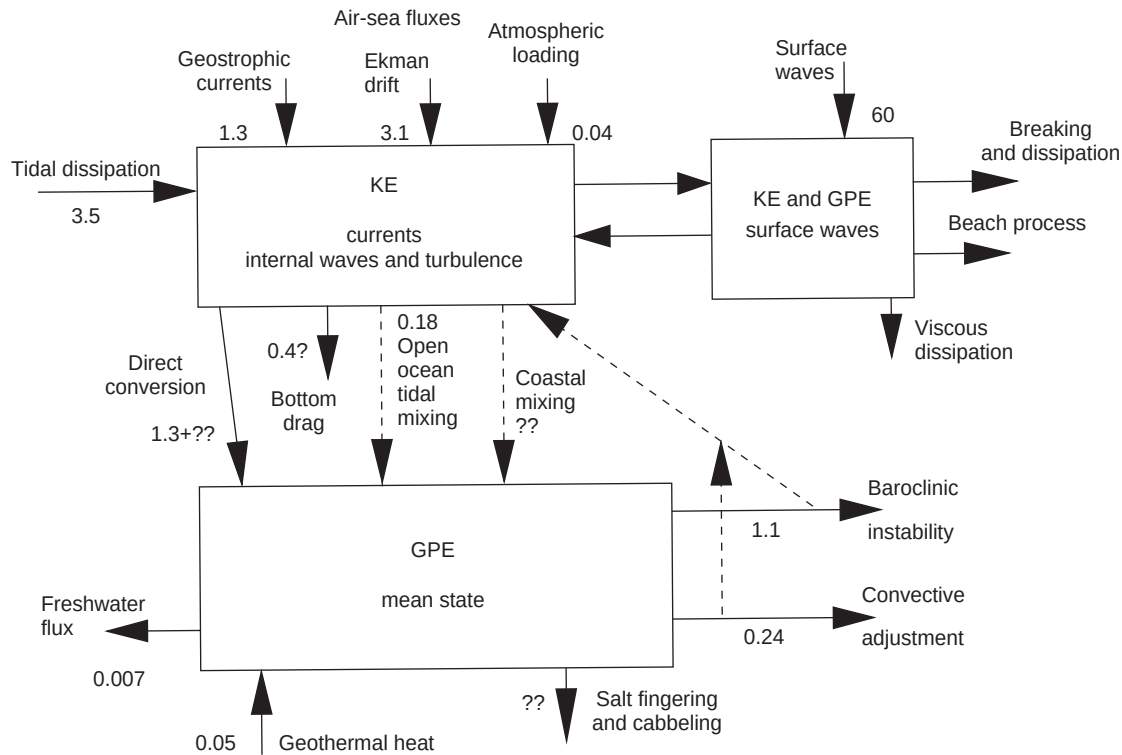


FIGURE 10 Mechanical energy balance for the world's oceans, unit in TW.

the ocean, but it is not clear how such energy is distributed and eventually dissipated in the oceans. On the other hand, there are two major sinks of gravitational potential energy, but it is not clear how energy is transported and supplied to such energy sinks. It is fair to say that most of the energy fluxes listed in Fig. 10 are accurate to the factor of 2 only. More accurate energy pathways and estimates require further study.

6. MAJOR CHALLENGE ASSOCIATED WITH ENERGETICS, OCEANIC MODELING, AND CLIMATE CHANGE

6.1 Scaling Laws Governing the Meridional Circulation

Scaling analysis shows that the meridional mass flux and heat flux are linearly proportional to the mechanical energy available for mixing. Thus, the amount of energy available for supporting mixing is what really controls the strength of circulation, so more energy available for mixing means stronger circulation.

6.2 Energetics and Numerical Simulation of Ocean Circulation

Since we do not well know the distribution of mechanical energy available for mixing, many oceanic general circulation models of the current generation are based on rather simple parameterization of subgrid mixing. A common practice in oceanic general circulation modeling is to choose a diapycnal mixing rate and let the models run; the diapycnal mixing rate remains the same all the time, even under different climatic conditions. In fact, no modelers have ever checked how much energy is required to support mixing and circulation and, most important, whether this amount of energy is really available.

An interesting and important question is how to parameterize mixing in climate-related models. Should we use the same mixing rate or same mixing energy when the climate is drifting? Clearly, both a fixed mixing rate and a fixed mixing energy represent the two extremes, and the reality is likely to be between these two assumptions.

It is speculated that when we have the new generation of the Oceanic General Circulation Model (OGCM), in which the mixing parameterization is more rational, climate variability induced by something like the CO_2 doubling may be different

from whatever we have learned based on existing models.

7. CONCLUSION

In summary, energetics of oceanic circulation is of capital importance in understanding ocean circulation and climate. Energetics of the oceanic circulation is one the most important research frontiers. There are many unanswered fundamental questions, and at this time, we do not have the lowest-order balance. Most flux terms cited here should be treated as preliminary estimates, which will certainly be improved in the near future. Since circulation in the oceans is primarily controlled by energy available for mixing, further study of energetics will bring about major breakthroughs in our understanding of the ocean circulation and climate.

SEE ALSO THE FOLLOWING ARTICLES

Aquaculture and Energy Use • Conservation of Energy Concept, History of • Desalination and Energy Use • Earth's Energy Balance • Fisheries and

Energy Use • Lithosphere, Energy Flows in • Ocean Thermal Energy

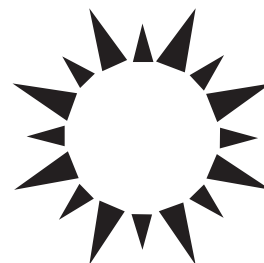
Further Reading

- Bryden, H. L., and Imawaki, S. (2001). Ocean heat transport. In "Ocean Circulation and Climate" (G. Siedler, J. Church, and J. Gould, Eds.), International Geophysical Series, pp. 455–474. Academic Press, New York.
- Garrett, C., and St. Laurent, L. (2002). Aspects of ocean mixing. *J. Oceanogr. Soc. Japan* **58**, 11–24.
- Gill, A. E., Green, J. S. A., and Simmons, A. J. (1974). Energy partition in the large-scale ocean circulation and the production of mid-ocean eddies. *Deep-Sea Research* **21**, 499–528.
- Huang, R. X. (1998). Mixing and available potential energy in a Boussinesq ocean. *J. Phys. Oceanogr.* **28**, 669–678.
- Huang, R. X. (1999). Mixing and energetics of the thermohaline circulation. *J. Phys. Oceanogr.* **29**, 727–746.
- Kagan, B. A., and Sundermann, J. (1996). Dissipation of tidal energy, paleotides, and evolution of the earth-moon system. *Adv. Geophysics* **38**, 179–266.
- Munk, W. H., and Wunsch, C. (1998). The moon and mixing: Abyssal recipes II. *Deep Sea Research I* **45**, 1977–2010.
- Trenberth, K. E., and Caron, J. M. (2001). Estimates of meridional atmosphere and ocean heat transport. *J. Climate* **14**, 3433–3443.
- Wunsch, C. (1998). The work done by the wind on the oceanic general circulation. *J. Phys. Oceanogr.* **28**, 2331–2339.
- Wunsch, C. (2001). Ocean observations and the climate forecast problem. In "Meteorology at the Millennium" (R. Pearce, Ed.), Royal Meteorological Society, pp. 233–245. Academic Press, New York.

Ocean Thermal Energy

DON E. LENNARD

Ocean Thermal Energy Conversion Systems Ltd.
Kent, United Kingdom



-
1. Definition and Basis of OTEC
 2. The Beginnings of OTEC
 3. Uses of OTEC Power—and Other Products
 4. Variations on OTEC Plants and Factors Affecting Them
 5. Technology of OTEC Plants
 6. Some OTEC Projects and Designs
 7. Cost, Economy, and Competitiveness
 8. Markets and the Way Ahead

Glossary

closed cycle When the working fluid of the heat engine cycle is fully contained and isolated, changing from liquid to gaseous phases and back again.

deep ocean water applications (DOWAs) Alternative products and processes that can be derived from an ocean thermal energy conversion plant.

exclusive economic zone (EEZ) The area/volume of sea/sea bed, and its resources, that can be claimed by a nation under the UN Convention on the Law of the Sea.

International OTEC/DOWA Association (IOA) The worldwide grouping for all aspects of ocean thermal energy conversion and deep ocean water applications.

ocean thermal energy conversion (OTEC) The technique of extracting energy from the vertical temperature difference in the oceans.

open cycle When seawater is the working fluid, is used in free flow, and is not recirculated.

The surfaces of the oceans capture huge amounts of solar energy—some thousands of times more than the energy consumed by our current civilization—with most of this being stored in the form of thermal energy in the surface layers of the oceans. Those surface layers do not mix freely or easily with the deeper waters, which are much colder. Ocean thermal energy conversion (OTEC) is based on the extraction of energy from the temperature difference existing between those warm surface waters of the

oceans in extensive tropical and subtropical areas and the deep waters in those same areas that flow from the polar regions.

1. DEFINITION AND BASIS OF OTEC

Ocean thermal energy conversion (OTEC) is based on the extraction of energy from the temperature difference existing between those warm surface waters of the oceans in extensive tropical and subtropical areas (approximately latitudes 25° north to 25° south) and the deep waters in those same areas that flow from the polar regions (predominantly the Antarctic). Surface temperatures can reach as high as 29°C in the Pacific, whereas the temperature at a depth of 1000 m will typically be 4°C, although this temperature can sometimes be found at depths as shallow as 700 m. Figure 1 shows the general scale and distribution of this thermal resource. The typical design case for an OTEC plant assumes a temperature difference of 20°C.

2. THE BEGINNINGS OF OTEC

Ocean thermal energy conversion (OTEC) has existed for more than 120 years, but its technological status now owes very much to developments in offshore activities associated with the oil and gas industries during the latter part of the 20th century—more particularly as exploitation of those fossil fuel resources has moved into deeper waters and harsher climates.

The OTEC power circuit is that of a standard heat engine cycle, but where the temperature difference is much less than that in the case of a steam engine or an internal combustion engine. It was in 1881 that the Frenchman Arsène d'Arsonval suggested using

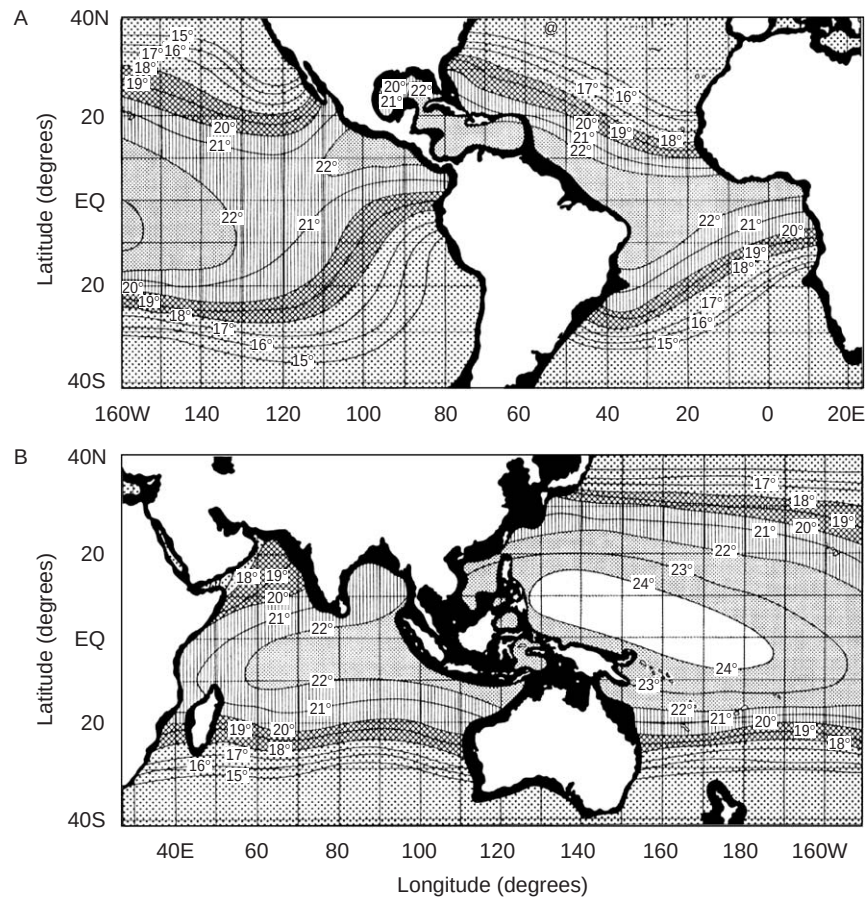


FIGURE 1 The ocean thermal resource: Worldwide distribution. Average temperature differences (in °C) between the surface and a depth of 1000 m. From U.S. Department of Energy/Ocean Data Systems Inc.

any working fluid having an appropriate vapor pressure at a temperature close to that of warm seawater. After cooling in a surface condenser, the liquid working fluid is circulated by pumping to the evaporator, where the fluid changes to a gaseous phase with a considerable increase in volume and pressure, passes through a turbine connected to an electrical generator, and is then recycled on through the condenser. This concept, known as closed cycle OTEC, is shown in Fig. 2. Working fluids, including ammonia, propane, butane, and freon, are suitable. Closed cycle OTEC requires the design and construction of large surface heat exchangers that are close to—some say beyond—the limits of currently available technology to enable the construction of modular closed cycle OTEC plants if the capacity is in excess of approximately 100 MW.

A second option was proposed by another Frenchman, George Claude, during the late 1920s. In Claude's OTEC concept, the working fluid is the vapor formed by the warm seawater itself when

boiled in an evaporator maintained at an appropriate low pressure. Water vapor is condensed either through a direct contact condenser by mixing with cold seawater or directly through a surface condenser. In the latter case, desalinated water is a by-product of the thermal process. In neither case is the condensed vapor reintroduced to the working fluid circuit. Claude's concept, known as open cycle OTEC, is shown in Fig. 3. The main technical difficulty for open cycle OTEC plants comes from the low vapor pressure of the working fluid; this requires very large turbines. Available technology suggests that the construction of modular open cycle OTEC plants with capacities of tens of megawatts is currently possible.

3. USES OF OTEC POWER—AND OTHER PRODUCTS

There are significant energy requirements in the tropical and subtropical areas where OTEC is most

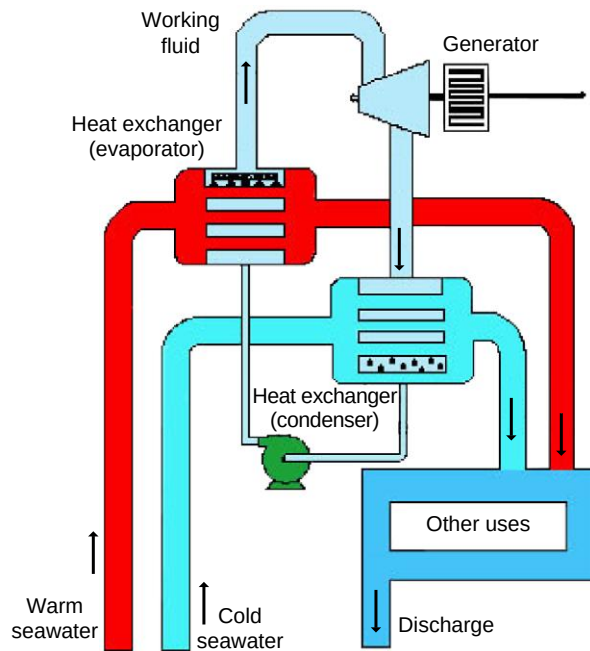


FIGURE 2 Closed cycle OTEC: Power circuit. From National Energy Laboratory of Hawaii Authority.

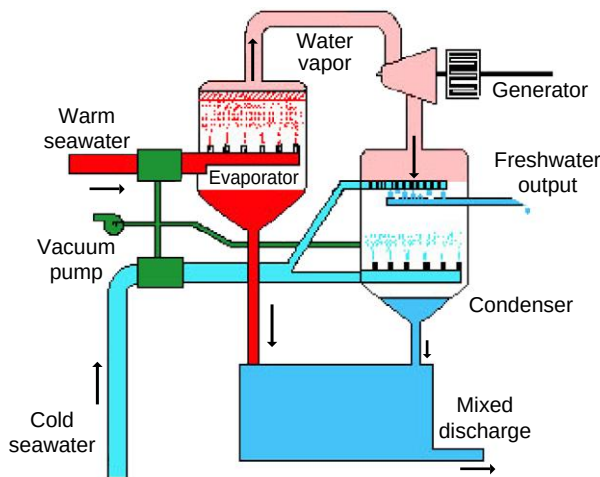


FIGURE 3 Open cycle OTEC: Power circuit. From National Energy Laboratory of Hawaii Authority.

attractive, but there also are a number of possible “by-products” from OTEC, the benefits from which are also considerable. Usually referred to as deep ocean water applications (DOWAs), examples of these include aquaculture and the production of potable water (Fig. 4 shows a flow chart for such a combination). Potable water can be an immediate product of open cycle OTEC where evaporated warm seawater is condensed, whereas in the case of a closed cycle OTEC plant, low-temperature distilla-

tion/multistage flash techniques can be used. Aquaculture and potable water production fit in well with the large quantities of both warm and cold water moved by an OTEC plant and can be incorporated separately or together. In the limiting case, all of the deliverable power can be used to drive pumps and other equipment for the desalination and aquaculture activities. The cold deep water for the OTEC plant is free of pathogens and rich in nutrients, and phytoplankton can convert the nutrient organic materials on which shellfish can then feed and grow. In addition, other fish and seaweed will grow at greatly enhanced rates. The cold water is also a source of pharmaceutical activities and products.

A further DOWA is use of the cold water, still cool after passing through the condenser, for agricultural purposes by passing the cool water through pipes buried in the soil, thereby enabling products of temperate climates, such as vegetables and tomatoes, to be grown in tropical zones—an application already occurring.

Other DOWAs for OTEC make use of the electrical power *in situ* to achieve, alternatively or additionally, air conditioning or energy-intensive products. For example, hydrogen may be produced by electrolyzing water, and that version of OTEC power can then be stored and transported to market—usually to industrialized nations outside the OTEC zones—as liquid hydrogen. Alternatively again, hydrogen may be produced as an intermediate product, being used in turn to produce ammonia. Currently, the use of ammonia fertilizers is determined in part by production capacity from natural gas. The use of such fertilizers in the developing world, much of it in the tropical and subtropical zones where OTEC processes are available, could make a major contribution to world food production.

Finally, mention should be made of the production of some metals from their ores—aluminum from bauxite is a good example—that requires considerable power input. The location of an OTEC plant onshore by the bauxite deposits, or transportation of the deposits to an offshore plant with refined metal being shipped back on the return journey, are other, probably longer term options for application of OTEC power.

Although the economic viability of these other processes varies during the first years of the 21st century, with some being economically attractive and others being less so, this range of possible uses of OTEC power gives flexibility to the ultimate uses of that power and gives rise to a diversified market. It is

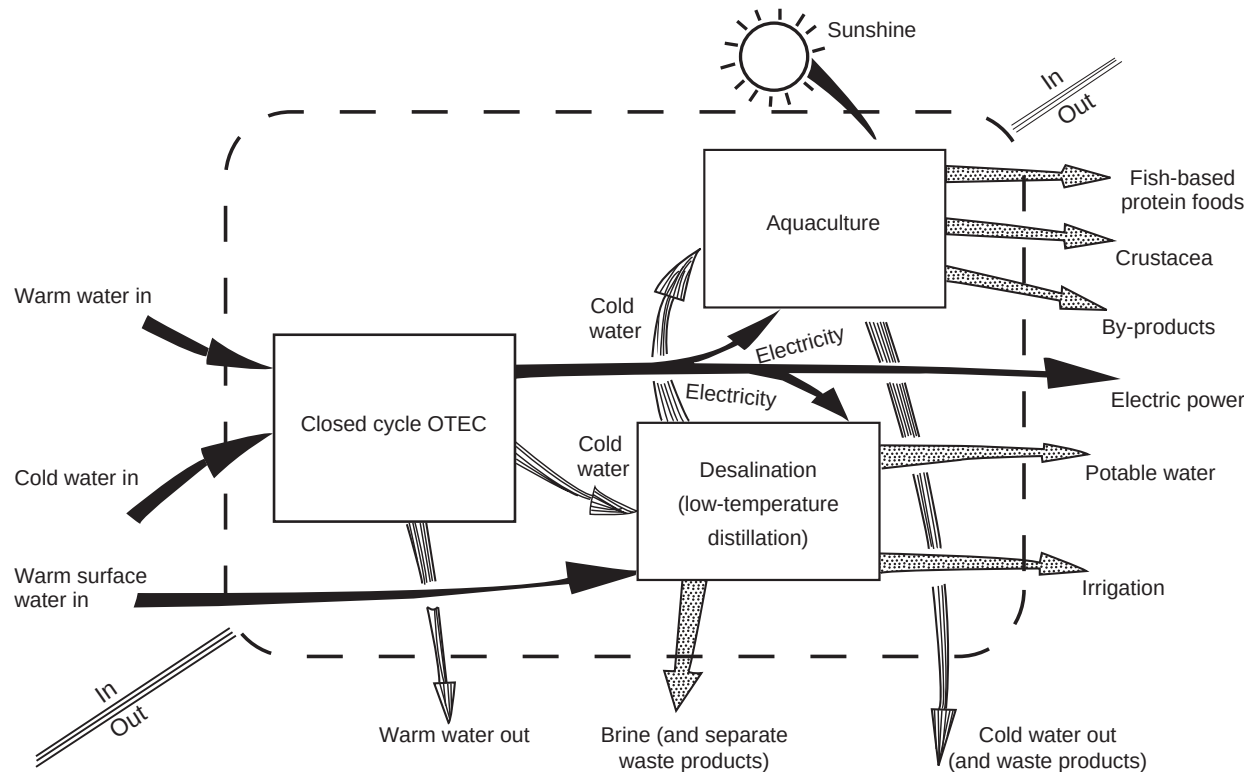


FIGURE 4 OTEC deep ocean water applications (DOWAs): Flow diagram of electricity, aquaculture, and potable water outputs.

this variety of markets, coupled with the technological developments of the offshore oil and gas activities, that encouraged a number of countries to become involved in OTEC developments during the final quarter of the 20th century, including Belgium, China, Finland, France, India, Indonesia, Ivory Coast, Jamaica, The Netherlands, Norway, Sweden, Taiwan, the United Kingdom, the United States, and the Soviet Union. Some of these developments are described later.

4. VARIATIONS ON OTEC PLANTS AND FACTORS AFFECTING THEM

Although based on well-established heat engine cycles, the OTEC energy extraction process has an overall practical efficiency of some 2.5% due to the relatively small operating temperature differential, which has to be compared with a figure of typically 30% or more for a fossil fuel power station. To obtain the 2.5% efficiency, the standard design case temperature difference of some 20°C is required, but clearly the larger the temperature difference, the

greater the efficiency of the plant. An increase of temperature difference of 1°C results in an increase in efficiency of approximately 10%. Because the cold water is at depths of up to 1000 m, and because the flow of water required per megawatt of power output is in the range of 4 to 8 m³/s, the scale of the plant will be apparent. Therefore, practical realization—engineering of the system—is a key issue in considering the future of OTEC and is a major element in its economics.

Although the thermal resource is particularly relevant to developing countries, there are many other factors to be considered before it can be said that a particular country or location is suitable for an OTEC installation. These include distance from shore to the thermal resource, depth of the ocean bed, depth of the resource, size of the thermal resource within the exclusive economic zone (EEZ), replenishment capability for both warm and cold water, currents, waves, hurricanes, sea bed conditions for anchoring, sea bed conditions for power cables of floating plants, present installed power and source, installed power per head, annual consumption, annual consumption per head, present cost per unit (including any subsidy), local oil, gas, and/or

coal production, scope for other renewables, aquaculture potential, potable water potential, and environmental impact.

Given satisfactory figures from these assessments, the forms of OTEC plant to be used, including combinations with other relevant DOWAs, are varied and include the following:

- Closed cycle and open cycle
- Floating but constrained to maintain station
- Land based
- Shelf/Tower based

Within the floating concept, a number of options are available, including the following:

- Ship-shape hull
- Semi-submersible
- Submersible
- Gued tower
- “Other options” being regularly developed in the offshore oil and gas sector

And then, there is the question of whether the OTEC plant is for one or more of the following:

- Generating electricity
- Aquaculture
- Desalination processes
- Other DOWAs

In the longer term, there can be “grazing” OTEC plants—plants that move freely with currents or under their own power—for use as a source of power to the following:

- Recover and process deep ocean minerals
- Produce liquid hydrogen, liquid ammonia, and the like for trans-shipment to other areas to be used in industrial processes and power generation
- Process at sea ores and other raw materials that are energy intensive in the extraction of the refined product

Such plants will be dependent for their operation on a broad agreement to, and application of, the Law of the Sea Convention.

It can be seen that the options for OTEC plants are considerable and that optimization for a particular site involves the review of many variables.

5. TECHNOLOGY OF OTEC PLANTS

Major parts of an OTEC plant involve the use of well-developed materials and equipment, many of these having also been tried, tested, and found to be

satisfactory in the type of marine environment where OTEC plants will operate, with predictable performance and economy. Because many of the components of an OTEC plant, whether land based or floating, are well-established technology, OTEC has an advantage over some of the other renewable energies. But there are exceptions. Four specific items—cold water pipe, heat exchangers, moorings, and power transmission to shore—have been the subjects of extensive research and development (R&D) programs to determine in detail the technical appropriateness and economy of preferred solutions. However, the availability of new materials and powerful analytical techniques are considerably easing remaining problems; and there is substantial scope for technology transfer from other industries, including not only offshore oil and gas but also aerospace, among others.

For the cold water pipe, the quantities of water involved and their momentum are substantial. To keep loads in the seawater circuit at acceptable levels, it is generally accepted that the velocity in the pipe should not exceed approximately 1.5 m/s, although some design studies allow for a figure as high as 4.0 m/s. The pipe must certainly be capable of repair during its working life, and because a number of potential sites for OTEC are in geographical areas subject to hurricanes, it should also have the ability to be structurally disconnected from the main OTEC hull on fairly short notice where a floating design is under consideration. The combination of these factors tends toward a pipe design that is not monolithic but rather has the ability to be sectioned.

Over the pipe length of 1000 m (or more for a land-based version), subsurface currents will have different magnitudes and directions. Therefore, calculations of loading will be dependent on wind, sea state and current conditions, and extensive analysis will be required to determine the critical load conditions. Much work on this has already been done, and representative practical tests have also been carried out in Japan and the United States.

For the heat exchangers, operational efficiency—of both the evaporator and the condenser—is fundamental to the economic success of the OTEC plant. Achievement of the plant’s operating efficiency of approximately 2.5% is highly dependent on the heat transfer rate in both the evaporator and the condenser. Also, at the low operating temperature differences in these items, the maintenance of optimum efficiency is dependent on minimizing biofouling and corrosion products on the surfaces of the heat exchanger materials. Because of the small

temperature difference, the heat exchanger surfaces must be very large to achieve practical power outputs, although for a 10-MW size the 50,000-m² surface area required is no bigger than the largest currently produced, particularly if modular units are used. Full commercial sizes of OTEC power plants, typically 25 to 40 MW for floating plants, will require correspondingly larger heat exchanger surfaces. Altogether, these requirements result in the heat exchangers being very expensive items. Some estimates place approximately 40% of the total cost for early OTEC plants against the heat exchanger system, whereas others estimate less than 25%. With a titanium–aluminum sandwich, estimated costs are lower, but this material combination is not yet available in a fully commercialized form. France, Canada, the United States, and the United Kingdom have undertaken extensive development programs for OTEC heat exchanger applications.

In the case of moorings, the floating variant of an OTEC plant, delivering electrical power to shore, must maintain station to provide a safe and secure path for the electrical power cable. Position fixing can be by mechanical means (e.g., cables, chains), by dynamic positioning (e.g., thrusters), or perhaps (in the future) by use of hydrodynamic forces.

It is again environmental forces from wind, current (which will vary in magnitude and direction with depth), and wave drift (the part of the wave spectra's momentum transferred to the floating plant) that give rise to substantial design problems. Hurricanes are not infrequent at many potential OTEC locations, and environmental loading must include their effect in terms of both wind and wave drift forces. Wave drift forces are often the major loading on conventional large floating structures, and there is a large range in the values of each of the forces when calculated from current theories. It follows that, with variation in calculated values, considerable work must be undertaken if sensible design values are to be obtained. The inefficiency of design that would follow from overestimates of environmental loads would have a very undesirable effect on the economy of an OTEC plant.

Although it is clear that specific values of mooring loads can be calculated only once the mooring system has been designed and vice versa, two related general points about the mooring can be made. First, it should locate the plant within a prespecified watch circle. In turn, the watch circle should be determined, at its maximum, by the strength and limit to freedom of response of the electric power cable and, at its minimum, by the limitations on load that can be

applied by the mooring system to the OTEC plant (e.g., hull, cold water pipe).

Existing "deep water" moorings make use of wire rope or chain, and the historical knowledge of these materials has made them strong contenders for OTEC application, although the weight of both wire rope and chain becomes excessive. The use of synthetic fiber ropes has been greatly extended during recent years for conventional marine applications, and these ropes are now strong contenders.

Finally, for power transmission, it should be noted that although the electrical generating system for an OTEC plant is essentially "state of the art," the particular environmental location and corresponding conditions do introduce constraints. The transmission for floating plants requires developments, to cater particularly to the riser section from sea bed to plant as well as to the sea bed cable running from the riser to shore and the grid, although the growth in offshore wind farms, albeit in shallower water, is providing relevant operating experience.

Apart from the required mechanical strength of cable for this application, its prime function is, of course, to transmit electrical power at a high efficiency from source to grid, and the distance offshore determines whether transmission is by direct or alternating current. In terms of efficiency alone, the alternating current (AC) option is preferred for shorter transmission distances (up to ~30 km), with losses at approximately 0.5% of input power per kilometer. For greater distances, the direct current (DC) is preferred, with losses much less at approximately 0.01%/km but with the additional inversion and conversion losses of approximately 2%. Distances as great as 400 km have been suggested as possible with the use of DC, but the loss of efficiency, as well as the cost of manufacture and installation of the cable, eliminates much if not all of the benefit of an OTEC installation at those distances.

For OTEC applications, with design depths initially specified at 2000 m, cable development itself is not anticipated to present problems, although long-term resistance to pressure at that depth will require validation. Laying of the cable is not expected to present insuperable difficulties either. The problem that will require particular attention is cable junctions at depth, including the junction with the riser cable. In addition, protection and repair procedures for the cable in shallower waters, where it is brought ashore and where anchors, trawl boards, and other hazards (natural as well as those connected with human activity), will present a greater variety of

problems than those in the deeper waters. Again, offshore wind experience is providing guidance.

6. SOME OTEC PROJECTS AND DESIGNS

There have been a number of designs, tests of components, and operations of small demonstration plants over the past 50 years or so, and some of these are briefly described in this section.

During the 1950s, France designed open cycle OTEC plants for two of its then overseas territories: Ivory Coast and Guadeloupe. During the late 1970s and on into the 1980s, designs for a 5-MW plant in Tahiti were completed, tests were undertaken on heat exchangers, and a range of cold water pipe designs were studied.

In the United States, also during the late 1970s, “mini-OTEC” was constructed on a moored barge and operated in Hawaiian waters with a 670-m cold water pipe. This successfully demonstrated the principles of OTEC, generating gross power of a little less than 100 kW. This was followed up with “OTEC-1,” operating on a converted tanker, specifically to check out the effectiveness of heat exchangers but also enabling cold water pipe and mooring details to be developed. Since then, Hawaii has maintained an active interest in OTEC, principally through the Natural Energy Laboratory of Hawaii Authority and the Pacific International Center for High-Technology Research, and during the 1990s it built and operated a land-based,

255-kW gross open cycle OTEC plant that produced net power. It is also in Hawaii that the majority of the practical DOWA activities have been developed, including pharmaceuticals, many on a fully commercial basis.

Japanese interest in OTEC developed through the 1970s, culminating in a land-based, 100-kW gross closed cycle OTEC plant on the island nation of Nauru in the Pacific, located very close to the equator with an excellent thermal resource, and a net output of more than 30 kW was achieved. Among other Japanese programs with industrial involvement, Saga University has played a major and ongoing role. At the turn of the 21st century, it was involved with the Indian proposal for a 1-MW floating OTEC plant that, after testing, is due for installation in an Indian Ocean island group.

European industry, through the international R&D organization Eurocean, completed a design for an OTEC/desalination/aquaculture plant during the 1980s, and at the same time The Netherlands completed a 250-kW design for installation in the Indonesian island of Bali.

Also during the 1980s, a U.K. combined industry–academic team design was proposed for a 10-MW floating closed cycle OTEC plant for island applications, including hydrographical surveys for the preferred location in the Caribbean where cold deep water comes close to the shore (Fig. 5).

During the 1990s, the European Union interest in OTEC and DOWAs took the form of a review of all options, bringing OTEC experts to Brussels, Belgium, from around the world. Although Europe itself has no home market for OTEC given that there is no

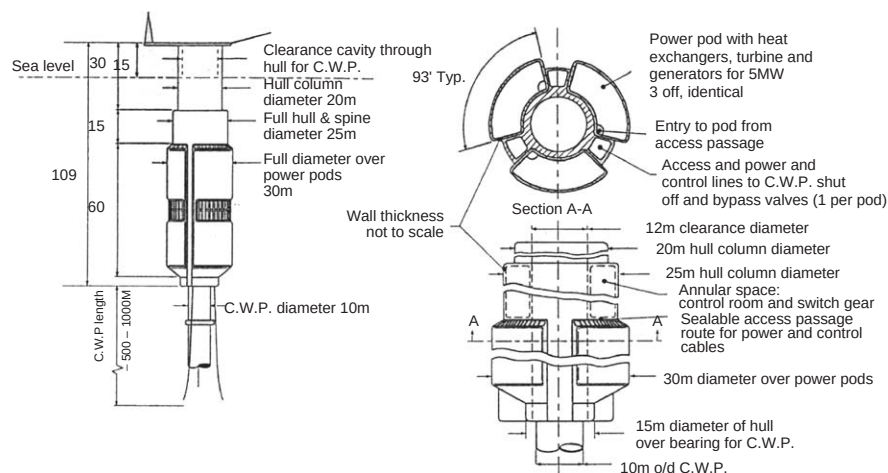


FIGURE 5 General arrangement of 10-MW closed cycle floating OTEC plant. All dimensions are in meters. From Ocean Thermal Energy Conversion Systems Ltd.

adequate thermal resource, there are a number of national dependencies, as well as many (mainly island) countries receiving aid from the European Union through the Lomé and Cotonou agreements, that are located where the OTEC resources are very strong.

Taiwan has a rapidly growing energy demand, increasing up to 10% per year, and for nearly a quarter of a century has developed ideas for OTEC, particularly for application off its east coast where plans for a 5-MW plant exist for a number of sites. During the mid-1990s, a 20-year master OTEC plan for Taiwan was produced, envisaging a 100-MW demonstration plant being developed, with a challenging total OTEC capacity for the country of 3200 MW in 20 years.

Information on all current developments in OTEC and DOWAs is available through the International OTEC/DOWA Association (IOA), which operates out of Taiwan and has a worldwide membership that includes most people involved in OTEC/DOWA activities.

7. COST, ECONOMY, AND COMPETITIVENESS

Because of the diffuse nature of most renewable energies, the size of these renewable energy generation systems is large when compared with fossil fuel generators, and the capital costs of the former are proportionately large as a result. Therefore, on the basis of capital cost alone, renewables (including OTEC) compare poorly with oil-fired power generation. Oil-fired plants would typically cost a few hundred dollars per installed kilowatt, whereas OTEC would cost a few thousand dollars per installed kilowatt. The case for cost "comparability" is crucial to the acceptability of OTEC. Fuel oil has a substantial price, whereas OTEC fuel is free. Maintenance costs for well-developed oil-fired plants are low, whereas maintenance costs for low-efficiency OTEC plants are relatively high. In addition, the high capital cost of OTEC means high total interest charges to be serviced in comparison with the interest charges for oil-fired plants.

All of these points must be incorporated into the financial calculation if an accurate and realistic cost comparison is to be made. Also, and almost uniquely, OTEC is a base load system. Output varies little whether the sun shines or it is nighttime because the thermal resource of the oceans is so massive and the

change between summer and winter is typically no more than 10%. Therefore, OTEC has advantages over other renewables that require buffer storage to achieve base load, with commensurate added costs.

A review of past national and international programs of work shows that a variety of assumptions were made, resulting in overlap between, for example, costs for floating OTEC plants and those for land-based ones, plants with closed working cycles and those with open ones, and simple designs using established technology and technologically advanced designs. Beyond confirming a reduction in costs with size, there is little in the way of conclusions to draw from these comparisons.

The following figures are based on the design of a 10-MW demonstration plant, shown as a general arrangement in Fig. 5. As a demonstrator, it has three 5-MW power pods, with the third pod for development or use if either of the two main power production pods has to be shut down at any time. The floating plant consists of a single cylindrical hull in concrete, heat exchangers in plate form constructed from a titanium-aluminium sandwich, a cold water pipe of 1000 m length in fiber-reinforced plastic, moorings in wire and chain, and transmission to shore over a distance of 10 km, resulting in a capital cost figure of \$94 million or \$9400/kW. The percentage costs of major components are shown in Table I.

The calculations for this 10-MW floating closed cycle OTEC plant, designed for a Caribbean or South Pacific island where the temperature difference varies between 23 and 21°C between summer and winter, show a generating cost of \$0.18/kWh using the 21°C temperature difference. If potable water is a by-product (highly desirable for both of these locations), the generating cost falls to \$0.11/kWh with the potable water costed at \$0.80/m³. Island costs for

TABLE I
Component Costs for 10-MW Floating OTEC Plant (Percentages)

Site-specific data	2
Heat exchangers	23
Cold water pipe	6
Moorings	5
Electrical transmission (sea bed and riser)	8
Pumps, turbines, generators, and control	13
Hull (including warm water circuit)	20
Installation and maintenance	5
Start-up and test	8
Contingency	10

potable water are in the range of \$0.40 to \$1.60/m³, so the figure used here is seen to be at the conservative end of the spectrum. Although there are islands where land fuel costs are three to four times those in developed countries, for the two islands considered here, the uplift is only 75%. So, instead of \$20/barrel, oil would be \$35/barrel. At that price, electrical generation from an oil-fueled plant would be costed at \$0.09/kWh. Therefore, the OTEC demonstrator plant is seen to be approaching cost competition with such a plant on these islands. General engineering experience suggests that by the 8th to 10th production plant, capital costs will be reduced and generating costs may be expected to fall by some 35%. If these reductions are achieved, OTEC will become competitive with oil-fired plants for many island locations. In these examples, no benefit has been given to OTEC for the “environmental plus” that it has in common with many other renewables. One recommendation of the 1992 Earth Summit in Rio de Janeiro, Brazil, was the introduction of a carbon tax for fossil fuels, but to date this has not been applied. If such a tax is brought into use, as may well be the case during the first decade of the new century, all renewables, including OTEC, will benefit further in terms of competitiveness with hydrocarbons.

In addition to these simple costing figures, it is essential that an operator or utility also sees this new technology as attractive. For the example given here, a notional return of 20.4%, corresponding to a real return of 14.7%, is calculated, and these are reasonably attractive returns. So, for both the consumer and the plant operator, OTEC is beginning to look attractive. But history shows that new technology has considerable difficulty in attracting financing for the first examples, and OTEC is unlikely to be an exception to this. Therefore, it seems essential that the first two or three plants of realistic size—approximately 10 MW if of the floating variety—will need to be funded by governments or international funding agencies such as the World Bank, the Asian Development Bank, or the European Bank for Reconstruction and Development. After that, and if the figures cited here are achieved, the financial business sector may be expected to step in and support further installations.

8. MARKETS AND THE WAY AHEAD

All surveys indicate that the world demand for energy will continue to increase considerably, and the

greater part of that increase is expected in developing countries. At the same time, it is anticipated that the percentage of “new” energies will also grow, from a near zero figure at the end of the 20th century to 6% by 2020. This translates into new energies of some 12,000 MW per year averaged over the period from 2000 to 2020. Capital costs for natural temperature difference equipment, due to the low efficiency, is on the order of \$5,000 to \$10,000/kW, some 10 times the capital cost for conventional power systems. Therefore, the funding of new energies equates to a total sum of \$60 billion to \$120 billion each year—very substantial business by any standards. Therefore, an important point to note is that this is, for the construction, operational, and financing sectors, an activity of very considerable interest. But this business will develop only if it is economically attractive to the utilities that will invest in and operate the OTEC plants. As indicated earlier, that situation is rapidly approaching.

Of course, regardless of the feasibility of OTEC-generated electricity, there must be a demand for it at a price it can meet or better. Superposition of a standard atlas onto the contours of Fig. 1 will illustrate which countries can benefit most from the OTEC resource, and it is apparent that, with significant exceptions, the great majority are developing nations rather than developed ones and that many of these are rather small islands (Table II). For this reason, many early OTEC plants will be small—typically less than 10 MW rather than on the order of 50 MW. The majority of island countries that do have adequate demand currently rely on imported oil for generating power and are paying high prices for this oil due to delivery charges. Therefore, in the shorter term, it is islands that provide the main market for OTEC plants, particularly those islands

TABLE II
Island Nations with High OTEC Resources

Bahamas
Cayman Islands
Fiji
Guam
Jamaica
Pacific Islands (Trust Territories)
Palau
Papua New Guinea
Seychelles
St. Lucia
Trinidad and Tobago

that have a growing power demand. Although the rate of growth of power demand in many independent developing countries should be dependent on the price of the power generated, surveys indicate that the greatest increase in power demand is in those developing nations, where incidentally potable water is also a high priority—an ideal combination for OTEC. Table II lists island nations where OTEC has particular application, and one of these intends to meet most of its power requirements—30 MW—from OTEC by 2015.

The growing acceptance of the 200-mile EEZ is also particularly relevant to oceanic island nations, although the benefit of that is likely to relate more to grazing OTEC plants that seem unlikely to materialize during the first quarter of the new century. Nevertheless, this potential application of OTEC is of considerable importance to island nations because it could be a substantial tool in enabling them to benefit from the resources within the EEZ.

The key to further progress, now that so much relevant R&D has been completed, is the construction of representative scale demonstration plants. An examination of island sites has indicated that floating OTEC plants of 10 MW (and smaller land-based plants) can produce electricity at prices that will be competitive with those of oil-fired generating systems for the best situations. However, the fully economic sizes for general application in the tropics and subtropics will be 25 to 40 MW for floating plants and 8 to 10 MW for land-based plants.

The way ahead is for the construction of demonstration-sized plants in one or more island

nations, financed through international funding agencies, to achieve cost-competitive, environmentally friendly base load power—and, probably, potable water as well.

SEE ALSO THE FOLLOWING ARTICLES

Desalination and Energy Use • Geothermal Direct Use • Geothermal Power Generation • Ground-Source Heat Pumps • Lithosphere, Energy Flows in • Ocean, Energy Flows in • Thermal Energy Storage

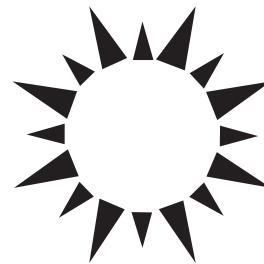
Further Reading

- Avery, W., and Wu, C. (1994). "Renewable Energy from the Ocean: A Guide to OTEC." Oxford University Press, Oxford, UK.
- Energy Commission. (1994). "Master OTEC Plan for the Republic of China." Ministry of Economic Affairs, Republic of China.
- Gauthier, M., Golmen, L., and Lennard, D. (2000). Ocean thermal energy conversion (OTEC) and deep ocean water applications (DOWA). In "New and Renewable Technologies for Sustainable Development" Conference Proceedings. European Community, Madeira Island, Portugal.
- Lennard, D. (2001). Ocean thermal energy conversion. In "Survey of Energy Resources." World Energy Council, London.
- Vadus, J. (1997). A strategy for OTEC commercialization. In "Proceedings of the 1997 International OTEC/DOWA Association," pp. 235–247. IOA, Singapore.
- Vega, L. (1994). Economics of ocean thermal energy conversion. In "Proceedings of the 1994 International OTEC/DOWA Association." IOA, Brighton, UK.

Oil and Natural Gas Drilling

J. J. AZAR

University of Tulsa
Tulsa, Oklahoma, United States



-
1. The Rotary Drilling Process
 2. Purpose of Drilling
 3. Rotary Drilling Requirements
 4. Rig Systems
 5. Drilling Fluids and Hydraulics
 6. Drill Bits
 7. Drill String
 8. Horizontal and Multilateral Well Drilling
 9. High-Angle/Extended Reach Drilling

Glossary

annulus The space between the outer wall of the drill string and the wall of the well bore; it provides a passageway for fluids and drilled rock cuttings to return to the surface.

blowout When a sudden intrusion of formation fluids into the well bore occurs and the preparation to prevent or control the kick fluid has failed.

development well drilling Drilling in an area where wells have been drilled before and geological information has been established.

draw-work Part of the rotary rig hoisting system that provides the power to move loads in and out of the wellbore.

drill bit A tool specially designed to efficiently make a hole in a rock formation by shearing, crushing, or shoveling.

drill string Consists of all the tubular connected by tool joints.

drilling fluid A liquid, gas, or gasified liquid that is used in drilling to perform several functions such as cuttings removal, cooling, and stabilization of well bore.

drilling hydraulics A branch of engineering that describes drilling fluids behavior while in motion.

exploration drilling Drilling in an area where the geological conditions are not known.

high-angle/extended reach drilling (HA/ERD) When the horizontal departure over the true vertical depth of the well is greater than 2.0 and the angle of inclination is more than 63° from vertical.

horizontal drilling Consists of drilling horizontally (approximately 90° from vertical) into the reservoir rock.

horsepower The rate of doing work or transferring energy that is equivalent to moving 33,000 pounds 1 foot per minute.

kick When formation fluid enters into a hole unintentionally; the fluid is what is called kick fluid.

marine riser Large diameter steel casing with buoy sections that serves as an extension of a well bore from sea bottom to rig surface area when drilling offshore.

multilateral drilling Consists of drilling several laterals from a main trunk drilled into a reservoir; the trunk may be vertical, horizontal, or directional.

rotary drilling A process of boring a hole into the ground by applying a force and right-hand rotation on a drill bit and circulating a drilling fluid from surface through a drill string and back to surface through the annulus.

Digging or boring holes into the earth subsurface in search of water and food dates back to prehistoric times. The tools used were very primitive in nature and included stones, sticks, and bones. Published records on hole digging in search of brine and possibly oil have been reported as early as 1500 BC. The term “drilling,” which during the early days simply meant boring a hole into the earth subsurface, today is defined as an elaborate “well construction” (Fig. 1). Around 1800, the art of drilling and its tools were developed to some extent, and it is known that by this time drilling for oil had been conducted in Japan, France, and Burma as well as other places around the world. Although oil was first discovered in America in 1814, the first well purposely drilled for commercial oil was the Drake well of 1859 (depth of 59 feet and estimated production of 20 to 30 standard barrels/day). While rotary drilling was being developed in other countries (mainly France), cable tool drilling (percussion-type tools) was being developed in the United States prior to the 1900s. Cable tool rigs were used to drill the most oil wells in the world by far until the 1920s. In the Corsicana oilfield in Texas, discovered during the 1890s, cable tool rigs were not successful in drilling the unconsolidated

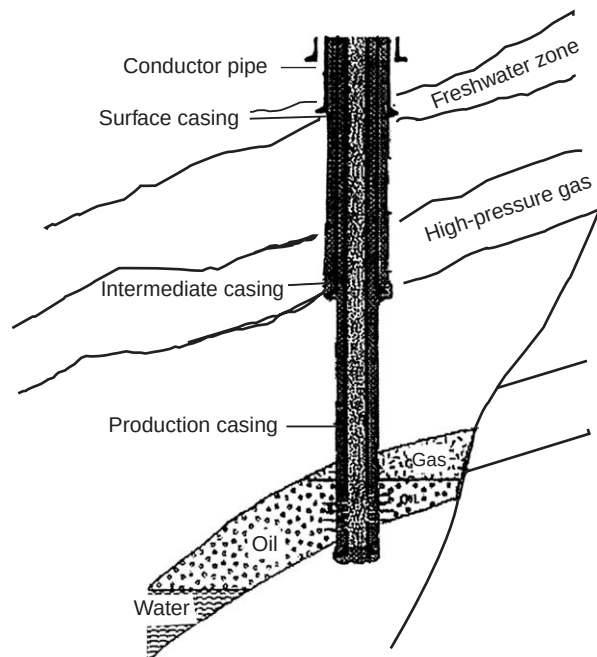


FIGURE 1 Well construction.

formations, thereby bringing wet rotary drilling rigs into the region. In 1901, the famous Lucas Spindletop well was drilled using a rotary rig system and was completed at an estimated oil production of 75,000 to 80,000 standard barrels/day at a total depth of approximately 1100 feet. This is the famous well that gave birth to the oil and gas industry, as we know it today, and has had the greatest impact on the economy and culture of the entire world since then. There have been many technological advances in drilling over the years. The most striking advances, which have occurred during the past decade or so, allowed the oil and gas industry to reach targets below and away from a surface location in excess of 38,000 feet (12 km) and moving farther

1. THE ROTARY DRILLING PROCESS

Whether drilling vertical, directional, horizontal, or multilateral wells (Fig. 2) for the exploitation of hydrocarbons, the elements needed to drill wells successfully and economically are the same but differ in requirements to achieve them. These elements, as shown in Fig. 3, include the following:

- Force acting downward on a rock-cutting tool called a drill bit

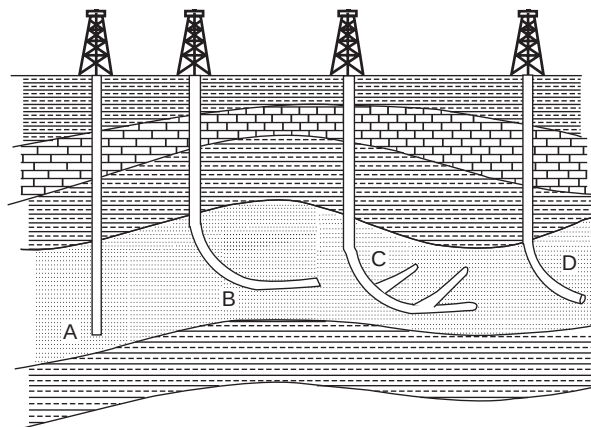


FIGURE 2 Well types: (A) vertical, (B) horizontal, (C) multi-lateral, and (D) directional.

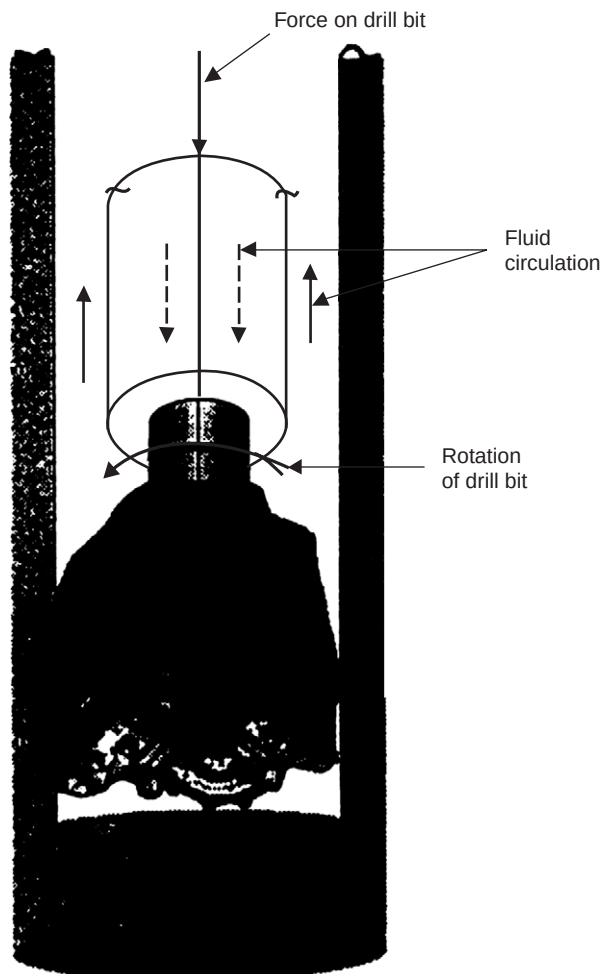


FIGURE 3 Elements required for drilling a hole.

- Right-hand rotation of the drill bit
- Circulating fluid (gaseous, liquid, or gasified liquid) from the surface through a tubular (drill

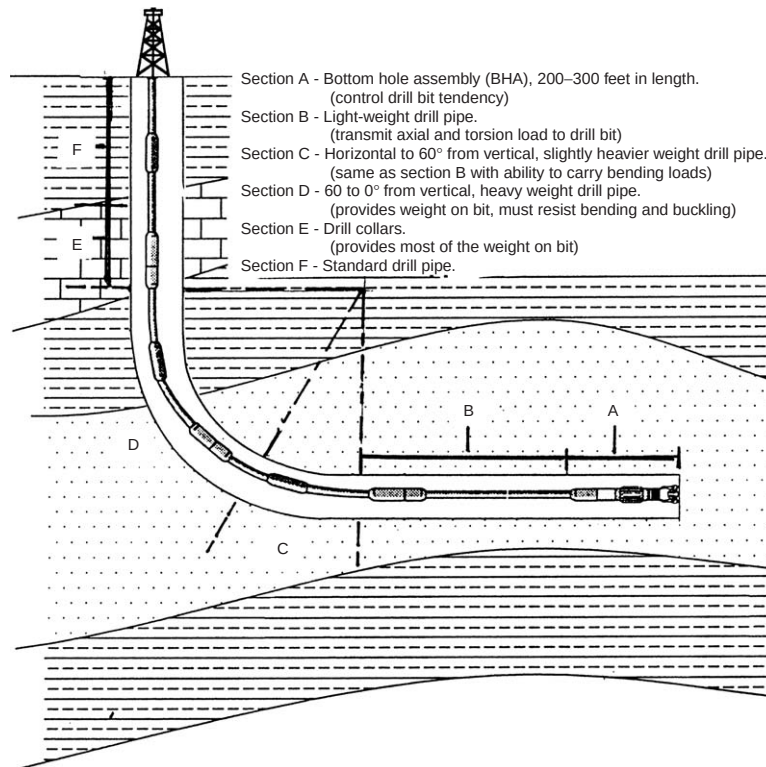


FIGURE 4 Drill string configuration in horizontal well drilling.

string) to hole bottom and back to the surface through an annular space

1.1 Force

The force on the drill bit is provided by slacking off tension on the drilling line through the draw-work part of the rig hoisting system. In vertical well drilling, this force is provided by the amount of certain length of heavy-weight pipes (drill collars) located directly above the bit that is allowed to go in compression as a result of slacking-off surface tension. When a well is being drilled at some inclined angle from vertical, heavy-weight pipes (e.g., drill collars, heavy-weight drill pipes) are no longer placed in the highly inclined sections; rather, they are placed in sections of lesser inclination or near vertical. The reason is that the heavier the tubulars in the highly inclined section of the hole above the bit, the greater the frictional forces on the drill string during sliding and/or rotation, possibly leading to the inability of advancing the drill bit to reach the intended target zone (Fig. 4).

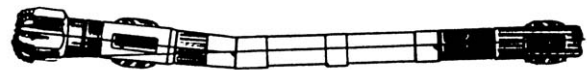


FIGURE 5 Positive displacement mud motors.

1.2 Rotation

Drill bit rotation may be induced from either a surface location or a subsurface location. From the surface, bit rotation can be provided through the use of conventional rotary table in conjunction with a kelly and kelly-drive system or a top drive motor. Subsurface rotation is provided through the use of down-hole mud motors (Fig. 5). For a given applied weight (force) on the drill bit and a desired rotary speed, the required rotary power is equal to the torque, mainly generated by the bit, multiplied by the corresponding rotary speed. However, when the well is inclined and bit rotation is provided

from the surface, the rotary power required is bit torque plus generated torque due to the frictional forces between contact surface of drill string and walls of the wellbore (hole) multiplied by the corresponding rotary speed. Therefore, in directional well drilling, rotary power requirement can be in excess of six times that in vertical well drilling, depending on hole inclination angle. As hole angle increases from vertical (0°) to horizontal (90°), the drag and torque will increase accordingly due to frictional forces. In the high inclined portion, including horizontal of the well, the effective weight of the drill string components will practically all go into friction drag and torque, and that represents undesirable drilling conditions. For excessive drag, advancing the bit may become the limiting factor in reaching the intended target. For excessive torque, the rig available rotary power may become the limiting factor in rotating the bit on bottom.

1.3 Circulation

While drilling, both heat and rock cuttings are being generated continuously. If the process of making a hole is to be conducted successfully, a fluid has to be circulated continuously from the surface to the bottom of the hole and back to the surface again to dissipate heat and remove rock cuttings. The various functions of drilling fluids are discussed later.

2. PURPOSE OF DRILLING

Drilling wells in an oil patch may be for the purpose of exploration, development, infill, injection, or reentry into existing wells. Exploration drilling is conducted to find and evaluate potential reserves in a reservoir rock. Development drilling is conducted to develop a field with proven reserves to its fullest extent of total production. Injection well drilling is conducted for the purpose of storing hydrocarbons (mainly gas), disposing of unwanted produced waters, or regaining the reservoir pressure by injecting water, steam, or gas into the reservoir. Reentry well drilling is conducted on existing wells for the purpose of deepening to lower pay zones, recompleting, or drilling laterals from the main trunk. Infill drilling is conducted to replace depleted wells or to drill additional wells so as to sustain production economics of the field.

3. ROTARY DRILLING REQUIREMENTS

Drilling a hole requires manpower and hardware systems. The manpower includes a drilling engineering group and a rig operating group. The first group provides engineering support for all drilling programs required to successfully and economically meet the objectives of the well being drilled. The second group is responsible for day-to-day operations of the rig and effective implementation of the drilling programs. It consists of tool pushers, drillers, and various rig crew personnel (e.g., roughnecks, derrickmen, enginemen). The hardware systems that make up a rig, whether for drilling offshore or onshore wells, are generally the same:

- Power generation system
- Hoisting system
- Rotary system
- Fluid circulation system
- Blowout prevention system
- Drilling data acquisition and monitoring system

Drilling rigs are classified as onshore (land) or offshore (marine) rigs. Their main features are mobility, safety, flexibility and well depth capability. The only major differences in offshore rigs are an additional system referred to as marine riser (extension pipe between rig floor and seabed) and in floating drilling a motion compensating system to keep the rig on location. Modern land rigs are built in units and skid-mounted so that they can be easily transported from one drilling site to another. Once on location, the rig components are assembled (rigging up) to drill the well. Offshore rigs are mounted on manmade surfaces that are sea bottom-supported or floating units. These units may be mobile or permanently fixed on location. The mobile units include semi-submersibles, jack-ups, and ships. Exploration drilling is always done using mobile units. Once a field has been economically proven in reserves, a stationary platform is permanently constructed to support further drilling, production operations, and living (Fig. 6).

4. RIG SYSTEMS

4.1 Power Generating System

The power system on rotary rigs is to provide sufficient power to allow all essential drilling operations to be conducted safely and efficiently. Depending on total well depth, rig power requirement may range

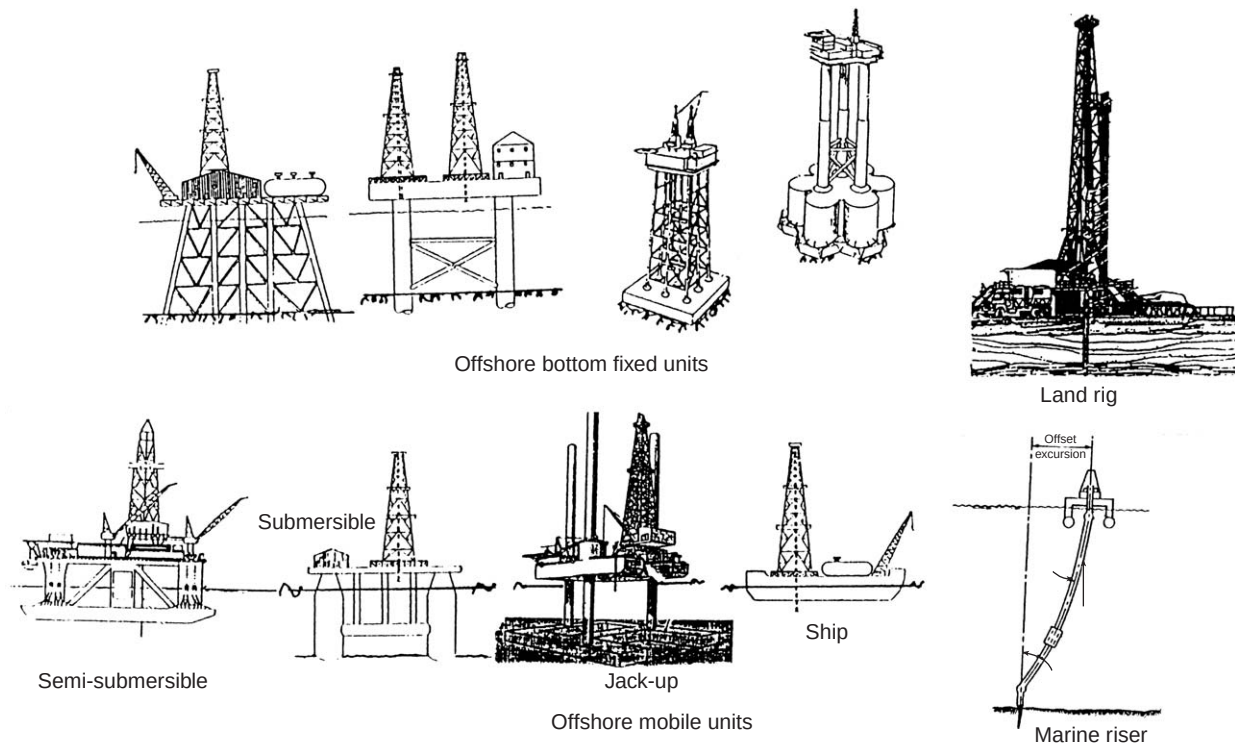


FIGURE 6 Various drilling rigs.

from few hundred horsepower to in excess of 5000 horsepower. Rig power is generated mostly by burning diesel fuel or sometimes on location by natural gas (if available). The power is transmitted to various rig systems (e.g., hoisting, circulation, rotary) by means of mechanical or electric drives. With mechanical drives, the power is taken directly from engines through chains or belts to particular rig systems. With electric drives, the engines drive electric generators to produce electricity, and electric motors (mainly the direct current type) are used to drive various rig systems.

4.2 Hoisting System

The primary function of the hoisting system is to lower and raise tools and equipment (e.g., drill string, casings, measuring tools) into and out of the wellbore during drilling, testing, and casings operations. The major components consist of the draw-work, crown block, traveling block, hook and swivel, drilling line, and elevator (Fig. 7).

The draw-work provides the hoisting and breaking power to enable movement of loads in and out of the hole. Its components include the drum, brake, transmission, and cathead. The blocks, which are pulley-type devices with drilling line strung between

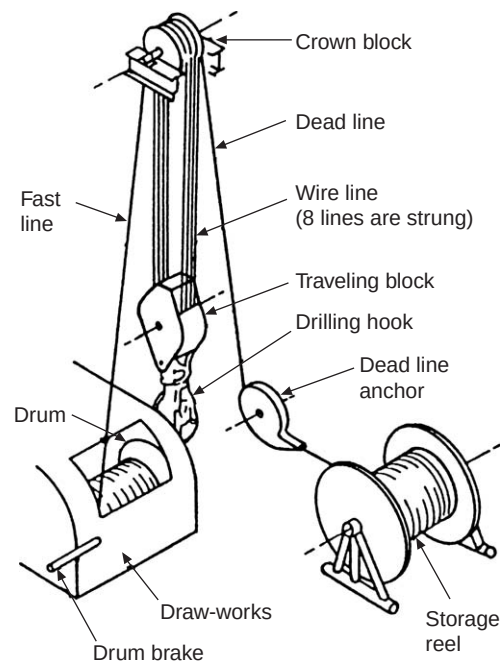


FIGURE 7 Hoisting system components.

them, provide a mechanical advantage that allows raising very heavy loads out of the hole at optimum tripping-out speed.

4.3 Fluid Circulating System

The function of the circulating system is to allow the movement of drilling fluids from the surface, through the drill string, to the bottom of the hole, and back to the surface through the annular space. The main components include mud pumps, gas/air compressors, the drill string, mud tanks and treatment equipment, annular space, and all surface tubular connections (Fig. 8). Pumps used in rotary drilling are of the positive displacement type. They consist of pistons with rods, cylinders (liners), and intake/discharge valves. Duplex pumps have two liners and move mud on the forward and backward strokes. Triplex pumps have three liners and move mud only on the forward stroke. Air/gas compressors are used only when drilling with gaseous or gasified fluids. The drill string consists of tubular sections connected together by threaded joints. It also includes the drill bit. The annulus is the space between the outer wall of the drill string and the wall of the hole and provides the passageway for the drilling fluids and the drilled rock cuttings to return to the surface. In offshore drilling, the marine riser is used to extend the wellbore from the sea bottom to the rig floor.

4.4 Rotary System

The rotary system includes all components that allow drill bit rotation on the bottom. These are the drill

string, swivel, rotary table, kelly, kelly bushing, and drive. Today, top drive motors, especially in offshore drilling, are becoming more popular and provide bit rotation instead of the conventional rotary table. They allow rotation and circulation not only during drilling but also during tripping operations. In addition, they allow full stand-length connections during drilling as compared with one length of drill pipe when using a rotary table. This adds more safety and higher drilling efficiency. In directional well drilling, down hole mud motors are sometimes used for bit rotation (Fig. 9).

4.5 Blowout Prevention System

The function of the blowout prevention system is to control the movement of kick fluids (formation fluids that enter the wellbore) during drilling, tripping, and casing operations. The system must allow (1) shutting the well at the surface, (2) safely removing kick fluids out of the wellbore, (3) replacing original drilling fluid with higher density fluids to prevent any further formation fluid intrusion, and (4) moving pipe in and out of the hole while under pressure (stripping operations). The basic components of the system include the blowout preventer stack or "BOP" (e.g., annular preventer, ram preventers, spools, internal preventers), the casing head, flow and choke lines and fittings, kill lines and connections, separators, and accumulators (Fig. 10).

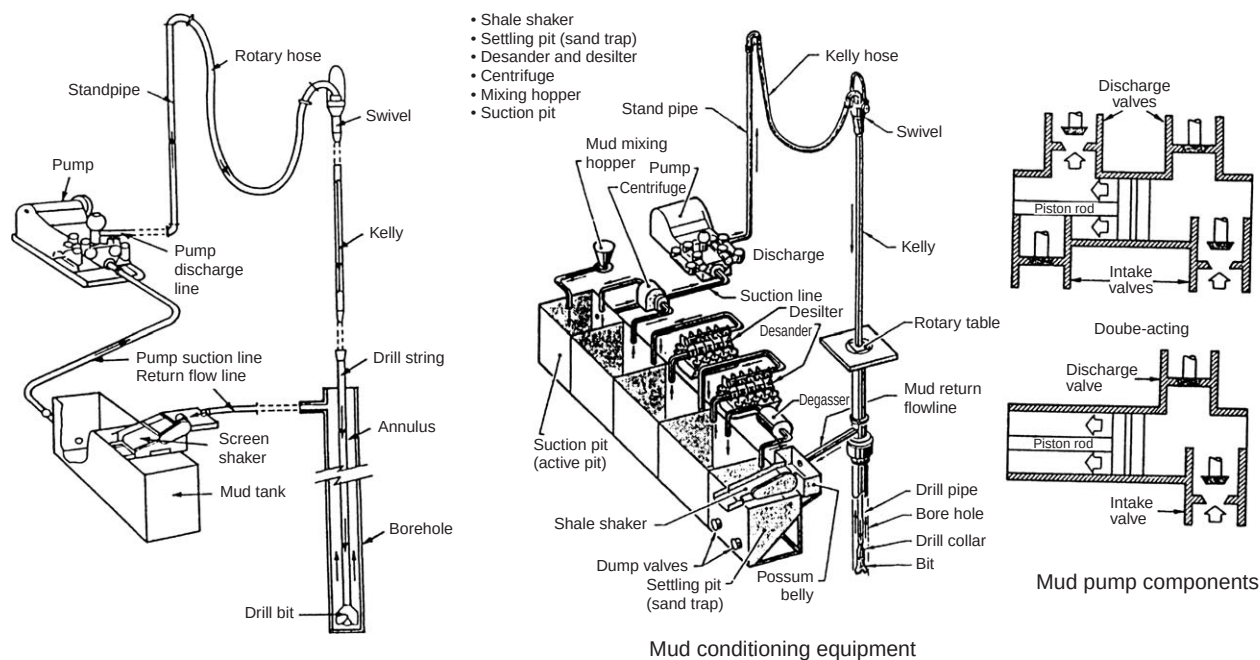


FIGURE 8 Fluid circulating system components.

The annular preventer, with reinforced rubber packing, will shut the annular space around any part of the drilling string in the hole regardless of shape or size. The ram preventers are of three types: pipe, shear, and blind rams. The pipe rams are like annular preventers; however, they have an opening that fits only around the body portion of drill pipe. The shear rams are two solid semicircular steel discs with a knife-like cutting edge that can totally shear (cut) the pipe into two parts and, hence, shut in the well above the lower part of the drill string left in the hole. The blind ram is similar to shear ram but with

no knife-like cutting edges; therefore, it is activated only if there is no drill string inside it, thereby causing the well to be shut in. Internal preventers are like gate valves in that they open if the flow of fluid is downward and shut in if the fluid reverses its flow direction. Drilling spools are casing-type fittings placed in the BOP stack to provide space between two consecutive pipe rams for temporary storage of drill pipe tool joints that do not pass through the pipe ram when it is closed during stripping operations. Spools also provide attachments for choke lines and kill lines.

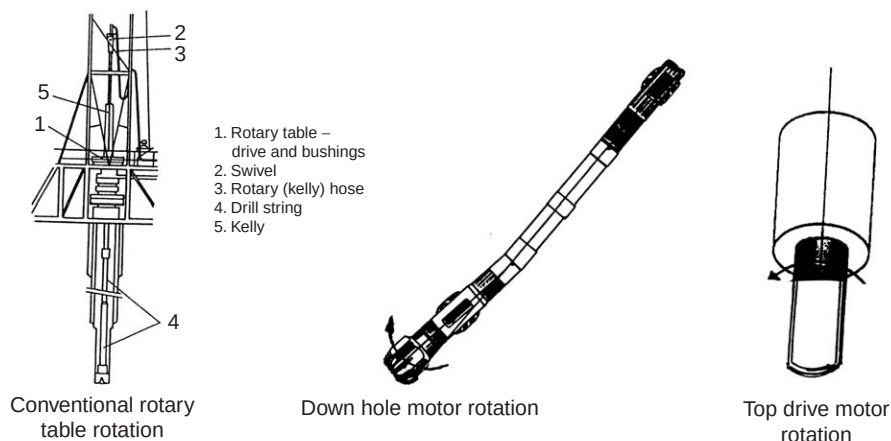


FIGURE 9 Rotary system.

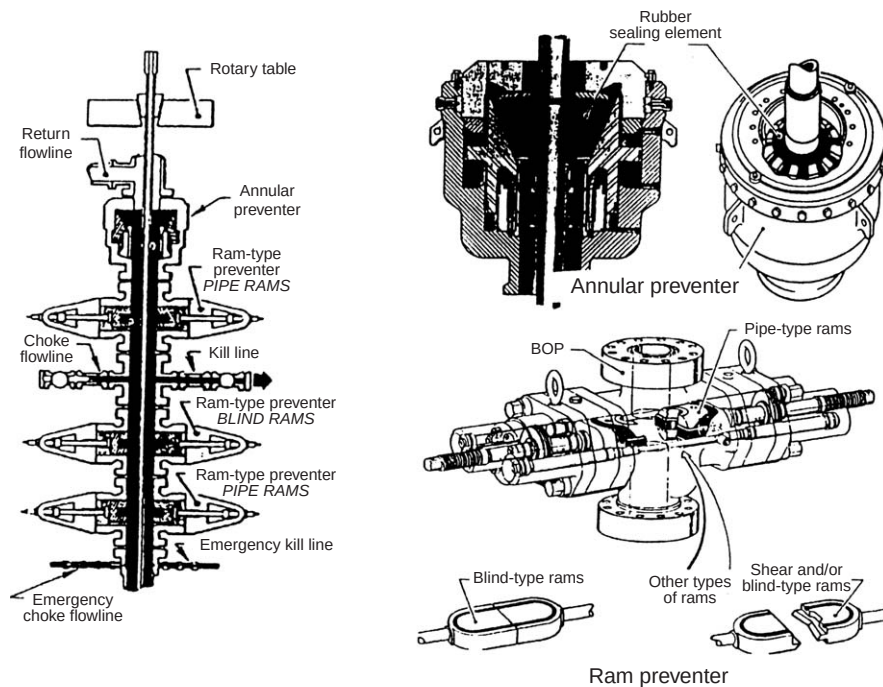


FIGURE 10 Blowout prevention system components.

4.6 Drilling Data Acquisition and Monitoring

The data acquisition and monitoring system consists of the devices (both surface and subsurface) used to monitor, analyze, display, record, and retrieve information regarding all aspects of the drilling operations. Parameters of interest include the rate of penetration (ROP) of drill bit into the rock; hook load of suspended equipment; hole depth; mud pump delivery pressure, flow rate, and speed (strokes per minute); torque; weight on bit (WOB); rotary speed and hoisting speed; mud properties; formation evaluation; and *in situ* bottom hole information (e.g., stresses, temperature, pressure). In the oil and gas industry, this system is regarded as the functional pulse of the drilling process. Costly drilling problems, such as well kicks, potential pipe sticking, lost circulation, excessive drag and torque on the drill string, and hole instabilities, can be minimized or prevented by a highly automated data acquisition and monitoring system.

5. DRILLING FLUIDS AND HYDRAULICS

5.1 Drilling Fluids

5.1.1 Functions

A drilling fluid is defined as a gaseous, liquid, or gasified liquid substance used in rotary drilling to perform any or all of the various functions required to successfully drill a usable hole for the production of oil or gas at the lowest overall well cost. These functions are (1) drilled cutting removal and bit cleaning, (2) containment of formation fluids, (3) hole stabilization prior to casing/cementing, (4) cooling and lubrication, (5) suspension of desired added solids, (6) reduction in casing and drill string suspended heavy weight, and (7) allowance in formation evaluation. The failure of a drilling fluid to perform any of its required specified functions can lead to very costly drilling problems and even a potential loss of the well. Therefore, the understanding and applications of each of the drilling functions are of the utmost importance. The following subsections provide a brief description of these functions.

5.1.1.1 Cleaning

Drill bit cleaning. For a drill bit to penetrate the rock efficiently, its cutting surfaces (e.g., teeth, compact diamond cutters, impregnated small dia-

mond matrix) have to be cleaned continuously to prevent clogging and balling. In general, the lubricity and level of hydraulics (amount of fluid flow rate) are responsible for this function.

Bottom hole cleaning. Removal of drilled cuttings from underneath the bit as soon as they are generated is critical for enhancing the ROP and, thus, reducing the cost of drilling. In addition to impeding the drilling rate, if these cuttings are not removed immediately from underneath the bit, they will be regrinded to very small fines, making their removal from the mud both difficult and expensive. Therefore, the drilling fluid must provide hydraulics at the bottom beneath the bit to create high enough cross-flow velocities to sweep generated cuttings efficiently.

Annular hole cleaning. To avoid very costly problems such as mechanical pipe sticking, high torque and drag on the drill string, running casings/cementing difficulties, logging difficulties, and slow drilling, the efficient removal of cuttings to the surface through the annular space is a vital function of the drilling fluid. Some of the factors that are known to affect hole cleaning include fluid annular velocity, mud rheology, hole inclination angle, drill string rotation, annular eccentricity, ROP, and the generated cuttings themselves (e.g., size, shape, density)

5.1.1.2 Containment of Formation Fluids In conventional drilling, the drilling fluid density must be in excess of the equivalent formation fluid pressure density so as to prevent any intrusion of formation fluids into the wellbore (kicks). For example, if it is known that the formation pressure is “P_{ff}” at a given well depth, a drilling fluid must be selected to have a density that will exert a higher pressure or “P_{df}.” This is referred to as overbalanced drilling. The pressure difference between P_{df} and P_{ff} is the overpressure in the wellbore and is referred to as the overbalance pressure.

5.1.1.3 Wellbore Stabilization Before drilling, the rock strength at some depth is in equilibrium with the *in situ* rock stresses and is free of contact with foreign fluids. However, while drilling, the balance between rock strength and the *in situ* stresses is disturbed. In addition, foreign fluids are introduced and an interaction process with the formation begins. This may lead to various types of wellbore instabilities such as hole closure, hole enlargement, hole fracturing, and hole collapse. Hole closure can cause excessive torque/drag, mechanical pipe sticking, and casing landing problems. Hole enlargement can cause hole deviation problems and difficulties in

cementing, logging, and removing annular cuttings. Wellbore fracturing can cause lost circulation, which in turn may lead to kick occurrence. Wellbore collapse can cause mechanical pipe sticking and possible total loss of the well. Therefore, to ensure wellbore stability, the density, chemical composition, and hydraulics of the drilling fluid must be selected and maintained properly during the drilling of a specific hole interval.

5.1.1.4 Cooling/Lubrication Considerable heat is generated from frictional contact of the rotating drill string and drill bit against the wellbore walls. This heat, if not dissipated continuously, can cause premature failure of drill bit and drill string components. Therefore, continuous circulation of a fluid during drilling to dissipate the heat is necessary. Lubrication of a drilling fluid is essential only in directional well drilling, specifically when far-reaching drilling is being conducted.

5.1.1.5 Aid in Formation Evaluation During drilling, it is important to identify the type of formation that is being drilled and to know the fluids it contains. The fluid in use must provide the means of allowing tools, such as logging tools, to obtain this information.

5.1.1.6 Suspension of Desired Solids When the density of a liquid must be increased by adding high-specific gravity solids, such as barite and iron oxide, the drilling fluid must suspend these solids under static conditions (so as not to allow settling) and dynamic conditions (so as not to allow sagging).

5.1.2 Selection

The governing criterion in the design of the drilling fluid program is to achieve the lowest overall cost per foot for the well. There are several considerations that enter into the selection of the drilling fluid to drill a specific hole interval. These include well type (e.g., exploratory well, development well), formation type (e.g., shale, anhydrite, salt, high temperature, lost circulation, producing/nonproducing), casing program, corrosion, rig, environmental impact, and water composition/availability.

5.1.3 Types

Drilling fluids are classified as gaseous, liquid, or gasified liquids. Gaseous drilling fluids include dry air/gas, mist, and foam. Liquid drilling fluids can be clear water (fresh or sea) or mud (water base or oil base). Gasified drilling fluids are mud with air/gas

mixed. Nitrogen is used most commonly, especially in underbalance drilling (drilling while producing).

5.1.4 Properties

The drilling fluid properties that have the most impact on the drilling efficiency are density, viscosity, fluid loss (loss of liquid phase into the formation), chemical composition, and amount of solids. In general, low density, low viscosity, and low solids are most desirable. Chemical composition is very important when drilling certain types of shale and when drilling the reservoir rock.

5.1.5 Additives

To make up a drilling mud or maintain it during drilling, several additives may be required: viscosity building (e.g., bentonite or attapulgite clay, polymers), viscosity reducing (e.g., phosphates, tannates, lignites, lignosulfonates, sodium polyacrylates), weighting (e.g., barite, iron oxide, dissolved salts), fluid loss reducers (e.g., bentonite, starch, polymers), emulsifiers (e.g., oil in water, water in oil), lost circulation materials (e.g., granules, fibrous, flakes), and special additives (e.g., flocculants, corrosion control, defoamers, pH, alkalinity).

5.2 Drilling Fluids Hydraulics

Hydraulics is an area of drilling that concerns the interrelated effects of viscosity, mud weight, and flow rate on the drilling fluid performance in motion of its various functions. Conditions that may be ideal for the performance of one function may be detrimental for another. Thus, the conditions selected must represent a compromise that will lead to the best outcome for a given drilling operation. In general, the hydraulics program design deals with the following field operations: (1) during drilling and kick control and (2) during tripping and casings. During drilling, rig hydraulic horsepower (flow rate times corresponding pressure) must be sufficient to allow fluid circulation at a desired flow rate so as to achieve those fluid functions that will lead to successful drilling. These functions include hole bottom cleaning (drill bit hydraulics) and annular hole cleaning (cuttings transport). Pump hydraulic horsepower that must be delivered is the sum of power related to friction pressure losses of the moving fluid throughout the circulating system and power due to accelerating the fluid across bit nozzles. The first step in hydraulic program design is to determine an optimum flow rate, Q_{opt} , and corresponding optimum nozzle size diameters that will

ensure optimum bottom hole cleaning, thereby leading to a higher rate of rock penetration. However, this optimum flow rate must be equal to or less than some maximum flow rate, Q_{max} , and equal to or greater than some minimum flow rate, Q_{min} . Limiting the maximum flow rate are the maximum available rig hydraulic power, the resulting equivalent circulating density (ECD) of the fluid in motion, and potential hydraulic hole erosion. Unlike static fluid columns where the column pressure is directly related to fluid density, in a moving fluid column, the pressure is related to fluid density and friction pressure losses. During drilling, this will lead to an ECD that may be high enough to exceed the formation fracture gradient and, therefore, cause lost circulation. This same situation can also arise during kick control operations. The annular fluid velocity needed to efficiently transport rock cuttings to the surface governs the minimum flow rate. Friction pressure losses in annular space during tripping and casings operations can cause an increase in mud weight while moving pipe downward into the fluid column (surge pressures) and a decrease while moving pipe out (swab pressures). In the former case, as during drilling, this may cause formation fracturing. In the latter case, the circulating mud pressure may get lower than the formation fluid pressure, and this can lead to kick occurrence.

6. DRILL BITS

The drill bit is a tool whose only function is to break rocks into fragments for the process of making a hole in the earth subsurface. Rocks can vary in strength from very soft to very hard, making it necessary to

have different types of bits that can be selected for a particular type of rock and strength.

6.1 Bit Types

Drill bits are grouped into two categories: roller cone and drag (Fig. 11). Roller cone bits are of two types: milled-tooth and insert. They are designed to drill rocks of various types and strengths. The milled-tooth bits are generally used when tooth wear is not a critical issue, whereas the insert bits with tungsten carbide inserts are used when tooth wear becomes the limiting factor in the life of the bit. Both the milled-tooth and insert bits have bearings wear that limits their operating lives. The rock-cutting mechanisms of roller cone bits use a shoveling (gouging) action in soft formations and a crushing (chiseling) action in hard formations. Drag bits are of two types as well: polycrystalline diamond cutters (PDCs) and diamond matrix body. The major differences from the roller cone bits are the rock-cutting mechanisms and a lack of bearings. The PDC bit cutting mechanism is rock shearing, whereas the diamond matrix is rock grinding. PDC bits are used in all nonabrasive rocks regardless of strength, whereas the diamond matrix bits are generally used in extremely hard abrasive formations. Drill bit selection is based mainly on past bit performance records. Past experience and learning are key factors in selecting the right bit for the right hole interval to be drilled.

6.2 Drill Bit Operating Parameters

The proper selection of bit operating parameters is of utmost importance in minimizing drilling cost.

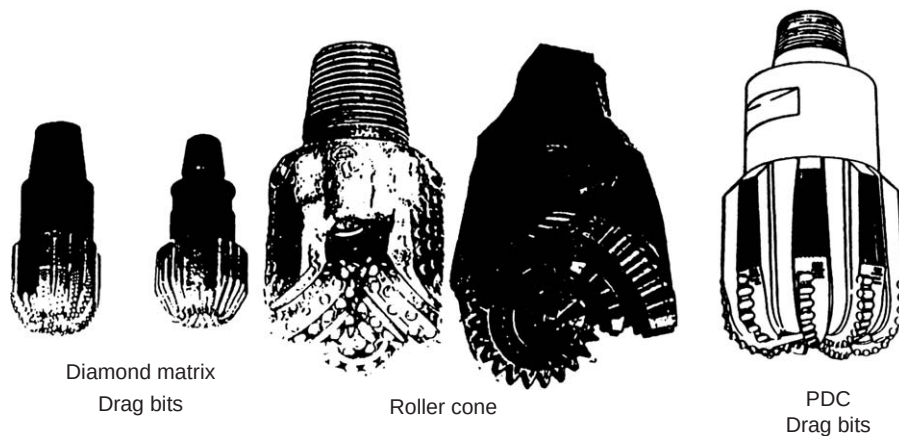


FIGURE 11 Various types of drill bits.

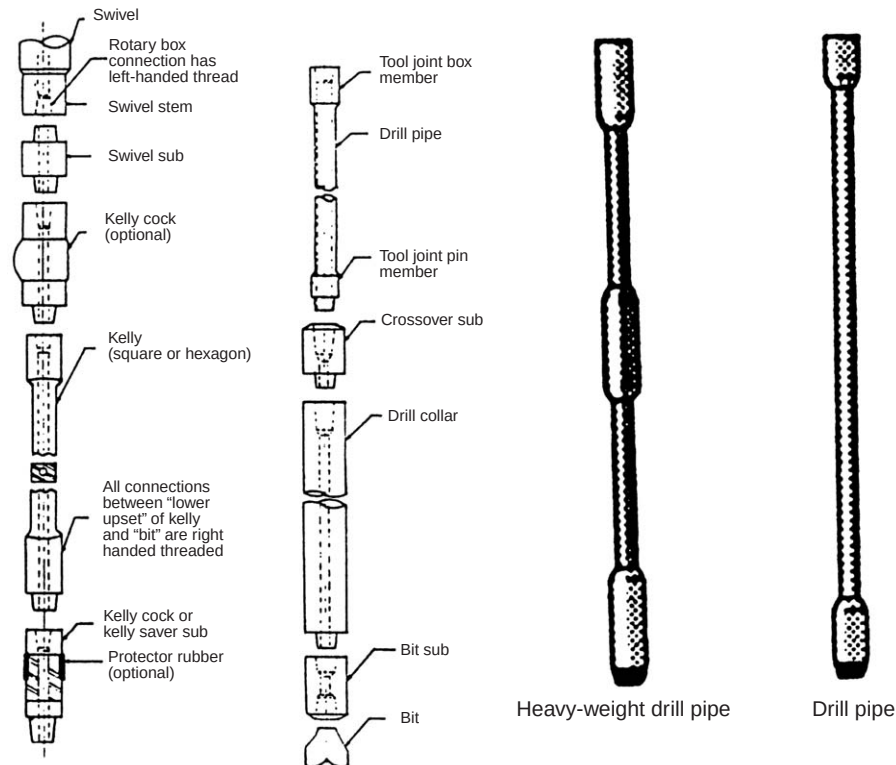


FIGURE 12 Drill string components.

These parameters include WOB, rotary speed in revolutions per minute (rpm), and the hydraulic parameters (flow rate and nozzle size). To achieve the lowest dollar per foot overall drilling cost, an optimum WOB and revolutions per minute with optimum bit hydraulics must be selected. In general, drag bits require lower weight and higher revolutions per minute than do roller cone bits. The hydraulic design is similar to both.

7. DRILL STRING

The drill string is an assemblage of tubular sections (approximately 30 feet in length) connected by threaded joints. It includes light- and heavy-weight drill pipe and very heavy drill collars. Its functions include the following:

- Provide and transfer weight to the drill bit
- Transfer rotation to drill bit if down hole motor rotation is not provided
- Provide a conduit for fluid circulation.

It must safely carry high axial loads, bending loads, torsional loads, and pressure, and it must resist buckling. The bottom hole assembly (BHA)

portion of the drill string is that part behind the drill bit that provides directional control and weight to the bit (Fig. 12).

8. HORIZONTAL AND MULTILATERAL WELL DRILLING

8.1 Definition

Horizontal drilling is a process of directing a bit to drill along a horizontal path oriented approximately 85 to 95° from a vertical direction, whereas multilateral drilling is a process of drilling lateral holes from a main trunk such as horizontal, vertical, or directional (Fig. 13).

8.2 Purpose

The interest in drilling horizontal and multilateral wells is attributed to the following major factors:

- Ability to orient the borehole to accommodate parallel or orthogonal hydraulic fractures
- Ability to improve overall hydrocarbon recovery
- Ability to orient the wellbore in the direction of maximum stability

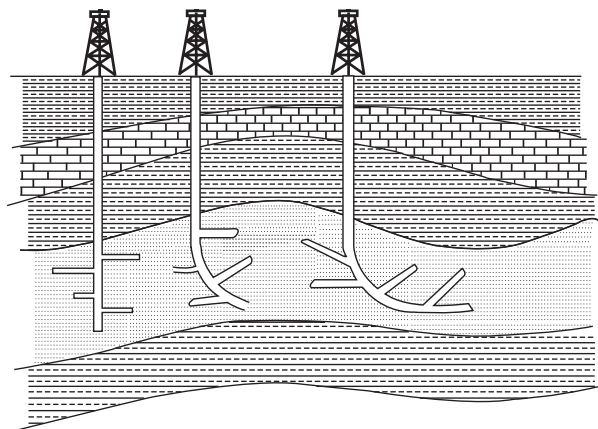


FIGURE 13 Multilateral wells.

- Ability to reduce the number of wells in the development of a whole field
- Ability to enhance primary and secondary production due to lower drawdown pressures.

8.3 Applications

The reservoir rocks that are potential candidates for horizontal and multilateral drilling applications include the following:

- Naturally fractured reservoirs
- Low permeability reservoir (< 1 md)
- Channel sand and reef reservoirs
- Coal bed methane reservoirs
- Reservoirs with potential water and/or gas coning problems
- Thin reservoirs
- Economically inaccessible reservoirs.

8.4 Drilling Methods

Horizontal well drilling consists of three stages: The first stage is a vertical hole section that starts at the surface and stays vertical down to a predetermined depth (kickoff depth). The second stage is a deviated section that begins at the kickoff depth and ends at an entry point into the reservoir at approximately 90° from vertical. The third stage is drilling the horizontal section into the reservoir (Fig. 14). The second stage (deviated section drilling) may be accomplished in one of four ways (Fig. 15):

Simple build curve. Continuously build at a constant build rate to reach the reservoir.

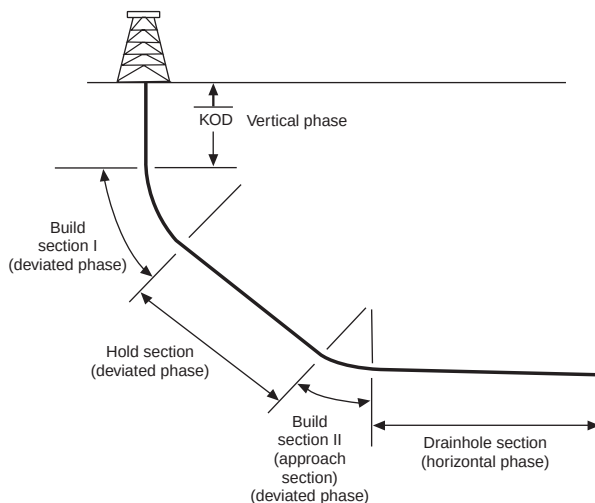


FIGURE 14 Horizontal drilling phases.

Ideal build curve. Continuously build at a constant build rate until a certain predetermined depth, and then change to another build rate to reach the reservoir.

Simple tangent build curve. Build at some preselected build rate until a predetermined hole inclination angle is reached, hold for a certain tangent length, and then build again at the same build rate as the first to reach the reservoir.

Complex tangent build curve. This is the same as the simple tangent except that the second build is at a different build rate than the first.

The drilling methods are identified by the amount of build rate angle used to drill the curved section. These include the following:

- Long radius drilling (build rate angle is $< 6^\circ/100$ feet of drilled section)
- Medium radius drilling (build rate angle is between 7 and $35^\circ/100$ feet of drilled section)
- Short radius drilling (build rate angle is between 1.5 and $3.0^\circ/1$ foot of drilled section)
- Ultra-short build curve (theoretically no build curve section)

The most commonly used methods in the field are the medium and long radius methods. In long radius drilling, the tools can be similar to those used in vertical well drilling. However, for better control of the preselected well path and the ability to use measurement while drilling (MWD) tools and logging while drilling (LWD) tools, down hole mud motors or rotary steering systems are used.

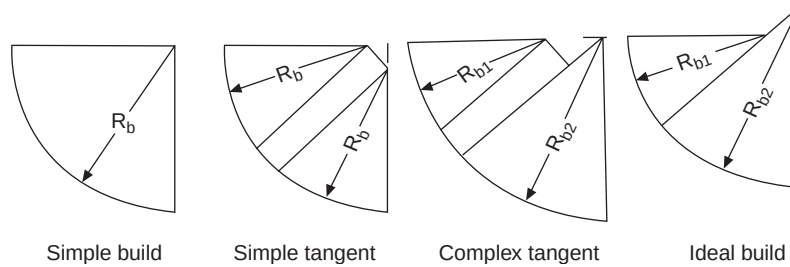


FIGURE 15 Build design curves.

9. HIGH-ANGLE/EXTENDED REACH DRILLING

9.1 Definition

High-angle (HA) wells are those having angles of inclination in excess of approximately 63° from vertical. Extended reach drilling (ERD) is where the ratio of well horizontal departure over its true vertical depth exceeds 2.0. Record horizontal reach with today's drilling technology has been reported in excess of 33,000 feet for a true vertical depth of 5000 feet.

9.2 Purpose

HA/ERD has unveiled worldwide opportunities to economically recover hydrocarbon reserves from single- or dual-field locations that might not have been possible otherwise. As in horizontal well drilling, it reduces the impact on the environment and overall cost significantly due to a reduction of the number of wells needed to develop an entire field or fields. The success of, and rapid increase in, HA/ERD activities is attributed to the tremendous amount of technology breakthroughs (engineering and tools), use of the learning process, effective teamwork, implementation of innovative ideas, preplanning, and real-time field monitoring of data while drilling.

9.3 Critical Issues in HA/ERD

Two of the most critical issues in HA/ERD are the associated inherent increase in torque/drag and the hydraulic requirement for drilled cuttings removal. In addition, there are other constraints that are related to drilling mechanics, field operating practices, logistics, and some formation characteristics. The following are brief descriptions of the concerns regarding these issues.

9.3.1 Torque/Drag

Efforts to reduce torque and drag, and therefore to increase target reach, have been in the proper selection of wellbore trajectory, increases in mud lubricity, use of rotary steerable systems, and development of nonrotating drill pipe protectors. Also of importance are models that enable engineers to predict torque and drag for effective preplanning in HA/ERD activities.

9.3.2 Top Drive Systems

Top drive systems are now available in the capacity range of 60,000 to 70,000 foot pounds and may soon reach higher capacities. The problem is matching the drill string torque strength with that of the top drive. There are some ways in which to increase drill string torque capacity such as reducing drill pipe tension capacity, and this in turn will allow an increase in tool joint strength, use of high-fiction thread compound that will allow higher torque with the same stresses, use of double-shoulder tool joints that will provide higher torsional capacity), and use of high-strength drill pipe (150,000–165,000 psi yield strength) that will provide higher torque capacity.

9.3.3 Wellbore Integrity

Successful HA/ERD is predicated on engineers' ability to properly manage the stability of the wellbore. Major factors that affect wellbore structural integrity are local area geology characteristics, *in situ* rock stresses, shale activity level, well inclination angle, and azimuth orientations. The primary means of managing wellbore stability are by use of proper mud density, use of mud filtrate that is compatible with that of the formation being drilled, proper selection of casing setting depths, advanced wellbore model analysis, real-time monitoring (both MWD and LWD) related to borehole closure, breakouts, enlargements, clay swelling, and pore pressure buildup. Data acquisition and data analysis are key factors in combating wellbore instability problems.

9.3.4 Drilling Fluids for HA/ERD

The key issues that need to be addressed in selection and design of the drilling fluids for HA/ERD are lubricity (high) to reduce frictional forces and proper viscosity (generally low shear rate rheology, high flow rates, drill string rotation) to enhance annular hole cleaning. In addition, the design of drilling fluids must have properties that will be least damaging to the pay zone and will provide the best wellbore stability for long open hole exposure. It must also have the least impact on the environment.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Petroleum • Natural Gas, History of • Natural Gas Transportation and Storage •

Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas: Offshore Operations • Oil Industry, History of • Oil Pipelines • Oil Recovery • Oil Refining and Products • Public Reaction to Offshore Oil

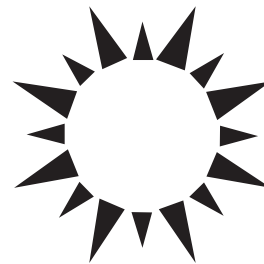
Further Reading

- Aadnoy, B. R. (1997). "Modern Well Design." Gulf Publishing, Houston, TX.
- Bourgoyne, A. T., Millheim, K. K., Chenevert, M. E., and Young, F. S. (1986). "Applied Drilling Engineering." SPE textbook series. Society of Petroleum Engineers, Richardson, TX.
- Lummus, J. L., and Azar, J. J. (1986). "Drilling Fluids Optimization: Practical Field Approach." PennWell Publishing, Tulsa, OK.
- Rabia, H. (1985). "Oilwell Drilling Engineering." Graham & Trotman, London.

Oil and Natural Gas: Economics of Exploration

EMIL ATTANASI and PHILIP FREEMAN

United States Geological Survey
Reston, Virginia, United States



-
1. Introduction
 2. Fundamental Concepts
 3. Discovery Process of Conventional Oil and Gas
 4. Exploration: Individual Firm Decisions
 5. Industry: Exploration and Public Policy
 6. U.S. Exploration: Market Failure and Regulation
 7. International Oil and Gas Exploration Provisions
 8. Trends in Exploration and Discovery
 9. Conclusions

Glossary

barrel of oil equivalent Approximate energy equivalence relations where one barrel of crude oil = 6 thousand cubic feet of gas or 1.5 barrels of natural gas liquids.

basin-centered accumulation A regionally extensive and typically thick zone or unit of hydrocarbon saturated low-permeability rock in the deep, central part of a sedimentary basin.

common property resources Resources not under the control of a single authority and where access is unrestricted or cannot easily be restricted.

continuous-type accumulation A regionally pervasive hydrocarbon accumulation that is not bounded (delineated) by a hydrocarbon-water contact that commonly occurs independent of structural and stratigraphic traps.

externality Situation where the welfare of an individual depends not just on his own actions but directly on actions of another independent economic agent, such as when pollution upstream affects the welfare of downstream water users.

inferred reserves Expected cumulative additions to proved reserves in oil and gas fields discovered as of a certain date, where such additions come by extensions, additions of new pools, or enhancing hydrocarbon flow to the well.

known recovery The sum of a field's past production and its current estimate of proved reserves. Known recovery is an estimate of field size.

market failure Inability of traditional markets to efficiently allocate the use of a resource.

proved reserves Estimated quantities of hydrocarbons that geologic and engineering data demonstrate with reasonable certainty to be recoverable from identified fields under existing economic and operating conditions.

user costs The value of all future sacrifices (or the value of opportunities foregone) associated with production and use of a particular unit of *in situ* resource now rather than in some future period.

Exploration is the search for undiscovered oil and gas resources that have development and production costs no higher than the expected costs associated with adding and producing reserves from known deposits. This article explains from theoretical and empirical perspectives the economic forces that govern oil and gas exploration. It discusses public policy issues associated with oil and gas exploration. The concluding section summarizes statistics relating exploration efforts to discoveries in the United States and areas outside of the North America.

1. INTRODUCTION

In many areas of the world outside of North America and Europe, natural gas is still not an economic commodity because there are no local natural gas markets. On a Btu basis, natural gas is almost four times more costly to transport than oil, so transporting gas to international markets can be prohibitively expensive. Most of the description of the exploration/discovery process applies to both oil and gas. At the heart of the discussion, however, is the assumption that gas is commercially valuable. In areas outside of North America and Europe, it can be expected that reporting of some gas discoveries and

reserves is incomplete because the gas discovered is not commercial.

The exploration decisions of firms in a competitive industry lead to a level of exploration where the marginal finding (discovery) costs are related to "user cost." User cost is defined as the value of all future sacrifices associated with production and use of a particular unit of *in situ* resources now, rather than in some future period. Marginal finding costs and exploration productivity are important leading indicators of the long-term trend in the industry's life cycle and in the future availability of oil and gas supplies. The theme of this article is that information generated by the exploration discovery process for oil and gas provides the basic data for both industry and society to assess future resource availability and for planning and formulating a reasoned energy policy.

2. FUNDAMENTAL CONCEPTS

"Proved reserves" are estimated quantities of hydrocarbons that geologic and engineering data demonstrate with reasonable certainty to be recoverable from identified fields under existing economic and operating conditions. Estimates of proved reserves are important because they are the leading indicators of short-term sustainability of oil and gas production. No more than 10 to 15% of proved reserves (and often much less) can be extracted annually to avoid reservoir damage. So proved reserve levels limit annual production to amounts well short of known recoverable resources.

The objective of oil and gas exploration is to identify resources that can be added to proved reserves. These resources should be of such a quality and quantity that the resource can be commercially developed immediately. Industry exploration consists of a variety of activities that include surveying of surface geology, processing and interpreting newly collected geophysical data, reprocessing and interpreting previously collected data, acquiring mineral rights and access, taking subsurface core samples, and finally drilling exploratory oil and gas wells.

Exploration wells drilled to find reserves are classified by risk level. Categories for exploration wells are extension or outpost wells, new pool tests (shallower or deeper pool), and new field wildcat wells. Even the standard infill development well is not without risk; some development wells are drilled that fail to make contact with the producing reservoir. Risk, or the probability of failure, on average tends to increase from infill development wells to exploratory

wells represented by extensions and outposts, new pool tests, and then to new field wildcat wells. Historically, an exploratory well is classified as a new field wildcat well when it is drilled at least two miles from the nearest productive accumulation. Other exploratory wells that find reserves through new pool tests and extension drilling commonly add reserves to fields already discovered.

Figure 1 is a schematic showing the different types of test wells (exclusive of new field wildcat wells). In the United States in the past two decades, the additions to reserves that are derived strictly from new field wildcat wells represent a relatively small proportion of annual additions to proved reserves. Most of the annual additions to proved reserves are the result of the drilling of well types represented in Fig. 1. The reserves found are credited to oil and gas fields that are already discovered and are recorded and published as extensions and revisions annually by the Energy Information Administration. The application of fluid injection programs and well stimulation also add reserves to discovered fields. These procedures facilitate flow of hydrocarbons fluids through the production well bore.

Estimates of sizes of oil and gas fields are commonly based on known recovery. A field's known recovery is defined as the sum of its past production and most recent estimate of proved reserves. When additions to proved reserves are credited to a field already discovered, its field size based on the estimate of known recovery is said to grow. Resource analysts call hydrocarbon resources that are expected to be added to the proved reserves of discovered fields inferred reserves. The movement of known hydrocarbon

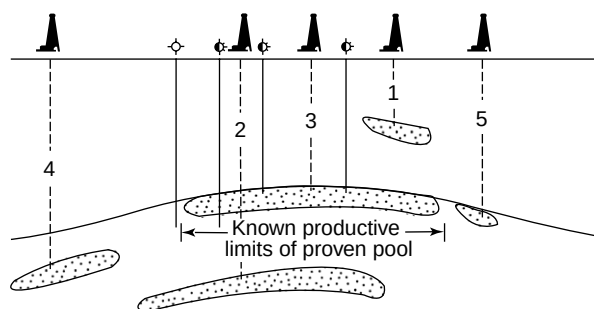


FIGURE 1 Schematic of wells leading to additions to reserves in discovered fields; (1) shallower pool test, (2) deeper pool test, (3) infill well, (4) new pool test, (5) extension or outpost. In practice, the operator or regulatory body may classify the accumulations penetrated by wells 1 through 5 as a single field or as more than one field. Recognition of the relationship among the accumulations could also be further complicated by the order in which the wells were actually drilled. Modified from Drew, 1997.

resources from the inferred to proved reserve category commonly requires exploratory drilling.

There may be some ambiguity whether new reserves are classified as either from new fields or from field growth (see Fig. 1). The U.S. Department of Energy defines a field as an area consisting of a single or multiple reservoirs all related to the same individual geologic structural feature and/or stratigraphic condition. There may be two or more reservoirs in the field that may be separated vertically or laterally by impermeable strata. From a practical standpoint, the definition of an oil or gas field is not exact. The assignment of pools to fields may be for convenience in regulation, or it may be an artifact of the discovery sequence of the pools, rather than for well-defined geologic reasons. Depending on the time sequence of discovery, accumulations shown in Fig. 1 may be assigned to a single field or to more than one field.

Figure 1 shows discrete oil or gas accumulations that are considered to be conventional and producible with conventional methods. Continuous-type oil or gas deposits are identified regional accumulations and are analogous to low-grade mineral deposits. An example is the Wattenberg continuous-type nonassociated gas accumulation that extends over an area of more than 1000 square miles in the Denver basin. The boundary of a discrete accumulation is usually defined by a hydrocarbon/water contact, but there is no such clear-cut boundary for a continuous-type accumulation (see Fig. 2). The geographic locations of continuous-type accumulations are generally known although their extent may be uncertain.

Continuous-type hydrocarbon accumulations commonly have reservoir properties that impede the flow of hydrocarbon fluids. Most of these accumulations require that extra measures be taken

to induce additional flow to the well bore in order to attain commercial production rates. These measures may include drilling wells with horizontal sections or fracturing the formation to create passageways for the resource to migrate to the well bore. Unconventional or continuous-type gas accumulations have received increasing attention by the U.S. domestic petroleum industry during the past decade. Basin-centered gas accumulations (Fig. 2) may be partially developed by the recompletion and stimulation of wells that have depleted shallower conventional pools. If new wells target the continuous-type accumulation, the drilling risk is dominated by failure to stimulate or complete the well so that commercial flow-rates are attained, rather than by missing hydrocarbon-charged sedimentary rocks. Most of the wells drilled in continuous-type accumulations are classified as development wells. In this article, only exploration for conventional discrete accumulations is discussed.

Outside of the United States and Canada, oil and gas exploration is still directed to discrete conventional accumulations. Nonetheless the *in situ* hydrocarbon resources in the continuous-type accumulations are substantial, and the costs incurred in drilling and producing such resources provide a backstop or upper bound to the costs that should be incurred in finding and developing discrete conventional accumulations.

3. DISCOVERY PROCESS OF CONVENTIONAL OIL AND GAS

Physical properties of the occurrence of oil and gas fields tend to cause regularity in the process of oil and gas discovery. The regularity provides the basis of predicting yields of future exploration from estimated undiscovered resources. Oil and gas deposits occur in sedimentary basins. Within a basin, discoveries are typically classified into sets of geologically similar deposits that are called petroleum plays. The petroleum industry uses the concept of the petroleum play, commonly identified in terms of a geologic formation or strata, as a basis to classify targets in a basin geologically.

The observed size-frequency distributions of discoveries in most petroleum plays and provinces are highly skewed. A very small proportion of the accumulations contain most of the discovered resource. Figure 3 shows the frequency-discovery size distribution of oil and gas fields discovered in the Permian basin of the United States through 1996, and Table I shows volumes of hydrocarbons

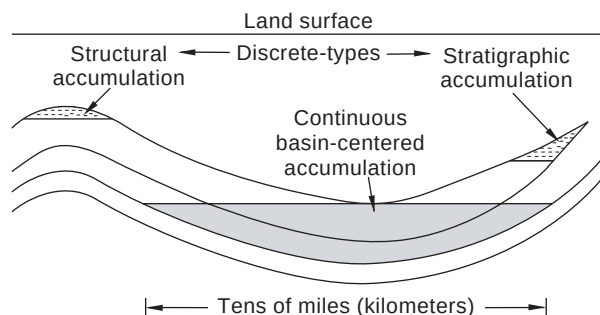


FIGURE 2 A schematic diagram that shows the geologic setting of a continuous-type accumulation in relation to discrete conventional accumulations in a structural trap and in a stratigraphic trap. Data from U. S. Geological Survey National Oil and Gas Resource Assessment Team, 1995.

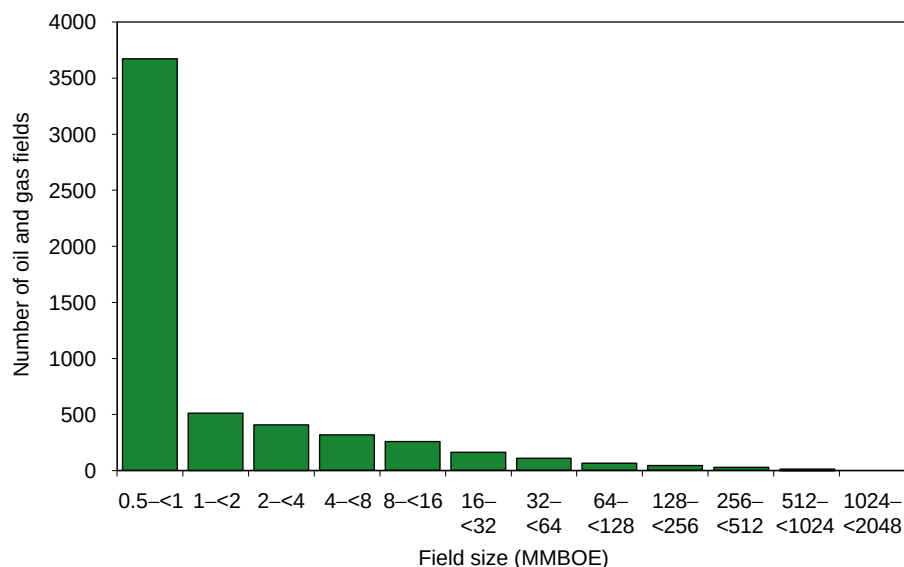


FIGURE 3 Histogram showing the size-frequency distribution of oil and gas fields in the Permian basin discovered through 1996. One barrel of oil equivalent = 6 thousand cubic feet of gas. MMBOE = millions of barrels of oil equivalent. Data are from the Energy Information Administration; field data are from 1998 issue of Oil and Gas Integrated Field File.

TABLE I

Proportion of Total Petroleum in the Permian Basin Contained in Various Size Classes of Fields

Size class ^a (BOE)	Fields		Petroleum	
	Number	Cumulative percentage	Percentage in size class	Cumulative percentage
2048–4096	1	0.02	4.86	4.86
1024–2048	4	0.10	12.14	17.01
512–1024	14	0.38	19.05	36.05
256–512	27	0.93	17.43	53.48
128–256	35	1.64	11.23	64.71
64–128	61	2.87	9.92	74.63
32–64	91	4.71	8.01	82.64
16–32	133	7.40	5.61	88.25
8–16	216	11.76	4.66	92.91
4–8	267	17.15	2.85	95.76
2–4	344	24.11	1.85	97.61
1–2	432	32.83	1.15	98.75
0.5–1	481	42.55	0.64	99.40
0.25–0.5	495	52.56	0.33	99.73
0.125–0.25	483	62.32	0.16	99.89
0.0625–0.125	451	71.43	0.08	99.97
<0.0625	1414	100.00	0.03	100.00

^aIn millions of barrels of oil equivalent (BOE) where 1 barrel of oil = 6 thousand cubic feet.

associated with each discovery size class. The Permian basin, located in Southwest Texas and Eastern New Mexico, is the most prolific oil-producing basin in the conterminous United States.

The largest 27 fields (or 0.55% of the fields) contain more than half of the hydrocarbons discovered to date, whereas the 1414 fields in the smallest size class (less than 60 thousand barrels of oil equivalent,

where one barrel of crude oil equals 6000 cubic feet of gas or 1.5 barrels of natural gas liquids) account of only 0.03% of the hydrocarbons.

The shape of the Permian basin discovery size frequency distribution is typical of most plays and basins; most of the hydrocarbons are found in a few very large accumulations or fields. Exploratory drilling success-ratios (successful wells to wells drilled) provide no information about the sizes of future discoveries. The large number of small accumulations determines success ratio levels. The largest field, containing nearly 2.6 billion barrels of the oil equivalent, is at least six orders of magnitude larger than the fields in the smallest size class. If the x-axis had not been in log base 2, the size classes for the smallest category and the largest category could not have been graphed without a break in the axis.

Assuming the surface expression of a deposit is roughly proportional to its volume, then even with random drilling the average discovery sizes would decline with equal increments of exploratory drilling, as the largest deposits are found early in the discovery process. For example, the surface area of the 6 billion barrel East Texas field covers more than 200 square miles. Any improvement over purely random drilling quickens the early discovery of large fields and accelerates the decline in the discovery rate. Figure 4 shows the progression in the average

discovery sizes (in barrels of oil equivalent) from 1920 to 1996 in 5-year increments for the Permian basin. The regularity of the discovery process allows the expected yields of future exploration to be computed with simple analytical models. The discovery rate (yield per exploratory well) is controlled by the sizes of fields discovered and the order of discovery.

The exploration histories of many of the basins in the United States follow predictable patterns. New-field wildcat wells were drilled in unproven plays at low but irregular rates. After a significant discovery is made, there is typically an influx of new prospectors and much higher levels of drilling; similar to the 19th century gold rushes. Drilling rates eventually weaken as returns deteriorate and exploration in other plays or basins becomes more attractive. As other plays in the basin are tested, then the process repeats itself following another significant discovery in an unproven play. An efficient exploration process, where the larger accumulations are found early, results from physical characteristics of the size distribution of oil and gas deposits in nature even when wells are randomly sited. If the industry is reasonably efficient in finding the largest and lowest cost accumulations early in the discovery process, then the past discovery sequence provides information useful in estimating the magnitude and

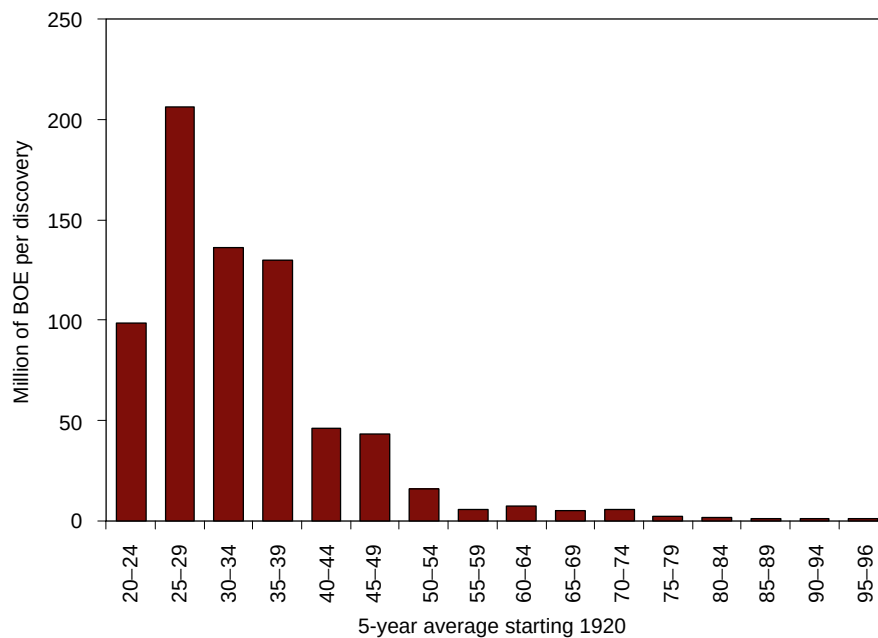


FIGURE 4 Average sizes of fields discovered in 5 year intervals in the Permian basin between 1920 and 1996. Field data are from the Energy Information Administration, 1998 issue of Oil and Gas Integrated Field File. 1 barrel of oil equivalent (BOE) equals 6 thousand cubic feet of gas.

characteristics of the undiscovered resources. For areas where field size is narrowly defined (such as the sum of only proved reserves plus past production), adjustments may be required to the estimated sizes of recent finds to allow for their full delineation so as not to overstate discovery decline in sizes.

In the United States, a relatively small proportion of the known sedimentary basins (or provinces) contain most of the hydrocarbons discovered to date.

Figure 5 shows the distribution of known recovery (past production and proved reserves of oil, gas, natural gas liquids) in fields discovered through 1998, expressed in barrels of oil equivalent for the U.S. onshore basins located in the conterminous United States. The distribution represents 57 onshore lower 48 provinces that were delineated and assessed by the U.S. Geological Survey for the 1995 National Assessment of Oil and Gas Resources. Table II lists

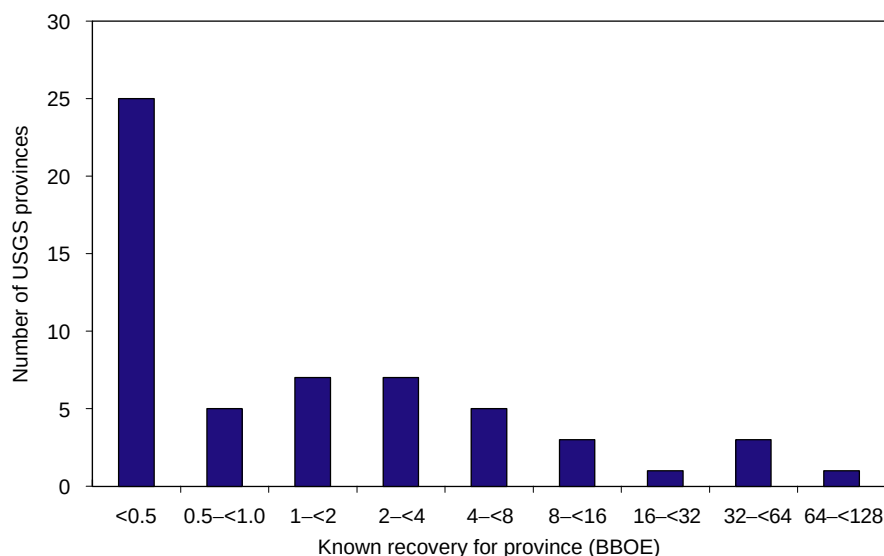


FIGURE 5 Histogram showing the cumulative discovered through 1998 hydrocarbon volumes-frequency distribution for U.S. basins through 1998. BBOE = billions of barrels of oil equivalent. Data from NRG Associates, 2000.

TABLE II

The Ten Leading U.S. Onshore Petroleum Provinces of the Conterminous 48 States

	Units ^a of oil and gas				Discovery date ^b
	(BBO)	(TCF)	(BBL)	(BBOE)	
1. Gulf Coast	25.8	265.8	8.3	75.6	1901
2. Permian Basin	34.4	102.3	6.5	55.7	1920
3. Anadarko Basin	5.0	157.1	5.6	34.8	1916
4. Miss-La. Salt Domes and E. Texas	16.5	84.6	3.5	32.9	1895
5. San Joaquin	15.4	13.2	0.8	18.2	1887
6. Appalachian Basin	3.4	43.0	0.0	11.0	1871
7. Los Angeles	9.0	7.5	0.4	10.5	1880
8. San Juan	0.3	45.5	1.5	8.8	1921
9. Bend-Arch Ft Worth	4.9	13.0	0.9	7.7	1902
10. Cherokee	6.3	2.3	0.0	6.7	1873

^aUnits abbreviations: BBO = billions of barrels of crude oil, TCF = trillions of cubic feet of gas, BBL = billions of barrels of natural gas liquids, BBOE = billions of barrels of oil equivalent. Gas and natural gas liquids conversion factors the following: 1 barrel of oil equivalent = 6 thousand cubic feet gas = 1.5 barrels of natural gas liquids.

^bDiscovery date refers to the discovery date of the first field of at least 10 million barrels of oil equivalent.

From NRG Associates, Significant oil fields in the United States 2000, database.

the known recovery for the ten leading onshore provinces and the discovery date of the first field having at least ten million barrels of oil equivalent. Four of the 57 provinces with the largest known recovery account for 63% of the petroleum discovered to date, and the top nine have more than 80% of the discovered petroleum. On a worldwide basis, relatively few provinces contain most of the hydrocarbons discovered to date.

Commercial petroleum was discovered in the United States in the middle of the 19th century. Just 50 years later, at the end of 1901, petroleum provinces containing more than half the known petroleum recovery (oil and gas) in the onshore lower 48 states had reported at least one discovery of at least 10 million barrels of oil equivalent. By the end of 1920, the basins that contained 94% of oil and gas discovered through 1998 had already been explored sufficiently, so that in each basin at least one field of 10 million barrels of oil equivalent had been discovered. The exploratory drilling through 1920, however, represented only a small percentage, less than 5%, of total exploratory drilling through 1998. Most onshore exploratory drilling since 1920 in the lower 48 states has been follow-up drilling in these and other less prolific basins.

Exploration access, distance from market, and technology are factors determining the order in which basins are explored. In the very early periods of the U.S. petroleum industry, search intensity was influenced by the costs of transporting and marketing petroleum. Some of the remote prolific basins were not explored as early as those closer to market areas.

4. EXPLORATION: INDIVIDUAL FIRM DECISIONS

Exploration expenditures are direct investments made by individual firms for the purpose of locating unidentified, but potentially commercial quantities of oil and gas. For an individual firm, exploration often begins with a hypothesis about the formation of hydrocarbon accumulations in a specific geologic setting. The process typically starts with a literature search to identify potential geographic locations where the hypothesis might be tested. The selection of a target area includes review of geologic maps, reconnaissance-type geochemical data and geophysical data compilation from the open literature and from commercial vendors. Fieldwork is often required to verify interpretations and to collect

additional data to identify specific drilling targets. During drilling, data are continuously collected and interpreted to maximize the chances of locating significant accumulations. With the location of hydrocarbons, additional geologic and economic analysis will determine chances of a commercial discovery.

The economic theory that explains how firms set the optimal level of exploration expenditures is complex and beyond the scope of this article. The discussion that follows is highly simplified, but readers primarily concerned with public policy and trends in exploration and discoveries may wish to skip it.

Economic determinants of exploration investment are typically studied in the context of a representative firm's planning strategy. Suppose we consider a representative firm that searches for and produces oil in several areas over several time periods. The optimal plan tells the firm how much oil to produce each time period and the level of investment expenditures in production facilities and exploration to make (both regionally and temporally) so as to maximize the present value of its net cash-flow stream over the planning period.

In general, the optimal strategy is to increase oil production (extraction per time period) until the marginal cash profit (revenue minus operating costs) is balanced with the present value of the user costs. Simply, user costs represent the opportunity costs associated with producing and using a unit of resource now rather than postponing it. Specifically, it consists of the sum of (1) the future value of a unit of production retained in the ground instead of produced and (2) the cost, in terms useful life of production facilities if the unit of resource is produced at the present time. The value of retaining the unit in the ground is linked to the ease of replacement of that unit. For example, the value of postponing production may be small if discovery costs are low, that is, if replacement is easy, or if future resource prices are expected to decline because a cheaper substitute can fill demand. Marginal discovery or finding costs are sometimes used to approximate user costs. The greater discovery costs to replace that unit, the greater the user costs that pushes the representative firm to postpone production of that unit.

Abstracting from effects of uncertainty for the moment, for each time period exploration expenditures should be set to where the marginal discovery costs of new reserves equals the marginal value of newly discovered reserves. This value, in turn, is related to the user costs associated with the extraction

rule discussed in the previous paragraph. If the firm can put value on additional variables related to, for example, its experience in an area, then it should try to include these benefits in its decision.

Exploration strategy is coordinated with production and investment strategies to ensure that overall costs are minimized and the present value of net income optimized for the entire planning period. For an individual firm operating in multiple areas, if its threshold hurdle rate (required rate of return on capital) cannot be achieved in the region, or if it has a host of other more attractive opportunities elsewhere, then for that area, the firm will withdraw exploration, sell off prospects, and deplete producing properties.

Uncertainty in the returns to exploration effort leads to a departure from the strategy described earlier. The magnitude of this effect depends on a firm's risk tolerances, how uncertainty affects marginal product of exploration, and the perceived risk of ruin. However, viewed at the industry level, there is a distribution of industry hurdle rates of return and risk tolerances. This distribution ensures that information generated by the exploration process—that is, marginal discovery costs—is not an artifact of a single firm's behavior, but is representative of the industry and thus has general value for assessing future industry prospects.

Though many authors have proposed elegant theoretical models of how firms make extraction, investment, and exploration decisions, there is still a surprising lack of empirical research that tests and verifies the hypotheses that have been generated.

5. INDUSTRY: EXPLORATION AND PUBLIC POLICY

At the industry level, production, investment, and exploration patterns are examined for their consistency with the representative firm's optimal strategy and in the context of certain public policy questions. For example, the government might institute incentives designed to encourage domestic oil and gas exploration effort in order to maintain domestic production capacity. In addition to direct tax incentives and subsidies for exploration, the other government policy instruments include access to minerals on public lands, research and development policy, import quota/tariff policy, and direct price controls. All of these instruments have, to some extent, been used during the history of the U.S. oil and gas industry.

The standard formulation of the economic theory of exhaustible resources, as proposed by economist Harold Hotelling, asserts that the net prices of *in situ* finite resources should increase at the rate of at least the current rate of interest. Hotelling's theory rests on the assumption of a fixed stock of reserves that can be produced at will. However, proved reserves of oil and gas can be augmented through exploration or through more intense development. Production rates of oil and gas wells are constrained or limited by reservoir pressures. It is not surprising that many empirical studies have failed to verify the naïve predictions of the Hotelling theory. When exploration is modeled at the industry level, the derived theoretical implications are rarely unambiguous.

A broader question relating to public policy is whether industry exploration costs and actions accurately indicate future prospects of oil and gas. If the finding costs accurately reflect the costs associated with replacement costs of current proved reserves, then trends in such costs could foreshadow future supply problems. Finding costs, however, are not directly observable from industry or government data sources. It may take years for new finds to be sufficiently developed so that their size can be accurately estimated. Though discovery costs are not directly observable, the locations, search intensity, and the apparent risks the industry accepts reflect its assessment of remaining domestic and worldwide oil and gas exploration prospects.

If it is assumed that each firm will equalize the marginal costs of additional reserves either through exploration, intensive development that adds reserves, and purchase, then one can infer from market transactions prices the approximate industry finding costs. Empirical studies of oil and gas property transactions values show a rather consistent pattern. Developed in-ground oil and gas reserve prices hover at about one-third wellhead prices or at half of net prices, which are the wellhead prices minus all operational costs including taxes.

For the individual firm, exploration effort can be expected to increase until marginal finding costs equal user costs. User cost represents the benefits of adding resources to reserves. Empirically, user costs are measured as the difference between the in-ground market value of the developed proved reserves and the development costs. User costs embody the future price expectations of market agents, just as the valuation of properties embodies future reserve price expectations.

For the U.S. petroleum industry, oil price expectations are set in world markets but natural gas price

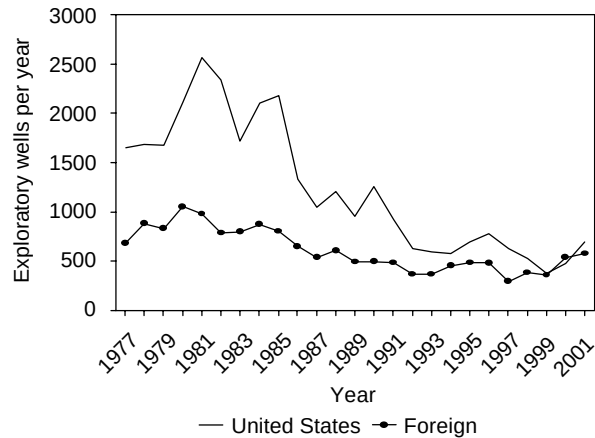


FIGURE 6 Number of total oil and gas exploration wells drilled by the group of Financial Reporting System (FRS) companies in the United States and foreign countries. Data from EIA, Performance profiles of major energy producers 2001, 2003.

expectations are set by the expected demand supply interaction within North America. The unprecedented increase in perceived U.S. user costs during the late 1970s and early 1980s led to record drilling levels and oil finding costs in the lower 48 states but failed to significantly increase proved reserves. User costs as well as exploration activity declined in the United States as world crude oil prices plummeted in 1986 to half the levels that prevailed in 1985.

The U.S. Energy Information Administration compiles and reports the financial accounts, called the Financial Reporting System (FRS), for the group of firms representing the largest domestic crude oil producers. Although there have been changes in the composition of the FRS firms over the years, these firms are the leading U.S. oil and gas producers. The report includes financial expenditures on exploration including numbers of wells drilled, oil and gas production by the country where the producing property is located, and investment expenditures.

Figures 6 and 7 show U.S. and foreign exploratory wells and expenditures reported for the FRS firms. During the early 1980s, both domestic exploratory drilling and exploration expenditures substantially exceeded exploration outside the United States. Since the late 1990s, domestic and foreign exploration drilling and expenditures by these firms have been about equal. The volume of crude oil production from foreign properties for the FRS firms has steadily increased, and in 1997 it exceeded oil production at U.S. properties. Within the United States since 1993, more than half the FRS firms' domestic exploration expenditures have targeted offshore prospects, principally in the deep water of the Gulf of Mexico. In

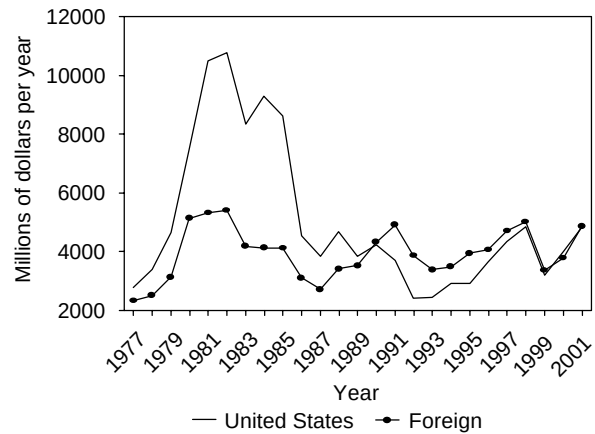


FIGURE 7 Total oil and gas exploration expenditures by the group of Financial Reporting System (FRS) companies in the United States and foreign countries. Data from EIA, Performance profiles of major energy producers 2001, 2003.

summary, the industry exploration trend demonstrated by the FRS group of firms (representing the industry leaders) has been to focus exploration effort in the deepwater offshore areas and countries outside of the United States where the opportunity to discover large accumulations is greater.

6. U.S. EXPLORATION: MARKET FAILURE AND REGULATION

The U.S. petroleum industry is a mature industry. U.S. crude oil production reached its peak in 1970, and gas production reached its peak in 1971. In 2000, the United States produced 5.86 million barrels per day of crude oil, or 8.7% of the world's production. It also produced 20.1 trillion cubic feet of gas, which represents about 23% of the world's gas production. During the history of oil and gas exploration in the United States, the combination of private ownership of mineral rights and the competitive market forces produced excesses in exploration and production that ultimately resulted in government regulation. Actions that were rational from an individual firm's perspective turned out to be very costly from society's standpoint. Although some of these inefficiencies are in the past, they are reviewed here to provide insight into the regulatory environment.

6.1 Externalities and Regulation

In the United States, mineral rights can be owned either privately and by the federal and state governments. Government-owned mineral rights are typically

transferred to the private sector for exploration and production. Several systems were used to transfer mineral rights to the private sector in the past. A comprehensive analysis of these systems with respect to economic efficiency and government's ability to capture economic rents is beyond the scope of this article. However, where significant hydrocarbons are expected, the method now used to transfer of oil and gas rights to the private sector is the sale of leases through a competitive bidding system.

Private ownership of mineral rights resulted in what economists refer to as market failures and led to governmental regulations to remedy these failures. Because an oil or gas pool can extend over several parcels, there could be a number of different property owners with a claim to the oil or gas in a common pool. This is the common property or pool externality. An externality occurs when the welfare of an individual depends directly not just on the individual's activities, but also on activities controlled by some other economic agent. For example, pollution generated upstream affects the welfare of downstream water users.

In the absence of regulation, oil and gas resources were subject to the rule of capture. Historically, the owner who first discovered the pool would start immediately to intensively drill and produce the resource. In the subsurface, oil and gas flows to where there is a reservoir pressure differential; so by quickly drilling production wells, the first producer could drain resources from neighboring owners. The neighboring owners must exploit their part of the pool as quickly as possible just to avoid losing resources to the early producer. This practice accelerated exploration and resulted in too much exploration and too rapid production of the resource. The accelerated production reduced the quantity of the resource that could ultimately be recovered from the pool.

Even when oil prices collapsed, production would not decline because each producer feared having one's own part of the pooled drained by a neighbor. In the leading oil-producing states, government regulators instituted pro-rationing that limited production for each well to an allowable rate per month and set minimum spacing for wells. Well-allowable rates were set to restrict oil supply in order to establish a floor on wellhead prices. For many states, pro-rationing remained in effect until the early 1970s, when the first Arab oil embargo occurred. State regulatory commissions and the federal government now require that production of newly discovered pools be unitized among the owners to avoid the common pool externality and to assure that

potentially recoverable resources are not lost by early overproduction.

Information is crucial to the competitive exploration process. During the period when the rule of capture rewarded early exploration, a system of scouting services was developed in the industry. The domestic and international scouting services report to subscribers land acquisition transactions, well-permitting applications, drilling activities, and post-drilling actions of firms exploring an area. The information generated by discoveries and dry holes is valuable to neighboring leaseholders and other exploration firms who have not taken the risk of actually drilling. The benefits from observing the outcome of drilling without bearing the risk is an externality known as an information spillover. In fact, some Canadian provincial governments lease tracts in a checkerboard fashion to maximize their benefits as the ultimate owner of the resource of such information spillovers.

With the general adoption of forced unitization of new discoveries during the last half of the 20th century, major integrated companies followed the strategy of taking large land positions around a prospect and farming out small parcels to an independent firm (operator) in exchange for drilling a risky wildcat well. In many cases, the independent operator has a lower cost of capital than the major firm, because the independent can fund the well through limited partnerships that provide tax benefits to individual investors. With this system, the firm that actually drilled the discovery well received only a small fraction of the actual discovery. For example, in the Denver basin from 1949 through 1974, the independent firms were credited with drilling discovery wells for fields representing 75% of the reserves but ended up owning only 21% of the reserves in those deposits. The remaining reserves were credited to the major companies because of their dominant land positions.

6.2 Federal Policies to Promote Exploration

The federal government has encouraged oil and gas exploration during most of the 20th century by favorable income tax treatment, specifically through a provision known as the percentage depletion allowance. Favorable tax treatment for petroleum investments enabled promoters and operators of independent firms to raise significant amounts of money by offering tax shelter plans to fund this

drilling. During the peak years of exploratory drilling from 1981 to 1982 in the United States, most of the exploration wells drilled were funded through such arrangements. This funding was reduced dramatically when the Tax Reform Act of 1986 significantly reduced tax rates and rescinded part of the tax code favorable to such tax shelter funds.

During the 1950s and the following decade, the federal government imposed quotas on imported crude oil. The stated purpose was to maintain domestic crude oil production capacity for national security reasons. The quotas had the effect of insulating domestic prices from lower world prices and providing the necessary incentives for exploration and production of domestic resources. Import quotas were rescinded as domestic oil production approached its peak in 1970 and it was clear that supply would not grow as rapidly as U.S. demand.

7. INTERNATIONAL OIL AND GAS EXPLORATION PROVISIONS

Since the early part of the 20th century, British, Dutch, and French oil companies have explored for and produced oil internationally, starting with their former colonies. After World War II, all major U.S. companies initiated international exploration to various degrees.

Outside the United States, national governments typically control ownership of mineral rights. The rights to exploration concession areas are negotiated with the central governments. Although some of the European countries with North Sea production follow an auction system, in some ways similar to the U.S. system, they still maintain tight control of resource development. Many leading oil and gas-producing countries require separate exploration and production agreements. For some countries, including those in the Middle East, Mexico, and Venezuela, petroleum exploration and production is almost entirely controlled by the state oil company, and foreign companies, if allowed at all, are used on a contract basis.

For countries that permit foreign participation, exploration concessions typically require a work commitment including investment in data collection and a specified number of wells drilled in exchange for the right to explore. The concession areas are sufficiently large so that a field will have a single owner, thereby eliminating common pool externalities. In addition, the concession contract will specify that all data collected by the concessionaire should

revert to the government for use by subsequent concessionaires. The agreement will also specify what rights, if any, the exploration firm might have in the event of a discovery. Once a discovery is made, a production concession is negotiated, often requiring the private company to share production with the state oil company or to assume the role of contract producer with payments taken in kind as a percentage of production.

Each government has a different set of contract provisions and production taxes designed to attain a specific set of policy goals. These goals may include training and employment of local workers, construction of infrastructure, and provision of social services to the local communities. Moreover, the governments typically reserve the right to tailor provisions on an individual basis. A description of these provisions is well beyond the scope of this article.

8. TRENDS IN EXPLORATION AND DISCOVERY

8.1 Drilling Activity

Summaries of the total numbers of wells drilled outside the United States are published by country in the journal *World Oil* and the American Petroleum Institute's *Petroleum Data Book*. Data show that the percentage of worldwide annual drilling represented by the wells drilled in the United States went from 80% in 1970 to 30% in 2000. In 1950, the United States accounted for 52% of the world's crude oil production, but by 2000 its share had fallen to less than 9%. The time profiles of new field wildcat wells (successful and dry) drilled each year since 1950 for each of three country groups are shown in Fig. 8. The drilling records for some countries are undoubtedly incomplete, but the graph illustrates the decline in U.S. new field wildcat drilling as well as the decline in the U.S. share of world new field wildcat drilling after 1981. The U.S. share declined from over 90% in 1950 to around 55% by the late 1990s.

8.2 Past Discoveries

There are no internationally accepted standards for reporting reserves or sizes of discoveries. In some countries such as the United States and Canada, reserve definitions are very conservative because they are tied to production facilities and economic conditions. In other countries, the definition of reserves is broader, reflecting the volumes of

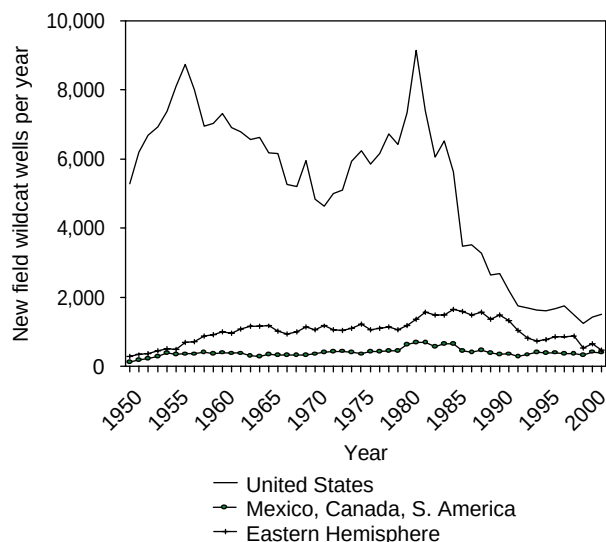


FIGURE 8 Time profile of new-field wildcat wells drilled in the United States, Eastern Hemisphere countries, and the group consisting of South America, Canada, and Mexico from 1950 to 2000. Data from I.H.S. Energy Group, International Petroleum Exploration Production Data Base, 2002, 2002 API Petroleum Data Book.

hydrocarbons that are technically recoverable without reference to being commercial. The discovery data presented in Figs. 9 and 10 must be viewed with this caveat in mind.

Figure 9 shows the annual oil discovery rate, averaged over 5-year intervals, that started in 1915 and annual production. The reported discoveries are in fields that can be produced with conventional production methods. The dominance of the Eastern Hemisphere is striking. It is likely that new discoveries will be revised upward as these fields are delineated and developed, so the sharp decline in overall discoveries is likely to be moderated. The peak levels of discovery from 1955 to 1965, however, will not be repeated. During this period, the Middle East accounted for about half of the discoveries worldwide. World production declined in response to the economic recession between 1980 and 1985, but it began to increase after 1986. It now appears to substantially exceed the past two decades of discoveries.

Figure 10 shows the natural gas discovery rate and production for the world. The Eastern Hemisphere discoveries dominate. The best years for discoveries of gas worldwide were from 1965 to 1970. During this period, the former Soviet Union accounted for two-thirds of all gas discoveries. During the 1970s, many of the largest gas discoveries were in the offshore areas of the North Sea and the Middle East. Figure 10 shows that gas production accelerated

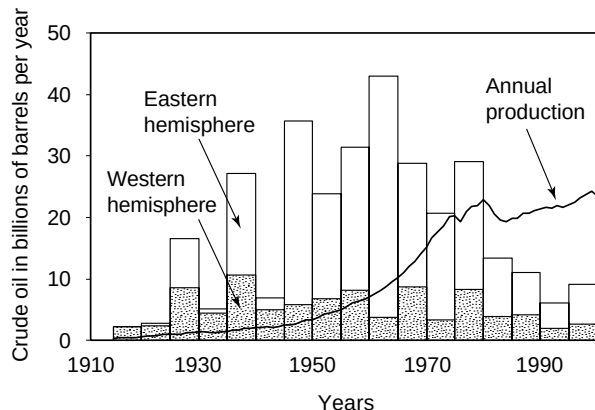


FIGURE 9 Estimated world annual crude oil discovery rate averaged over 5-year periods and annual production. Bottom section of bar is Western Hemisphere and top is Eastern Hemisphere. From U.S. Energy Information Administration, I.H.S. Energy Group, International Petroleum Exploration and Production Data Base, 2002, and Canadian Association of Petroleum Producers, *Statistical Handbook*, 2002.

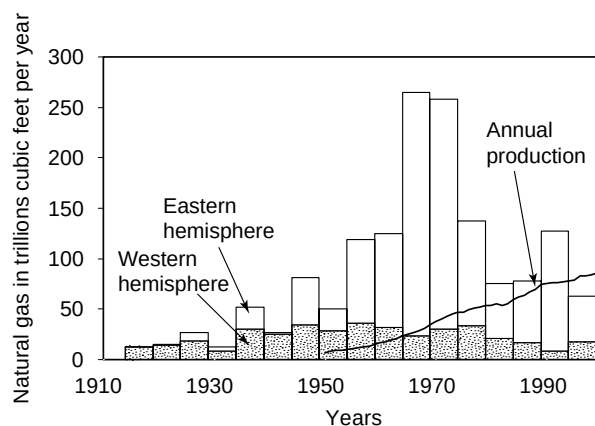


FIGURE 10 Estimated world annual natural gas discovery rate averages over 5-year periods and annual production. Bottom section of bar is Western Hemisphere and top is Eastern Hemisphere. From U.S. Energy Information Administration, I.H.S. Energy Group, International Petroleum Exploration and Production Data Base, 2002, and Canadian Association of Petroleum Producers, *Statistical Handbook*, 2002.

during the late 1980s. Estimates of the size of gas reserves are probably underreported because gas is not commercially marketable in many parts of the world. Consequently, the reported discovery rates shown in Fig. 10 are probably underestimated.

9. CONCLUSIONS

This discussion has identified the determinants of oil and gas exploration, considered the effects of

mineral property ownership on exploration rates, and examined empirical data showing the worldwide oil and gas discovery rates. Because of the way oil and gas occurs in nature, the statistics generated by the discovery process provide the basis for predicting the potential future availability of resource supply from conventional accumulations. As discovery rates inevitably decline and discovery costs increase, resources that are brought to market will, in all likelihood, come increasingly from oil and gas accumulations that are unconventional.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas Drilling • Oil and Natural Gas Exploration • Oil and Natural Gas Leasing • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil-Led Development: Social, Political, and Economic Consequences • Oil Price Volatility • Petroleum Property Valuation

Further Reading

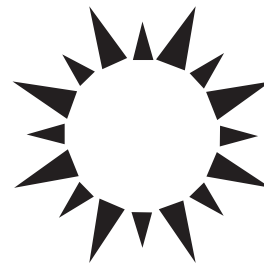
- Adelman, M. A. (1992). Finding and development costs in the United States 1945–1986. In “Advances in the Economics of Energy and Resources” (J. Maroney, Ed.), pp. 11–58. JAI Press, Greenwich, CT.
- Adelman, M. A. (1993). Modeling world oil supply. *Energy J.* **14**, 1–32.
- Adelman, M. A., De Silva, H., and Koehn, F. (1991). User cost in oil production. *Res. Energy* **13**, 217–240.
- Adelman, M. A., and Watkins, G. C. (1997). The value of United States oil and gas reserves: Estimation and application. In “Advances in the Economics of Energy and Resources” (J. Moroney, Ed.), pp. 131–183. JAI Press, Greenwich, CT.
- Attanasi, E. D., and Root, D. H. (1995). Petroleum reserves (oil and gas reserves). In “Encyclopedia of Energy Technology and the Environment,” pp. 2264–2277. John Wiley and Sons, New York.
- Bohi, D. R., and Toman, M. A. (1984). Analyzing nonrenewable resource supply. Resources for the Future, Washington DC.
- Devarajan, S., and Fisher, A. (1982). Exploration and scarcity. *J. Pol. Econ.* **90**, 1279–1290.
- Drew, L. J. (1990). “Oil and Gas Forecasting.” Oxford University Press, New York.
- Drew, L. J. (1997). “Undiscovered Petroleum and Mineral Resources: Assessment and Controversy.” Plenum, New York.
- Energy Information Administration (2002). U.S. Crude oil, natural gas, and natural gas liquids reserves, 2002, Annual Report. DOE/EIA-0216(03).
- Energy Information Administration (2003). Performance profiles of major energy producers, 2001, January, DOE/EIA00206(01) at <http://www.eia.doe.gov/emeu/perpro/perfpro2001.pdf>.
- Hotelling, H. (1931). The economics of exhaustible resources. *J. Pol. Econ.* **39**, 137–175.
- Klett, T. R., Ahlbrandt, T. S., Schmoker, J. W., and Dolton, G. L. (1997). Ranking of the world’s oil and gas provinces by known petroleum volumes. U.S. Geological Survey Open File Report 97–463, one CD-ROM. Washington, DC.
- McDonald, S. L. (1994). The Hotelling principle and in-ground values of oil reserves. *Energy J.* **15**, 1–17.
- Securities and Exchange Commission (1981). Regulation S-X Rule 40–10, Financial Accounting and Reporting Oil and Gas Producing Activities, Securities and Exchange Commission Reserves Definitions, Bowne & Co. Inc., March 1981, New York.
- Society of Petroleum Engineers (1987). Definitions for oil and gas reserves, *Journal of Petroleum Technology*, May 1987, pp. 557–578.
- U.S. Geological Survey Assessment Team (1995). 1995 National Assessment of United States Oil and Gas Resources. U.S. Geological Survey Circular 1118. Washington, DC.
- Yergin, D. (1991). “The Prize.” Simon & Shuster, New York.

Oil and Natural Gas Exploration

MARLAN W. DOWNEY

Southern Methodist University

Dallas, Texas, United States



-
1. General Characteristics of Oil and Natural Gas Exploration
 2. Fundamental Concepts Guiding Exploration for Oil and Natural Gas
 3. Initiation of Exploration
 4. Regional Studies
 5. Prospect Definition
 6. Search Techniques for Oil and Natural Gas
 7. Assessing Risk in Exploration
 8. The Results: Evaluating Exploration Effectiveness
 9. Summary

Glossary

field A commercially valuable subsurface accumulation of hydrocarbons.

permeability A measure of the transmissibility of fluids in the void space in a rock.

petroleum A naturally occurring mixture of liquid hydrocarbons.

porosity Percentage of void space in rock volume.

reservoir A rock layer with substantial porosity and permeability.

wildcat An exploration well (supposedly one drilled so far from civilization that the drillers can hear the cries of wildcats).

Exploration for oil and natural gas is a search for buried treasure, on a huge scale, for commercial purposes. Exploration is a business-oriented search for new hydrocarbon assets. Geology is a science; geophysics is a science; exploration is a business. It is not difficult to find traces of oil and natural gas; it is very difficult to find sufficient quantities of hydrocarbons to make a profit.

1. GENERAL CHARACTERISTICS OF OIL AND NATURAL GAS EXPLORATION

A romantic image of exploration is one of hardy drillers risking fortunes on the luck of a wildcat well. The large risks and the financial dangers are still very much a part of modern exploration, but all the tools of modern technology are coupled with the strengths of business analysis to quantify and diminish the risks of exploration.

Exploration accomplishments come with high costs and high risks. Some idea of the risk of exploration investments, relative to other investments, may be gained by understanding that banks do not loan money for exploration investments. Hardy investors in exploration oil and gas projects are effectively lenders-of-last-resort, and such investors demand the potential for high returns to balance the exceptional risks. When a company drills a \$20 million “dry hole,” a well that fails to find significant oil or gas, there is no salvage value and no optional use for the dry hole: It is a \$20 million loss.

In a typical year, approximately 10 billion barrels of oil and 70 trillion cubic feet of natural gas are discovered in the world. Assuming a value of \$25 per barrel for oil and \$3.00 for 1000 cubic feet of gas, it can be seen that \$460 billion dollars worth of new assets are discovered each year by oil and gas exploration. The rewards for exploration success can be huge, but costs are high and failure is common.

The exploration task is far greater than discovering new subsurface reservoirs of hydrocarbons; it must be that of discovering those hydrocarbons in an economically efficient manner. Successful oil and gas exploration owes as much to efficient business practices as to technical excellence.

The first step in exploration is always an innovative idea about where oil or gas might be found. The pioneering geologist Wallace Pratt once said, “Oil is first found in the minds of men.” It is hoped that such

ideas arise from fundamental knowledge of the rock layers in the earth, their suitability for generating and retaining oil and gas accumulations, and an analysis of the clues available to suggest that proper conditions are actually present in a particular area. Oil and gas accumulations in the earth are extremely valuable. The discovery of new accumulations (fields) generates great wealth. Oil and gas fields are very difficult to find, and exploration—a systematic search for new oil and gas accumulations—depends on proper use of technology, acquisition of new data, and the deductive analysis of earth clues in a manner reminiscent of a detective.

2. FUNDAMENTAL CONCEPTS GUIDING EXPLORATION FOR OIL AND NATURAL GAS

All oil and much natural gas are created by deep burial and heating of organic matter contained within sedimentary rocks. As this organic matter is converted to oil and gas by burial and heating in the earth, oil and gas are forced out of the organic-rich sedimentary layers and expelled into open fractures or porous transmissive zones. These fractures and transmissive zones in the earth provide paths for hydrocarbon migration. The oil and gas may be retained and trapped in subsurface reservoirs or may continue to leak upward to the surface of the earth. Such leaks of hydrocarbons to the surface have been forming tar pits, oil lakes, and gas seeps for millions of years.

If the migrating hydrocarbons encounter an impervious layer of rock, they will be prevented from escaping to the surface. If the impervious layer roofs a rock layer with porosity, the migrating petroleum may fill the pore space of the reservoir and displace some of the water originally filling the pores of the rock. If the roofing layer and the associated reservoir rock have been folded in a domal shape, the hydrocarbons may fill the container formed by the domed surface. The oil, natural gas, and water in the reservoir strata have distinctly different densities. Gas floats on oil; oil floats on water. In a domal structure, then, the lighter gas will occupy the pore spaces in the uppermost part of the structure and be underlain by the oil, which in turn is underlain by water.

The two most fundamental search elements in exploration for oil and gas are (i) a knowledge of where oil and gas have been generated and provided to

migration routes and (ii) the location of specific traps that could interrupt migration and retain petroleum.

3. INITIATION OF EXPLORATION

The exploration process can start with a new idea, new information, or a new area becoming available for search. If exploration is initiated because a new international search area becomes available for leasing and drilling, it is quite likely that the new search area will be leased to whomever bids the most for the right to explore. In such new areas, most bidders will have a similar set of information obtained from the host country. A fascinating area of operations research derives from the competitive aspects of bidding under uncertainty. An insight into this rich area of decision making is given by understanding that any winning bidder (the one who bid higher than anyone else) often bids more than the lease is worth. The more bidders, the more likely that the winning bidder has overbid. If 20 companies, with similar knowledge, submit bids for the undrilled area, what is the chance that the winning bidder (who values the undrilled area higher than 19 other knowledgeable bidders) has still managed to bid less than the true value of the lease? This chance ranges between slim and none. In the oil and gas business, this situation of winning the bid (but paying too much) is described by Capen as the winner's curse.

When exploration is initiated because of exclusive access to new information, the company with the new information has some remarkable advantages in finding and valuing new exploration projects. This competitive advantage is one of the reasons that large companies spend huge amounts of money in acquiring new data for their exclusive use and advantage.

Sometimes, exploration is initiated simply because a team or unique individual has extracted a novel concept, a new interpretation, from generally available data. Such individuals or teams are rare and extraordinarily valuable; they spawn new directions for exploration and give incredible competitive advantages to their employers. Such great oil finders are like master chess players, seeing more deeply, more comprehensively, than others.

4. REGIONAL STUDIES

Exploration for oil and gas in the earth can be usefully subdivided into two scales of effort: regional

evaluation and prospect definition. In the first effort, data are gathered to indicate whether specific regions of the earth are likely to contain hydrocarbons. In the second effort, an attempt is made to decide exactly where to drill to find oil or gas, within the general region that appears favorable for containing subsurface accumulations of hydrocarbons.

Regional studies attempt to answer the following questions: Where are the organic-rich layers in the subsurface? What is their richness and character? Where in the earth have they been buried to sufficient temperatures to generate oil and gas? What was the time at which the source rocks were heated and provided petroleum? What traps were available to retain the generated hydrocarbons at the time the hydrocarbons were generated and migrating?

Geologists sometimes use the phrase “petroleum systems analysis” to describe the type of studies that must be made to determine whether a region should be expected to contain accumulations of hydrocarbons. This concept of the petroleum system, pioneered by Magoon and Dow, recognizes that accumulations of hydrocarbons in the earth require a special combination of unique circumstances before hydrocarbons can be generated and be available at the right time to fill thick porous reservoirs contained within traps. If the hydrocarbons are not available, the reservoirs and traps will be barren. If the porous reservoirs are absent, the traps will only contain hydrocarbons in rocks unsuitable for production. If the traps are not present, the hydrocarbons will move through the porous reservoir and continue to the surface of the earth. All of the necessary parameters must be present simultaneously at the time of generation and accumulation and persist until the present day.

One simple test for the presence of a working petroleum system is a documentation of a well in the area that flowed hydrocarbons to the bore hole. Although this flow of hydrocarbons may be trivial from a commercial standpoint, it is of great significance in indicating that a working petroleum system is likely. From such an observation, the presence of hydrocarbons trapped in a reservoir is demonstrated and not merely guessed at or hoped for.

In the final stage of regional studies, the geologist and the geochemist may make a mass balance calculation of the total amount of hydrocarbons expected to have been generated from the petroleum source rocks in an area to compare with the amount and type of hydrocarbons already discovered in an area. If these regional studies are promising, the task of the explorationist can shift to a search for specific

“prospects” suitable for drilling and evaluation in the region.

5. PROSPECT DEFINITION

5.1 Prospect Varieties

There are two general varieties of hydrocarbon traps: structural and stratigraphic. However, many combinations of these trap types are known.

5.1.1 Structural Traps

Structural traps involve folds of subsurface rock layers that provide interruptions to the continued upward migration of oil and gas. If hydrocarbons are migrating in an uninterrupted porous reservoir, beneath a layer of sealing rock, they may move hundreds of miles laterally and up to the surface, unless collected along the way by a concave fold (anticline) in the earth layers. (In an analogous manner, on the surface of the earth, depressions in the earth's surface collect the flow of streams to provide the accumulations of water we know as lakes.) Each concave-downward trending fold attracts the hydrocarbons, which are buoyant and floating on the subsurface waters in the reservoirs. As each concave-downward fold is filled to capacity with the migrating hydrocarbons, the excess hydrocarbons may spill from an edge of the confining surface and continue to move upward.

Structural traps have two significant attractions to explorers: The structural fold is readily located by various techniques, and the structural fold perturbs the moving hydrocarbons and may act to focus and collect the buoyant hydrocarbons.

Another common type of structural trap, the fault trap, is found where alternating layers of reservoirs and seals are broken by faults that cause the reservoir layers to become offset against sealing layers on the up-dip side. These fault traps are important containers of hydrocarbons.

5.1.2 Stratigraphic Traps

Stratigraphic traps occur in great varieties and combinations. A large proportion of the world's oil and gas is contained in stratigraphic traps. Essentially, a stratigraphic trap results from a lateral change in the subsurface rock strata—for example, where a porous and permeable rock layer gradually becomes nonporous and impermeable in an up-dip direction. Hydrocarbons migrating up-dip in the porous strata will collect at the place where the

hydrocarbons are impeded by the seal. Stratigraphic traps are more complex and more difficult to find than structural traps, and they require special understanding of such reservoir characteristics as the lateral extent of the reservoir and its variance laterally and vertically. Whether hydrocarbons are trapped at the up-dip edge of a reservoir depends on the pore throat openings characterizing the up-dip barrier and on the buoyancy properties of the migrating hydrocarbons. The buoyancy pressure of the hydrocarbon column is dependent on the density difference between the hydrocarbon and water and on the height of the hydrocarbon column. If increasingly more hydrocarbons migrate into a stratigraphic trap, increasing the height of the hydrocarbon column and its buoyancy pressure, the hydrocarbons may eventually be forced into the fine pore throats of the up-dip barrier and move through it.

As Downey pointed out, it is easy to find conditions in which tiny quantities of hydrocarbons are entrapped because small accumulations lack buoyancy pressure to overcome even minor barriers. It becomes progressively more difficult to trap larger accumulations of hydrocarbons because large columns of hydrocarbons exert large buoyancy forces against the pore throats of the potential up-dip seal. Seal risk becomes more important as larger fields are sought.

6. SEARCH TECHNIQUES FOR OIL AND NATURAL GAS

6.1 Geophysical Prospecting Techniques

6.1.1 Seismic Surveys

Prospects (hypothesized accumulations) are most commonly located and described by geophysical techniques. The most useful of these geophysical techniques involves penetrating the rock layers with acoustic waves and recording, at the surface, the reflection of these acoustic waves from subsurface rock interfaces. These seismic surveys require immense compute power and technical sophistication, but the general approach is not unlike the use of sonar to find schools of fish or to depict the form of the sea bottom beneath a boat. Generally, the structural form, strata thicknesses, and the general characteristics of the subsurface rocks can readily be determined from information provided by seismic investigations.

In the past, seismic data were obtained by using linear tracks, providing two-dimensional (2D) data; modern seismic work favors collection and analysis of data in three dimensions, giving unrivaled detailed pictures of the subsurface. Of course, the cost to acquire 3D data is much more than that of the 2D data set, and careful analysis must be made to ensure that the value obtained from the 3D data set justifies its costs. In addition, a 3D set may be repeated over time to monitor changes in seismic reflection character. Such 3D data sets are called 4D sets and can be very useful for review of *in situ* changes in field characteristics. The depletion of hydrocarbons in various segments of a field can be monitored over time and extraction efficiencies maximized.

In favorable circumstances, the acoustical data collected from the seismic work may directly indicate whether the subsurface layers contain oil and gas. The discovery of seismic techniques (bright spots) for direct detection of underground oil and gas by M. C. Forrest of Shell Oil in the late 1960s revolutionized exploration.

6.1.2 Gravity and Magnetic Surveys

Gravity surveys provide measurements of variations in the earth's gravity at a number of locations in a region. These gravity variations represent changes in the density of the rock column under the measuring site and are helpful as a quick, inexpensive way of recognizing the presence of thick sections of sedimentary rock. Gravity surveys are most helpful in regional studies, but in special cases, such as for areas where prospects are created by large, buoyant salt domes, gravity can be very useful to define the distribution and structural attitude of the salt masses, which are less dense than the adjacent rocks.

Magnetic surveys are also helpful in regional studies to provide an estimate of depth to magnetic basement. On the prospect scale, magnetic investigations can readily reveal areas of igneous intrusion; the igneous dikes and sills appear as strong magnetic anomalies.

6.2 Surface Geochemical Techniques

Surface measurements of hydrocarbons are often a useful supplement to other exploration search techniques. The presumption is that tiny leaks of hydrocarbons are occurring above subsurface oil and gas accumulations, and with proper collection and analysis of soils and soil gases these microseeps can be recognized and used to pinpoint the subsurface origin of the leaks. The problem with the proper

utilization of surface geochemical techniques is not one of adequate sensitivity but one of differentiating hydrocarbons leaking from buried accumulations from hydrocarbons abundantly generated by plants and living organisms and hydrocarbons provided by surface contamination.

6.3 Geologic Techniques

Geological techniques in exploration have changed greatly during the past 50 years. Originally, the geologist searched surface exposures for oil and gas shows and mapped folds of surface rocks. Later, the geologist emphasized subsurface data such as recognition of oil and gas shows from rock cuttings from wells, identified the reservoir characteristics of subsurface layers, and integrated geophysical data with electric well surveys. The modern geologist has added expertise in organic geochemistry, geophysics, petrophysics, and economics. One of the greatest values of a modern petroleum geologist is his or her ability to visualize the subsurface in three dimensions using a knowledge of rock facies, environments of deposition, and sequence stratigraphy so as to reason from negative data (where oil is not) where oil should be.

6.4 Dowsing Techniques

Dowsing and other prospecting techniques calling on supernatural powers or divine insights are of absolutely no value. Dowsers say that they can feel emanations from oil and gas fields, underground water currents, and deposits of precious metals. There are many practitioners, devotees, and believers, but there is no science. No such technique has ever been shown to work at levels better than pure chance in any controlled trial. The Randi Institute maintains a \$1 million prize to anyone who can demonstrate such remarkable powers in controlled circumstances. It remains uncollected.

6.5 Exploration Drilling Techniques

Drilling is costly and is generally reserved to test the best of the exploration prospects. Drilling of the exploratory test wells, the wildcats, is typically the most expensive procedure in the exploration cycle. Drilling represents the ultimate test of the oil finders' hypothesis that the mapped prospect may contain hydrocarbons.

Onshore drilling is generally accomplished with a moveable drilling rig. The drilling rig has three basic

components: the derrick, which provides height for hoisting the drill pipe as the well is drilled; the powered rotary table, which clasps and turns the drill pipe with its attached drill bit; and the "mud" pumps, which provide power to force the mud through the drill pipe and drill bit. The drilling mud is really a sophisticated mixture of expensive chemicals designed to create a temporary skin for the walls of the newly drilled hole, act as coolant to the drill bit, and to provide adjustable pressure control for restraining any encountered earth pressures. The mud is carefully weighted with additives, such as barium carbonate, so that the weight of the column of mud in the drill pipe is sufficient to hold down any earth pressures encountered and prevent a blowout.

The procedure for drilling a hole into the earth involves rotating a dangling hollow drill pipe, which has a perforated bit at the end. As the upper end of the drill pipe is held firmly within the rotary table on the drilling floor, the drill bit on the down-hole end of the drill pipe rotates and grinds on the rocks on the bottom of the hole. Mud is pumped at high pressure down the hollow pipe and out the orifices in the bit at the bottom of the hole. The mud jets out of the orifices in the drill bit and assists in the drilling process. On the return trip up the annulus outside the drill pipe, the mud serves to carry out the rock fragments created by the drilling process. These rock chips are screened from the mud system at the surface, rinsed, and examined by the well-site geologist for rock characteristics and oil shows. The cleansed mud is carefully checked to ensure that its physical and chemical properties are exactly what is required, and then it is recirculated down the drill pipe.

As the drill bit becomes worn, the entire drill pipe must be hoisted from the hole; each 90-ft section of pipe is unscrewed from the upper end of the dangling drill pipe and stacked in the derrick. When the bottom end of the pipe is pulled onto the rig floor, the old bit is removed, a new bit is added, and the 90-ft sections of pipe are sequentially reattached to the upper end of the drill string as the bit is lowered to the bottom to continue drilling. In the colorful language of the oil fields, this long and difficult procedure is called "making a trip."

If the well has been drilling at 15,000 ft below the derrick floor, the round trip involves 332 separate connection activities by the roughnecks, the skilled drilling team.

The fundamental mechanics of drilling an oil well are more akin to rotating a weight on a string than pressing on a brace and bit because the miles-long

drill pipe has enormous weight but little compressive strength. At specific planned depths, the drill pipe and bit are hoisted out of the hole and long lengths of steel casing are lowered and set in the hole. Cement is pumped down the hollow casing until it flows out the bottom of the casing and up the annulus between the casing and the earth layers. After the cement cures, it bonds the outside of the casing to the bore hole walls. The steel casing can now protect the drilling operation from influxes of fluids or rocks caving from previously penetrated layers of rock. The drill pipe and bit are lowered into the hole again and drilling is recommenced, with the upper portion of the hole “cased off” and protected. This casing procedure may be repeated many times during the drilling of the well, with each set of casing nestling inside the others.

Offshore drilling utilizes the same basic drilling procedures as those of onshore drilling but involves a number of special and expensive modifications. Offshore drilling requires a secure place to set the drilling rig as well as space for the staff, equipment, and supplies.

In waters shallower than 600 ft, it is common to use a large floating barge, called a jack-up, with long legs on each corner. The jack-up is towed into position with the legs elevated into the air. After the jack-up is positioned over the prospect, the legs are jacked down into the sea bottom, and the barge with the drilling rig is lifted and supported above the water on the legs. After drilling is completed, the legs are jacked up, the barge is let down to float on the sea, and the barge is towed to another location.

In deep waters (1000–12,000 ft), very large quantities of oil and gas must be discovered by the explorers to pay for the immense costs of drilling. Large deep-water drilling vessels can cost more than \$500 million to construct. Ordinary expenses to operate the deep-water drilling program can easily be \$250,000 a day. Costs to drill wildcat wells in deep water have sometimes exceeded \$100 million per well.

The drilling rig for deep water is installed on a specially designed mobile vessel, providing crew quarters and containing all necessary equipment for months of work. Such drilling vessels may have numerous thrusters for lateral positioning and be capable of automatically adjusting their position to keep the drilling rig directly over the drilling target. The drill bit and drill pipe are lowered thousands of feet to the sea floor, and drilling is commenced toward targets that may be tens of thousands of feet below the sea floor.

7. ASSESSING RISK IN EXPLORATION

The major reason for failure in exploration is misunderstanding risk. Most exploration ventures fail. Proper selection of exploration projects can lessen the average risk in the portfolio. Selection of exploration investments involves an estimation of the exploration costs to determine success, an estimation of the probability of success, and the value of the project, given success. It is generally understood that proper estimation of the probability of success is one of the most important elements in maintaining a high-quality portfolio of exploration projects. Probabilities of success for various exploration ventures in a portfolio can range over three orders of magnitude. These estimations of probability of success must be made in a consistent manner, with geologic and mathematical rigor.

A number of approaches to estimating probability of success are utilized by various groups. In all cases, the historical wildcat success rates in a particular province should be compiled for comparison to any estimates of future success. Past history does not define future probabilities of success, but historical knowledge acts to constrain wildly optimistic forecasts.

The most useful general approach is outlined by Capen and is easy to understand because it mimics a board game with a “wheel of chance” and a spinner. For each prospect risk analysis, the wheel of chance is divided into two segments, whose areas denote the historical probability that a wildcat would or would not expect to encounter an accumulation of hydrocarbons in the region being analyzed. If one-fourth of the wildcats in the region have encountered some sort of accumulation of hydrocarbons, one-fourth of the wheel of chance would be noted as “oil and gas,” and three-fourths of the wheel would be identified as “dry hole.” One time in four, the spinner, representing chance, would land on the quarter circle called oil and gas. Although this example indicates that one wildcat out of four might be expected to find an oil and gas accumulation, remember that most oil and gas accumulations in the earth are very small.

Oil and gas accumulations have a nearly log-normal frequency distribution in the earth. As a result, we can divide the quarter circle area of oil and gas with a log scale along the quarter circle circumference, and estimate the differing likelihood that the spinner might alight on the segment (probability) of 500 million barrels versus the segment (probability) of a 10-million barrel or smaller field (Fig. 1).

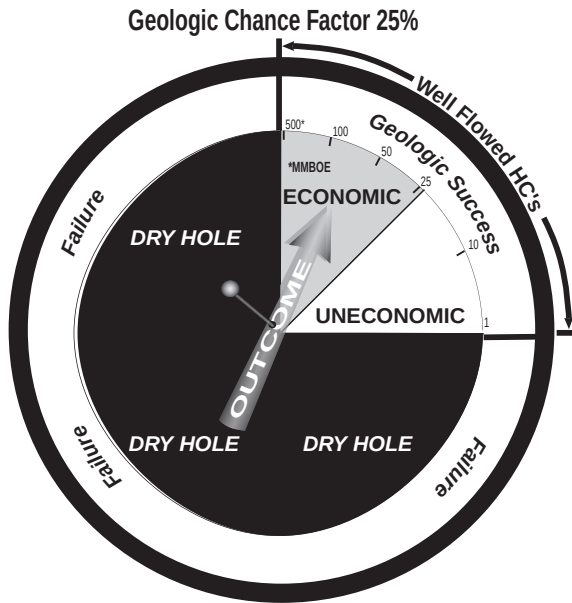


FIGURE 1 "Wheel of chance" for assessing probability of success in an exploration drilling venture; after Capen (1992).

Lastly, we can determine the minimum size of accumulation that can be commercially developed. If we are interested in making money in oil and gas exploration, we must analyze the probability of finding a field at least as large as the minimum developable size. If our minimum developable field size is 10 million barrels of oil, then we need to determine the probability that our next wildcat will encounter at least 10 million barrels of oil. Using the probability of commercial success, the expected cost of discovery, and the present value worth of the target, we can estimate the risk discounted present value worth of the exploration project.

7.1 Contractual Risk in Exploration

At the inception of exploration, an agreement must be negotiated that provides for the recovery of costs, taxation, and the share of the production retained by the explorer. These agreements or production contracts typically extend for the life of the production. In all cases outside the United States in which significant exploration finds have been made, the contract has been unilaterally altered in favor of the host country at some time during the life of the contract. Even within the United States, modifications to previous oil and gas contracts with state and federal governments are common. This certainty of some degree of contract abrogation is a fact of life and another risk to be assessed by the oil and gas investor.

8. THE RESULTS: EVALUATING EXPLORATION EFFECTIVENESS

8.1 Reserve Replacement

Large oil and gas companies can be considered as powerful search engines for the discovery of new assets each year. A large oil company that produces 500 million barrels of oil each year to supply its refinery and marketing system should find at least 500 million barrels of new oil each year or its asset base will diminish and the value of its shares may fall. The ratio of new reserves found to those produced is called the reserve replacement ratio, and it is closely monitored as one indication of the efficiency of the exploration process for individual companies.

8.2 SEC Finding Costs

Each publicly traded company in the United States is required to provide a yearly report to the Securities and Exchange Commission (SEC) that includes specific data on current-year cost of finding oil and gas. These company reports to the SEC are greatly relied on by analysts but can be very misleading. This is because the SEC finding cost reports mandate a comparison between the amount of oil and gas volumes "booked" each year by each company compared to the amount of money that each company spent in the current year on exploration. The problem for the unwary analyst is that the volumes booked in the current year may have been found many years ago with past exploration expenditures; the exploration expenditures of the current year may have found volumes that will not be booked as producible reserves until several years in the future. The oft-quoted SEC exploration finding cost, dollars spent on exploration in the current year divided by volumes booked in the current year, is an unreliable and often meaningless number. Exploration efficiency is extremely important to know and very difficult to decipher from such published information.

8.3 Internal Measures of Finding Costs

The internal finding costs of a company are determined by comparing each year's exploration expenditures to the volumes thought to have been discovered during the same year. These internal exploration finding costs, properly created and regularly monitored, are a very useful monitor of exploration efficiency.

8.4 Financial Measures of Exploration Efficiency

The perfect method for evaluating exploration efficiency might seem to be a simple comparison of the money spent each year on exploration versus the present value worth of the petroleum assets discovered each year. An indication of the problems with this method is suggested if we rephrase the financial measure more precisely and state, “a comparison of the money spent each year on exploration versus an estimate of the present value worth of petroleum assets discovered.” Much of the apparent rigor of the financial measure disappears when it is realized that the financial measure depends on guesses as to oil and gas prices many years in the future, which are multiplied by guesses as to the amount and timing of the future production.

8.5 The “Gold” Standard: Risk-Discounted Expectation versus Actual

The best technique for monitoring exploration efficiency (the gold standard) is a cumulative comparison of the risk-discounted volume expectation of each wildcat versus the actual volume outcome. Exploration efficiency can be easily monitored by a cumulative plot of barrels predicted to be found (and barrels actually found) recorded against the wildcat well sequence. This graph provides two cumulative curves: the volumes expected to have been found and the volumes actually found. Such a plot also illustrates the need for an adequate number of wildcat chances to

obtain success. Over any significant period of time, the sum of the predicted volumes must match the actual outcomes. This gold standard is a rigorous test that compares what the exploration team promised it was going to find versus actual outcome over time (Figs. 2 and 3).

8.6 Finding and Development Costs

One of the most useful measures of the performance of an oil company over a long period is given by the company’s average costs to find and produce oil and gas fields on a per barrel basis. Companies with low finding and development costs have a considerable advantage in the marketplace. Prices received by the producer for its oil and gas may increase or decrease, but companies that are able to deliver oil and gas at low finding and development costs are always the strongest.

Current finding and development costs represent the outcomes of decisions and actions made many months, perhaps many years, earlier. Low finding and development costs reflect well on the company and, to a lesser degree, on the current management.

8.7 Reserve Appreciation

A final caution in evaluating exploration efficiency is more subtle to understand and more powerful in its effect: What volumes are actually being discovered by current exploration? Numerous “hindcasting” studies of exploration have demonstrated that the oil volumes thought to have been

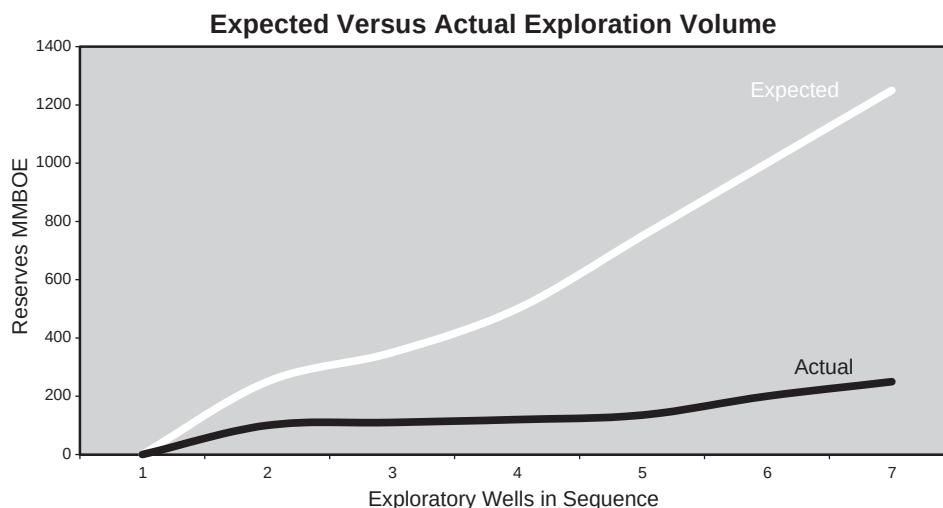


FIGURE 2 Volumes predicted to be discovered versus volumes actually discovered; unsuccessful exploration portfolio.

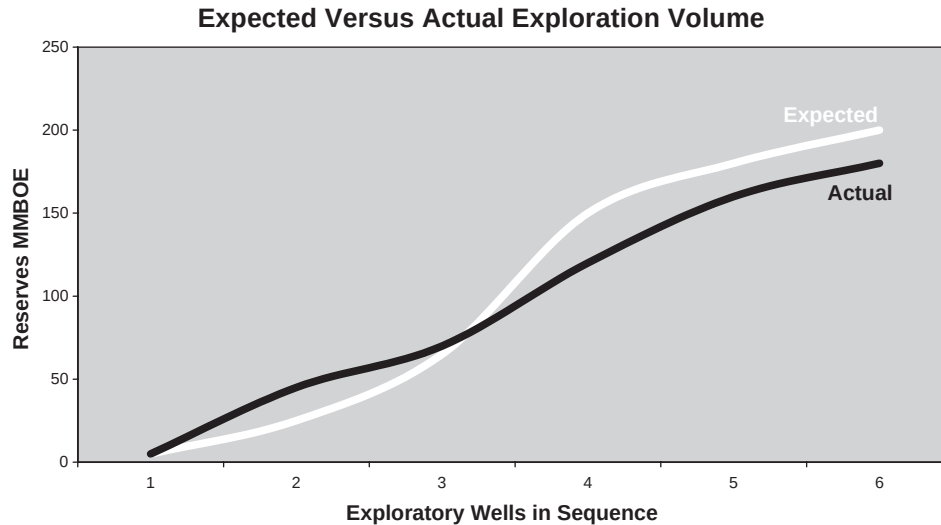


FIGURE 3 Volumes predicted to be discovered versus volumes actually discovered; improved portfolio.

found by exploration continue to increase with time and field production, until the ultimate production from oil fields is seen to dwarf the early, ultraconservative estimates made at the time the fields were discovered. How great is this underestimation of exploration discoveries as a fraction of ultimate oil recovery? On an overall basis, the early reserve estimations appear to more than triple after long-term production gives a revised picture of ultimate oil reserves. (Techniques for determining natural gas volumes are far more accurate and early exploration estimates appear to understate ultimate gas reserves by only 15–20%). Why is this understatement of new reserves so prevalent? Two explanations may suffice: one economic and one technical.

At the time of first development of a newly discovered field, huge sums of money must be invested before there is any production or payback. A 100-million barrel oil field may require \$200–800 million for drilling wells and providing facilities for production. Estimates of field size, made at this early time for the purpose of justifying development costs, must be very conservative because company management must guarantee sufficient reserves to cover the projected development expenses. Such early estimates should properly be considered as being 95% certain of at least the volumes, rather than expectation volumes.

Second, once the wells are drilled, the field facilities are in place, and costs recovered, incremental volumes become available from secondary, higher cost hydrocarbon-bearing horizons, whose volumes often dwarf

those of the originally productive horizons. The ultimate effect of this general understatement of discovery volumes is to underestimate the value of the assets discovered each year by exploration.

8.8 Size Distribution of Oil and Gas Fields

In any sufficiently large area where oil and gas fields occur, the size of the individual oil and gas fields will be found to conform to a near log-normal size distribution. In lay terms, this means that there will be very few large fields, a substantial number of medium fields, and a great number of very small fields. This remarkably consistent outcome is a result of some subtle mathematics. Each oil field depends on the simultaneous occurrence of a number of parameters, such as the amount of hydrocarbon charge available, suitable reservoirs to contain the hydrocarbons, a trap configuration to hold the hydrocarbons, a seal to confine the hydrocarbons, and sufficient permeability to allow the hydrocarbons to flow to the surface at commercial rates. The log-normal field size distribution, ubiquitous in nature, is understandable as a result of multiplication of the probability distributions describing the critical parameters.

9. SUMMARY

Finding traces of oil and gas is not difficult; making money finding oil and gas is very difficult. Exploration

is the business of efficiently finding oil and gas so that the new discoveries can be developed and produced. Successful exploration requires a fundamental knowledge of how oil and gas are generated and how they migrate and accumulate. A number of search techniques are available to explorers and the proper use of these technologies is the responsibility of the technical staff in the oil companies. In addition, oil and gas exploration must be conducted in a cost-efficient manner so that the exploration product, the field discoveries, will be highly profitable and will properly reward the oil company for its assumption of large financial risk.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil

and Natural Gas Drilling • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Leasing • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Petroleum Property Valuation • Risk Analysis Applied to Energy Systems

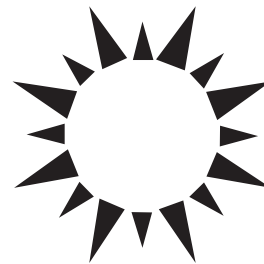
Further Reading

- Attanasi, E. D., and Root, D. H. (1994). The enigma of oil and gas field growth. *Am. Assoc. Petroleum Geol. Bull.* **78**, 321–333.
- Capen, E. C. (1992). Dealing with exploration uncertainty: Business of petroleum geology, AAPG Treatise of Petroleum Geology. American Association of Petroleum Geologists, Tulsa, OK.
- Downey, M. W. (1984). Evaluating seals for petroleum accumulations. *Am. Assoc. Petroleum Geol. Bull.* **68**(11), 1752–1763.
- Forrest, M. C. (2000, May). “Bright” ideas still needed persistence. *Am. Assoc. Petroleum Geol. Explorer*, 4–12.
- Magoon, L. B., and Dow, W. G. (1994). The petroleum system: From source to trap, AAPG Memoir 60. American Association of Petroleum Geologists, Tulsa, OK.

Oil and Natural Gas Leasing

DENNIS D. MURAOKA

California State University, Channel Islands
Camarillo, California, United States



-
1. Current and Historical Acreage and Production from Federal Oil and Gas Leases in the United States
 2. The Legal Framework for Federal Oil and Gas Leasing
 3. Federal Oil and Gas Leasing Procedures
 4. The Effectiveness of Federal Oil and Gas Leasing Procedures
 5. State and Local Government and Environmental Concerns Regarding Federal Oil and Gas Leasing

Glossary

bonus payment A one-time, nonrefundable cash payment made by a lessee to the lessor at the time a lease is issued. It is usually the bid variable for competitive leases.

competitive lease A lease issued in an area believed to contain crude oil or natural gas. Such leases are auctioned to the highest qualified bidder.

economic rent The difference between the total revenue generated by a lease and the total necessary cost needed to generate this revenue.

noncompetitive lease A lease issued for lands not believed to contain crude oil and natural gas.

outer continental shelf The shallow portion of the continental margin that extends from the coastline to the more steeply inclined continental slope. These offshore lands fall under the jurisdiction of the federal or state governments.

primary lease term The time period specified in a lease wherein the lessee has the exclusive right to explore for crude oil or natural gas.

rental payment An annual payment by the lessee to the lessor determined on a per acre basis from the time the lease is issued to keep the lease in force.

royalty payment A payment in value or in kind by a lessee to the lessor calculated as a percentage of the gross income from a lease.

transfer payment A payment that redistributes wealth from one sector or group within the economy to another.

Federal, state, and local governments own and manage over 30% of the surface area of the United States. In addition, government agencies oversee vast

offshore lands adjacent to the U.S. coastline. These holdings place vast quantities of crude oil, natural gas, and other natural resources under public stewardship. Rather than develop these resources as public sector enterprises, government agencies have typically transferred subsurface mineral rights to the private sector while retaining the ownership of the public lands. The contract used to convey these rights is a lease. In the case of a crude oil and natural gas lease, the government (the lessor) conveys to individuals or businesses in the private sector (the lessees) the right to explore for and extract the oil and gas from beneath the government lands. In return, the lessees agree to compensate the government with a series of payments described in the lease agreement. The leasing procedures described below pertain to federal oil and gas leases. State and local government leasing procedures vary widely. Nevertheless, they include the same or similar components as their federal counterpart.

1. CURRENT AND HISTORICAL ACREAGE AND PRODUCTION FROM FEDERAL OIL AND GAS LEASES IN THE UNITED STATES

As of December 31, 2000, the federal government had contracted with the private sector for oil and gas production from federal lands in 31 states (Table I). There were 23,844 producing or producible leases covering an area of 12.6 million acres of federal and Native American lands. The average federal onshore lease covered 530 acres. In addition, the federal government had lease contracts for submerged lands in the Gulf of Mexico and lands offshore from California and Alaska. There were 1833 leases offshore covering 8.7 million acres of offshore land. The average offshore lease was 4749 acres—much larger than the average onshore lease.

TABLE 1

Federal Onshore and Offshore Oil and Gas Leases as of December 31, 2000

Onshore	Number of producing and producible leases	Acreage leased (acres)
Alabama	24	13,822
Alaska	36	64,178
Arizona	15	69,096
Arkansas	192	90,092
California	323	77,478
Colorado	2193	1,893,447
Florida	3	3476
Illinois	8	1581
Kansas	446	122,597
Kentucky	61	35,481
Louisiana	223	79,222
Maryland	2	34,941
Michigan	70	78,800
Mississippi	94	44,245
Missouri	1	200
Montana	1756	883,473
Nebraska	25	37,275
Nevada	37	23,124
New Mexico	6464	3,989,351
New York	4	1009
North Dakota	572	320,969
Ohio	167	30,472
Oklahoma	2579	272,268
Pennsylvania	60	24,680
South Dakota	76	34,777
Tennessee	7	2446
Texas	207	110,678
Utah	1931	1,129,060
Virginia	15	11,119
West Virginia	159	149,693
Wyoming	6094	3,017,958
Onshore subtotal	23,844	12,647,008
Offshore		
Alaska	4	20,481
California	43	217,668
Gulf of Mexico	1786	8,467,355
Offshore subtotal	1833	8,705,504
Onshore and offshore total	25,677	21,352,512

Source. Reprinted from U.S. Department of the Interior (2001). "Mineral Revenues 2000," Minerals Management Service, Herndon, VA.

Federal oil and gas leases contribute over one-third of the crude oil and natural gas produced in the United States. Federal leases produced 689 million

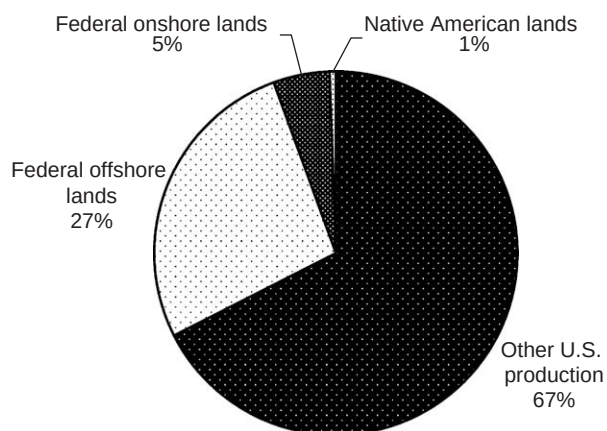


FIGURE 1 Crude oil production in the United States in 2000.

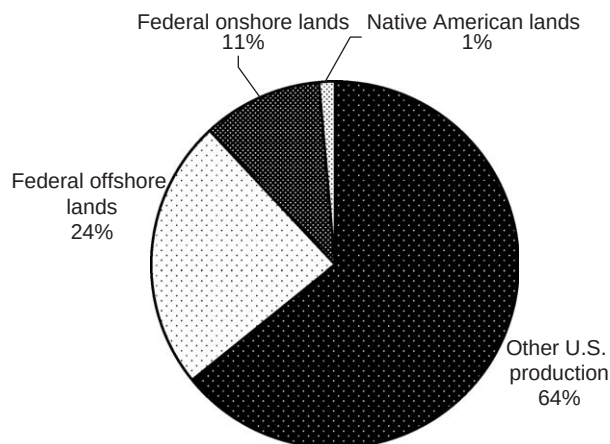


FIGURE 2 Natural gas production in the United States in 2000.

barrels of oil or approximately 32.4% of the total production in the United States in 2000 (Fig. 1). Similarly, federal leases produced 7.1 billion MCF (thousand cubic feet) of gas or approximately 35.5% of the United States total production of gas in 2000 (Fig. 2). To put these production totals into perspective, the population of the United States in 2000 was 281.1 million people. Thus, in the year 2000, federal leases produced approximately 2.5 barrels of oil and 25 MCF of gas for every person in the United States.

The cumulative production from federal oil and gas leases is enormous. From 1920 to 2000, the production of crude oil from federal leases totaled 23.7 billion barrels. The total value of this production was \$273.2 billion. Natural gas production from federal leases totaled 199.1 billion MCF with a total value of \$305.5 billion.

2. THE LEGAL FRAMEWORK FOR FEDERAL OIL AND GAS LEASING

The push to develop federal lands for oil and gas production followed shortly on the heels of World War I. The United States found itself in an energy shortage in 1919 that would extend into the 1920s. In response to this shortage, Congress passed the Mineral Leasing Act (Table II). President Woodrow Wilson signed this landmark legislation into law in 1920. Although the legislation has since been amended many times, many of the provisions of the original law are still in effect today. The Mineral Leasing Act of 1920 authorized the Department of the Interior to lease certain federal onshore lands for the production of oil and gas and other hard minerals. In addition, the law established many of the terms, conditions, and procedures used to issue federal oil and gas leases. The Mineral Leasing Act

(for Acquired Lands) of 1947 allowed for the leasing of additional federal lands including national forests and national grasslands.

Offshore oil and gas development began in 1896 with the drilling of the first offshore well, which was drilled in the tidelands adjacent to Summerland, California (near Santa Barbara). The possibility of mining offshore lands sparked a decades-long debate as to who controlled these lands, with both the federal government and the governments of the adjacent states claiming ownership. The issue was finally resolved with the passage of the Submerged Lands Act of 1953. This statute assigns the rights to offshore lands from the coastline and seaward for 3 nautical miles (1 nautical mile is approximately 1.15 miles) to the adjacent state except for Texas and the west coast of Florida. These two states hold the rights to the submerged lands from the coastline and seaward for 9 nautical miles. The federal government

TABLE II

Important Legislation Affecting Federal Oil and Gas Leasing

Mineral Leasing Act	Originally passed in 1920 and later amended, this law authorizes the leasing of federal lands for the development of crude oil, natural gas, coal, sulfur, phosphate, potassium, and sodium
Mineral Leasing Act (for Acquired Lands)	Originally passed in 1947 and later amended, this law authorizes the leasing of federal-acquired lands; acquired lands include national forests, national grasslands, etc.
Submerged Lands Act	Passed in 1953, this law grants jurisdiction to the submerged lands from the shoreline seaward for 3 nautical miles (3.45 miles) to the coastal states, except for the coasts of Texas and Florida, where jurisdiction extends 9 nautical miles (10.35 miles)
Outer Continental Shelf Lands Act	Originally passed in 1953 and later amended, this law authorizes the Department of the Interior to lease submerged lands under federal jurisdiction for mineral development
Clean Air Act	Passed in 1970 and later amended, this law regulates the emission of air pollutants from industrial activities
Marine Mammal Protection Act	Passed in 1972, this law provides legal protection for marine mammals and their habitats
Endangered Species Act	Passed in 1973, this law provides legal protection for endangered species and their habitats
Clean Water Act	Passed in 1977, this law regulates the discharge of pollutants into the surface waters of the United States
Federal Oil and Gas Royalty Management Act	Passed in 1982, this law ensures that royalties from federal leases are properly collected and accounted for
National Fishing Enhancement Act	Passed in 1984, this law allows offshore platforms to be used as artificial reefs
Oil Pollution Act	Passed in 1990, this law prohibits the discharge of oil into federal waters and ensures that owners and operators have sufficient resources for oil spill cleanup should such discharges occur
Energy Policy Act	Passed in 1992, this law sets the primary lease term for onshore oil and gas leases at 10 years
Deepwater Royalty Relief Act	Passed in 1995, this law allows the Department of the Interior to grant royalty relief and mandates royalty relief under certain conditions for Gulf of Mexico leases in water depth of 658 ft (200 m) or greater.
Federal Oil and Gas Royalty Simplification and Fairness Act	Passed in 1996, this law improves the management of royalties from federal leases.

Source. Reprinted from the U.S. Department of Interior, Minerals Management Service, U.S. Offshore Milestones and Bureau of Land Management Web site "Minerals and Mining Laws and Regulations," available at www.ut.blm.gov/infocenter/info&gleasing.html.

holds the rights to the offshore lands extending beyond for a distance to 200 miles offshore. These submerged lands are known as the “outer continental shelf.” As a practical matter, the suitability of these lands for oil and gas production is limited by drilling technology. The record for deepwater drilling is a well drilled in 9687 ft of water in the Gulf of Mexico in 2001. Continued advances in drilling technology will allow for exploration at greater and greater water depths.

The Outer Continental Shelf Lands Act of 1953 authorized the Department of the Interior to lease outer continental shelf mineral rights to the private sector. This law, as amended, specifies the terms, conditions, and procedures for federal offshore leasing.

Federal leasing progressed rapidly following the enactment of the Outer Continental Shelf Lands Act, first in the Gulf of Mexico and then offshore from California. In 1968, the federal government issued leases in the Santa Barbara Channel. On January 28, 1969, the fifth well drilled from a platform in the Santa Barbara Channel blew out and spilled crude oil for the next 11 days. Media coverage of the resulting oil spill outraged the nation. It is widely believed that this outrage was one of several factors that led Congress to enact a series of environmental laws, many of which affect onshore and offshore oil and gas operations. These laws include the Clean Air Act of 1970, the Marine Mammal Protection Act of 1972, the Endangered Species Act of 1973, the Clean Water Act of 1977, and later, the Oil Pollution Act of 1990.

There has long been concern that the federal government has not received all of the revenue that it is due from its leasing activities (see Section 4.2.3). The Federal Oil and Gas Royalty Management Act of 1982 and the Federal Oil and Gas Royalty Simplification and Fairness Act were enacted to ensure that the government accurately accounted for and collected all payments from lessees for federal oil and gas leases.

3. FEDERAL OIL AND GAS LEASING PROCEDURES

The terms of federal oil and gas leases depend on factors such as whether the lease is competitive or noncompetitive and whether the lease is for onshore or offshore lands. Nevertheless, each federal lease contains certain common elements including the term of the lease, the method for granting the lease, and the types of payments made by lessees. These elements are described below. This description is followed by discussion of how these elements are

put together to form the various types of federal oil and gas leases.

3.1 The Primary Lease Term

When the federal government issues an oil and gas lease, it conveys to the lessee the exclusive right to explore and develop the leased land for a predetermined period of time, usually 5 to 10 years. This time period is called the “primary lease term.” If a lease begins production during the primary lease term, the lease term is extended for as long as production continues. It should be noted that during the primary lease term, the lessee is typically not obligated to begin exploration and may at any time surrender the lease to the government.

3.2 Methods for Granting Federal Oil and Gas Leases

When more than one party expresses an interest in a lease tract, how does the federal government determine which party will be awarded the tract? In these circumstances, the rights to the lease are auctioned to the highest qualified bidder. Such is the case with all onshore leases where there is a high likelihood that oil and gas will be discovered in commercial quantities and with all offshore leases. For the vast majority of leases, an upfront, cash payment called the bonus payment is the bid variable. Auctions for onshore leases are conducted using oral bidding. Offshore leases are auctioned using sealed bidding.

3.3 Types of Payment for Federal Oil and Gas Leases

There is a series of payments made by lessees to the federal government for the exclusive right to explore and develop a lease tract. These payments include bonus, rental, and royalty payments.

3.3.1 Bonus Payments

A bonus payment is a one-time, nonrefundable, cash payment made by a lessee to the government at the time a lease is issued. It is typically the bid variable at competitive auctions; that is, bidders make their offers in the form of the bonus payment that they will make if they are the successful bidder. Bonus payments are nonrefundable. In the event that a tract is dry or does not contain oil or gas in commercial quantities and the lease is surrendered

to the government, the bonus payment is not refunded. The Department of the Interior accepted bonus payments of \$52.2 million for onshore leases during fiscal year (FY) 2000. There were 2311 leases issued, with an average bonus payment of \$22,626. Offshore leasing activity was limited to the Gulf of Mexico in FY 2000. There were two lease sales that included 553 lease tracts and 2.9 million acres of submerged lands. The Department of the Interior accepted \$441.8 million in bonus payments for these tracts. The average bonus payment was \$798,912 per tract. Total onshore and offshore bonus payments exceeded \$494 million in 2000.

3.3.2 Rental Payments

Lessees make annual rental payments to the federal government to retain their leases. The amount of the annual rental payment is determined by federal law and is calculated on a per acre basis ranging from as little as \$0.25 per acre to as much as \$5 per acre. For offshore leases, rental payments cease when production begins. For onshore leases, rental payments continue until the leases are surrendered. Rental payments for onshore oil and gas leases added \$44.5 million to government coffers in FY 2000. Rental payments for federal offshore leases totaled \$207.8 million. Total onshore and offshore rental payments exceeded \$250 million in 2000.

3.3.3 Royalty Payments

In addition to bonus and rental payments, lessees also typically make payments that vary directly with production. The most widely used payment of this type is the royalty.

The term “royalty” harkens back to an era of powerful monarchies. It is, quite literally, “a payment to the crown” and it was a payment to the monarch for using his or her land. In modern usage, royalty payments are payments by lessees to the government calculated as a percentage of gross output. These payments are generally made in monetary terms, but a royalty may be paid in kind. Royalty rates are set by federal law and are typically one-eighth (12.5%) to one-sixth (18.67%) of production. Royalty payments from federal oil and gas onshore oil and gas leases totaled \$967.8 million in calendar year 2000. Offshore oil and gas lease royalties totaled \$4.1 billion over the same period. The combined total onshore and offshore oil and gas royalties exceeded \$5 billion in calendar year 2000. These payments exceeded the receipts from bonus and rental payments by a ratio of 6 to 1.

A minimum royalty payment is sometimes included in a lease contract. As noted earlier, once production begins, lessees make royalty payments to the federal government. A minimum royalty is closely related to a rental payment in that it is determined on a per acre basis. When a minimum royalty payment is included in the lease contract, the lessee must pay the actual royalty payment based on gross production or the minimum royalty payment calculated on a per acre basis, whichever is greater.

A sliding scale royalty is included in some lease contracts. With a sliding scale royalty, the royalty rate varies directly with the level of production; that is, as the production from a lease increases, the royalty rate increases and vice versa. In the case of leases with low production, this royalty relief allows the leases to continue to produce longer than they would have been able to had a higher fixed royalty remained in effect. Similarly, the Deepwater Royalty Relief Act of 1995 provides royalty relief for deepwater (and, therefore, high exploration and development cost) leases.

3.3.4 Typical Federal Lease Procedures and Terms

The federal government uses different combinations of the terms and procedures described above when leasing federal lands for oil and gas production. The leasing terms and procedures for onshore public domain lands and offshore lands are summarized in Table III and are described below.

3.3.4.1 Onshore Public Domain Leases The leasing of onshore public domain lands begins when private firms and individuals express an interest in leasing the lands. The private sector “nominates” lands in which it has an interest and then the Department of the Interior determines whether to issue leases for the nominated lands. In making this decision, the Department of the Interior evaluates nominated lands in terms of its multiple-use objectives for these federal lands. Should the government decide to lease the lands, it advertises the lease sale. An exact description of the lands to be leased and the terms of the lease are included in the advertisement. For federal onshore leases, an oral auction is conducted to determine who shall win the rights to the mineral rights to the lands—the customary bid variable is a cash bonus bid. Only citizens of the United States or corporations organized under the laws of the United States may bid for federal oil and gas leases. Foreign citizens may hold lease interests only as stockholders in U.S. corporations.

TABLE III

Summary of Federal Oil and Gas Lease Procedures and Terms

Oil and gas leases on offshore lands	
Method of award	Competitive sealed bidding
Customary bid variable	Bonus payment
Customary royalty rate	Flat rates of 12.5% or 16.67% in amount or value of production
Rent and minimum royalty	Rent \$3 to \$5 per acre Minimum royalty \$3 to \$5 per acre postdiscovery Rent \$10 per acre for drainage leases Minimum royalty of \$10 per acre for drainage leases
Lease term	Primary lease term 5 to 10 years (8 to 10 years in water depths of 400 m or more); continued if producing in commercial quantities
Size of lease	2500 to 5760 acres unless a larger area is needed for a production unit
Oil and gas leases on onshore lands: Competitive leases	
Method of award	Competitive oral bidding
Customary bid variable	Bonus payment
Customary royalty rate	Leases issued from 5/3/45 to 12/22/87: oil 12.5 to 25% in amount or value of production, depending on production per well per day for the calendar month Leases issued after 12/22/87: flat rate of 12.5% in amount or value of production
Rent	Leases issued before 9/2/60: \$0.25 to \$1 per acre Leases issued from 9/2/60 to 12/22/87: \$2 per acre Leases issued after 12/22/87: \$1.50 per acre for the first 5 years, \$2 per acre for subsequent years
Lease term	Primary lease term 10 years; continued if producing in commercial quantities
Size of lease	Leases issued before 12/22/87: 640 acres or less Leases issued after 12/22/87: maximum 2560 acres for lower 49 states and maximum 5760 acres in Alaska
Oil and gas leases on onshore lands: Noncompetitive leases	
Method of award	Noncompetitive, by application
Customary bid variable	Not applicable
Customary royalty rate	Flat rate of 12.5% in amount or value of production
Rent	Leases issued before 9/2/60: \$0.25 to \$1 per acre Leases issued from 9/2/60 to 2/1/77: \$0.50 per acre Leases issued from 2/1/77 to 12/22/87: \$1 to \$2 per acre for the first 5 years, \$2 per acre for subsequent years Leases issued after 12/22/87: \$1.50 per acre for the first 5 years, \$2 per acre for subsequent years
Lease term	Primary lease term 10 years; continued if producing in commercial quantities
Size of lease	640 acres minimum, 10,240 acres maximum within 6-mile square

Successful lessees must make annual rental payments of \$1.50 per acre for the first 5 years of the lease and \$2.00 per acre thereafter. A fixed royalty rate of one-eighth of all production (12.5%) is also required. The primary lease term for onshore leases is 10 years as required by the Energy Policy Act of 1992 and may be extended for as long as production continues. It is interesting to note that bidders are not allowed to conduct exploratory drilling on the lands prior to the auction. Bidders must instead rely on interpretations of seismic surveys and other geophy-

sical tools to evaluate the oil and gas potential of a lease tract. A lease issued in this fashion is a competitive lease.

In some instances, the prospects for oil or gas discovery may be low and the lease sale will attract no bidders. In this case, the lease is issued on a first-come, first-served basis and the lease is deemed "noncompetitive." The terms for a noncompetitive onshore lease are the same as those of a competitive onshore lease except that there is no cash bonus payment. The Federal Onshore Oil and Gas Leasing

Reform Act of 1987 requires that before lands can be leased on a noncompetitive basis, they must first be offered on a competitive basis.

3.3.4.2 Federal Offshore Leases The terms and procedures used to issue offshore leases are similar to those of competitive onshore leases, but with some important differences. As with onshore leasing, the offshore leasing process begins with nominations from the private sector. The customary bid variable is the cash bonus; however, unlike onshore leases, offshore leases are issued using a sealed bid auction. The rental rate for offshore leases is \$3 to \$5 per acre unless the lease is designated as a drainage lease, in which case the annual rental rate is \$10 per acre. A drainage lease is a lease issued for lands containing a known oil field. Successful lessees in this case will “drain” the known oil field. With offshore leases, rental payments cease once production begins. There is a minimum royalty payment that is calculated using the same formula as the rental payment. A fixed royalty rate of one-eighth (12.5%) or one-sixth (16.67%) is required from production. The primary lease term is 5 to 10 years in shallow waters and 8 to 10 years in waters exceeding 400 m in depth. Except for the largest oil companies, firms may bid jointly on offshore leases and leases are transferable to other qualified firms. The Outer Continental Shelf Lands Act requires the issuance of all offshore leases on a competitive basis.

4. THE EFFECTIVENESS OF FEDERAL OIL AND GAS LEASING PROCEDURES

Economic theory suggests that the effectiveness of federal oil and gas leasing procedures should be judged using the following criterion: Federal oil and gas leasing procedures should maximize the value of the public lands and collect this value for the American people.

This criterion has several important implications for economic behavior. First, to maximize the value of the land, only those oil and gas exploration and production activities will be undertaken that are such that the additional economic benefits of the oil and gas produced (or marginal benefits) exceed the additional economic costs needed in their production (or marginal costs). Next, these activities will be optimally timed. There are many alternative time

paths for oil and gas exploration and production for any given lease. Each alternative time path should be evaluated in terms of its net present value using an appropriate social discount rate and the alternative that yields the highest net present value should be undertaken. By foregoing those activities whose marginal costs are not justified by their marginal benefits, and by the optimal timing of these activities, the value of the land is maximized. Assuming that all of the benefits and costs of oil and gas exploration, development, and production are borne by competitive lessees, and assuming that social and private discount rates closely approximate one another, the industry behavior will maximize the value of the lands to society. If a portion of the benefits or costs is external to the firm, or if private and social discount rates diverge, then industry behavior will maximize profit, but may maximize the value of the lands to society. The case of external costs is discussed in Section 5. Finally, the leasing procedures should collect any surplus value generated by the leases for the American people. This surplus value, also known as economic rent, is the difference between the total revenue generated by the lease and the total necessary costs needed to generate this revenue. Included in the total cost is a normal rate of return on investment for the lessee. In economic theory, a normal rate of return on investment is properly thought of as a cost because such a return could have been earned in an alternative investment (it is an opportunity cost). Any administrative and compliance costs associated with the leasing procedures reduce the value of the leases and, therefore, reduce the economic rent. Thus, an ideal leasing procedure holds administrative and compliance costs to a minimum. In summary, leasing procedures should maximize the economic rent from public lands and collect this rent for the American people.

4.1 The Effectiveness of the Primary Lease Term

Federal oil and gas leases are issued with a primary lease term to encourage lessees to conduct exploration activities expeditiously. To the extent that the primary lease term alters the timing of exploration, it reduces the economic rent that can be derived from a lease. To maximize economic rent, the primary lease term should be lengthened so that it does not alter lessee behavior or should be eliminated altogether.

4.2 The Effectiveness of the Different Types of Payment for Federal Oil and Gas Leases

There are four types of payments made by lessees to the government for the mineral rights on public lands: bonus, rental, royalty, and profit-share payments. Although the lessees view these payments as costs of production, in a social sense they are not economic costs like expenditures for labor and equipment to drill an exploratory well. They are transfer payments that redistribute income from the private to the public sector. As such, it would be ideal if these payments did not alter lease operating decisions.

4.2.1 The Effectiveness of Bonus Payments

Such is the case with bonus payments. A bonus payment is a one-time payment made at the time a lease is issued. Hence, it does not affect future marginal exploration and development decisions. The exploration, development, and production decisions that maximize the value of the lease to the lessee also maximize the value of the lease to society.

4.2.2 The Effectiveness of Rental Payments

Rental payments affect marginal decisions and alter the productivity and life of a lease. It is interesting to note that the effects of rental payments differ for onshore versus offshore leases. For onshore leases, rental payments continue for the life of the lease. They are viewed as a marginal cost of operating a lease for one more period and can be avoided by surrendering a lease. As such, they may lead to the abandonment of a lease that would be profitable in the absence of the rental payment. In certain circumstances, a lease may be abandoned as “dry” when in fact it would be developed in the absence of a rental payment. The value of the output lost to premature abandonment or nondevelopment of a lease is a loss to society and wastes a valuable natural resource.

For offshore leases, rental payments cease when production begins. As such, a rental payment can be avoided by accelerating exploration. In both onshore and offshore leases, rental payments alter the behavior of lessees and reduce the economic rent that can be derived from the federal lands.

Administrative and compliance costs associated with rental payments are not as small as those associated with bonus payments, but still are not high. This is because the amount of the annual rental payment is not recalculated each year and does not vary with production.

4.2.3 The Effectiveness of Royalty Payments

Like rental payment, royalty payments are perceived as additions to the marginal cost of production. As such, they also can lead to the premature abandonment or nondevelopment of a lease that should otherwise be developed, reducing the economic rent from the lease. Reducing the royalty rate as production decreases can mitigate the adverse effect on total production. The Outer Continental Shelf Lands Act (as amended) and the Deepwater Royalty Relief Act include royalty rate reductions for offshore leases.

Royalty payments also alter the timing of exploration and production activities by lessees. With other factors held constant, businesses in general and lessees in particular will realize revenues as early as possible and delay costs for as long as possible. Because lessees perceive royalty payments as marginal costs, they delay production.

Unlike bonus and rental payments, royalty payments are contingent on production. The administrative and compliance costs associated with bonus payments are negligible and those associated with rental payments are small. Such is not the case with royalty payments because they depend on actual production. For the government to ensure that it is receiving the full amount of the royalty payments it is due, it must monitor actual production. Concern over the accuracy of royalty payments led to the passage of the Federal Oil and Gas Royalty Management Act of 1982 and the Federal Oil and Gas Royalty Simplification and Fairness Act of 1996.

4.3 The Effectiveness of Different Methods for Granting Federal Oil and Gas Leases

The bonus payment is the bid variable for federal offshore leases and the auction is conducted using sealed bidding. This combination is highly effective in maximizing and collecting economic rent for leased lands. As noted earlier, bonus payments do not affect decisions at the margin and, therefore, do not alter the time path for exploration and production, thereby maximizing the economic rent. With a sealed bid auction, bidders submit their bids in a sealed envelope. On the day of the auction, the bids are opened and the lease is awarded to the highest qualified bidder. The bidders need not be present to win and may not bid again if their initial bids are not high enough to win the auction. At the time of the auction, bidders are unaware of the number of competing bids. Because of this, they prepare bids that approximate the economic rent. Indeed, concerns have been raised that this

method of leasing is too successful and that winning bids will systematically come from a bidders who are too optimistic about the prospects for the lease. These “lucky” bidders have bid too much and will find that they are operating their leases at a loss. This phenomenon is called the “winner’s curse.” Bidders are aware of this phenomenon and can be expected to adjust their bids accordingly.

Oral bid auctions are used to issue onshore leases. Oral bidding is also effective in collecting economic rent if there are large numbers of bidders and if the bidders do not collude with one another. With an oral bid auction, the winning bid is just above the valuation of the second high bidder. That is, the winning bidder must outbid the nearest competitor, but may not need to bid the full valuation of the lease.

Experiments have been conducted with alternate bid variables including the royalty rate and profit-share rate. These bid variables neither maximize economic rent nor collect it and should be avoided.

5. STATE AND LOCAL GOVERNMENT AND ENVIRONMENTAL CONCERNS REGARDING FEDERAL OIL AND GAS LEASING

The production of oil and gas from federal lands yields large net benefits to the nation through the

revenue it generates for government coffers and through the oil and gas it provides for the American economy. Despite these benefits, federal oil and gas leasing does not enjoy universal support. The most vocal critics of the federal leasing program are state and local governments and environmental groups. Their opposition stems from the fact that not all of the costs of federal oil and gas leasing operations are internalized by the lessees.

Lessees bear the lease acquisition costs and the costs of the labor and capital needed for lease exploration and development, but may not bear the cost of local public services including providing police and fire protection, building and maintaining the transportation infrastructure, and providing public education. Whereas the benefits of a federal lease projects are spread across the entire nation, these external costs are borne disproportionately by those citizens living near the projects.

The opposition from environmental groups to federal oil and gas leasing is based on two interrelated concerns. First, lessees do not bear the external cost resulting from the air, water, and noise pollution generated by lease operations. To the extent that the lessees do not internalize these costs, environmental concerns are well founded. Second, public lands often have multiple conflicting uses and it may be difficult to determine which use yields the highest value to the nation. For example, certain public lands may be used to produce oil, gas, or

TABLE IV

Selected Formulas for Disbursements of Revenues from Federal Oil and Gas Leases

Oil and gas leases on offshore lands	
States	27% of royalty, rent, and bonus revenues to the affected state
Special Purpose Accounts	\$150 million annually to the Historic Preservation Fund
U.S. Treasury	The remainder of the revenue to the General Fund of the U.S. Treasury
Public domain lands	
States	50% to the state in which the lease is located except Alaska 90% to the state of Alaska if the lease is located in Alaska 40% to the Reclamation Fund except if the lease is located in Alaska in which case no revenues are directed to the Reclamation Fund
Special Purpose Funds	
U.S. Treasury	10% to the General Fund of the U.S. Treasury
National forests	
States	25% to the state(s) in which the forest is located
Special Purpose Funds	None
U.S. Treasury	75% to the General Fund of the U.S. Treasury

Source. Reprinted from U.S. Department of the Interior, Minerals Management Service (2001). “Mineral Revenues 2000,” Table 30, pp. 97–98.

wood products, for recreation, or as wildlife preserves. The value of the lands in oil and gas production or timber production is relatively easy to determine because of the markets for these end products. The value of the lands in recreational uses or as wildlife preserves is more difficult to determine. In fact, there is no general agreement that the value of these and other ecosystem services can be measured in monetary terms. Techniques have been developed to estimate the value of public lands in these circumstances, but these techniques are in their infancy and the estimates they generate are subject to debate. Indeed, disagreements between the oil and gas industry and environmentalists on the appropriate use of public lands can be traced to disagreements as to the value of the lands in alternate uses.

Local governments and environmental groups remain active in their opposition to federal oil and gas leasing. They have enjoyed some successes in limiting federal oil and gas leasing, as evidenced by the fact that very few offshore leases have been issued in the past decade and the repurchase by the federal government of leases offshore from Florida in 2002.

A possible solution to the concerns of state and local governments regarding the provision of public services may be found in the way that the federal government disburses oil and gas lease revenues. A summary of revenue disbursement formulas is found in Table IV. These formulas are set by federal law. State and local opposition to federal oil and gas leasing will lessen if a greater percentage of lease revenues are directed to the affected geographic area.

Environmental concerns may be taken into account by using environmental laws to internalize the external costs of federal oil and gas leasing. Federal leasing should take place only when the benefits of the leasing exceed all economic costs (including environmental costs). Unfortunately, some environmentalists and environmental groups would like to halt federal oil and gas leasing regardless of the benefit to the nation. Such a prohibition on leasing reduces the welfare of the nation.

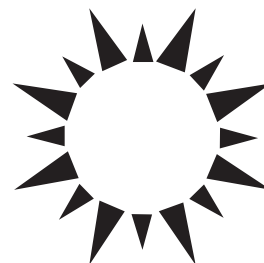
SEE ALSO THE FOLLOWING ARTICLES

Markets for Natural Gas • Markets for Petroleum • Natural Gas, History of • Oil and Natural Gas Drilling • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Exploration • Oil Industry, History of • Oil Price Volatility • Oil Recovery • Oil Refining and Products • Petroleum Property Valuation • Strategic Petroleum Reserves

Further Reading

- McDonald, S. L. (1979). "Leasing of Federal Lands for Fossil Fuel Production." Published for Resources for the Future by the Johns Hopkins University Press, Baltimore, MD.
- Mead, W. J., *et al.* (1985). "Offshore Lands: Oil and Gas Leasing and Conservation on the Outer Continental Shelf." Pacific Institute for Public Policy Research, San Francisco, CA.
- Minerals Management Service, U.S. Department of the Interior (2001). "Mineral Revenues 2000." Herndon, VA.
- Minerals Management Service, U.S. Department of the Interior (2001). "Outer Continental Shelf Oil and Gas Leasing Procedures," Outer Continental Shelf Report MMS 2001-076. New Orleans, LA.

Oil and Natural Gas Liquids: Global Magnitude and Distribution



THOMAS S. AHLBRANDT

U.S. Geological Survey

Denver, Colorado, United States

-
1. Introduction
 2. Worldwide Distribution
 3. Impact of Distribution
 4. Data Sources

Glossary

continuous accumulations Petroleum that occurs in extensive reservoirs and is not necessarily related to conventional structural or stratigraphic traps. These accumulations of oil and/or gas lack well-defined down-dip petroleum/water contacts and thus are not localized by the buoyancy of oil or natural gas in water.

conventional accumulations Petroleum that occurs in structural or stratigraphic traps, commonly bounded by a down-dip water contact, and therefore affected by the buoyancy of petroleum in water.

crude oil A mixture of hydrocarbons that exists in a liquid phase in natural underground reservoirs and remains liquid at atmospheric pressure after passing through surface separation facilities. Crude oil may also contain some nonhydrocarbon components; referred to as oil in this article.

cumulative production Volumes of oil and natural gas liquids that have been produced.

endowment The sum of cumulative production, remaining reserves, mean undiscovered recoverable volumes, and mean additions to reserves by field growth.

field A contiguous area consisting of a single reservoir or multiple reservoirs of petroleum, all grouped on, or related to, a single geologic structural and/or stratigraphic feature.

future petroleum The sum of the remaining reserves, mean reserve growth, and the mean of the undiscovered volume. Cumulative production does not contribute to the future petroleum. The terms future oil, future liquid volume, or future endowment are sometimes used as

variations of future petroleum to reflect those resources that are yet to be produced.

liquid petroleum A term used to represent crude oil combined with natural gas liquids.

natural gas liquids (NGL) Those portions of reservoir gas that are liquefied at the surface in lease separators, field facilities, or gas processing plants. Natural gas liquids are heavier homologs of methane (including ethane, propane, butane, pentane, natural gasoline, and condensate).

oil field A field producing oil and gas is termed an oil field when the petroleum contained within has a gas/oil ratio (GOR) of less than 20,000 cubic feet per barrel. If the GOR > 20,000 cubic feet per barrel, the field is called a gas field.

remaining reserves Recoverable oil and NGL volumes that were originally present and have not yet been produced.

reserve The estimated quantities of petroleum expected to be commercially recovered from known accumulations relative to a specified date, under prevailing economic conditions, operating practices, and government regulations. Reserves are part of the identified (discovered) resources and include only recoverable materials.

reserve (field) growth The increases of estimated petroleum volume that commonly occur as oil and gas fields are developed and produced.

resource A concentration of naturally occurring solid, liquid, or gaseous hydrocarbons in or on the earth's crust, some of which is currently or potentially economically extractable.

undiscovered resources Resources inferred from geologic information and theory to exist outside of known oil and gas fields.

Although their abundance and potential shortages are debated, under most energy projections, fossil fuels will remain the dominant source of energy for the remainder of this century. Their distribution, quantities, and availability will thus continue to be of

paramount importance for world economies. Policy decisions made in the United States and elsewhere depend on detailed knowledge of world petroleum resources. Therefore, petroleum resources are periodically reassessed, not just because new data become available and better geologic models are developed but also because many nongeologic factors determine which part of the crustal abundance of petroleum will be economic and acceptable over some foreseeable future. Misunderstanding of the reserve and resource terminology and the nature of resources allow those with fixed points of view to argue vehemently over the size of remaining resources.

1. INTRODUCTION

Detailed petroleum information was published in a world assessment of petroleum resources conducted by the U.S. Geological Survey (USGS) in its World Petroleum Assessment 2000 (WPA). This assessment provides estimates of the quantities of conventional technically recoverable oil, gas, and natural gas liquids (NGL) outside the United States that have the potential to be added to reserves. Many individuals and groups, such as the International Energy Agency and the Energy Information Administration of the Department of Energy, use this report as the baseline for world petroleum resources. This assessment is also the basis for estimates of reserve growth and undiscovered resources for this article. Reserves and cumulative production data are derived from the IHS Energy database.

1.1 Resource Classification

This article classifies liquid petroleum resources into four categories: cumulative production—those volumes of oil and NGL that have already been produced; reserves—volumes of liquids remaining in already discovered conventional oil fields that are known with reasonable certainty; reserve growth (also known as field growth)—those volumes in already discovered conventional oil fields that are not known with certainty but can potentially be added to reserves in the future; and undiscovered resources—volumes of oil and NGL that have not yet been discovered. These volumes are estimates and subject to change through time in light of new geologic and engineering data as well as changing economic conditions. Additionally, as oil or NGL are discovered or produced, volumes from one category may shift to another category (e.g., from reserves to

cumulative production). Many terms are used to identify categories and subcategories of reserves and resources; such terms are not always used consistently and must be considered cautiously. In this article, only the four categories defined are used and their sum is the endowment.

Oil resources in continuous accumulations, such as heavy oils, bitumens, tar sands, and oil shales, are generally classified and evaluated separately. Such continuous resources were identified, but not formally assessed, in USGS WPA 2000; they are not included in this article. This article considers only conventional liquid petroleum.

1.2 Uncertainty of Estimates

Estimates of oil and NGL are subject to considerable uncertainty. This is true for all categories, including estimates of past production. The uncertainty of these volumes leads to probabilistic approaches to estimating reserves, reserve growth, and undiscovered resources. The World Petroleum Congress/Society of Petroleum Engineers/American Association of Petroleum Geologists issued a reserve and resource classification system that recognized this uncertainty in all categories, including reserve estimates. Their classification uses the terms proved, probable, and possible: A 90% certainty is assigned to proved reserves, a 50% certainty to probable reserves, and a 10% certainty to possible reserves. In this article, the means of the probabilistic estimates are used.

Recent studies demonstrate that reserve growth is a worldwide phenomenon and must be considered in any analysis of world petroleum resources. As with other estimates, for reserve growth, the mean estimates are used here.

2. WORLDWIDE DISTRIBUTION

Remaining reserves and cumulative production data as of 1/1996 were obtained from the IHS Energy database. Two major components of potential petroleum resources for the world were assessed in USGS WPA 2000: undiscovered conventional resources and reserve (or field) growth. The assessed areas include 128 of the 937 geologic provinces defined by the USGS WPA 2000, and the U.S. provinces assessed by the USGS in 1995 and the MMS in 1996. This study includes more than 95% of the conventional petroleum resources of the world (Fig. 1).

Potential additions to reserves from reserve growth were estimated at the world level only, but for this

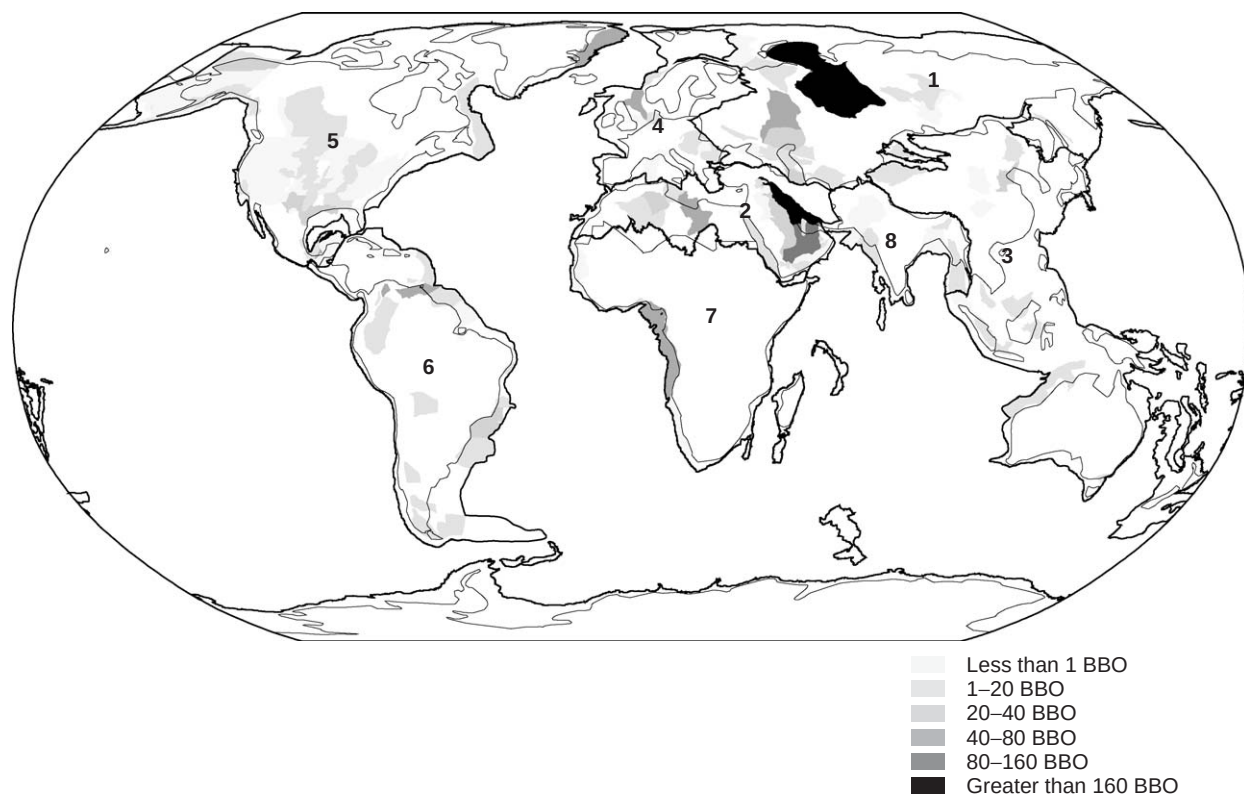


FIGURE 1 Conventional oil resources of the world [billions of barrels (BBO)] as assessed by the USGS World Petroleum Assessment (2000). Darker shading indicates greater resource volume. Data include cumulative production and remaining reserves from IHS Energy as of January 1996 and the mean estimate of undiscovered conventional oil resources of the world as assessed by the USGS World Petroleum Assessment Team in 128 provinces of the world outside the United States, including the U.S. province assessments as assessed by the USGS in 1995 and the Minerals Management Service in 1996. USGS regions: 1, Former Soviet Union; 2, Middle East and North Africa; 3, Asia Pacific; 4, Europe; 5, North America; 6, Central and South America; 7, sub-Saharan Africa and Antarctica; 8, South Asia.

article they have been allocated to the regional level (Fig. 2; Tables I and II). Reserve growth for the world was first estimated by the USGS in WPA 2000, but it has also been estimated by British Petroleum, the International Energy Agency, and individual researchers. A hierarchical assessment system was used in the USGS WPA 2000, with the foundation being the assessment unit, a subdivision of a total petroleum system, which was the formally assessed entity. These basic estimates were then allocated to the appropriate total petroleum systems, provinces, region, and finally aggregated to the world level.

Approximately 3 trillion barrels of oil (TBO) is estimated to be the world's petroleum endowment of recoverable oil and NGL; this endowment includes cumulative production, remaining reserves, reserve growth, and undiscovered resources (mean estimate). This volume includes the U.S. estimates made by the USGS in 1995 for both the onshore and state waters and by the Minerals Management Service (MMS) in

1996 for the federal offshore U.S. World oil reserves are currently 1.1 TBO but were 0.96 TBO as of January 1996; world consumption is approximately 0.028 TBO per year according to the IHS database. Half of the world's undiscovered liquid petroleum potential is offshore. Arctic basins are the next frontier, with approximately 25% of undiscovered petroleum resources (Fig. 1).

The estimate of oil endowment of 3 TBO is comparable to 1998 a estimate by British Petroleum of 3.3 TBO at the mean and the 2002 estimate by J. S. Edwards of 3.7 TBO at the median. The USGS mean reserve growth estimate for the world is 667 billion barrels of oil (BBO). For the world's oil fields in 1998, British Petroleum estimated reserve growth to be 500 BBO, and the International Energy Agency estimated it to be between 418 and 1136 BBO.

This section discusses the worldwide distribution of conventional oil and NGL resources. Much of the data for this section are from the USGS

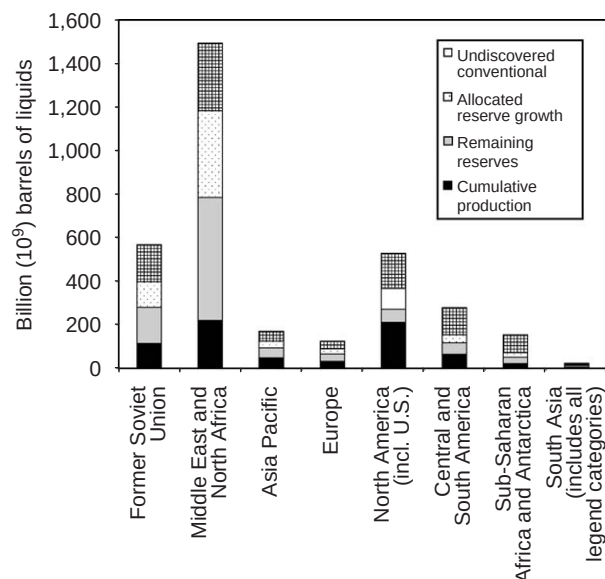


FIGURE 2 Conventional liquid petroleum endowment (cumulative production, remaining reserves, mean estimate of reserve growth, and mean estimate of undiscovered resources) for the eight regions of the world as defined by the USGS World Petroleum Assessment (2000).

WPA assessment, the USGS 1995 U.S. national assessment, the MMS federal offshore assessment, the U.S. Energy Information Administration, the International Energy Agency, and the British Petroleum statistical review.

Table I shows the volumes of conventional oil and NGL for the eight USGS world regions. The geographic distribution of the world's oil endowment for USGS-defined provinces within the eight regions is shown in Fig. 1. Table II combines the regional oil and NGL data into liquid petroleum resources in each region. The regions are discussed in decreasing order of their total resource volume. Because reserve (field) growth was estimated by the USGS in 2000 at the world level only, this regional analysis includes the first discussion of reserve growth by the world's regions, and Tables I and II and Fig. 1 give a complete analysis of the regional oil and/or liquid petroleum endowment.

Information presented in Figs. 1–3 and Tables I–V supports several observations. First, liquid petroleum resources are unevenly distributed throughout the world. Three regions contribute more than 77% of the world's liquid petroleum endowment; the Middle East/North Africa contributes the greatest volume (44.7%), followed by the former Soviet Union (17%) and North America (15.8%). Second, four regions of the world export most of the liquid petroleum

(Middle East/North Africa, the former Soviet Union, Sub-Saharan Africa, and Central and South America in descending order) to the remaining five regions. Third, North America, the oldest producing province, still produces nearly 20% of the world's liquid petroleum. Fourth, Europe is the next to last ranked region, with only 3.8% of the world's liquid petroleum endowment, and it is a substantial net importer of these resources. Finally, reserve growth may significantly contribute to the endowment in three regions: the Middle East/North Africa, the former Soviet Union, and North America (Fig. 2).

2.1 Middle East and North Africa

With 44.7% of the world's liquid petroleum endowment and 58.8% of the remaining liquid resources, this region will continue be the largest source of oil and NGL. Recent discoveries in the Arabian Peninsula, the Gulf of Suez, and the Algerian basins confirm the high liquid petroleum potential of the region. For example, discoveries in the Trias/Ghadames Basin in the past 5 years have more than doubled the volume of all previous discoveries in the province. Multi-billion barrel discoveries have been made in other basins in this region, including the Azadegan field in Iran, the Kra Al Maru field in the northern Arabian Gulf, many satellite fields to Ghawar field in Saudi Arabia, and significant (1 billion barrels or larger) discoveries in the Sirte Basin in Libya. Detailed reserve growth studies demonstrate very large potential for discovered but undeveloped reserves, particularly in Iraq and Iran. With 312 BB of undiscovered liquid petroleum and reserve growth potential of 398 BB (Table II), this region will continue to be preeminent for liquid petroleum resources (Tables III and IV).

2.2 Former Soviet Union

This region has 17% of the world's liquid petroleum endowment. Recent oil discoveries in the Caspian region demonstrate the high liquid petroleum potential of this region—for example, in the North Caspian Basin, the Kashagan field, for which estimates range well above 20 BB, and the Azeri-Chirag-Gunashli complex in the Caspian, for which estimates are at least 5 BB. Frontier provinces in the offshore Arctic of the former Soviet Union (FSU), such as the Laptev Sea, may have significant petroleum potential but are unevaluated. More mature provinces, including the West Siberian,

TABLE I

World Oil Resources and NGL Resources by Region (in Million Barrels)^a

Region ^b	Oil					NGL				
	Cumulative production	Remaining reserves	Allocated reserve growth	Undiscovered resources	Endowment	Cumulative production	Remaining reserves	Allocated reserve growth	Undiscovered resources	Endowment
1. Former Soviet Union	112,106	152,173	104,537	115,985	484,801	1801	14,753	9,071	54,806	80,431
2. Middle East and North Africa	218,644	528,620	363,141	229,882	1,340,287	1413	35,376	21,750	81,747	140,286
3. Asia Pacific	47,722	39,980	27,465	29,780	144,947	1022	4,906	3,016	15,379	24,323
4. Europe	31,235	27,572	18,941	22,292	100,040	649	6,400	3,935	13,667	24,651
5. North America (including U.S.) ^c	208,832	59,143	95,341	153,491	516,807	1980	1,462	899	7,853	12,194
6. Central and South America	64,096	50,272	34,535	105,106	254,009	486	2,016	1,239	20,196	23,937
7. Sub-Saharan Africa and Antarctica	22,437	26,049	17,895	71,512	137,893	10	2,932	1,803	10,766	15,511
8. South Asia	5,131	7,071	4,858	3,580	20,640	24	467	287	2,604	3,382
Total (including U.S.)	710,203	890,880	666,712	731,628	2,999,423	7385	68,312	42,000	207,018	324,715

^a All estimates are mean estimates. Oil and NGL resource totals from U.S. Geological Survey; cumulative oil production and remaining reserves from IHS Energy as of January 1996. Allocated mean reserve growth from U.S. Geological Survey estimates; allocated (for non-U.S.) proportional to mean volumes of remaining reserves.

^b USGS region and code.

^c For U.S. portion of North America, oil and NGL estimates are from USGS (1995) and MMS (1996).

TABLE II

World Petroleum Liquids (Oil and NGL Combined) Resources by Region^a

Region ^b	Million barrels						%					
	Cumulative production	Remaining reserves	Allocated reserve growth	Undiscovered resources	Endowment	Future resources	World cumulative production	World remaining reserves	World reserve growth	World undiscovered resources	World future resources	World's endowment
1. Former Soviet Union	113,907	166,926	117,502	170,791	569,126	455,219	15.87	17.40	16.10	18.20	17.32	17.01
2. Middle East and North Africa	220,057	563,996	398,421	311,629	1,494,103	1,274,046	30.67	58.80	54.58	33.20	48.48	44.66
3. Asia Pacific	48,744	44,886	31,504	45,159	170,293	121,549	6.79	4.68	4.32	4.81	4.63	5.09
4. Europe	31,884	33,972	23,581	35,959	125,396	93,512	4.44	3.54	3.23	3.83	3.56	3.75
5. North America (including U.S.) ^c	210,812	60,605	96,240	161,344	529,001	318,189	29.38	6.32	13.18	17.19	12.11	15.81
6. Central and South America	64,582	52,288	37,061	125,302	279,233	214,651	9.00	5.45	5.08	13.35	8.17	8.35
7. Sub-Saharan Africa and Antarctica	22,447	28,981	20,364	82,278	154,070	131,623	3.13	3.02	2.79	8.77	5.01	4.61
8. South Asia	5,155	7,538	5,326	6,184	24,203	19,048	0.72	0.79	0.73	0.66	0.72	0.72
Total (including U.S.)	717,588	959,192	730,000	938,646	3,345,426	2,627,838	100.00	100.00	100.00	100.00	100.00	100.00

^a All estimates are mean estimates. Oil and NGL resource totals from U.S. Geological Survey; cumulative oil production and remaining reserves from IHS Energy as of January 1996. Allocated reserve growth from U.S. Geological Survey estimates; allocated (for non-U.S.) proportional to volumes of remaining reserves.

^b USGS region and code.

^c For U.S. portion of North America, oil and NGL estimates are from USGS (1995) and MMS (1996).

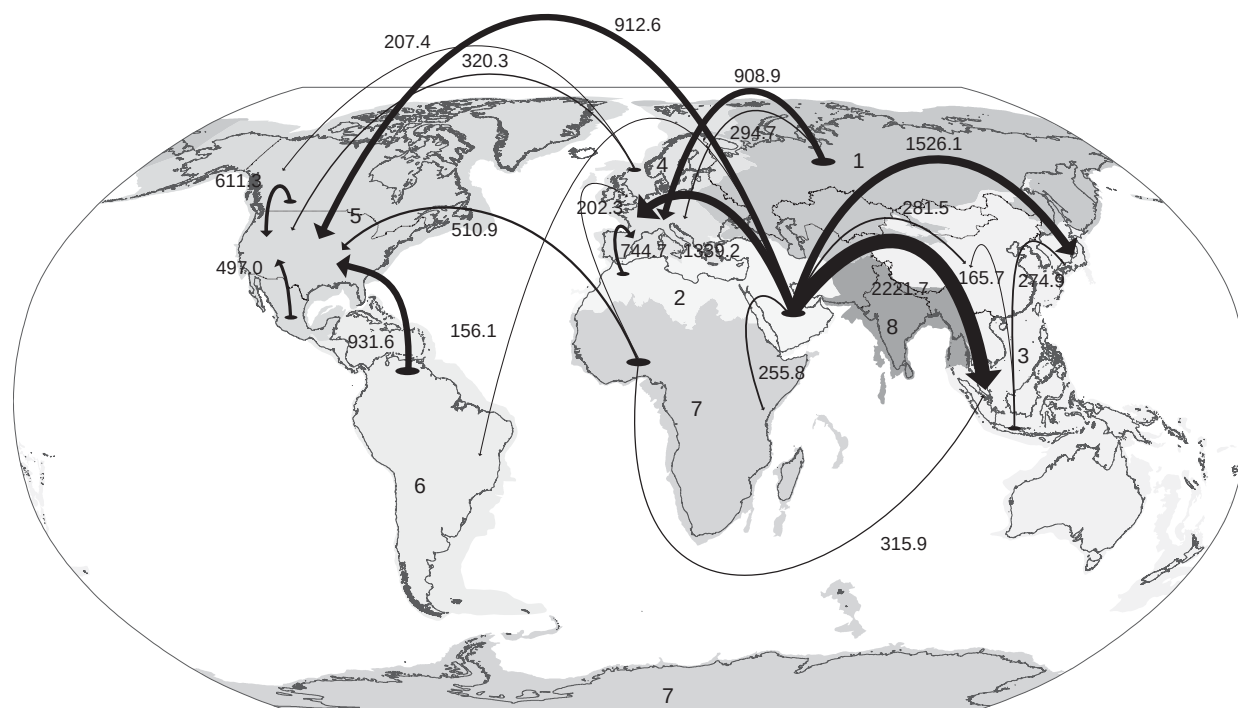


FIGURE 3 Major oil trade movements in million barrels of oil for 2000. Modified from British Petroleum statistical review to unit measurements in millions of barrels assuming 33 API gravity. USGS regions: 1, Former Soviet Union; 2, Middle East and North Africa; 3, Asia Pacific; 4, Europe; 5, North America; 6, Central and South America; 7, sub-Saharan Africa and Antarctica; 8, South Asia.

TABLE III

World Oil Production in 2000^a

Region ^b	Oil production	Oil consumption	% world production	% world consumption	% surplus or deficit
1. Former Soviet Union	2,978	1,340	10.63	4.83	+ 5.80
2. Middle East and North Africa	9,877	2,292	35.24	8.26	+ 26.98
3. Asia Pacific	2,662	6,655	9.50	23.97	-14.47
4. Europe	2,562	5,522	9.14	19.89	-10.75
5. North America (including U.S.)	5,588	8,684	19.94	31.27	-11.33
6. Central and South America	2,547	1,874	9.09	6.75	+ 2.34
7. Sub-Saharan Africa and Antarctica	1,520	469	5.42	1.69	+ 3.73
8. South Asia	296	930	1.06	3.35	-2.29
Total (including U.S.)	28,029	27,767	100.00	100.00	

^a Year 2000 production and consumption from U.S. Energy Information Administration. All volumes in MMBO. NGL data not available.

^b USGS region and code.

Timan-Pechora, and Volga Ural Basins, still have significant oil potential. The East Siberia Basin contains the discovered but undeveloped Yurubchen-Takhoma oil field (~10 BB) and very large heavy oil deposits, neither of which were formally

assessed in 2000 by the USGS. This region will remain a significant source of liquid petroleum resources as demonstrated by the estimated undiscovered resource of 171 BB and reserve growth of 117 BB (Table II).

TABLE IV

Top 20 Producers of Liquid Petroleum (Crude Oil and Natural Gas Liquids), 2000^a

Country	Million barrels per day
Saudi Arabia	9.12
United States	9.06
Russia	6.71
Iran	3.78
Mexico	3.48
Norway	3.32
China	3.25
Venezuela	3.14
Canada	2.76
Iraq	2.59
United Arab Emirates	2.57
United Kingdom	2.55
Kuwait	2.25
Nigeria	2.15
Brazil	1.54
Indonesia	1.51
Libya	1.47
Algeria	1.44
Oman	0.94
Qatar	0.87

^aData from the Energy Information Administration (2001).

2.3 North America

The major oil-producing regions in Canada, Mexico, and the United States are relatively mature areas in terms of exploration and development compared to much of the rest of the world. Despite this maturity, North America still produces 20% of the world's liquid petroleum, second only to the Middle East/North Africa, which provides 35.2%. North America has 15.8% of the world's liquid petroleum endowment (which includes all resource categories for oil and NGL), but it has only 12.1% of its future liquid petroleum potential (which excludes cumulative production). These facts indicate that North America has utilized a significant portion of its liquid petroleum endowment. Since North America is the largest consumer of liquid petroleum with 31.3% of world consumption, this region will remain a major importer of liquid petroleum. However, significant new discoveries are still being made, particularly offshore in the Gulf of Mexico and on the North Slope of Alaska. For example, Thunder Horse, Atlantis, and Green Canyon 640 fields are new discoveries in excess of a half billion barrels in the

TABLE V

Top 20 Consumers of World Petroleum Liquids, 2000^a

Country	Million barrels per day
United States	19.70
Japan	5.53
China	4.78
Germany	2.77
Russia	2.50
Brazil	2.16
South Korea	2.15
Canada	2.07
France	2.02
Mexico	1.99
India	1.99
Italy	1.87
United Kingdom	1.72
Spain	1.46
Saudi Arabia	1.30
Iran	1.08
Indonesia	1.04
Australia	0.86
The Netherlands	0.85
Taiwan	0.85

^aData from the Energy Information Administration (2001).

Gulf of Mexico. Multiple discoveries in excess of several hundred million barrels on the North Slope of Alaska may provide significant additions to reserves. In Canada, the eastern offshore continental shelf remains prospective for petroleum, as evidenced by new discoveries, mostly gas, but the multi-billion barrel Hibernia field indicates the presence of significant liquid petroleum resources. Despite these successes, the North American region imports 10.8% more liquid petroleum than it produces and will probably continue to import. Canadian heavy oil and tar sand deposits in Alberta will be critical to hydrocarbon supply in the region. It is estimated that 10% of North America's oil production will come from Alberta's oil sands by 2005. They are estimated to contain 1.6 trillion barrels of oil, of which 311 billion barrels are recoverable. These sands are considered to be a continuous deposit and as such were not assessed by the USGS in 2000, and they are not included in the figures or tables in this article.

Northeast Greenland was included in this region and as such added considerably to the potential of North America, although the potential in Greenland is hypothetical and there is an accessibility risk of developing resources due to its hostile environment.

Numerous geologic structures that occur offshore are thought to be on the western side of the North Sea producing area. These structures may contain oil, as evidenced by oil accumulations exhumed onshore in Greenland. The same geologic settings are prolific liquid petroleum provinces in the North Sea.

With undiscovered resources of 161 BB and potential reserve growth of 96 BB, this region remains an important producer of liquid petroleum (Tables II and IV).

2.4 Central and South America

With 8.4% of the world's liquid petroleum endowment, and consuming only 6.8% of the world's liquid petroleum, this region is a petroleum supplier to the remainder of the world. The Atlantic offshore of South America has become a prolific hydrocarbon producer. The success of the offshore areas of Brazil in the Campos Basin, with oil discoveries greater than 1 BBO at Roncador and Barracuda fields and a 600-million barrel discovery in the Santos Basin, demonstrate the significant liquid petroleum resources offshore. Elsewhere along this Atlantic margin, large undiscovered liquid petroleum resources were estimated for offshore areas of Guyana and Surinam. The discovery of hydrocarbons in deep water of the south Atlantic (approximately 4000-m water depths) in the Campos Basin make both the offshore area of South America and counterpart basins along the western margin of Africa one of the most active and successful exploration areas. This significant potential is reflected in the estimate of 125 BB of undiscovered liquid petroleum resources and another 37 BB of potential reserve growth of liquid petroleum (Table II).

Although not assessed, enormous heavy oil deposits are known in the Orinoco heavy oil belt. These resources are estimated at more than 1 trillion barrels of oil in place, comparable to the Alberta oil sands. The recoverable amount from the Orinoco heavy oil belt varies depending on the gravity of the oil, depth, and location.

2.5 Asia/Pacific

With only 5.1% of the world's liquid petroleum endowment and a huge population to support, the Asia/Pacific region is the second largest consumer of liquid petroleum (24%) but produces only 9.5% of it, resulting in the largest percentage deficit (−14.5%) of liquid petroleum of any region of the world (Tables III and IV). Overall, estimates for

undiscovered liquid petroleum resources in this region have been reduced from previous estimates to 45 BB, with 31 BB of potential reserve growth (Table II). Much of this reduction resulted from the individual reassessment of Chinese provinces. Extensive exploration in China has achieved only modest success. China has been a net oil importer since 1993, and the consequent reduced likelihood of adding substantial new reserves will make China increasingly competitive elsewhere for petroleum in the world. Modest increases of undiscovered oil were estimated for the northwest shelf of Australia and deep-marine reservoirs in the Kutei and Sabah offshore basins of Indonesia and Malaysia.

2.6 Sub-Saharan Africa and Antarctica

With only 1.7% of the world's liquid petroleum consumption and 4.6% of the world's endowment, this region is an important supplier of liquid petroleum (Tables III–V). Continued success in the Niger Delta, as well as in deep-water reservoirs especially offshore of Cameroon and Angola, make this region particularly important for liquid petroleum. Many new oil discoveries of more than a half billion barrels of oil have been made offshore of western Africa. The undiscovered liquid petroleum resource estimate of 82 BB and the estimated 20 BB of potential reserve growth demonstrate the significant potential volumes of this region (Table II). Examples of discoveries in excess of 1 BB include Girassol and Dalia fields in Angola. In terms of liquid petroleum export potential, this region ranks third behind the Middle East/North Africa and the FSU (Table III).

Antarctica was not assessed by the USGS in 2000 because it will probably not be accessible for petroleum exploration, due to international agreements, in the next few decades.

2.7 Europe

The liquid petroleum endowment of Europe ranks next to last at 3.8%, and Europe has the second largest production versus consumption deficit (−10.8%). In the North Sea, the region has benefited from the latest exploration and development technology in developing relatively young petroleum provinces. With fields performing much longer and at better rates than once anticipated, Europe remains heavily dependent on foreign supplies of liquid petroleum on a percentage basis and will likely have a continuing reliance on foreign supplies (Tables III–V).

With 36 BB of undiscovered liquid petroleum resources and 24 BB of potential reserve growth (Table II) this region will continue to be an important provider of liquid petroleum, as demonstrated in Table IV.

2.8 South Asia

The South Asia region is next to last in liquid petroleum consumption and has the smallest endowment (0.7%; Table II). It produces approximately 1% of the world's liquid petroleum but consumes 3.3%, thus running a deficit (−2.3%) that is fourth largest in the world (Table III). This region generally has relatively low oil potential but significant gas potential. The liquid petroleum potential resides mostly in the Bombay High, Irrawaddy, and Assam provinces. Recent discoveries in the thrust belt of Pakistan provide optimism but will not meet the increasing liquid petroleum needs of this region. The large population, the need for economic growth, and the current deficit consumption will require considerable and increasing imports of liquid petroleum. The estimated undiscovered liquid petroleum resources of 6.2 BB and reserve growth potential of 5.3 BB are modest.

3. IMPACT OF DISTRIBUTION

As can be seen in Figs. 1 and 2 and Table II, the largest liquid petroleum endowments are concentrated in a few areas, primarily the Middle East, the FSU, and, to a lesser degree, North and South America, North Africa, and sub-Saharan Africa. This uneven distribution of liquid petroleum results in considerable trading of oil as demonstrated in Fig. 3 (data for 2000 to be consistent with the resource data). Generalized oil trade movements depicted in Fig. 3 demonstrate the dominance of the Middle East and FSU, with significant contributions from South America and West Africa. For the United States and Europe, significant interregional trading is required to meet the enormous liquid petroleum consumption. These two regions account for most of the oil trading; that is, North America and Europe consume 51.2% (North America, 31.3%; Europe, 19.9%) but only produce 29% of the resources, ensuring the continuation of large amounts of imports (Table III). Tables IV and V demonstrate that some large producing and consuming nations rank high on both lists, such as the United States, Russia, China, Canada, the United Kingdom, Brazil, and Saudi

Arabia. The tables also show that many small producers are large consumers, such as Japan, South Korea, Taiwan, Germany, France, Italy, Spain, and India. This imbalance of supply and consumption drives an increasing degree of world liquid petroleum trading, as shown for 2000 in Fig. 3.

The USGS estimates that future liquid petroleum supplies for the world are significant (Tables I and II). According to 2001 IHS Energy data, world oil reserves are at all-time highs (more than 1.1 trillion barrels). Annual oil consumption is approximately 0.7% of the endowment (28 BBO/year). The world has used approximately 23% of its total oil endowment of approximately 3 TBO. In summary, liquid petroleum remains relatively abundant and will continue to be a dominant source of energy along with natural gas and coal throughout this century. However, these resources are unevenly distributed geographically, and the successful transition to other forms of energy will require a continuing global trading environment and continued investment to access the world's undiscovered and reserve growth resources of liquid petroleum.

Various groups have projected the time at which peak oil production might occur using the USGS data as a reference case. For example, the Energy Information Administration estimates a world oil peak in approximately 2036–2040. However, in 2002, Albert Cavallo suggested a 2015–2020 peak for oil production from non-OPEC countries. Cavallo's analysis suggests even greater dependence on the relatively localized conventional oil resources. Increasingly, natural gas resources, only 7% of which are utilized outside the United States, will be developed and produced. Additionally, development of heavy oil in the Orinoco and tar sand deposits in Alberta is already occurring and may increase. Other remote heavy oil deposits, such as those of the East Siberian Basin in the FSU, can be developed once they are viewed as economic.

4. DATA SOURCES

Information about liquid petroleum resources is dynamic and commonly proprietary. Fortunately, several sources of information are easily accessible. Much of the data on discovered resources in this article was obtained from the U.S. Energy Information Administration. Their Web site (www.eia.doe.gov) provides extensive data on resources and use for all energy sources (not just petroleum) for the United States and the rest of the world. The data for reserve

growth and undiscovered resources are derived primarily from the USGS's 2000 world assessment. The USGS also has a Web site (energy.usgs.gov) containing extensive data for the United States and the rest of the world.

Additional information can be found in several petroleum industry magazines (e.g., *Oil and Gas Journal* and *World Oil*) that publish annual statistical summaries. BP also produces an annual statistical summary and has a Web site (www.bp.com/centres/energy) with much data.

SEE ALSO THE FOLLOWING ARTICLES

Markets for Petroleum • Nationalism and Oil • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil Industry, History of • Oil-Led Development: Social, Political, and Economic Consequences • Petroleum Property Valuation • Strategic Petroleum Reserves

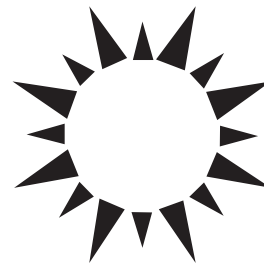
Further Reading

- British Petroleum (2001). BP Statistical Review of World Energy. BP, London. [Available at <http://www.bp.com/centres/energy>].
- Cavallo, A. J. (2002). Predicting the peak in world oil production. *Nat. Resour. Res.* 11, 187–195.
- DeGolyer, and MacNaughton. (2000). "Twentieth Century Petroleum Statistics," 56th ed. DeGolyer & MacNaughton, Dallas, TX.
- Edwards, J. S. (2002). Twenty-first century energy: Transition from fossil fuels to renewable, nonpolluting energy sources. In "Sustainability of Energy and Water through the 21st Century" (L. C. Gerhard, P. P. Leahy, and V. J. Yannacone, Eds.), Proceedings of the Arbor Day Farm Conference, pp. 37–48. Kansas Geological Survey/American Association of Petroleum Geologists (AAPG) and the AAPG Division of Environmental Geosciences, Lawrence, KS/Tulsa, OK.
- Energy Information Administration (2001). "International Energy Outlook 2001." U.S. Department of Energy, Washington, DC. [Available at <http://www.eia.doe.gov>].
- IHS Energy (2001). "Petroleum Exploration and Production Database." IHS, Denver, CO. [Database available from IHS, P.O. Box 740619, Houston, TX 77274–0619]
- International Energy Agency (1998). "World Energy Outlook." International Energy Agency/Organization for Economic Cooperation and Development, Paris.
- International Energy Agency (2000). "World Energy Outlook." International Energy Agency/Organization for Economic Cooperation and Development, Paris.
- International Energy Agency (2001). "World Energy Outlook—2001 Insights." International Energy Agency/Organization for Economic Cooperation and Development, Paris.
- Masters, C. D., Attanasi, E. D., and Root, D. H. (1994). In "World Petroleum Assessment and Analysis: Proceedings of the 14th World Petroleum Congress," pp. 529–541. Wiley, London.
- Minerals Management Service (1996). "An assessment of the undiscovered hydrocarbon potential of the nation's outer continental shelf, OCS Report MMS 96–0034." Minerals Management Service, Washington, DC.
- Schollnberger, W. E. (1998, November). Projections of the world's hydrocarbon resources and reserve depletion in the 21st century. *Houston Geol. Soc. Bull.*, 31–37.
- U.S. Geological Survey (1995). "National assessment of United States and gas resources," Circular 1118. U.S. Geological Survey, Washington, DC.
- U.S. Geological Survey World Energy Assessment Team (2000). U.S. Geological Survey world petroleum assessment 2000—Description and results, Digital Data Series DDS-60. U.S. Geological Survey, Washington, DC. [Available at <http://energy.usgs.gov>]

Oil and Natural Gas: Offshore Operations

RON BAKER

University of Texas
Austin, Texas, United States



-
1. History
 2. Scope
 3. Exploration
 4. Drilling Rights
 5. Drilling Rigs
 6. Drilling a Well
 7. Formation Evaluation
 8. Development Drilling
 9. Well Completion
 10. Production and Workover
 11. Oil and Gas Transportation
 12. Summary

Glossary

bit The cutting or boring element used in drilling oil and gas wells. The bit consists of a cutting element and a circulating element. The circulating element permits the passage of drilling fluid and utilizes the hydraulic force of the fluid stream to improve drilling rates.

Christmas tree The control valves, pressure gauges, and restrictions (chokes) assembled at the top of a well to control the flow of oil or gas after the well has been drilled and completed.

drill collar A heavy, thick-walled tube, usually steel, placed between the drill pipe and the bit in the drill stem to provide weight on the bit.

drilling fluid A fluid (either gas or liquid) used in rotary drilling. Drilling fluid serves many functions—for example, lifting cuttings out of the wellbore and carrying them to the surface, keeping the hole from caving in, controlling pressure found in underground rock formations, and cooling and lubricating the bit and drill stem.

drill pipe Heavy seamless tubing used to rotate the bit and circulate the drilling fluid.

drill stem All members of the assembly used for rotary drilling from the surface to the bottom of the hole. Includes drill pipe and drill collars.

glycol dehydrator A processing unit that removes all or most of the water from gas. A typical glycol unit consists of an absorber, in which wet gas is put into contact with glycol to remove the water (glycol absorbs the water), and a reboiler, which heats the wet glycol to remove the water from it so the glycol can be reused.

jackup drilling rig A mobile bottom-supported offshore drilling structure with legs that support the deck and the hull. When positioned over the drill site, the bottom of the legs rests on the seafloor. A jackup is towed to position with its legs up; once the legs are firmly positioned on the bottom, the deck and hull height are adjusted and leveled.

mud The liquid usually placed in steel tanks, or pits, on a rig, which is circulated through the wellbore during rotary drilling operations. In addition to bringing cuttings to the surface, mud cools and lubricates the bit and drill stem, holds formation pressures in check, and prevents the hole from caving in.

platform rig An immobile offshore structure from which development wells are drilled and produced. Platform rigs may be built of steel or concrete.

semisubmersible drilling rig A floating offshore drilling unit that has pontoons and columns that, when flooded, cause the unit to submerge in the water to a predetermined depth.

wellhead The equipment installed at the surface of a wellbore. A wellhead includes such equipment as the casinghead and tubinghead, from which special hangers and pressure-tight seals suspend the well's casing and tubing.

Extracting oil and gas from layers of rock that lie beneath oceans and seas presents special problems to oil producers that they do not face on land sites. This article examines some of the special conditions the marine environment imposes on finding, producing, and transporting oil and gas. The petroleum industry produces oil and gas from special layers of rocks called reservoirs. Like a multilayered cake, additional

beds of rock lie above and below these reservoirs. And, like the frosting on a cake, a relatively thin layer of ground sometimes covers the rock layers. On the other hand, the “frosting” may not be dry land; it may be water instead. Since oceans and seas cover about three-fourths of the earth, it is no surprise that water also covers rock layers.

1. HISTORY

Edwin Drake drilled the first oil well in the United States in 1859 on a piece of land near Titusville, Pennsylvania. It was only 38 years later, in 1897, that another enthusiast drilled the first offshore well in the United States. He drilled it off the coast of Southern California, immediately south of Santa Barbara in the town of Summerland.

In addition to a pleasant, sunny climate, Summerland had numerous springs. These springs did not, however, produce water; instead, crude oil and natural gas bubbled from them. One person, H. L. Williams, was knowledgeable about extracting oil from the earth. So he drilled several wells in the vicinity of the springs. These early wells were successful and, as a result, he and others drilled many more in the area.

After drilling several wells, they noticed that those nearest the ocean were the best producers. Eventually, they drilled several wells on the beach itself. At this point, however, the Pacific ocean stymied them. Experience convinced them, however, that more oil lay in the rock formations below the ocean. The question was how to drill for it.

Williams came up with the idea of building a wharf and erecting the drilling rig on it. His first offshore well, drilled from a wharf made of wooden pilings and timbers, extended about 300 ft (90 m) into the ocean. On the end of the wharf, Williams erected a drilling rig and used it to drill the first offshore well in the United States. As expected, it was a good producer and before long the entrepreneurs built several more wharves. The longest stretched more than 1200 ft (nearly 400 m) into the Pacific.

2. SCOPE

Offshore activities take place in the waters of more than half the nations on Earth. No longer do primitive, shore-bound wooden wharves confine offshore operators. (Operators are the oil companies that produce oil and gas.) Instead, they drill wells from

modern steel or concrete structures. These structures are, in many cases, movable. What is more, they can float while being moved, and often while drilling. Further, offshore rigs can drill in waters that are more than 10,000 ft (more than 3000 m) deep and as far as 200 miles (over 300 km) from shore.

Drilling and producing oil and gas wells are important phases of offshore operations, but the scope goes further. Offshore operations also include exploring—looking for likely places where oil and gas may exist in the rock formations that lie beneath the surface of the oceans, seas, gulfs, and bays. In addition, offshore operations include transporting oil and gas—moving them from their points of production offshore to refineries and plants on land.

3. EXPLORATION

Deep reservoirs buried in rock layers beneath an ocean do not indicate their presence on the surface. Explorationists cannot therefore find them by direct observation. To locate likely rock formations that contain hydrocarbons—oil and gas—companies often employ seismic surveying. Seismic surveying provides a considerable amount of information about subsurface formations. And modern computing power reveals reservoirs that the technology of the past could not locate. Besides powerful computers to interpret the findings, seismic surveying depends on the fact that rock layers reflect sound waves. That is, sound bounces off the layers like an echo. Oil seekers use this reflection to draw a sort of underground picture.

As mentioned before, rocks occur in layers, one on top of the other, much like the layers of a cake. In offshore seismic exploration, a boat trails a special noisemaker that generates a loud, low-frequency sound. The loud, low-frequency sound waves go down through the water and into the rock layers. They then continue traveling downward through the many layers for thousands of feet (meters).

Where one kind of layer ends and another begins, the boundary reflects some of the sound waves. The reflected sound waves travel back up through the layers and into the water. In the water, sensitive detectors called hydrophones, which the boat also tows, pick up the sound waves. The reflections from shallower layers arrive at the hydrophones sooner than the reflections from deeper layers.

The arrival times of each reflection indicate the depth of the layers, giving explorationists a cross-sectional view of the subsurface. Expert interpreters

using powerful computers examine a seismic section and determine whether a geological structure may contain hydrocarbons. Geologists call this cross section a seismic section (Fig. 1).

4. DRILLING RIGHTS

An oil company that wishes to drill for and possibly produce hydrocarbons must obtain the rights to do so. In most cases, the company must obtain these rights from the country in whose waters the oil and gas activity will take place. In the United States, most offshore drilling occurs in the Outer Continental Shelf, or OCS. The Outer Continental Shelf starts at the point offshore where state ownership of the water and of any minerals (such as hydrocarbons) under the water ends. For instance, the state of Texas owns the mineral rights from shore to a distance of 9 miles (14.4 km). Louisiana, on the other hand, owns the mineral rights from shore to a distance of 3 miles (4.8 km).

The U.S. Department of the Interior, Minerals Management Service (MMS), regulates the OCS. The OCS ends where international waters begin or where treaties with other countries, such as Mexico and Canada, establish the end of U.S. jurisdiction. Thus, to explore in OCS waters of the United States, the oil company or companies (sometimes several compa-

nies form partnerships) must obtain the right to do so from the federal government. Companies usually obtain this right by bidding on offshore blocks, or tracts, offered for sale by the government at various times. The company or companies that bid the most money are the ones most likely to win the right to drill for and produce any hydrocarbons discovered on the block.

In countries other than the United States, the system is similar, except that most other nations own all the mineral rights from the shoreline out to international waters or the waters of bordering countries. For example, in the North Sea, which is bordered by the United Kingdom, Norway, Denmark, Germany, and the Netherlands, the area is divided into sectors. Each country controls its own sector. In this area, the country in whose sector the drilling site lies grants the oil companies concessions to drill.

5. DRILLING RIGS

Once an oil company obtains the right to drill a wildcat, or exploratory, well to confirm the existence of hydrocarbons in a rock layer, it must then select some type of drilling rig. More often than not, the oil company will use a mobile offshore drilling unit, which is shortened to MODU and is pronounced "moe-doo" (Fig. 2). Rig owners can move mobile

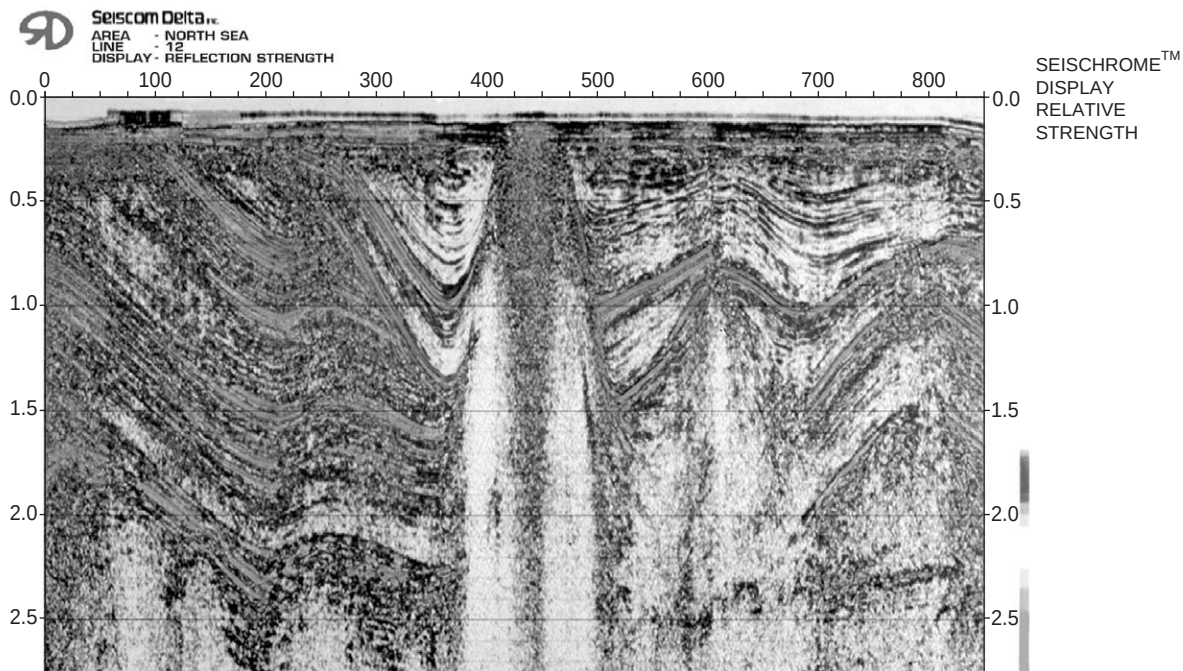


FIGURE 1 A seismic section gives a view of subsurface formations.

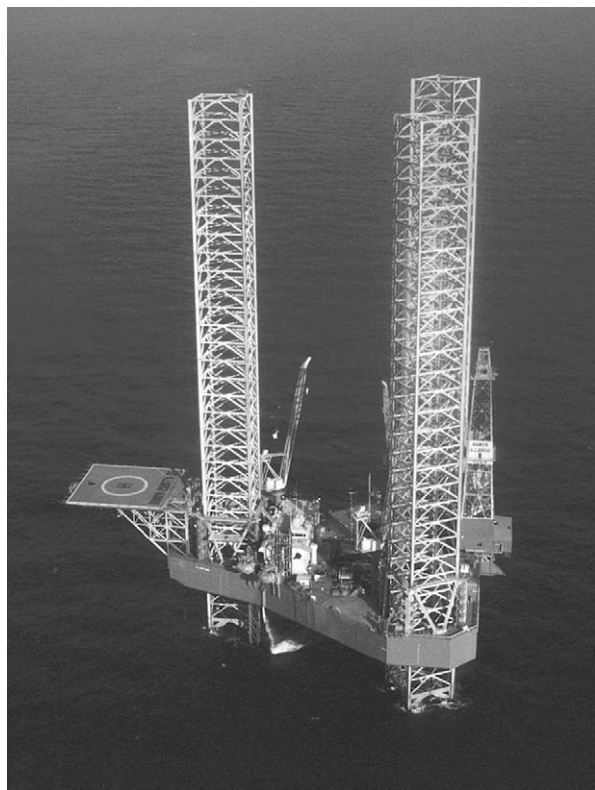


FIGURE 2 A mobile offshore drilling unit (MODU).

offshore drilling units from one drill site on the water to another. A rig has to be mobile because, after it finishes drilling one exploratory well, a crew has to move it to another site—perhaps nearby, perhaps far away—to drill another. Oil operators use two basic types of MODUs to drill most offshore wildcat wells: bottom-supported units and floating units. Bottom-supported units include jackups. Floating units include semisubmersibles and drill ships (Fig. 3).

5.1 Jackups

Jackups are popular because they provide a stable drilling platform. They are stable because part of their structure firmly contacts the bottom of the ocean. They can also drill in relatively deep water: up to 350 ft (100 m) deep. What is more, towboats can easily move a jackup from one location to another.

Most jackups have open-truss legs. In this design, manufacturers construct the legs from tubular steel members. They then crisscross the members to make very strong, but relatively lightweight, legs. Open-truss legs look somewhat like the towers that carry high-voltage electric lines across the countryside. The legs of a jackup pass through openings in a barge



FIGURE 3 A drill ship.



FIGURE 4 A special ship carries a semisubmersible rig.

hull. Three or four legs are common, but some designs call for more or fewer. The deck of the barge serves as the platform for drilling equipment and other machinery. When a crew moves a jackup, it raises, or jacks, the legs up out of the water so that the barge floats. With the legs completely out of the water, the rig movers can transport the entire rig to the drilling location. Usually rig movers tow the rig.

Sometimes, large ships with flat decks move jackups. Crew members submerge the ship so that its deck is below the water. They then maneuver the rig onto the submerged ship's deck. With the rig in place, crew members pump out the water from the ship, which allows it to float. Especially for long-distance moves, ships can carry the jackup at speeds faster than boats can tow it. Ships can also carry semisubmersibles (Fig. 4).

Once the jackup rig is on location, crew members jack the legs down until they come into firm contact with the seafloor. Then they jack up the barge hull onto the legs until it is clear of the water's surface and well above high waves.

5.2 Floating Units

Floating rigs include drill ships and semisubmersibles. Generally, oil companies most often use drill ships to drill wells in the deep, remote waters of the world. Semisubmersibles can drill in seas too rough for drill ships. Drill ships have a ship-shaped hull that allows the drilling contractor (the rig owner that the oil company hires to drill) to store a large amount of supplies and equipment on board the vessel. Thus, a drill ship can stock up in port, sail to the drilling location, and drill for long periods without being restocked.

Semisubmersibles have large pontoon hulls and big columns, which the rig operator submerges to a depth below the water's surface. Submergence makes semisubmersibles the most stable of floating drilling units. Even though semisubmersibles cannot carry as much drilling supplies as a drill ship—they do not have a hull for storage—drilling companies often have to use them to drill in rough seas.

6. DRILLING A WELL

A writer once described a drilling rig as a portable hole factory. He pointed out that the sole purpose of a rig was to make holes in the ground, and since drilling contractors had to drill holes at different locations, the rig had to be movable. The writer wasn't too far off the mark. A rig—be it large or small, on land or offshore—has one main job: drilling wells.

6.1 Bits and Drilling Fluid

The job of a rig and the people who work on it is to put a drill bit in the earth and turn it. A bit is a hole-boring device (Fig. 5). Pressed very hard against the ground and turned, or rotated, it makes a hole. The teeth on the bit grind and gouge the rock into small pieces. These pieces of rock, or cuttings, must be moved out of the way so the bit teeth can be constantly exposed to uncut rock. To move cuttings away from the bit, the rig pumps a special liquid called drilling fluid, or mud, down to the bit. Most often, drilling mud is a mixture of water, special clays, and certain minerals and chemicals.

Special openings in the bit called nozzles eject drilling mud out of the bit with great speed and pressure. These jets of mud lift the cuttings off the bottom of the hole and away from the bit. With the cuttings out of the way, they do not interfere with the bit teeth's ability to drill. Mud also carries the cuttings up the hole and to the surface for disposal.



FIGURE 5 A bit has teeth (cutters) to penetrate a formation and make a hole.

The mud pump in the circulating system moves drilling mud down the hole through special pipe called drill pipe and drill collars, which are called the drill string. Mud then shoots out of the bit through jets, picks up the cuttings, and carries them to the surface. On the surface, special equipment removes the cuttings. Then, the mud pump recirculates the clean mud back down the hole.

6.2 Rotating System

Part of keeping the bit drilling involves rotating it, so the rig needs a rotating system. Commonly, offshore rigs use a top drive to rotate the bit. Top drives have a very powerful built-in electric motor (some have more than one motor). The motor or motors rotate a drive shaft. Crewmembers make up the top of the drill string to the drive shaft. When the driller activates the top drive's motor and drive shaft, the drive shaft rotates the attached drill string and bit.

6.3 Power System

Any drilling rig needs power—power to actuate the mud pumps, to hoist the drill string, and to run all

the machinery on the rig. Several large diesel engines usually provide the power. A diesel engine is an internal-combustion engine, which means that it runs because a mixture of fuel and air burns inside the engine. Electricity transmits engine power—that is, the diesel engines put out mechanical power. The mechanical power turns electrical generators. The generators produce electricity, which heavy-duty cables carry to powerful electric motors on the equipment needing power. Drillers control the transfer of power with controls and switches from their position on the rig floor.

6.4 Hoisting System

Another major system on the rig is the hoisting system, which is a very large block-and-tackle system (Fig. 6). It consists of the drawworks, or hoist; a mast or a derrick; the crown block; the traveling block; and wire rope. It suspends the drill pipe and drill collars in the hole. (Drill collars are extra heavy pipe that is used to put weight on the bit to make it bite into the formation and drill.)

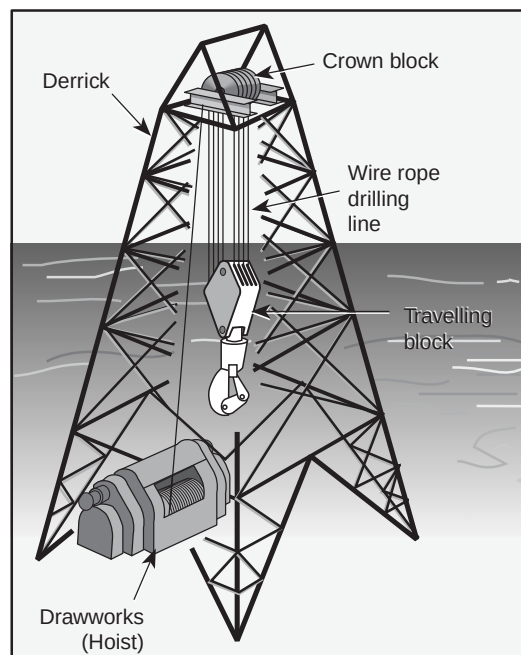


FIGURE 6 The hoisting system consists of several components, all of which serve to suspend and move drilling equipment in and out of the hole.

7. FORMATION EVALUATION

Whether the oil company drills an exploration well from a bottom-supported or a floating unit, it must evaluate prospective formations—that is, the company must determine whether the well has struck oil or gas in sufficient quantities to justify further reservoir development. Because the cost of developing, or exploiting, offshore reservoirs is so high, the quantity of oil or gas must be large. In fact, companies sometimes abandon reservoirs discovered offshore even though they contain a fair amount of oil and gas. Evaluation involves measuring certain properties of a potential reservoir rock and obtaining samples of the rock and any fluids that may be in it.

7.1 Well Logging

Well logging is an evaluation method in which a logging crew lowers a special tool, a sonde, into the well and then pulls it back up. As the sonde passes the formations on its way up the wellbore, it senses and measures electrical, radioactive, and acoustic (sound) properties of the rocks. Measuring and recording these properties provides a good deal of information about the rocks. The sonde transmits its measurements via conducting cable to the surface, where computers record them. Experts analyze the

resulting recording, the log (Fig. 7). Analysis involves careful study of the curves on the log. The curves indicate whether hydrocarbons exist in the formations investigated and to what extent.

7.2 Coring

Coring is another method of formation evaluation. The operator—the oil company— obtains one type of core sample by lowering a sidewall sampler into the borehole. At the desired interval, the driller activates the sampler to fire several small cylinders into the wall of the hole. The cylinders penetrate a short distance into the rock and cut a small core. When crewmembers start to remove the tool from the hole, cables attached to the cylinders pull them from the wall of the hole, and the cores come up the hole with the tool. Geologists can then examine the small cores and analyze them for indications of the character of the formation.

7.3 Drill Stem Testing

A drill stem test (DST) is a method of formation evaluation. The test allows fluids in the formation being evaluated to flow into the drill stem and to the surface. Pressure-recording devices in the drill stem test tool measure and record pressure in the well while it is both flowing and not flowing (shut in).

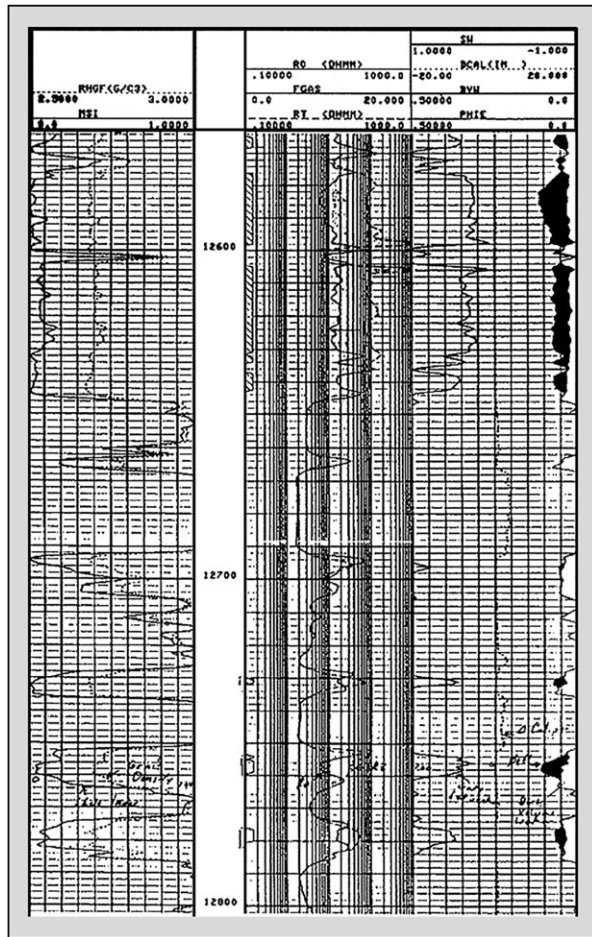


FIGURE 7 Curves on this well log reveal information about formations penetrated by the hole.

A DST can also obtain actual fluid samples from the formation. From detailed analysis of the pressure recordings and of the fluids, the operator can infer a great deal of information about the reservoir.

8. DEVELOPMENT DRILLING

If tests on an exploratory well prove favorable, the oil operator usually drills and evaluates several additional wells in or near the reservoir. These additional wells are appraisal wells. The operator drills appraisal wells to confirm further that the reservoir contains enough hydrocarbons to justify the enormous expense of developing it. If the appraisal wells reveal that the reservoir contains enough hydrocarbons, then development drilling may occur. Development drilling is the drilling of several wells into a reservoir to extract hydrocarbons discovered by exploratory wells and confirmed by appraisal wells.

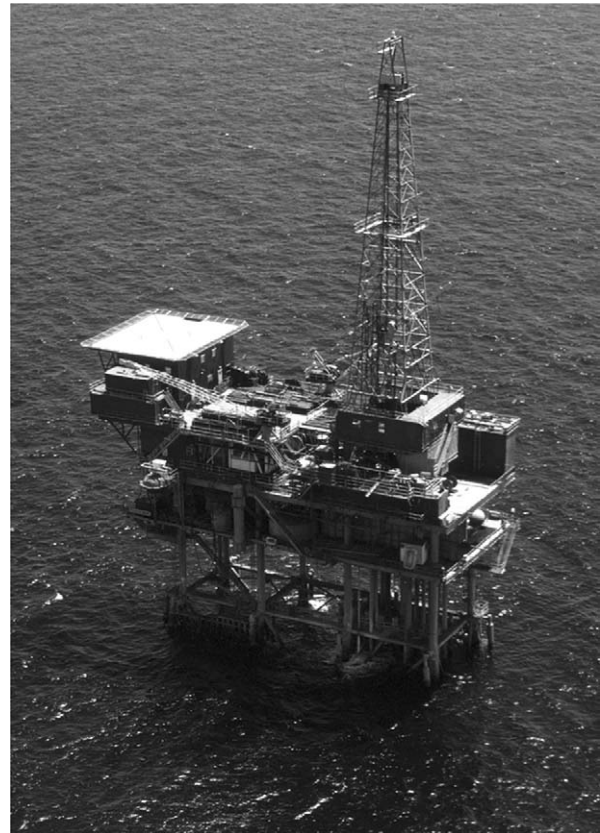


FIGURE 8 A self-contained platform.

Operators use many kinds of rigs to drill development wells. A common type in U.S. waters is the platform rig. Operators also use mobile offshore drilling units to drill development wells, the same rigs they often employ to drill exploratory and appraisal wells.

8.1 Drilling Platforms

Oil companies, or operators, drill a large number of offshore development wells from totally fixed, self-contained platforms. The operator drills, completes, and produces all the development wells from the platform, thus eliminating the need for rig mobility. Modern platforms house all the drilling equipment, production equipment, crew quarters, offices, galley, and recreation rooms (Fig. 8). Manufacturers build some of steel and some of concrete.

A steel-jacket platform consists of a jacket—a tall, vertical section made of tubular steel members—which is the foundation of the platform. Piles driven into the seabed support it. The jacket extends upward so that the top rises above the waterline. Additional sections on top of the jacket provide

space for crew quarters, a drilling rig, and all the equipment needed to drill. The height of the jacket depends on the water's depth. Some of the tallest stand in water over 1000 ft (300 m) deep.

In the North Sea, rig builders make concrete gravity platforms from steel-reinforced concrete. They construct them in deep, protected waters near shore. They then float and tow to the site in a vertical position (Fig. 9). Some have three or four tall concrete caissons, or columns, that resemble smokestacks. At the site, the operator floods the caissons, and the platform submerges to come to rest on bottom.

One end of the caissons rests on the seafloor. The rig builders install drilling equipment, crew quarters, and production equipment on a platform on the other end, well above the waterline. Concrete platforms weigh a tremendous amount, so builders do not have to pin them to the bottom with pilings. The force of gravity alone keeps them in place.

8.2 Mobile Offshore Drilling Units

In some cases, an operator uses a MODU, such as a jackup, to drill development wells. For example, the operator can erect a relatively small platform jacket, only large enough to provide space for drilling and production wellheads on the site. In such cases, the drilling contractor places the MODU next to the jacket and drills and completes the wells on it. After the wells are drilled, the contractor moves the MODU away and, if necessary, the operator erects an additional platform nearby to house production equipment and quarters for the production crew. A system of two relatively small platforms and a MODU can reduce costs below those of a single large drilling and production platform. Operators can also use mobile rigs to drill development wells into a reservoir,



FIGURE 9 Boats tow a concrete platform to a site in the North Sea.

perhaps into parts of it that are not accessible from a platform located over another section. They drill these satellite wells one at a time—that is, the MODU drills one well and then moves to drill other wells at different locations over the reservoir. This method of drilling a single development well and then moving the rig is practical only with MODUs.

Sometimes it is not economically efficient to drill several single development wells at various points into a reservoir. In such cases, operators can use MODUs with subsea templates. A subsea template is an arrangement of special equipment the operator places on the seafloor. The template allows the operator to drill, complete, and produce several wells from the relatively small point on the seafloor that the template covers. The operator places the MODU, usually a floater, on the water's surface over the template, and drills and completes several wells on the template.

9. WELL COMPLETION

After operators drill a development well into the reservoir and run and cement the final string of casing, they must complete the well. The operator must provide a way for the reservoir fluids to enter the well and reach the surface.

9.1 Perforating

To open the well to the reservoir, the operator can perforate the casing and cement that line the well drilled into the reservoir. That is, a perforating company can make several small holes in the casing and cement. Hydrocarbons flow from the reservoir into the well through the perforations.

To perforate a well, a perforating operator lowers a special perforating gun to a point opposite the producing formation. The operator fires the gun, and several small but intense explosions penetrate the casing and cement and travel a short distance into the reservoir. These small explosions are shaped charges—explosives that release a high-energy charge of gases and particles. These gases and particles travel in one direction and penetrate casing, cement, and formation. Reservoir fluids then enter the well through the perforations.

9.2 Tubing and Packers

Usually, operators run a special string of pipe inside the casing. This special pipe is tubing. Reservoir fluids flow up the well through the tubing that is inside the

cemented casing. Operators produce most wells through tubing instead of directly through the casing. They produce through tubing for several reasons. For one thing, the crew does not cement the tubing in the well. As a result, when a joint of tubing fails, as it almost inevitably will over the life of the well, the operator can easily replace the failed joint. Since casing is cemented, it is very hard to replace. For another thing, tubing allows the operator to control the well's production by placing special tools and devices on or in the tubing string. These devices allow operators to produce the well efficiently. Casing does not provide a place to install tools or devices required for production. In addition, the operator usually installs safety valves in the tubing string. These valves automatically shut off the flow of fluids from the well if damage occurs at the surface. Finally, tubing protects casing from the corrosive and erosive effects of produced fluids. Over the life of a well, reservoir fluids tend to corrode metals with which they are in contact. Producing fluids through tubing, which the operator can easily replace, preserves the casing, which is not so easy to repair or replace.

Crewmembers often install a packer near the bottom of the tubing above the producing zone. A packer is a length of pipe through which well fluids flow. The packer has gripping elements that latch onto the inside wall of the casing to help anchor the packer. Rubber sealing elements on the packer, when expanded, provide a pressure-tight seal between the tubing and casing. The packer forms a fluid-tight seal between the tubing and the casing. Thus, when reservoir fluids flow into the well, the packer forces them to flow into the tubing.

9.3 Surface and Subsea Completions

Completing a well also involves the control of fluids as they flow from the top of the well. The operator can place the devices that control fluid flow either above the surface of the water or onto the seafloor. In a surface completion, the operator places the equipment that controls the flow of hydrocarbons from the well above the waterline. Operators usually complete the wells they drill from a platform on the platform itself. In a subsea completion, the operator places the equipment that controls the flow of hydrocarbons from the well on or below the seafloor.

In a subsea completion, the operator locates the collection of special valves and fittings that controls the flow of hydrocarbons from the well—the so-called Christmas tree—at the top of the well on the seafloor. A subsea Christmas tree can be wet or dry. If

wet, the operator installs it on the wellhead where it remains exposed to the surrounding water. If dry, the operator installs the Christmas tree in some type of housing that isolates it from the water, often at atmospheric pressure.

Divers may maintain and repair wet trees; sometimes, the operator utilizes remotely operated vehicles (ROVs) to make repairs. Operators can maintain some dry trees with small submarines called submersibles or diving bells to gain access to the housing around the Christmas tree. Workers enter the housing without any special clothing or underwater breathing equipment.

10. PRODUCTION AND WORKOVER

Offshore production involves a wide range of techniques and equipment. The operator uses the techniques and equipment to get oil and gas out of the reservoir and into a transportation system. As produced fluids flow up the well and to the surface, the operator has to separate oil, gas, and water, and must handle each according to its individual requirements. As a result, a typical production platform may have a large amount of equipment and several people to successfully extract, treat, and move oil and gas to shore for refining and processing.

10.1 Reservoir Drive Mechanisms

Natural pressure in the reservoir forces oil and gas to flow from a reservoir into a well and to the surface. The main sources of this pressure are gas, water, or both in combination. Four kinds of natural drive mechanisms occur with oil and gas: dissolved-gas drive, gas-cap drive, water drive, and combination drive (Fig. 10).

Dissolved-gas drive is present when all or almost all of the hydrocarbons in the reservoir are liquid in its undrilled state. When the operator drills a well into such a reservoir, the well creates an area of reduced pressure. It is somewhat like jamming a straw into a bottle of liquid and sucking on the straw. Sucking on the straw reduces pressure and liquid flows up the straw. Similarly in a reservoir, reducing the well's pressure causes some of the lighter liquid hydrocarbons to become gaseous. The gas comes out of solution and, as it does, it drives oil into the well and to the surface.

Gas-cap drive results when a reservoir contains so much gas that all of it is not dissolved in the oil. Because gas is lighter than oil, it usually rises to the

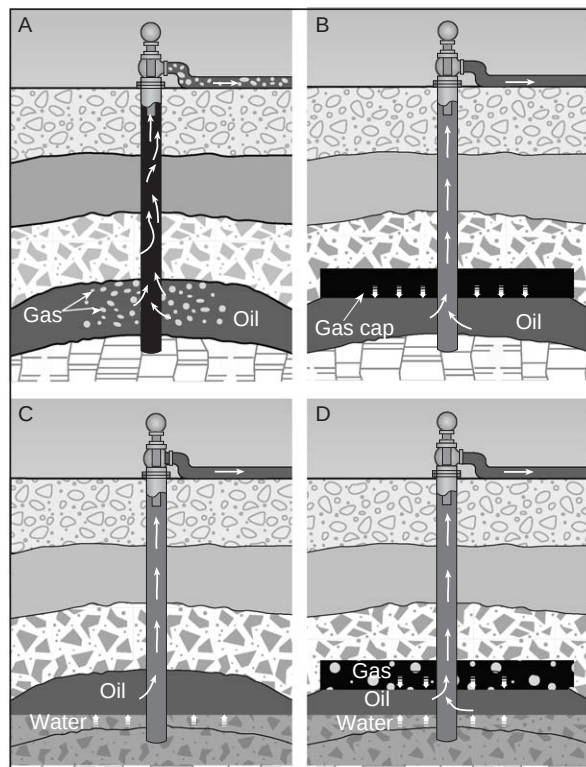


FIGURE 10 Dissolved gas drive (A), gas-cap drive (B), water drive (C), and combination drive (D).

top of the reservoir and forms a cap over the oil. When the operator completes a well into such a reservoir, the gas cap expands. The expanding gas drives oil into the well and to the surface. Water drive occurs when large amounts of water—usually salt water—lie beneath the oil in a reservoir. When the operator drills a well, the water in the reservoir moves toward the well and drives oil ahead of it. Combination drive exists when both gas and water provide energy to drive oil to the surface. Gas—in the form of a cap or dissolved in the oil or a combination of the two—and water lying below the oil work to lift oil to the surface. Combination drive is the most common drive mechanism.

10.2 Handling Oil, Gas, and Water

Oil and gas seldom reside in a reservoir by themselves. Water and other impurities like sediment usually coexist with them. As operators produce the reservoir, these impurities come to the surface with the oil and gas. Operators must separate and remove the foreign material from the hydrocarbons. They must also separate the oil and gas from each other before they can transport them to shore for refining

and processing. Operators must separate and remove impurities—especially sediment and water (S&W)—because the impurities take up space in a tanker or pipeline that valuable hydrocarbons could otherwise occupy. Furthermore, water can corrode practically any steel surface, including that of a tanker compartment or a pipeline.

Operators have to separate oil from gas primarily because each has different handling requirements. For instance, the operator usually has to put gas under additional pressure to put it into a pipeline or special tanker for transport to shore facilities. Because oil is a liquid, it cannot be compressed, and therefore must be pumped into a pipeline or tanker. To separate and handle oil, gas, and water, operators use several techniques and a variety of equipment. They usually locate the equipment on the same platform from which they drilled the wells. But once they start producing the wells, the platform becomes a production platform and operators use the derrick for such purposes as raising and lowering equipment to work over the well. Because it's not being used strictly for drilling at this time, the derrick can be used to run special tools into wells that are producing but that need cleanout or repair.

In a subsea completion, the operator routes oil, gas, and water into flow lines that run from the wells to a production riser. Production flows up the riser to a floating buoy. An oil tanker can then tie up to the buoy, take on the produced fluids, and transport them to shore for separation and treatment. In some cases, the operator moors a processing and storage ship to the buoy. Equipment on the ship separates, treats, and stores the reservoir fluids. When oil fills up the ship's tanks, the ship transports it to a receiving terminal on shore. Operators sometimes use this system in deep waters where erecting a platform is not economically feasible.

In shallower waters, operators can send production from subsea wells to a platform. Once on the platform, equipment separates the fluids and off-loads the oil onto a tanker or injects it into a pipeline.

Water produced with oil and gas occurs in three ways: (1) as free water, (2) combined with oil in an emulsion, or (3) mixed with gas as vapor. Free water is water not tightly bound to the oil and gas. It readily settles out if the operator gives the produced fluids a chance to remain stationary for a short period in a tank or vessel. Thus, one type of equipment producers sometimes use on an offshore production facility is a free-water knockout (FWKO). A free-water knockout is usually a cylindrical vessel. The operator pipes well fluids that

contain relatively large amounts of free water after they leave the Christmas tree into the FWKO. An FWKO can be either vertical (standing upright) or horizontal (lying on its side). Vertical designs take up less space but cannot handle as much volume as horizontal ones. Horizontal vessels expose the fluids inside them to more surface area. The design selected depends primarily on the volume, or quantity, of the fluids being produced and the amount of space available on the platform.

When well fluids go into a free-water knockout, the free water, since it is heavier than oil or gas, falls to the bottom of the vessel. The operator then removes it and disposes of it properly. If significant amounts of gas are present in the fluid stream, the manufacturer can equip the knockout to handle the gas. In such cases, the gas rises to the top of the knockout. The producer removes the gas and sends it to other equipment on the platform for further handling. Oil generally winds up in the middle. The operator pipes the oil to equipment on the platform for further handling.

Sometimes, water tightly binds to the oil in the produced fluids. Usually, this water spreads out (disperses) in the oil as small droplets that cannot readily settle out. This combination of oil and water is an emulsion. The operator specially treats emulsions to remove the emulsified water from the oil. To treat an emulsion, producers use special vessels—heater-treaters. They pipe the emulsion into the heater-treater and use chemicals, electricity, heat, or all three to break the emulsion. Breaking the emulsion means making the water droplets coalesce (merge) and settle out of the oil. Usually, the operator adds chemicals to the emulsion and pipes it into the heater-treater. Heater-treaters, like free-water knockouts, can be either vertical or horizontal. Chemical companies formulate the chemicals to make the small water drops merge into large drops. Large water drops fall from the oil to the bottom of the treater where the operator removes them.

Electricity can also cause small water drops to merge and fall to the bottom of a treater. If an emulsion, usually with emulsion breaking chemicals in it, passes over a high-voltage electric grid in the treater, the water droplets merge and fall. Electrostatic treaters use electricity to treat, or break, an emulsion.

Heat reduces the viscosity, or thickness, of oil. Since water can fall out of a heated, thin oil more easily than it can out of a cool, thick oil, operators sometimes use heater-treaters in the emulsion-breaking process. In such cases, a fire tube, which is an enclosed gas flame, heats the emulsion while it is in

the vessel. Operators often use chemicals and electricity in addition to heat.

10.3 Separation of Oil and Gas

To separate oil and gas from one another, the producer usually employs a separator. Most separators are vertical or horizontal, although a spherical (ball-shaped) design is available. All provide an enclosed space into which the operator pipes produced fluids. Given a chance to remain stationary for a short period, the force of gravity makes the fluids separate. Some are two-phase separators, which means that only oil and gas or emulsion and gas separate inside them. Others are three-phase separators, which means that oil (or emulsion), gas, and water separate inside them. If only relatively small amounts of free water are in the produced fluids, or if some free water remains after the fluids pass through a free-water knockout, then the producer may use a three-phase separator.

As separation occurs inside the separator, the gas goes to the top of the vessel. The operator pipes it to additional gas-handling facilities on the platform—to a glycol dehydrator, for example. If water separates out, operators remove it from the bottom and dispose of it. They remove oil from the middle (or bottom, if they are using a two-phase separator) and transport it to shore if the oil does not require further treatment. Usually, however, water is emulsified in the oil, so the operator pipes this emulsion to treaters elsewhere on the platform. After the treaters break out most of the emulsified water, the operator transports the treated oil to shore.

10.4 Artificial Lift

As natural reservoir pressure produces hydrocarbons, this natural pressure declines. Eventually, the pressure dwindles to the point that it can no longer lift oil to the surface. Natural reservoir pressure usually depletes long before the operator is able to extract all the hydrocarbons from the reservoir. So operators turn to artificial lift to produce at least some of the remaining hydrocarbons. In artificial lift, the operator applies some form of external energy to the well. The applied energy must reduce pressure at the bottom of the well. It must reduce the bottom-hole pressure enough so that the small amount of remaining reservoir pressure can force more reservoir fluids into the well and lift them to the surface. Offshore, the most common form of artificial lift is gas lift. In gas lift, compressors inject natural gas into

a well. This gas lowers pressure at the bottom of the well. Gas is a very lightweight, or low-density, fluid. A low-density fluid like gas does not exert as much pressure as a liquid. Injecting gas to the bottom of the well forces most of the liquids out of the well so that a mixture of liquid and gas remains. This liquid-gas mixture is lighter than the original liquid column and thus exerts less pressure on the reservoir at the bottom of the well. Since reservoir pressure is now higher than pressure inside the well, fluids from the reservoir flow into the well and to the surface.

10.5 Improved Recovery Techniques

Artificial lift cannot produce all of the hydrocarbons that reside in a reservoir. In spite of every effort, a fair amount of oil is left behind even after artificial lift is applied. The laws of physics, the behavior of fluids in a reservoir, and the nature of the reservoir itself all combine to make it impossible to extract every bit of oil in a reservoir. In some reservoirs, up to 75% of the hydrocarbons originally in place are left behind. On the average, an oil company can produce only about 33% of the total amount of oil in the reservoir. In an attempt to recover at least some of the remaining oil, companies can apply improved recovery, which is the application of various fluids, chemicals, and heat to a reservoir. Improved recovery techniques, if successful, can remove additional amounts of oil from the reservoir that is left behind by natural and artificial lift.

10.6 Well Servicing and Workover

As wells produce and start aging, downhole equipment wears out, sand from the reservoir gets into the well and restricts fluid flow, gas-lift valves malfunction, corrosion eats away metal parts, and other problems occur. In some cases, the operator hires a well service and workover contractor, a company that specializes in offshore well repair, to repair or maintain the well. In other cases, the operator turns to the company's own employees who have the expertise and equipment to do it themselves. In either case, the goal of well servicing and workover is to fix whatever is wrong and restore production to a normal rate or to a rate as near normal as possible.

Operators and contractors use several techniques, tools, and methods to repair or maintain wells. On large, self-contained platforms, the operator often retains the derrick used for drilling the original wells and uses it for servicing those same wells. For instance, if part of the tubing string fails—maybe

corrosion eats away part of it—a repair crew has to pull the entire string out of the well. The crew then replaces the damaged portion and run the string back in. Having the original derrick in place makes pulling and running tubing relatively easy. On the other hand, if the platform is small and has no derrick or if a satellite well needs to have its tubing pulled, then the operator may move in a mobile workover rig. Both jackup and semisubmersible workover rigs are available for servicing and workover.

Special wireline tools can run, manipulate, and pull much of the equipment operators place in the well when they complete it. Wireline is strong, fine wire that spools onto a reel. Repair crews can attach various tools and devices to wireline and lower or pull it from the well in one continuous movement. Using wireline is much faster than running and pulling a tubing string. Since wireline is one continuous strand, crewmembers do not need a derrick to run tools and equipment on it.

For servicing subsea completions, repair crews can pump special tools down the flow line and into the tubing string to do a specific job. After they complete the job, they circulate the tool back to the surface. Such pumpdown tools are often called through-the-flow-line tools, or TFL tools for short. TFLs have simplified remedial work on subsea completions. Also, operators use them on surface completions as well.

11. OIL AND GAS TRANSPORTATION

After operators produce, separate, treat, and, in some cases, store oil and gas on an offshore production facility, they must ultimately send them to shore. Once on land, refineries and processing plants make products that practically everyone in the world uses. Companies normally transport oil and gas in two ways: in pipelines or in tanker ships. A pipeline is a series of connected pipes through which a company sends oil or gas. One end of the pipeline originates at a production facility and the other end terminates at a facility on shore. Pumps or compressors inject the oil or gas into the pipeline.

Building an offshore pipeline is a complex operation. It often involves the laying of several miles (kilometers) of large-diameter pipe on the seafloor under adverse conditions. Usually, the pipeline construction company coats all but the ends of the pipe with concrete. Concrete protects the pipe, helps prevent corrosion caused by seawater, and

weights the pipe so that it will stay on bottom. The construction crewmembers leave the pipe ends free of concrete so they can weld the joints together. They then coat the welded joints to prevent them from corroding. If the pipeline carries crude oil, the terminal on shore will probably consist in part of several large storage tanks. From these tanks, land pipelines send oil to refineries. A gas pipeline generally terminates at a facility capable of processing the gas. Gas processing involves recovering the heavier hydrocarbon components of natural gas, such as propane and butane. After the processing plant breaks down the gas into its components, it then pipelines the components to other plants for additional processing or to consumers.

An offshore pipeline construction company can lay an offshore pipeline in several ways. Commonly, companies will use a special barge, a lay barge. It floats on the water's surface, or, if the water is rough, it floats in a semisubmerged state, much like a semisubmersible drilling rig. Think of a lay barge as a floating offshore assembly line. Each step in the operation—welding, inspection, wrapping, laying—occurs at a station on the barge as a special machine conveys the pipe along the barge's length. Ideally, once the operation starts, it won't stop until the line is completely laid. Unfortunately, bad weather and unforeseen events can bring things to an abrupt halt and increase what are already very high costs.

A lay barge carries a large number of pipeline joints. It also carries the personnel and equipment necessary to weld the joints together, inspect them for integrity, apply a protective coating on the welded joints, and place the pipe on the seafloor. Further, as the welding machines join each joint, the barge has to be able to move forward so that it lays the pipe in a steady and continuous motion with as few delays as possible. Since the lay barge usually cannot carry enough pipe to complete the job, supply boats shuttle pipe to the barge from a shore base.

Sometimes, it is not feasible for a company to build a pipeline to carry oil to shore. For example, if an area's production is not very large, the expense of building a pipeline could be more than the income received from the sale of oil. Or building a pipeline at a particular site could be physically impossible. Whatever the reason, when companies cannot pipeline oil to shore, they often use oil tanker ships instead. In the case of an offshore gas field, the operator usually must build a pipeline because no other practical way exists to transport gas. In a few cases, however, a company may build an offshore gas-liquefaction facility—where the facility cools natural

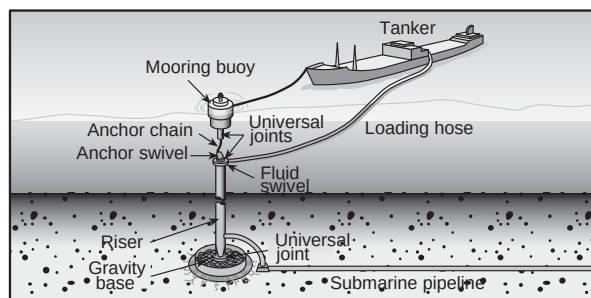


FIGURE 11 Single point buoy mooring (SPBM) system.

gas to an extremely low temperature and liquefies it. The company can then transport the liquid gas to shore via tankers specially designed to handle liquefied natural gas (LNG). On the other hand, a company can produce oil from a platform or subsea wells and pipe it to a special mooring system. A tanker ship can tie up to the system, take on oil, and then transport the oil to an onshore terminal for unloading.

The most common type of mooring system is single-point buoy mooring (SPBM) (Fig. 11). An SPBM system is an installation in the sea where a tanker can take on a crude oil cargo. It is called a single-point system because it provides only one point to which the tanker ties up, or moors. Oil to the SPBM system comes from wells on a platform or from subsea wells.

12. SUMMARY

Offshore operations cover a broad range of activities that include exploration, exploration drilling, development drilling, production, workover, and transportation. Offshore exploration includes the use of seismic surveys. Seismic survey methods yield evidence of the presence of hydrocarbons in formations that lie below the oceans and seas of the world. Should exploration reveal the likelihood of hydrocarbons at a certain site, then drilling a well becomes necessary. Often, operators use a mobile offshore drilling unit—a MODU—to drill the well. Because companies can evaluate a well after they drill it, they can find out for sure whether oil and gas reside in a potential reservoir. Evaluation techniques such as logging, drill stem testing, and coring can confirm that the exploration well has penetrated a hydrocarbon-bearing formation.

If the wildcat well shows what appear to be ample quantities of oil and gas, a company may drill appraisal wells. Appraisal wells further confirm that

the reservoir holds enough hydrocarbons to justify the enormous expense of developing the reservoir. If commercial amounts of oil or gas do exist in a reservoir, the company must develop the reservoir. It may move a self-contained platform to the site and erect the platform. The platform may be a rigid steel or concrete structure. Usually, the operator drills several development wells from the platform. On the other hand, the operator may place a subsea template on the seafloor, and drill the wells from a MODU. Whether from a platform or on a template, the operator drills and completes several wells to produce hydrocarbons from the reservoir.

Producers may drill satellite wells where they cannot get to the reservoir from a platform or a template. Satellite wells and wells drilled on a template are often subsea completions—that is, producers install the Christmas trees on or below the seafloor. Wells that operators complete on platforms are usually surface completions: they install the Christmas trees on a platform deck above the waterline.

The production phase of offshore operations includes the installation of special pieces of equipment to separate and treat crude oil, natural gas, and water. In most cases, operators install the equipment on a platform—usually the same platform they drilled the wells on. Sometimes, however, they install the equipment on a special processing and storage ship. They most often use such ships with subsea completions.

Operating companies may have to put wells produced from platforms and from special ships on artificial lift. They will most likely use gas lift, which allows wells to produce after natural drive depletes. In some cases, applying additional recovery techniques to extract still more hydrocarbons from an aged and dying reservoir may be possible.

Once a company completes and puts development wells on production, it will need to service and repair the wells to keep them producing at the desired rate. In many instances, operators can use wireline and pumpdown repair methods, but in other instances, they may have to move in a semisubmersible or jackup well service and workover unit.

Finally, a company may build a pipeline to transport the oil and gas to refineries or plants for the extraction of useful products. Or perhaps tanker ships will shuttle oil to shore. If so, the operator will install special mooring systems. Offshore operations occur all over the world because it is beneath the oceans and seas where the petroleum industry will find most of the world's future hydrocarbons.

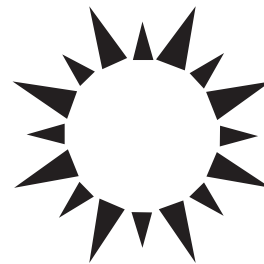
SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Spills, Environmental Impact of • Gas Hydrates • Natural Gas, History of • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Drilling • Oil and Natural Gas Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Industry, History of • Oil Pipelines • Oil Recovery • Public Reaction to Offshore Oil

Further Reading

- Baker, R. (1983). "Oil and Gas: The Production Story." Petroleum Extension Service, University of Texas at Austin, Austin, TX.
- Baker, R. (2001). "A Primer of Offshore Operations." Petroleum Extension Service, University of Texas at Austin, Austin, TX.
- Gray, F. (1995). "Petroleum Production in Nontechnical Language." PennWell, Tulsa, OK.
- Hosmanek, M. (1984). "Pipeline Construction." Petroleum Extension Service, University of Texas at Austin, Austin, TX.
- Hyne, N. J. (1995). "Nontechnical Guide to Petroleum Geology, Exploration, Drilling, and Production." PennWell, Tulsa, OK.
- Johnson, D. E., and Pile, K. E. (1988). "Well Logging for the Nontechnical Person." PennWell, Tulsa, OK.
- Kennedy, J. L. (1993). "Oil and Gas Pipeline Fundamentals." PennWell, Tulsa, OK.
- Link, P. K. (2001). "Basic Petroleum Geology." Society of Petroleum Engineers, Richardson, TX.
- Van Dyke, K. (1996). "A Primer of Oilwell Service, Workover, and Completion." Petroleum Extension Service, The University of Texas at Austin, Austin, TX.
- Van Dyke, K. (1997). "Fundamentals of Petroleum." Petroleum Extension Service, University of Texas at Austin, Austin, TX.

Oil and Natural Gas Resource Assessment: Classifications and Terminology



T. R. KLETT

United States Geological Survey
Denver, Colorado, United States

-
1. Introduction
 2. Commodities Considered in Oil and Natural Gas Resource Assessments
 3. Classification of Oil and Natural Gas Accumulations
 4. Assessment Units
 5. Exploration and Discovery History
 6. Assessment Concepts
 7. Parameters for Classification of Oil and Natural Gas Resources
 8. Classification of Oil and Natural Gas Resources

Glossary

accumulation One or more reservoirs of petroleum that have distinct trap, charge, and reservoir characteristics. For purposes of assessment, an accumulation is treated as a single entity. An accumulation may encompass several fields or be equivalent to a single field or reservoir; alternatively a field may be equivalent to one or several accumulations.

assessment An evaluation or appraisal of the quantity, quality, and composition of technically or economically recoverable petroleum or in-place petroleum that provides information on which an estimate of value of the resources on land and in offshore areas can be made.

field An area consisting of a single reservoir or multiple reservoirs of petroleum, all grouped on, or related to, a single geological structural and (or) stratigraphic feature.

petroleum A collective term for crude oil, natural gas, natural gas liquids, and tar.

reserves The estimated quantities of petroleum of a specified date, expected to be commercially recovered from known accumulations under prevailing economic conditions, operating practices, and government regulations. Reserves are part of the identified (discovered) resources and include only recoverable quantities.

reserve growth That part of the identified resources over and above measured reserves, estimated to be added to existing fields and reservoirs through delineation of additional in-place oil and natural gas, improved recovery efficiency, and revisions resulting from recalculation of viable reserves under dynamically changing economic and operating conditions.

resource A concentration of naturally occurring solid, liquid, or gaseous hydrocarbons in or on the earth's crust, some of which is, or potentially is, economically extractable.

The United States Geological Survey (USGS) periodically assesses petroleum resources of the world. An assessment of petroleum resources is an evaluation or appraisal of the quantity, quality, and composition of technically or economically recoverable petroleum or of in-place petroleum that provides information on which an estimate of value of the resources on land and in offshore areas can be made. The goal of USGS assessments is to provide impartial, scientifically based, societally relevant petroleum-resource information essential to the economic and strategic security of the United States. Resources must be continuously reassessed in light of new geologic knowledge, progress in science and technology, and shifts in economic and political conditions. Assessments of resources may involve either discovered quantities (reserves) or undiscovered resources. Assessed petroleum resource volumes are estimates, not measurements. In resource-assessment studies, appraisal is synonymous to, and used interchangeably with, assessment. Appreciation refers to reserve growth, the increases in reserves of fields or accumulations through time with continued development and production.

1. INTRODUCTION

Society's dependence on petroleum, as both a fuel and a feedstock for materials used in industry and agriculture (such as plastics, medicine, and fertilizer), requires knowledge of oil and natural gas availability. Availability can be considered in terms of geology and economics. Geologic availability concerns the quantity and quality of petroleum and is the most fundamental characteristic that governs commercial use. Assessment of geologic availability of petroleum requires basic knowledge of the geology.

Societal concepts of the magnitude of petroleum resources commonly are based on economic rather than geologic availability. However, economic availability is evaluated within the frame of petroleum estimated to be geologically available. Several categories of petroleum volumes in terms of geologic availability exist and evaluations of economic availability may be very different depending on which is used. Because major political and economic decisions are based on the economic availability of petroleum, the classification system and terminology used to report estimated volumes of petroleum, therefore, must be precisely defined.

This article provides classifications and terminology used by the USGS for the assessment of petroleum resources. First, an approach to petroleum resource assessments is presented, followed by descriptions of petroleum commodities, spatial geologic units (assessment units), petroleum accumulations, and supporting exploration and discovery data. Assessment concepts used by the USGS are then presented, followed by classifications of estimated petroleum resources. A comparison of resources and reserves definitions is made between the older joint U.S. Bureau of Mines and U.S. Geological Survey (USBM/USGS) classification system and the newer classification system of the Society of Petroleum Engineers and World Petroleum Congresses (SPE/WPC) and Society of Petroleum Engineers, World Petroleum Congresses, and American Association of Petroleum Geologists (SPE/WPC/AAPG), which is representative of the industry sector. Other organizations and agencies in the United States, such as the Potential Gas Committee, the Department of Energy's Energy Information Administration, and the Department of Interior's Minerals Management Service, have similar classifications for resources and reserves, but are not included in this article.

The classifications and terminology used in this article are intended to represent standard usage by the U.S. oil and natural gas industry, the resource-

assessment community, and the USGS. The USGS adopted many of the definitions given by the American Petroleum Institute. Nevertheless, the classifications and terms in this article do not represent an exhaustive compilation from all sources and other professionals, agencies, and organizations may define some terms used in this article differently.

2. COMMODITIES CONSIDERED IN OIL AND NATURAL GAS RESOURCE ASSESSMENTS

Petroleum, as used in this article, is a collective term for crude oil, natural gas, natural gas liquids (NGL), and tar. The petroleum commodities considered in USGS assessments are crude oil, natural gas, and NGL, all of which are typically produced from the subsurface through wells. Petroleum commodities produced by other means require different assessment methodologies. Such commodities have not been appraised in recent USGS assessments; they include natural gas dissolved in geopressured brines, resources in tar deposits, oil shales, and gas hydrates.

Crude oil, according to the American Petroleum Institute (API), is defined as a mixture of hydrocarbons that exists in a liquid phase in natural underground reservoirs and remains liquid at atmospheric pressure after passing through surface separation facilities. Although oil is mainly a mixture of complex hydrocarbon compounds, having carbon/hydrogen ratios ranging typically from 6 to 8, it may also contain some nonhydrocarbon components, such as sulfur, nitrogen, oxygen, vanadium, and nickel. Oil has specific gravities ranging typically from 0.76 (55° API gravity) to 1.00 (10° API gravity). API gravity, defined by the API, is a measure of the density of oils.

The volumes reported as crude oil generally include (1) liquids technically defined as crude oil, (2) small amounts of hydrocarbons that exist in the gaseous phase in underground reservoirs but are liquid at atmospheric pressure after being recovered from oil well gas in lease separators, and (3) small amounts of nonhydrocarbons produced with the oil.

The American Petroleum Institute defines natural gas as a mixture of hydrocarbon compounds and small quantities of various nonhydrocarbons existing in the gaseous phase or in solution with oil in natural underground reservoirs under reservoir temperature and pressure conditions and produced as a gas under standard temperature and pressure conditions [60°F or 15°C and 1 atm (14.7 lbs/in.² or 101.325 kPa)].

Natural gas is principally methane, but may contain ethane, propane, butanes, and pentanes, as well as certain nonhydrocarbon gases, such as carbon dioxide, hydrogen sulfide, nitrogen, and helium.

Natural gas can be associated with or dissolved in oil accumulations (associated/dissolved gas or wet gas) or not associated with any liquid hydrocarbons (nonassociated gas or dry gas). Associated gas is free natural gas, commonly known as gas-cap gas, which overlies and is in contact with crude oil in the reservoir. Dissolved gas is natural gas in solution with crude oil in the reservoir under reservoir temperature and pressure conditions. Associated/dissolved gas is commonly reinjected into the reservoir to maintain a pressure drive for oil production and may therefore not be an economically feasible resource in a near-term assessment time frame. Nonassociated gas is free natural gas that is not in contact with crude oil in the reservoir.

Natural gas liquids, as defined by the American Petroleum Institute, are those portions of reservoir gas that are liquefied at the surface in lease separators, field facilities, or gas processing plants. Natural gas liquids are the heavier homologues of methane (including ethane, propane, butane, pentane, natural gasoline, and condensate) and are in a liquid phase under surface conditions [that is, standard temperature and pressure conditions; 60°F or 15°C and 1 atmosphere (14.7 lbs/in.² or 101.325 kPa)]. The specific gravity of NGL is generally less than 0.76 (greater than 55° API gravity).

3. CLASSIFICATION OF OIL AND NATURAL GAS ACCUMULATIONS

Assessed petroleum commodities may be categorized with respect to accumulations as (1) oil in oil accumulations, (2) natural gas in oil accumulations (associated/dissolved natural gas), (3) NGL in oil accumulations, (4) natural gas in gas accumulations (nonassociated natural gas), and (5) liquids (both crude oil and NGL) in gas accumulations.

An accumulation is one or more reservoirs of petroleum that have distinct trap, charge, and reservoir characteristics. For purposes of assessment, an accumulation is treated as a single entity. An accumulation may encompass several fields or be equivalent to a single field or reservoir; alternatively, a field may be equivalent to one or several accumulations. The USGS identifies two major types of petroleum accumulations based on geology: conventional and continuous (Fig. 1). Conventional accumulations are associated with structural or stratigraphic traps, commonly bounded by a down-dip water contact, and therefore affected by the buoyancy of petroleum in water. Continuous accumulations are areally extensive reservoirs of petroleum not necessarily related to conventional structural or stratigraphic traps. These accumulations lack well-defined down-dip petroleum/water contacts and thus are not localized by the buoyancy of oil or natural gas in water. Examples of continuous accumulations include “tight gas reservoirs,” coal bed

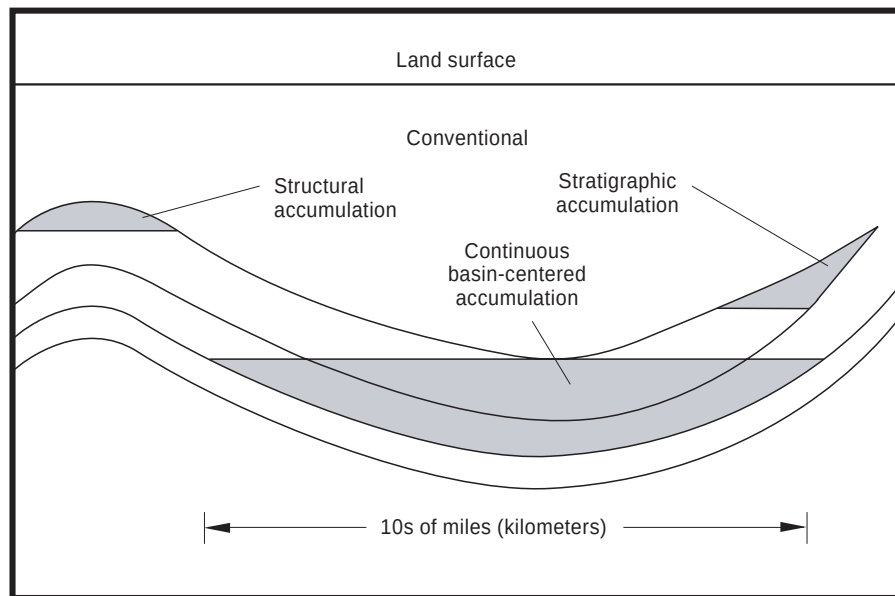


FIGURE 1 Conventional and continuous accumulations. Modified from Schmoker (1996).

gas, oil and gas in shale, oil and gas in chalk, basin-centered gas, and shallow biogenic gas.

A continuous accumulation can be treated as a collection of petroleum-containing cells. The USGS defines a cell as a subdivision or area within a continuous accumulation having dimensions related to the drainage areas of wells (not to be confused with finite-element cells as used in reservoir modeling). Cells differ from wells by having areal properties, whereas wells are considered areally as points and a cell may contain more than one well. Two types of cells are recognized: tested cells (those evaluated by drilling) and untested cells. Tested cells are categorized as productive, having at least one well with reported production, and nonproductive, evaluated with drilling but where none of the wells have reported production. Untested cells have not been evaluated by wells.

The term unconventional has been loosely used interchangeably with continuous when describing these accumulations. Continuous accumulations are defined by geologic criteria, whereas unconventional accumulations are characterized by a variety of standards, such as arbitrary matrix-permeability limits, special regulatory status, the need for unusual (at the time) engineering techniques, or difficulty of physical access (for example, polar regions or deep water).

Oil and gas accumulations, both conventional and continuous, are treated separately in the USGS assessment process. Gas/oil ratios (GOR) are calculated for each accumulation to identify or categorize the major commodity (oil or gas). An oil accumulation is defined as having a GOR less than 20,000 ft³ of gas/barrel of oil; a gas accumulation is defined as having a GOR equal to or greater than 20,000 ft³ of gas/barrel of oil. Coproduct ratios (NGL/natural gas ratio; and petroleum liquids/natural gas ratio) are also calculated to aid in the assessment of those associated resources on an accumulation level. Petroleum liquids are undifferentiated oil and NGL.

Accumulations can contain various production elements having spatial characteristics that include fields, reservoirs or pools, and wells.

A field is an area consisting of a single reservoir or multiple reservoirs of petroleum, all grouped on, or related to, a single geological structural and (or) stratigraphic feature. Individual reservoirs in a single field may be separated vertically by impervious strata or laterally by local geologic barriers. When projected to the surface, the reservoir(s) within the field can form an approximately contiguous area that may be circumscribed. Common field classifications include discovered (located by exploratory drilling),

under development (undergoing methods or additional drilling to bring into production), producing, and abandoned.

The volume of petroleum produced within a given year is called annual production. The successive sum of annually produced volumes is the cumulative production to that date. Recoverable oil and gas volumes that have not yet been produced are called remaining reserves. Remaining reserves and just "reserves" are used interchangeably; reserves include only recoverable quantities. The recoverable volume is the sum of cumulative production and remaining reserves. The recoverable volumes are typically well known within a specified uncertainty range and provide a measure of the size of a field (field size). These recoverable volumes may also be described as known recoverable volumes (those reported at any given time) and total (ultimate) recoverable volumes.

A reservoir, as defined by the American Petroleum Institute, is a porous and permeable underground formation containing an individual and separate natural accumulation of producible petroleum, which is confined by impermeable rock or water barriers. A reservoir is typically a single hydraulically connected unit with a single natural pressure system. Pool is an older term for reservoir, but is still in use by some. A pool, however, may consist of more than one reservoir in certain situations.

Wells are drilled to find, delineate, and produce petroleum from accumulations. Some wells are used to inject fluids to enhance productivity of other wells in the accumulation. Wells may be classified with respect to drilling goals. Such classifications are called initial classifications and are (1) exploratory wells, which include new-field wildcat, new-pool wildcat, deeper pool test, shallower pool test, outpost or extension tests; (2) development and injector wells; (3) service wells, which include observation wells; and (4) stratigraphic tests. A final classification of status may be assigned to a drilled well, which includes oil well, gas well, multiple completions, junked or abandoned, and dry hole. Additionally, a classification of petroleum content is commonly recorded for wells, which includes oil, gas, oil and gas, dry, oil shows, and gas shows.

The volumes of petroleum from producing wells are reported as annual and cumulative production, as for fields. However, the total volume of petroleum estimated to be recovered from wells is called the estimated ultimate recovery. The estimated ultimate recovery of a well is typically calculated by extrapolation of its past production performance.

4. ASSESSMENT UNITS

In order to organize, evaluate, and delineate areas and rock volumes for assessment, the USGS developed a hierarchical scheme of geographic and geologic units. This scheme consists of regions, geologic provinces, total petroleum systems, and assessment units. Regions serve as organizational units and geologic provinces are used as prioritization tools. Assessment of undiscovered resources is carried out at the level of the total petroleum system or the assessment unit.

A geologic province is an area, having characteristic dimensions of hundreds of kilometers, that encompasses a natural geologic feature (for example, a sedimentary basin, thrust belt, or accreted terrane) or some combination of contiguous geologic features. Geologic province boundaries are delineated along natural geologic boundaries, although in some places they may be located arbitrarily (for example, along specific water-depth contours in the open oceans).

Total petroleum systems and assessment units are delineated within each geologic province considered for assessment. It is not necessary for the boundaries of total petroleum systems and assessment units to be entirely contained within the geologic province.

The USGS identifies “total” petroleum systems for the assessment process, rather than the narrowly defined petroleum systems delineated by known or discovered accumulations or production. The definition of a total petroleum system is similar to that for a petroleum system, except that undiscovered accumulations are included in total petroleum systems. The total petroleum system includes all genetically related petroleum that occurs, or is estimated to occur, in shows and accumulations, both discovered and undiscovered, which have been generated by a pod or by closely related pods of mature source rock. Particular emphasis is placed on the similarities of petroleum fluids within total petroleum systems. Total petroleum systems exist within a limited mappable geologic space, together with the essential mappable geologic elements necessary for the accumulation of petroleum (source, reservoir, seal, and overburden rocks). These essential geologic elements control the fundamental processes of generation, expulsion, migration, entrapment, and preservation of petroleum within the total petroleum system.

In areas or rock volumes where more than one source rock exists and significant mixing of petroleum occurred, identification of individual total petroleum systems may not be possible or practical

for the assessment processes. In such areas or rock volumes, a composite total petroleum system may be defined. The concept of composite total petroleum systems is applied when more than one source rock is assumed to charge the accumulations.

Assessment units are identified to organize the assessment process and serve as the basic area or rock-volume entities for resource appraisal. An assessment unit is a mappable part of a total petroleum system in which discovered and undiscovered oil and gas accumulations constitute a single relatively homogenous population such that the methodology of resource assessment is applicable. An assessment unit might be equivalent to a total petroleum system, but total petroleum systems may be subdivided into two or more assessment units such that each assessment unit is sufficiently homogenous in terms of geology, exploration considerations, accessibility, and risk to be assessed individually.

Some USGS assessments identify plays to organize the assessment process and serve as the basic area or rock-volume entities for resource appraisal. A play is a set of known or postulated oil and gas accumulations sharing similar geologic, geographic, and temporal properties, such as source rock, migration pathway, timing, trapping mechanism, reservoir rocks, and hydrocarbon type. A play may differ from an assessment unit in that it can include one or more assessment units. More commonly, an assessment unit includes one or more plays. Other USGS studies use prospects as the basic units for assessment. A prospect is a location where an accumulation is expected to exist and geologic information and economics justify the drilling of a wildcat or test well.

The terms conventional and continuous are used to describe assessment units and plays and are assigned primarily to describe their respective accumulations.

A subdivision of continuous accumulations or continuous assessment units is the “sweet spot.” A sweet spot is an area within a continuous accumulation where production characteristics are relatively more favorable than elsewhere.

5. EXPLORATION AND DISCOVERY HISTORY

Petroleum resource assessments are based, partly or wholly, on analysis of the process by which accumulations within a given area are discovered (discovery

process). With conventional accumulations, typically, the largest and shallowest accumulations are found relatively early in the exploration history, followed by the discovery of smaller accumulations. The smaller accumulations found later in the history typically require more time and exploration effort, including the number of exploratory wells drilled for discovery.

Exploration- and discovery-history data are used to model the discovery process, by which the number and sizes of yet to be found (undiscovered) accumulations are estimated. These data are graphed and analyzed statistically. Graphs of accumulation-size distribution and of the number of discovered accumulations with respect to size class portray the sizes of past discoveries. The difference between the discovered size distribution and an assumed distribution for the parent population of accumulations can provide an estimate of the number of undiscovered accumulations for each of the size classes.

Graphs of accumulation sizes with respect to year of discovery or amount of exploration effort provide indications of success and discovery maturity and their extrapolation allows inferences concerning numbers and sizes of remaining (undiscovered) accumulations. Success commonly can be described as one of two ratios: (1) the number of discoveries per number of years of exploration and (2) the number of discoveries per number of exploratory wells drilled. During initial exploration, success ratios may increase with time due to the experience gained by exploration. After initial exploration, the success ratio typically decreases with time as an area becomes more mature with respect to exploration (discovery maturity). However, new exploration-play concepts can offset or reverse a decreasing trend in success ratio.

6. ASSESSMENT CONCEPTS

The USGS incorporates the concepts of forecast span, minimum accumulation size or minimum recovery per cell, and probability of occurrence, or risk, in the assessment process.

In making assessments, the USGS considers only those quantities of oil and gas resources that have the potential to be added to reserves within a specified forecast span. Petroleum resources can be ranked to provide a visual representation of practicability of extraction in the form of a triangle or pyramid (a concept of J. K. Gray). A relatively small volume of high-quality resources is present near the top, with increasing volumes of progressively lower quality and less practicable resources advancing toward the base (Fig. 2). For an assessment, a slice can be drawn through the pyramid at some quality level, which defines the resource of oil or gas that is estimated to become economic within the foreseeable future. The position of the slice can be regarded as the forecast span, whereby resources below the slice are not assessed. The slice may move lower into the pyramid on successive assessments, as originally less practicable and poorer quality resources are deemed potentially economic over time.

Oil and natural gas from undiscovered conventional accumulations and the development of cells within continuous accumulations must exceed or equal a selected minimum recoverable volume. A minimum recoverable volume is selected to avoid inclusion of large numbers of volumetrically insignificant accumulations or cells. In the time frame of the assessment, volumes of oil and natural gas from very small accumulations or cells are unlikely to contribute significantly to the total resources.

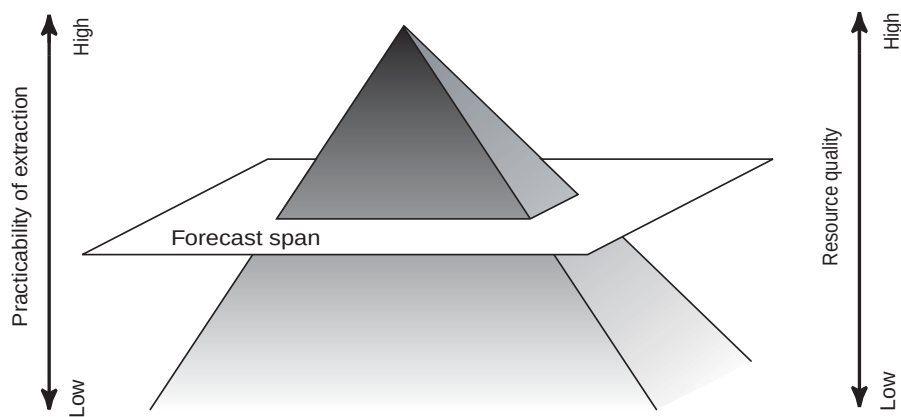


FIGURE 2 Pyramid of undiscovered petroleum resources.

Risking elements are used by the USGS to address the question of the probability of occurrence of at least one undiscovered accumulation of minimum volume or greater, or cell of minimum total recovery (or estimated ultimate recovery) or greater, somewhere in the area or rock volume being assessed that has the potential to be added to reserves in the time frame of the assessment. Probabilities for the occurrence of adequate charge, adequate rocks (reservoirs, seals, traps, etc.), and adequate timing of geologic events (burial, maturation, migration, trap formation, uplift, etc.) define geologic risk. The probability that essential petroleum exploration and production activities will be possible in the specified forecast span of the assessment, at least somewhere in the area or rock volume, is also assigned, defining access risk. Each element in this risking structure applies to the area or rock volume as a whole and is not equivalent to the percentage of the area or rock volume that might be unfavorable in terms of geology or access.

Estimates of oil and natural gas resources are termed conditional if oil and natural gas are assumed to be present in the subject area or rock volume to be assessed. Conditional estimates do not incorporate the risk that the area or rock volume may be devoid of oil or natural gas. If a risk that oil and natural gas may not be present is incorporated in the resource estimates, then these are termed unconditional or risked estimates. Statistically, the unconditional (risked) mean may be determined through multiplication of the mean of a conditional (unrisked) distribution by the related probability of occurrence.

7. PARAMETERS FOR CLASSIFICATION OF OIL AND NATURAL GAS RESOURCES

A resource is a concentration of naturally occurring solid, liquid, or gaseous hydrocarbons in or on the earth's crust, which is, or potentially is, economically extractable. Resources may be classified in terms of economic recoverability, geologic certainty, and uncertainty of the resource quantity.

The 1980 USBM/USGS classification system is based on two parameters whereby resources are classified by feasibility of economic recovery and degree of geologic certainty or assurance (originally called degree of uncertainty of existence) (Fig. 3). The SPE/WPC/AAPG have proposed a new system whereby petroleum resources may be classified by three parameters: feasibility of economic recovery, degree of geologic assurance, and degree of uncertainty of estimated resource quantities (Fig. 4). Although some differences exist between the two systems, the classification schemes are comparable.

Resources include reserves and all other petroleum accumulations that may eventually become available—including known accumulations that are not recoverable under current economic conditions or technology, or unknown accumulations of varying degrees of richness that may be inferred to exist, but not yet discovered. Therefore, resources can be classified in terms of geologic assurance as discovered (identified) and undiscovered.

According to the USBM/USGS, identified resources are those whose location, grade (for mineral

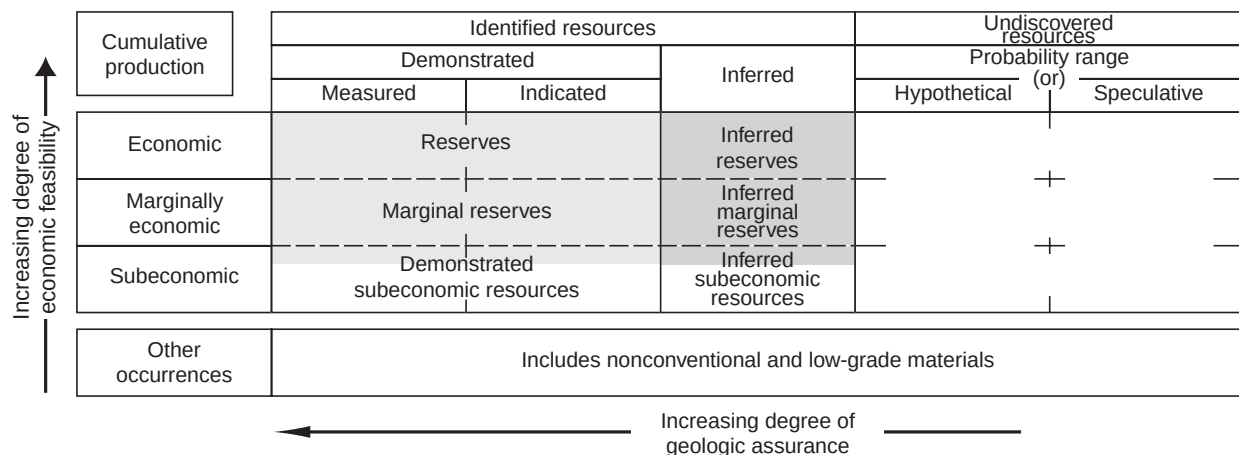


FIGURE 3 Diagram showing the resource classification system of the USBM/USGS. The terminology used in this classification system is from McKelvey (1972, 1973), but modified by Brobst and Pratt (1973) and adopted with minor changes for joint use by the USBM/USGS (1976, 1980).

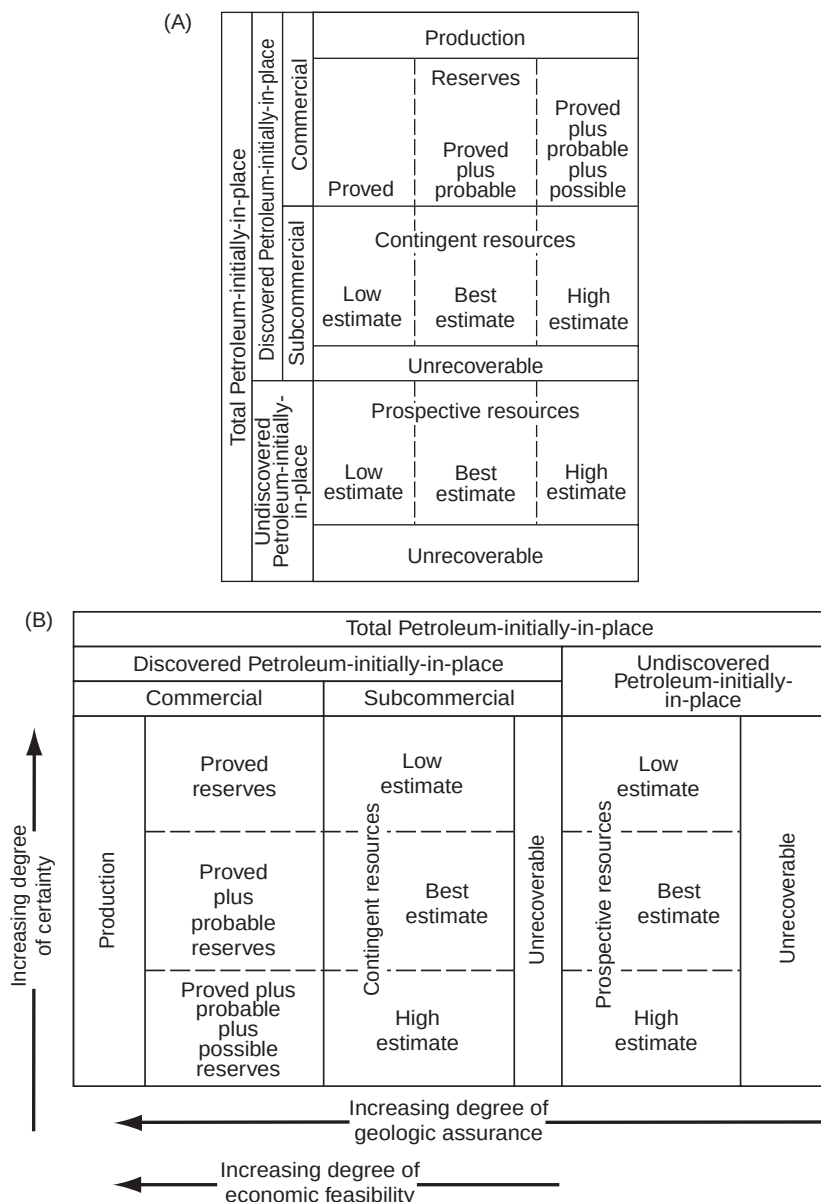


FIGURE 4 (A) Diagram showing the resource classification system of the SPE/WPC/AAPG. Reproduced from the Society of Petroleum Engineers, World Petroleum Congresses, and American Association of Petroleum Geologists (2001), "Guidelines for the Evaluation of Petroleum Reserves and Resources," with permission. Copyright Society of Petroleum Engineers. (B) Same diagram as in A, but slightly modified for comparison with Fig. 3.

resources), quality, and quantity are known or estimated from specific geologic evidence. Undiscovered resources are those that are inferred from geologic information and theory and comprise accumulations that are separate from identified resources [that is, existing outside of known oil and (or) natural gas accumulations]. Similarly, the SPE/WPC/AAPG define identified resources as discovered petroleum-initially-in-place. This is the quantity of petroleum that is estimated, on a given date, to be

contained in known accumulations, plus those quantities already produced therefrom. In addition, undiscovered petroleum-initially-in-place is defined as that quantity of petroleum that is estimated, on a given date, to be contained in accumulations yet to be discovered.

Resources are also classified in terms of feasibility of economic recovery as economic or commercial and subeconomic or subcommercial (Fig. 3). A third classification, marginally economic, had been used

by the USGS. The USBM/USGS classified both discovered (identified) and undiscovered resources in terms of economic recoverability. In contrast, the SPE/WPC/AAPG classify only discovered resources in terms of economic recoverability (Fig. 4).

A degree of uncertainty is typically reported for the estimated quantities of discovered and undiscovered resources that are potentially recoverable. The uncertainty of estimated resource quantities may be expressed probabilistically, either as a range or single-value statistic such as a mean, mode, median (P_{50}), or some other percentile.

8. CLASSIFICATION OF OIL AND NATURAL GAS RESOURCES

Discovered (identified) and undiscovered resources are incorporated in the classification systems of both the USBM/USGS and the SPE/WPC/AAPG. The USBM/USGS divide discovered (identified) resources in terms of geologic assurance into measured, indicated, and inferred classes (Fig. 3). Demonstrated resources are the sum of measured and indicated resources.

The estimated economically recoverable portion of discovered (identified) resources is classified as reserves. The part of the discovered (identified) resources from which reserves are estimated is called the reserve base (Fig. 3, shaded and striped areas). Discovered resources, in the SPE/WPC/AAPG classification, that are subeconomic or subcommercial are called contingent resources. This classification system also includes all produced petroleum and identified unrecoverable petroleum as discovered resources. The USBM/USGS system does not include the amount of past production as part of the existing identified resource. However, the produced volume of petroleum may be used to better understand the amount of resources.

Reserves are that part of the identified (discovered) resources that could be economically extracted or produced at the time of determination and include only recoverable quantities. Similarly, the SPE/WPC and SPE/WPC/AAPG define reserves as the estimated quantities of petroleum expected to be commercially recovered from known accumulations relative to a specified date. The American Petroleum Institute adds that reserves are anticipated to be commercially recoverable under prevailing economic conditions, operating practices, and government regulations. Uncertainty of reserve estimates depends on the amount of geologic and engineering data and

interpretations available at the time of the estimate. Based on the level of uncertainty, reserves are defined in classes of (1) proved (most certain), (2) probable, and (3) possible. Although the measured, indicated, and inferred classes of reserves are assigned to reflect geologic assurance, these classes have been loosely interchanged with, respectively, the proved, probable, and possible classes.

Proved petroleum reserves are the estimated quantities of crude oil, natural gas, and NGL, which geological and engineering data demonstrate with reasonable certainty to be recoverable in future years from known reservoirs under existing economic and operating conditions, that is, prices and costs prevailing at the time the estimate is made. The SPE/WPC and SPE/WPC/AAPG define proved reserves similarly, but with the addition that the estimate is made from a given date forward and under current economic conditions, operating methods, and government regulations.

Reserves are considered proved if economic producibility of the reservoir is supported by either actual production or conclusive formation tests. (See Fig. 5.) According to the SPE/WPC and SPE/WPC/AAPG, there should be at least a 90% probability that the quantities of oil or natural gas actually recovered will equal or exceed the estimate. Proved reserves (informally called 1P reserves) are typically reported for fields in the United States.

Proved reserves can be categorized as (1) developed—those expected to be recovered from existing wells; or (2) undeveloped—those expected to be recovered from new wells, deeper wells, well recompletions, or on installation of production or transportation facilities.

Probable reserves are those unproved reserves that are more likely than not to be recovered. Probable reserves should have at least a 50% probability that the quantities actually recovered will equal or exceed the sum of estimated proved plus probable reserves (Fig. 4). Volumes of proved plus probable reserves are greater than those of proved reserves alone. Proved plus probable reserves (informally called 2P reserves) are typically reported for fields in countries other than the United States.

Possible reserves are those unproved reserves that have at least a 10% probability that the quantities actually recovered will equal or exceed the sum of estimated proved plus probable plus possible reserves. Proved plus probable plus possible reserves (informally called 3P reserves) are not used in USGS assessments.

In contrast to reserves, contingent resources are estimated to be potentially recoverable from known

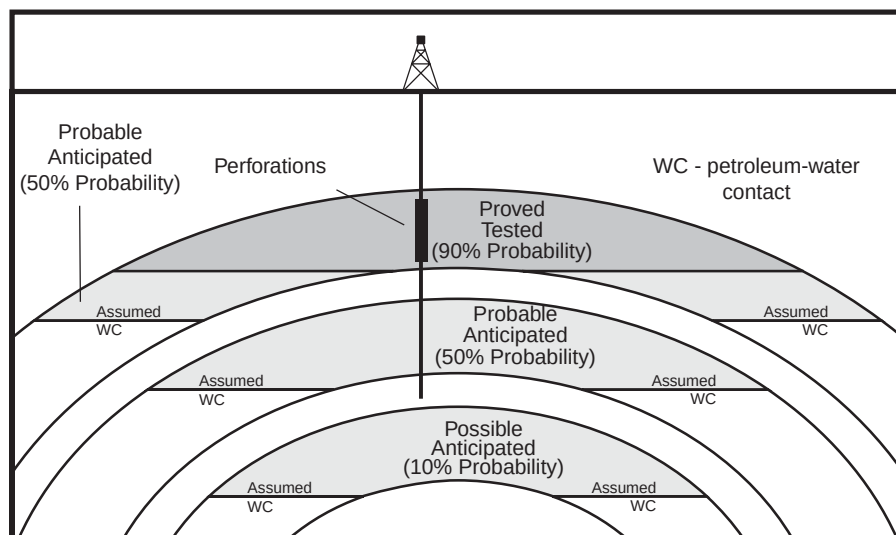


FIGURE 5 Proved, probable, and possible reserves.

(discovered) accumulations, but are not considered to be commercially recoverable.

Successive estimates of the total crude oil, natural gas, and natural gas liquids to be recovered in fields and reservoirs generally increase through time with continued development and production, the result commonly being additions of reserves. These additions are directly related to increases in the total size (cumulative production plus remaining reserves) of the field or reservoir. Reserve growth (also called field growth, reserve appreciation, and ultimate recovery appreciation) is therefore that part of the identified resources over and above measured reserves, estimated to be added to existing fields and reservoirs. Reserve growth occurs for a variety of geologic, engineering, operational, and economic reasons, including the following: (1) delineation of additional in-place oil and natural gas, including addition of new reservoirs and extensions; (2) improved recovery efficiency; and (3) revisions resulting from recalculation of viable reserves under dynamically changing economic and operating conditions.

The USBM/USGS divided undiscovered resources into hypothetical and speculative classes to reflect geologic assurance (Fig. 3). Hypothetical resources are undiscovered resources that may be reasonably expected to exist under geologic conditions analogous to those in known producing districts or regions. Speculative resources are undiscovered resources that may exist elsewhere, in districts or regions with no discovered resources. In the SPE/WPC/AAPG classification, undiscovered resources are divided into prospective resources, which are con-

sidered to be potentially recoverable, and resources that are unrecoverable (Fig. 4).

The USGS assesses the quantities of oil, gas, and NGL that are technically recoverable and have the potential to be added to reserves within some specified future time frame. The USGS defines technically recoverable resources as resources in accumulations producible using current recovery technology and in 1995 the U.S. Geological Survey National Oil and Gas Resource Assessment Team reported that "these are oil and natural gas resources that may be produced at the surface from a well as a consequence of natural pressure within the subsurface reservoir, artificial lifting of oil from the reservoir to the surface, and the maintenance of reservoir pressure by fluid injection. These resources are generally conceived as existing in accumulations of sufficient size to be amenable to the application of existing recovery technology."

USGS assessments consider three types of technically recoverable resources: (1) undiscovered conventional accumulations of oil and natural gas; (2) additions of oil and natural gas from untested cells within continuous accumulations; and (3) the potential future additions to reserves of known conventional accumulations by reserve growth. Each of these resource types requires a different technique for evaluation and assessment.

The oil and natural gas from undiscovered conventional accumulations are clearly equivalent to the undiscovered resource classification, whereas oil and natural gas from development of continuous accumulations may be equivalent to both discovered

and undiscovered classes (Figs. 3 and 4). Oil and natural gas from field or reserve growth would include inferred resources using the USBM/USGS classification (Fig. 3) or possible reserves and contingent resources using the SPE/WPC/AAPG classification (Fig. 4).

Although based on the best geologic and historical information and theory available, assessed petroleum volumes are unknown quantities, not measurements, and therefore should be expressed with probability distributions of uncertainty.

Acknowledgments

The definitions of the terms crude oil, natural gas, natural gas liquids, field, reservoir, reserves, and proved reserves are from "Standard Definitions of Petroleum Statistics" (1995), 5th ed. Reproduced courtesy of the American Petroleum Institute.

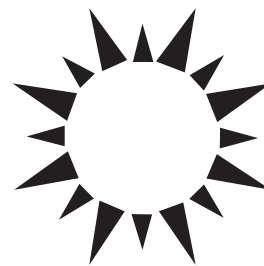
SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Geological Methods • Oil and Natural Gas Resource Assessment: Production Growth Cycle Models • Petroleum Property Valuation

Further Reading

- Ahlbrandt, T. S., and McCabe, P. J. (2002). Global petroleum resources: A view to the future. *Geotimes* 47, 14–18.
- American Petroleum Institute. (1995). "Standard Definitions for Petroleum Statistics." 5th ed. American Petroleum Institute, Washington, DC.
- Attanasi, E. D. (1998). "Economics and the 1995 Assessment of United States Oil and Gas Resources." U.S. Geological Survey Circular 1145. U.S. Geological Survey, Washington, DC.
- Brobst, D. A., and Pratt, W. P. (1973). Introduction. In "United States Mineral Resources" (D. A. Brobst and W. P. Pratt, Eds.), pp. 1–8. U.S. Geological Survey Professional Paper 820. U.S. Geological Survey, Washington, DC.
- Dolton, G. L., Carlson, K. H., Charpentier, R. R., Coury, A. B., Crovelli, R. A., Frezon, S. E., Khan, A. S., Lister, J. H., McMullin, R. H., Pike, R. S., Powers, R. B., Scott, E. W., and Varnes, K. L. (1981). "Estimates of undiscovered recoverable conventional resources of oil and gas in the United States." U.S. Geological Survey Circular 860. U.S. Geological Survey, Denver, CO.
- Gautier, D. L., Dolton, G. L., Takahashi, K. I., and Varnes, K. L. (eds.). (1996). "1995 National Assessment of United States Oil and Gas Resources—Results, Methodology, and Supporting Data." U.S. Geological Survey Digital Data Series DDS-30, Release 2 (CD-ROM).
- Lahee, F. H. (1944). Classification of exploratory drilling and statistics for 1943. *Am. Assoc. Pet. Geol. Bull.* 28, 701–721.
- Levorsen, A. I. (1967). "Geology of Petroleum." W. H. Freeman, San Francisco, CA.
- Magoon, L. B., and Dow, W. G. (eds.). (1994). "The Total Petroleum System—From Source to Trap." American Association of Petroleum Geologists Memoir 60. American Association of Petroleum Geologists, Tulsa, OK.
- Magoon, L. B., and Schmoker, J. W. (2000). The total petroleum system—The natural fluid network that constrains the assessment unit. In "U.S. Geological Survey World Petroleum Assessment 2000—Description and Results (U.S. Geological Survey World Energy Assessment Team, Eds.), U.S. Geological Survey Digital Data Series DDS-60 (CD-ROM).
- Masters, J. A. (1979). Deep basin gas trap, western Canada. *Am. Assoc. Pet. Geol. Bull.* 63, 152–181.
- McCabe, P. J. (1998). Energy resources—Cornucopia or empty barrel? *Am. Assoc. Pet. Geol. Bull.* 82, 2110–2134.
- McKelvey, V. E. (1972). Mineral resource estimates and public policy. *Am. Scientist* 60, 32–40.
- McKelvey, V. E. (1973). Mineral resource estimates and public policy. In "United States Mineral Resources" (D. A. Brobst and W. P. Pratt, Eds.), pp. 9–19. U.S. Geological Survey Professional Paper 820. U.S. Geological Survey, Washington, DC.
- Schmoker, J. W. (1996). Method for assessing continuous-type (unconventional) hydrocarbon accumulations. In "1995 National Assessment of United States Oil and Gas Resources—Results, Methodology, and Supporting Data" (D. L. Gautier, G. L., Dolton, K. I. Takahashi, and K. L. Varnes, Eds.), U.S. Geological Survey Digital Data Series DDS-30, Release 2 (CD-ROM).
- Schmoker, J. W. (1999). "U.S. Geological Survey Assessment Model for Continuous (Unconventional) Oil and Gas Accumulations—The 'FORSPAN' Model." U.S. Geological Survey Bulletin 2168. U.S. Geological Survey, Washington, DC.
- Society of Petroleum Engineers and World Petroleum Congresses. (1997). "Petroleum Reserves Definitions." Society of Petroleum Engineers, Richardson, TX. Available on the World Wide Web at www.spe.org.
- Society of Petroleum Engineers, World Petroleum Congresses, and American Association of Petroleum Geologists. (2001). "Guidelines for the Evaluation of Petroleum Reserves and Resources." Society of Petroleum Engineers, Richardson, TX. Available on the World Wide Web at www.spe.org.
- U.S. Bureau of Mines and U.S. Geological Survey. (1976). "Principles of the Mineral Resource Classification System of the U.S. Bureau of Mines and U.S. Geological Survey." U.S. Geological Survey Bulletin 1450-A. U.S. Geological Survey, Washington, DC.
- U.S. Bureau of Mines and U.S. Geological Survey. (1980). "Principles of a Reserve/Resource Classification for Minerals." U.S. Geological Survey Circular 831. U.S. Geological Survey, Washington, DC.
- U.S. Geological Survey National Oil and Gas Resource Assessment Team. (1995). "1995 National Assessment of United States Oil and Gas Resources." U.S. Geological Survey Circular 1118. U.S. Geological Survey, Washington, DC.
- U.S. Securities, Exchange Commission. (2002). Regulation S-X. U.S. Securities and Exchange Commission SEC 1887 (08-2002).

Oil and Natural Gas Resource Assessment: Geological Methods



DONALD L. GAUTIER

U.S. Geological Survey
Menlo Park, California, United States

1. Introduction

2. Assessment Methodology

Glossary

play A volume of rock having common geologic attributes with respect to oil and gas accumulations.

reserves That part of a resource that can be produced under existing conditions of technology and economics.

resources The natural occurrence of potentially useful hydrocarbon compounds.

trap The geometric arrangement of porous and permeable reservoir rocks and permeability barriers to hydrocarbon migration that results in an oil and/or gas accumulation.

1. INTRODUCTION

Since the beginning of the petroleum age, the utility of oil and gas has inflamed the desire to understand and quantitatively predict the resources yet to be found. Alongside the search for petroleum, the sciences of geology, geophysics, and organic geochemistry have grown to maturity in the effort to predict petroleum ahead of the drill. Consequently, the properties of petroleum are understood in great detail. The geological requirements for large oil and gas accumulations are quite specific and, as a result, resources are highly uneven in their distribution. Of the hundreds of sedimentary basins in the world, 5 of them contain almost one-half of the known petroleum and fewer than 50 basins contain well over 90% of the known oil and gas (Fig. 1).

Geological methods of resource assessment are based on the simple understanding that petroleum is a natural component of the earth and that its

quantity, quality, and distribution are the result of geological processes. Geological methods of resource assessment attempt to use the scientific underpinnings of geology and the empirical results of the search for petroleum to predict the oil and gas accumulations yet to be found.

1.1 Geological Controls on the Distribution of Oil and Gas Resources

Nearly all petroleum geologists agree that oil is formed from the thermo-chemical transformation of sedimentary organic matter. John Hunt states this view succinctly: "The overwhelming geochemical and geological evidence from both sediment and petroleum studies of the past few decades clearly shows that most petroleum originated from organic matter buried with the sediments in sedimentary basins."

Knowing the organic origin of oil and gas has led to extensive investigations of the properties of

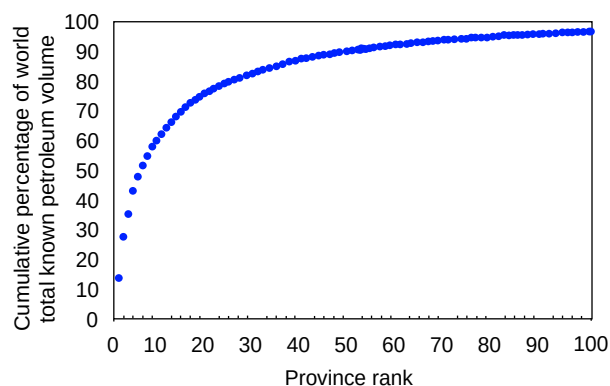


FIGURE 1 Cumulative percentage of world total known petroleum volume by top 100 ranked geologic provinces. Reprinted from U.S. Geological Survey World Assessment Team (2000).

carbon compounds in sedimentary sequences and to the concept of organic sedimentary facies. Recognition that areas of highly successful petroleum exploration (so-called fairways) are geographically related to places where particular types of organic-matter-rich sediments have been buried and heated (petroleum kitchens) has resulted in increased success in explorations.

Although an adequate concentration of organic matter is necessary, that alone is not sufficient for the existence of exploitable hydrocarbon accumulations. Oil generation occurs only from sedimentary rocks with certain types and concentrations of sedimentary organic matter under particular conditions of temperature and burial (the petroleum source rock). When the source rock attains the temperature required, hydrocarbon fluids and gases are squeezed out by compaction and expelled into adjacent permeable beds, such as porous limestone or sandstone.

Oil and gas have much lower density than the water that ordinarily fills the pore networks of sedimentary rocks. As a result, hydrocarbons that have been generated and expelled from the heated organic matter migrate upward under the effects of buoyancy. Most hydrocarbons generated from sedimentary organic matter probably migrate to the surface of the earth, where they form natural oil or gas seeps and are lost to erosion and oxidation. Less commonly, the geologic environment through which the hydrocarbons are moving forms a barrier to migration. A permeability barrier that slows or halts petroleum migration is called a seal and the porous rocks that contain petroleum are reservoirs. Taken together, the geometric relationship of reservoirs and seals is the trap. Structural traps, stratigraphic traps, hydrodynamic traps, and diagenetic traps are four well-known types of barriers to migration. A trap is the geometric arrangement of porous and permeable reservoir rocks and permeability barriers to hydrocarbon migration that results in an oil and/or gas accumulation.

Unless unusual conditions exist, oil, gas, and water are vertically arrayed in the trap in order of their buoyancy, with oil overlying water (the oil–water contact) and gas overlying oil (the gas–oil contact). The pay zone (the rocks from which oil and gas can be produced) may extend from the lowest part of the oil–water contact to the top of the gas accumulation. The volume of oil and gas in the trap (the size of the accumulation) depends on the availability of hydrocarbons, the properties of the reservoir, the dimensions of the trap, and the

permeability of the seal that slows further migration. Giant oil fields are found in close proximity to high-quality source rocks, where large geologic structures have an effective seal and where thick reservoirs with high porosity are present. Petroleum geologists and geophysicists are devoted to the search for such accumulations.

1.2 Resources to Reserves—The Intersection of Geology, Technology, and Economics

Despite the fundamental importance of understanding the geological distribution of oil and gas resources, a great deal of additional knowledge and effort is necessary for successful (profitable) production. Many petroleum geologists can cite interesting but frustrating anecdotes of “geological success” in which months or years of research and expensive exploratory drilling resulted in the discovery of hydrocarbons as predicted by geological considerations, but where technological requirements or market forces made production impossible. The petroleum accumulation may be in water too deep, or it may consist of hydrocarbons that are too viscous to flow to the wellbore, or numerous other possibilities may render the accumulation impossible to produce profitably.

For the purposes of resource analysis, a determination of the geological distribution of hydrocarbon molecules in the earth’s crust is insufficient. An attempt must be made to estimate that part of the resource that can be actually utilized. In a geologically based resource assessment, it is essential to distinguish between resources and reserves:

Resources are the natural occurrence of potentially useful hydrocarbon compounds.

Reserves are that part of a resource that can be produced under existing conditions of technology and economics

Resources are geological, but reserves are that subset of the resource that exists at the intersection of geology, available technology, and favorable economic conditions. In order to convert resources into reserves, three conditions must be met. First, the hydrocarbons must be present in geologically defined and identifiable accumulations. Second, technology must be available through which such accumulations can be exploited. Third, the economic situation of the accumulation must be such that the application of technology can result in profitable sale of the produced commodity. The principal goal of

geologically based resource assessments is the prediction of the sizes and numbers of such potentially developable accumulations in advance of their discovery.

A well-known example of the difference between resources and reserves comes from the Green River Basin (GRB) of Wyoming. Geological investigations of the GRB suggest that enormous quantities of natural gas may be present in thick sedimentary sequences containing gas-prone coals and carbonaceous mudstones (the source rocks) and low-permeability, nonmarine sandstones (the reservoirs). The total volume of natural gas resources in the GRB has been estimated to exceed 5000 trillion ft³ (TCF). This volume of natural gas is roughly equivalent to the total proved natural gas reserves of the world! However, initial resource assessments suggested that less than 120 TCF of this huge resource is recoverable using available technology, and of the technically recoverable resources, only a few TCF can be produced economically. In this example, the volumetric difference between resources and reserves ranges over three orders of magnitude.

In as much as hydrocarbon accumulations have been successfully explored and developed, a record of the conversion of geological resources to exploitable reserves exists in the form of drilling records, production histories, and reported field sizes. In combination with geological interpretation, analysis of exploration and production data provides a basis for prediction. Geological insight tempered with exploration and production experience facilitates and underlies all geological approaches to resource assessment.

2. ASSESSMENT METHODOLOGY

A wide variety of geologically based methodologies have been employed over the years to estimate quantities of undiscovered accumulations. In each case, geological information in some combination with exploration history (as opposed to direct extrapolation of drilling statistics, well production profiles, or economic projections) has been used to estimate future additions to oil and gas reserves. In all geologically based assessment methods, drilling and production data are viewed in an explicitly geological context. Geological understanding is necessary for resource prediction.

Geologically based methodological approaches can be grouped into a few basic themes:

1. Direct assessment and geological consensus techniques;
2. Trap-oriented assessments, including prospect appraisal, deposit simulation, and play analysis; and
3. Source rock/fluid geochemistry-oriented approaches, including petroleum system and material balance methods.

Each of these themes is typically used to some degree in the geologically based approaches and they are usually used in combination with one or more other techniques. In addition, geological information is more or less necessary to many numerically based, historical performance and areal or volumetric yield methods of assessment.

2.1 Direct Assessment and Geological Consensus Methods

Direct assessment is a subjective approach, in which knowledgeable individuals (experts) directly provide their views of the resource potential of a specified place or subject. Typically, a group of experts is assembled and these experts independently or jointly review the geological information available for a region, basin, prospect, or other subject of interest to the assessment. This review may include previous assessments, including nongeological assessments. Each expert arrives at an individual opinion or position concerning the resource potential. This opinion may be expressed as a simple point estimate or as a range of possibilities. Individually determined estimates can be statistically combined into a final probability density function that captures the full range of opinions held by the group.

In the geological consensus approach, which is a variant of the direct assessment method, the assembled group may review the individual estimates and make revisions where necessary. Depending on the rules established for the assessment, each member of the group may be required to agree to reach a consensus value, which is often achieved only after vigorous debate. Alternatively, final results may reflect the average of the group's opinions or they may agree that outlying high and low estimates will be omitted from the final results.

This direct, subjective approach is sometimes, perhaps derisively, called the Delphi method, after the famous Greek oracle located near the ancient city of the same name. The title rightly emphasizes the highly subjective nature of this approach, within which little or no documentation is required and

precisely reproducible results may be impossible to achieve. Direct assessment by experts is thus only as good as the expertise of the individuals involved. As noted by White and Gehman, “one must know how expert are the experts in order to assess the assessment.”

However, when the experts are indeed expert, the method readily captures their views. Direct assessment has the great advantage, when compared to other methodologies, of being applicable to any and all levels of data availability. An expert group can equally well assess a frontier province with little or no data or a highly developed and explored play, where thousands of wells and fields have been documented. The direct assessment method also has the advantage of being a rapid, relatively simple (and therefore inexpensive) procedure, one that is particularly effective for capturing the knowledge of highly divergent positions or organizations. The direct assessment approach has been used to great effect, for example, by the National Petroleum Council and by the Department of Energy Oil Resources Panel, when it was desired to represent the views of a wide range of experts from diverse organizations to form a group position on various issues.

2.2 Trap-Oriented Assessment

In this approach, which may include prospect appraisal, deposit simulation, and play analysis, specialists focus on anticipated geological properties of individual accumulations or collections of accumulations. Resource estimates are derived directly from geological concepts and principles, commonly by means of standard equations employed by petroleum engineers to evaluate discovered accumulations. This approach is commonly data- and labor-intensive, relying on seismic geophysical data to resolve individual prospects to be recommended for drilling. Alternatively, a more regionally based perspective may be used to predict the undiscovered resources in collections of prospects. In any case, the goal is to integrate the best available earth-science information and then to derive volumetric estimates from that information. In the play or prospect approach, the conditions of oil and gas entrapment are emphasized. In particular, the properties of potential reservoir rocks, the timing of formation of structural features that might serve as traps, and the efficiency of sealing lithologies are regarded as crucial to predicting volumes of undiscovered resources.

Many resource assessments have been based on the play and prospect concept. Although other interpretations exist, most resource analysts agree that a play is a volume of rock having common geologic attributes with respect to oil and gas accumulations. Common attributes usually include reservoir rock, trapping mechanism and geometry, and seal properties, as well as maturation and migration timing and geographic extent. Common, specific source rocks for hydrocarbon generation may or may not be essential to play definition.

Play analysis is the systematic attempt to estimate undiscovered resources within a well-defined play. Although in practice play-based assessments vary considerably in approach and results, most share the goal that variations in results should be the consequence of availability and quality of geological information rather than of the methodology.

Play analysis is derived directly from the exploration methods used by petroleum companies, but as an assessment technique, play analysis originated with methodology developed by the Geological Survey of Canada in the early 1970s. The approach has been applied in many guises since.

Many common features of play analysis can be illustrated by an example from a U.S. Geological Survey study of Arctic National Wildlife Refuge (ANWR) Area 1002 in Alaska.

For the ANWR assessment, plays were defined based on geological conditions; in most cases, the stratigraphic relationships of reservoir rocks and postulated trapping mechanisms were sufficient to distinguish among plays. For each play, the number of prospects (potential oil or gas accumulations) was estimated from interpretation of seismic data and geologic inference. The uncertainty in the number of prospects was evaluated, as was the range of values for each of the various geologic characteristics in the analysis. The link between the geological observations in related wells and rock outcrops and the geophysical measurements was provided by standard equations developed by the petroleum engineers working on the project.

Using Monte Carlo simulation (a commonly used numerical probabilistic technique), the numbers of prospects and their characteristics were combined with the range of (simulated) oil and gas accumulations. This simulation leads to the name “deposit simulation approach.” In the ANWR assessment, the geologic prospect characteristics were as follows: area of closure (the map view of the boundaries of the trap), reservoir thickness, reservoir porosity, water saturation, trap fill, and the distribution of

oil versus gas, along with some expected properties of trapped hydrocarbon fluids (such as sulfur content or oil density).

The initial step, as outlined in Fig. 2, was to define plays. In play definition, geologists specify the geologic properties of the play in terms of reservoirs, seals, and traps. Following play definition, the number and size of potential petroleum accumulations were estimated for each play as a probability distribution, based on the range of expected values for geologic characteristics such as reservoir thickness or porosity. These distributions were limited by a chosen minimum field size (the size of the smallest potentially viable accumulation in

the play). The resulting distributions were then subjected to a risking procedure designed to weigh the likelihood that geologic conditions were favorable to generate at least one accumulation larger than the minimum field size. Geologic risk was evaluated at both the play and the prospect levels and the resulting distributions and risks were combined to make an estimate of in-place petroleum resources for each play. Next, an estimated recovery factor (the fraction of oil in the trap that can actually be extracted by specified types of technology) is applied to the in-place petroleum resources of each play to calculate the technically recoverable petroleum.

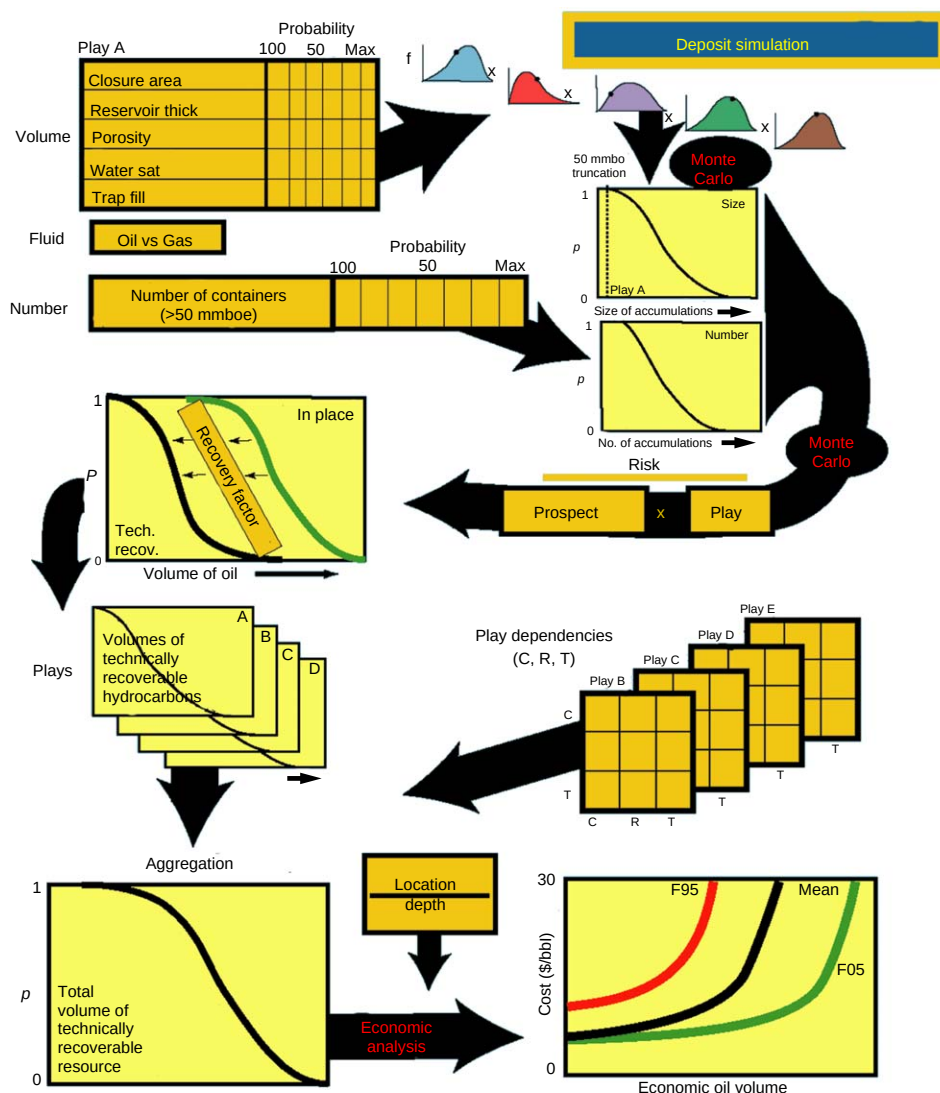


FIGURE 2 Flow chart showing the steps followed in the U.S. Geological Survey assessment of the Arctic National Wildlife Refuge 1002 Area. Reprinted from Bird (1999).

Estimates for each play are aggregated to determine the total resources of the area of assessment. Such aggregates are sensitive to geologic dependencies between the plays. For example, if two plays share the same source rock, inadequate charge in one play might indicate inadequate charge in all plays sharing that source rock. Such dependencies need to be considered and accounted for in the prospect simulation.

This sort of play and prospect simulation approach can be applied to frontier areas with little or no drilling data (as is the case with ANWR) or can be applied to areas of extensive exploration, where well data serve to reduce the uncertainty associated with the numbers and properties of known and undiscovered accumulations.

2.3 Source Rock Geochemistry and Material Balance Methods

Source rock geochemical and material balance methods of resource assessment are an attempt to quantitatively estimate volumes and distributions of hydrocarbons based on physical and chemical principles. In contrast to deposit simulation or play analysis, which emphasize traps and reservoirs, the source rock geochemical/material balance approach focuses first on the properties and processes of the petroleum source rocks and the related fluid system.

Although organic geochemistry has been used in the analysis of petroleum and petroleum products since the earliest days of the industry, systematic applications of geochemical methods to resource evaluation have come about only since the development of laboratory equipment suitable for the routine analysis of kerogens (specific organic components of sedimentary rocks, some of which yield hydrocarbons during thermal maturation) and oils. Such work began in the laboratories of major oil companies during the 1960s and 1970s. At that time, researchers were confident that they finally were beginning to understand the fundamental conditions of oil generation, expulsion, and migration.

2.4 Source Rock-Based Approaches: Material Balance Methodology

In the material balance approach, the volumes of hydrocarbons available to charge traps (the charge risk) are estimated through a series of estimates and

calculations. J. T. Smith and his colleagues at Shell's Bellaire Research Center near Houston, Texas, developed one of the earliest versions of the approach in the 1970s. The Shell group based their calculations on their definition of the petroleum system. The petroleum system was defined to include all related thermally mature source rocks, migration paths, reservoirs, seals, and hydrocarbons. The material balance calculation followed an explicit sequence of steps:

1. The description and definition of the petroleum system, based on concepts and data of the relationships between postulated source rocks and oils and gases inferred to have been generated from those source rocks.
2. A fetch area (G) of hydrocarbon generation is estimated for the defined petroleum system.
3. Primary migration efficiency (PM), the fraction of generated hydrocarbons that actually are expelled from the source rock, is estimated for the system.
4. The fraction of expelled hydrocarbons that moves through permeable migration conduits of the petroleum system is evaluated. This fraction is usually referred to as the secondary migration (SM).
5. An explicit estimate is made of loss (L) along the migration pathway because secondary migration incurs significant loss of hydrocarbons due to residual saturation along the migration route.
6. The charge available to the traps of the petroleum system is then calculated:

$$\text{Charge} = G(PM)(SM)L.$$

Advanced versions of the basic material balance approach have been some of the most complex and data-intensive methods ever applied to resource analysis. For example, the Prospect Appraisal by Quantitative Calibration System used by Royal Dutch/Shell in The Hague during the 1980s employed a Bayesian, calibrated geologic model. In this system, input parameters include charge, structure, reservoir properties, and retention (seal efficiency). The model follows a scheme of Bayesian updates and uses equations derived from geologic calibration studies, which in turn relied on statistical analysis of extensive data sets from Shell's worldwide exploration experience. The model was thus conceptually based on material balance, but was broadly experimental and empirical in its calibration and application.

These geochemical mass-balance approaches have the advantage of relying on concepts of physical geological processes, which, although subject to error, are likely to be less influenced by preconceptions than more subjective methods such as a Delphi assessment. The mass-balance method is scientifically elegant in its attempt to base resource evaluation on scientific principles. Geochemists and research-oriented geologists often favor this approach because, in principle, uncertainty in resource estimation can be systematically reduced through scientific investigation.

The principal disadvantages of the material balance approach are that it is time-consuming, data-intensive, expensive, and difficult to verify. Material balance studies, by definition, are state-of-the-art, with scientific updates constituting a fundamental part of the assessment process. Even after decades of research, uncertainties concerning types and distributions of organic matter and the physics of thermal maturation are great. Models of secondary migration are highly uncertain due to complexities in mapping fluid conduits and determining flow properties. Uncertainties surrounding the thickness, depth, porosity, seal competence, and timing of traps are as great in this approach as in the trap-based methodologies. For all these reasons and more, verifiable geochemical mass balance assessments remain a desirable, but as yet unattainable goal.

2.5 Resource Appraisal by Petroleum System

Some organizations have sought to combine the strengths of the geochemical material balance concepts with the empirical drilling and discovery data of the play- and prospect-based approaches. A widely (but by no means universally) adopted definition is as follows:

The petroleum system is a natural system that encompasses a pod of active source rock and all related oil and gas and which includes all the geologic elements and processes that are essential if a hydrocarbon accumulation is to exist.

The petroleum system as defined by Magoon and Dow is thus similar to the definition used at Shell in the 1970s. For each system, the geographic and stratigraphic extent of the fluids and source rocks is mapped and the maps provide the framework within which plays or assessment units can be evaluated. The petroleum system becomes the primary con-

ceptual driver for resource analysis and evaluation of charge risk becomes a critical concern. Once the petroleum system is defined and subdivisions of assessment units or plays are identified, resource assessment can proceed in a manner similar to that of the prospect simulation or play analysis methods. By directly relating exploitable accumulations to source rocks and their burial history, the play or prospect analysis is made more scientifically grounded in those geological processes responsible for petroleum formation. Despite sharing many common features with play analysis, the petroleum system approach unequivocally places emphasis on the source rock-hydrocarbon fluid relationship and all analyses are based on this pairing.

2.6 Significance and Use of Geological Methods in Historical Performance, Volumetric Yield, and Discovery Process Modeling Methods

Historical performance, volumetric yield, and discovery process modeling methods of resource analysis depend largely on the evaluation of drilling statistics, discovery rates, productivity rates, and known field size distributions. Although they do not incorporate geological information directly, each method explicitly or implicitly relies on geological information for success.

In the historical performance approach, historical data of drilling, discovery, and production may be mathematically fitted to various logistic or growth functions from which past performance can be extrapolated into the future. Probably the most famous of all oil and gas assessment methodologies is the historical performance model developed by M. King Hubbert, who extrapolated discovery and production data to more or less successfully predict the year of maximum U.S. oil production (1970). Hubbert's methods have had a resurgence of popularity as they have been applied to evaluations of global oil and gas data to predict future world oil production. The historical performance methods do not directly incorporate geological information and it is important to note that such techniques cannot be used in frontier areas where no exploration or production has occurred. Moreover, historical performance methods are really adequate only in areas in their later stages of exploration and development or in regions where a geologic analogue of the historical data can be reliably applied.

Discovery process modeling is a sophisticated approach to resource assessment, in which numbers and sizes of accumulations are extrapolated to future discoveries. Most discovery process models are based on numbers of wells drilled and areas remaining unexplored. As with other historical performance methods, discovery process modeling works best in areas of well-defined, consistent geology and a reasonably long history of exploration and development.

In the areal or volumetric yield methods, resources are calculated as the product of the land surface area or volume of rock under assessment, multiplied by a yield factor determined from analogous areas or rock volumes. An attempt is made to determine the quantities of oil yet to be found in untested areas or rock volumes, based on the volumes previously found in geologically similar settings. The results of such studies are typically expressed in terms of volumes of oil or gas per unit area or per unit volume. As with the historical performance models, yield methods do not directly utilize geologic information but rely on geological considerations in defining areas to be assessed. Areal or volumetric yield methods were used by some of the pioneering scientists in the field of resource assessment. Examples of early works include the studies of Weeks, Zapp, and Hendricks. Early approaches included worldwide average yields expressed in terms of barrels of oil per square mile of sedimentary rock, applied over the sedimentary cover of the entire planet.

Although such broadly generalized approaches are rarely used alone, the concept of undiscovered quantities of hydrocarbons per unit of untested rock remains fundamental to geologically based methods. The success of the areal or volumetric yield approach depends critically on the selection of meaningful geologic analogues; the method works only to the degree that explored and unexplored areas are comparable. The problem, of course, is that no two geologic provinces are identical.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Classifi-

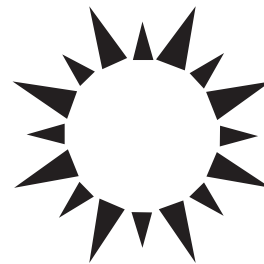
cations and Terminology • Oil and Natural Gas Resource Assessment: Production Growth Cycle Models • Petroleum Property Valuation

Further Reading

- ANWR Assessment Team (1999). "The Oil and Gas Resource Potential of the Arctic National Wildlife Refuge 1002 Area, Alaska," U.S.G.S. Open File Report 99-34. U.S. Geological Survey. [2 CD-ROMs]
- Attanasi, E. D. (1998). "Economics and the 1995 National Assessment of United States Oil and Gas Resources." U.S. Geological Survey Circular 1145, U.S. Geological Survey.
- Bird, K. J. (1999). Assessment overview. In "The Oil and Gas Resource Potential of the Arctic National Wildlife Refuge 1002 Area, Alaska" (ANWR Assessment Team, Eds.), U.S.G.S. Open File Report 99-34, pp. AO1-56. U.S. Geological Survey. [2 CD-ROMs]
- Brekke, H. (1989). Play analysis of the Norwegian North Sea: Appendix 12. In "Proceedings of the Sixth Meeting of the Working Group on Resource Assessment, Bangkok, Thailand, September 1989," United Nations Committee for Co-ordination of Joint Prospecting for Mineral Resources in Asian Offshore Areas (CCOP), CCOP Technical Secretariat.
- Campbell, C. J., and Laherrer, J. (1998). The end of cheap oil. *Sci. Am.* 278, 78-83.
- Deffeyes, K. (2002). "Hubbert's Peak: The Impending World Oil Shortage." Princeton University Press, Princeton, NJ.
- Demaison, G. J., and Huizinga, B. J. (1991). Genetic classification of petroleum systems. *Am. Assoc. Petroleum Geol. Bull.* 75, 1626-1643.
- Dolton, G. L., Bird, K. J., and Crovelli, R. A. (1987). Assessment of in-place oil and gas resources. In "Petroleum Geology of the Northern Part of the Arctic National Wildlife Refuge, Northeastern Alaska," US Geological Survey Bulletin 1778 (K. J. Bird and L. B. Magoon, Eds.), pp. 277-298. U.S. Geological Survey.
- Gautier, D. L., Dolton, G. L., Varnes, K. L., and Takahashi, K. I. (1995). "National Assessment of Oil and Gas Resources," U.S. Geological Survey Digital Data Series 30. U.S. Geological Survey. [CD-ROM]
- Hunt, J. M. (1996). "Petroleum Geochemistry and Geology." 2nd ed. Freeman, New York.
- Law, B. E., Spencer, C. W., Charpentier, R. R., Crovelli, R. A., Mast, R. F., Dolton, G. L., and Wandrey, C. J. (1989). Estimates of gas resources in over pressured low-permeability Cretaceous and Tertiary sandstone reservoirs, Greater Green River Basin, Wyoming, Colorado, and Utah. In "Gas Resources of Wyoming: 40th Wyoming Geological Association Field Conference, Casper, Wyoming, 1987" (J. L. Eisert, Ed.), pp. 39-62.
- Magoon, L. B., and Dow, W. G. (1994). "The Petroleum System—From Source to Trap." American Association of Petroleum Geologists Memoir 60, American Association of Petroleum Geologists, Tulsa, OK.
- National Petroleum Council. (1991). "U.S. Arctic Oil and Gas." National Petroleum Council, Washington, DC.
- Oil Resources Panel. (1993). "An Assessment of the Oil Resource Base of the United States," U.S. Department of Energy Publication DOE/BC-93/1/SP. U.S. Department of Energy.
- Sluijk, D., and Nederlof, M. H. (1984). Worldwide geological experience as a systematic basis for prospect appraisal. In "Petroleum Geochemistry and Basin Evaluation," American

- Association of Petroleum Geologists Memoir 35 (G. Demaison, and R. J. Murris, Eds), pp. 15–26. American Association of Petroleum Geologists.
- Ulmishek, G. F., and Klemme, H. D. (1990). “Depositional Controls, Distribution, and Effectiveness of World’s Petroleum Source Rocks.” U.S. Geological Survey Bulletin 1931, U.S. Geological Survey.
- U.S.G.S. World Assessment Team. (2000). “U.S. Geological Survey World Petroleum Assessment 2000—Description and Results,” U.S. Geological Survey Digital Data Series DDS-60. U.S. Geological Survey. [4 CD-ROMs]
- White, D. A. (1988). Oil and gas play maps in exploration and assessment. *Am. Assoc. Petroleum Geol. Bull.* 72, 944–949.

Oil and Natural Gas Resource Assessment: Production Growth Cycle Models



JEAN LAHERRERE

Association for the Study of Peak Oil and Gas
Uppsala, Sweden

-
1. Quality of Discovery and Production Data
 2. How Many Models?
 3. Cases
 4. Other Models and Ultimate Assessments
 5. Conclusions

Glossary

creaming curve Displays the cumulative discoveries versus the cumulative number of exploratory wells (new field wildcats).

cumulative production The addition of all production for a given area at a given date, from the start of production to this date.

Hubbert curve A bell-shaped curve representing annual production versus time.

parabolic fractal A display in a size rank (by decreasing size), log-log format showing a curved plot.

production growth cycle Where production starts at zero and goes up and down to end back at zero again, but possibly with several peaks.

proven reserves Those quantities of petroleum that, by analysis of geological and engineering data, can be estimated with reasonable certainty to be commercially recoverable from a certain date forward from known reservoirs and under current economic conditions, operating methods, and government regulations.

reserve additions For a given year, the addition in the reserve reporting by means of adjustments, revisions of past estimates (net of sales and acquisitions of fields), and new field discoveries.

ultimate recovery The final cumulative production when production is completely abandoned.

The goal of oil and gas modeling is to forecast future production, and the aggregation of this future production represents reserves. Oil has to be

discovered before it is produced. Discovery (i.e., reserves) and production need to be studied together. A model has to be represented by the simplest equation that best fits the past discovery and production, starting when the volume was significant. The value of a model is mainly the value of the data. Some modeling may be ignored, not because of the model but rather because the data are poor or incomplete. There is no theoretical justification for any model; only the fit between the model and the past data justifies its value.

1. QUALITY OF DISCOVERY AND PRODUCTION DATA

Oil production has several problems in terms of definition. Oil is not produced; rather, it is extracted, except when it is synthesized from coal or natural gas by chemical reactions. The term “oil” may represent crude oil only (including lease condensate for some countries such as the United States), totaling approximately 65 million barrels/day (25 billion barrels/year) of the world supply, or the demand for all liquids (including synthetic oil, condensate and natural gas plant liquids, and refinery gains), totaling approximately 75 million barrels/day (28 billion barrels/year). Oil can be reported in volume (cubic meters or barrels) or in weight (tonnes or metric tons), making the world total difficult to estimate when density is not well known or is ignored. Furthermore, some oil is lost (e.g., 2 billion barrels during the Gulf War), stolen, or unaccounted for when “cheating” occurs due to quotas. The inaccurate reporting by the International Energy Agency (IEA) in 1998 led to 600 million “missing

barrels” between supply and demand, giving a false impression of an abundance of oil and a low price of \$10/barrel. Only scout companies (as petrologistics) can provide fairly accurate volume estimates of the oil being shipped.

Models need to study the past reserves with time. Reserves are, in fact, the addition of all future production until the end. The production of today was part of the reserves of yesterday. It is imperative that reserves and production represent the same product. In reality, they do not. The most often used databases are the *Oil & Gas Journal* (OGJ) published every December, the *BP Statistical Review* (BP) published every June, *World Oil* (WO) published every August, and the Organization of Petroleum Exporting Countries (OPEC) (Table I).

For the world’s oil, the discrepancies are striking. In addition to unrealistic figures, BP specifies that tar sands are not included in their reserves but that they are included in their production. Most published

data are incoherent and/or badly defined. It is possible that the authors of these publications do not understand the importance of accuracy and/or do not want to define what is measured because they do not know.

Oil discovery has to be studied carefully. Unfortunately, the reporting of reserves to the public is a political act, and operators are reluctant to tell the truth if doing so can damage their images. Operators in the United States have to comply with Securities and Exchange Commission (SEC) rules and report only proven reserves, thereby omitting probable reserves, whereas the rest of the world reports both proven and probable reserves (close to the mean or expected value). The SEC shuns the probabilistic approach for known reserves, whereas the 2000 U.S. Geological Survey (USGS) largely uses probabilistic simulation to evaluate the undiscovered reserves. The volume of proven reserves is left to the appreciation of what is a “reasonable certainty to exist,” and this can vary between 51 and 99%. (The same definition of “reasonable certainty of no harm” is used by the U.S. Food and Drug Administration [FDA] to allow the sale of new products.) Figure 1 plots the remaining oil reserves from political sources (being current proven) published by OPEC, BP, OGJ, and WO; these reserves are completely different from the technical reserves that are the present mean values backdated to the year of discovery and are confidential. Statistically, the mean (expected) values do not grow despite the fact that some values will increase and some will decrease. On the contrary, proven reserves are very conservative and, thus, are

TABLE I

Common Databases for Oil Production and Reserves

	Reserves (billions of barrels)	Production (millions of barrels/day)
OGJ end 2002	1 212.880 852	66.042 7
BP end 2002	1047.7	73.935
OGJ end 2001	1 031.553 477	63.694 8
BP end 2001	1050.3	74.350
WO end 2001	1 017.763 1	67.700 444
OPEC end 2001	1 074.850	65.498

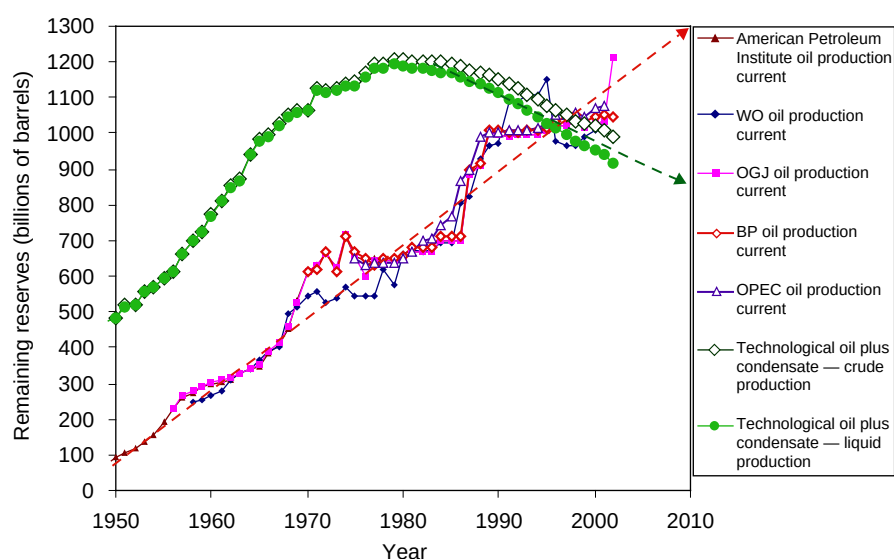


FIGURE 1 World remaining oil reserves from political and technical sources.

designed to generate growth, and this pleases most everyone (e.g., operators, bankers, shareholders). Unfortunately, although during the past quarter-century the U.S. reserve additions have come mainly from revisions of past discoveries as positive revisions were twice as large as negative revisions, the U.S. Department of Energy's (DOE) 2001 annual report indicated that negative revisions were larger than positive ones, with the Minerals Management Service (MMS) in the Gulf of Mexico reporting the same things since 1998. Only in 1990 did DOE publish "mean" estimates of annual discoveries from 1900 to 1988. These values were used, as were the annual new discoveries up to 2001, to get the present U.S. mean values by combining them with MMS reserve growth. It is unfortunate that such mean value data known by DOE are kept confidential. It is also unfortunate that the United States uses an old method to evaluate its reserves rather than using the approach used by the rest of the world.

The technical reserves decline since 1980 indicates that discovery has been lower than production since that time, but in fact the political reserves jumped by 300 billion barrels during the second half of the 1980s when OPEC quotas were decided on the reserves. OPEC countries have increased their reserves by 50% during some years (except in the neutral zone owned 50/50 by Saudi Arabia and Kuwait because the two owners disagree on the year of increase) when in fact no significant discoveries occurred during those years. Exxon-Mobil publishes world oil discoveries that stand in very close agreement with the technical reserves displayed in the Fig. 1, declining since 1980 as annual discovery became smaller than annual production.

Recently, OGJ increased Canada's reserves by 175 billion barrels to include the tar sands (extracted mainly by mining) that were previously excluded as unconventional because the Orinoco extra-heavy oils are now being produced in very conventional ways. BP, which follows OGJ for OPEC members, refused to include this large increase for the tar sand reserves even though the oil from the tar sands was included in BP production.

2. HOW MANY MODELS?

There is no theory to fully explain the best model to assess oil and gas discovery and production. The best approach is to deal with the most homogenous data covering the most natural area. In the past, exploration basins were explained in terms of tectonics.

Today, they are described as "petroleum systems" in terms of generation of hydrocarbons because the most important factor is the source rocks that have generated oil and gas. Gathering fields by large petroleum systems gives much better results than does gathering fields by country. However, exploration activities depend strongly on the country (e.g., opening of blocks, fiscal terms). It is better to deal with large areas such as continents. The data should also be considered over the full life of exploration; too many studies consider only the past decade or so. The best way is to study the past and extrapolate toward the future with the best simple model. However, finding a good fit for a particular set of data does not mean that this model is necessarily the solution; perhaps another model will provide a better solution. Comparing oilfield distributions can be done with different models such as log-normal and parabolic fractal. The Gulf of Mexico, Niger delta, and Saharan Triassic petroleum systems were compared in a study. When comparing with a log-normal model, the Gulf of Mexico is different from the Niger delta and Saharan Triassic (which are similar). This model considers mainly the frequency of activity, and the Gulf of Mexico is explored intensively in search of small fields, whereas in Niger and the Sahara only large fields are drilled. When comparing with a parabolic fractal model, the Gulf of Mexico and Niger are similar in that they are dispersed habitats with many large fields of similar size, whereas the Sahara is a concentrated habitat with few giant fields. Finding a solution with one model should not keep one from looking at other models to find another solution given that nature is not linear and often involves several solutions.

The range of production and discovery models is large (e.g., creaming curve with hyperbola, cumulative production and discovery with logistic curve, annual discovery and production with normal curve or derivative of logistic (Hubbert curve), parabolic fractal for field size rank in a log-log display, log-normal distribution, stretched exponential). Each case is particular and cannot be generalized. Instead of speaking about the strengths and the weaknesses of each model, many cases should be shown to determine exactly how good the model is.

One of the most famous models for oil production is the Hubbert model, which is a bell-shaped curve that peaks at its midpoint. Hubbert was a brilliant geophysicist (for Shell and USGS) but had a short temper. He is well known by geologists for his theory on hydrodynamics, that is, tilting the water contact in oilfields when the aquifer is strongly moving. He is

also well known in the oil industry for his forecast in 1956 that U.S. oil production would peak in 1970. His theory was rejected when published but was acknowledged later when the U.S. production did in fact peak in 1970. Buzz Ivanhoe created a Hubbert Center for petroleum supply studies at the Colorado School of Mines and issues a quarterly newsletter.

2.1 Hubbert Curve: One Peak?

Hubbert, in a famous 1956 paper, contended that oil has to be found first before it is produced and that the production pattern has to be similar to the discovery pattern. He stated that production starts from zero, rises to a peak, and declines back to zero, but he did not give any equation except the bell-shaped curve. He said that the curve will likely be symmetrical. His Hubbert curve was obviously drawn by hand with an abacus, and he measured the volume by the surface below his curve, using as a unit a square in billions of barrels in the corner of the graph. He did not show any graph with several peaks, although he did not dismiss such a possibility.

Hubbert's 1956 forecast was based on the assessment that the U.S. oil ultimate could be 200 billion barrels, giving a peak in 1970 (but also 150 billion barrels). It was not until the 1980s that he wrote that his curve was the derivative of the logistic function. The logistic function was introduced in 1864 by the Belgian mathematician Verhulst as a law for population growth. The equation for the cumulative

production (CP) for an ultimate U is

$$CP = U / \{1 + \exp[-b(t - tm)]\},$$

where tm is the inflexion point (corresponding to the peak time for the annual production). A population with a constant growth rate (e.g., bacteria) displays an exponential growth until a point where the population reaches the limit of resources and starts a constant decline to stabilize at a constant level. A constant growth is impossible in a limited universe. Bacteria doubling each half-hour, without being constrained by the food resource, will reach the limit of the solar system in a week and the limit of the universe in 11 days.

The derivative of the logistic curve is a bell-shaped curve very close to a normal (Gauss) curve. Its equation for the annual production P , peaking at a value Pm for the time tm , is as follows:

$$P = 2Pm / \{1 + \cosh[-b(t - tm)]\}.$$

When plotting the annual/cumulative production in percentage versus the cumulative production, if the production follows a derivative of logistics, the plot is linear. Using the mean values for the continental United States (48 states) as indicated previously, the plot is nearly linear from 1937 to 2001. The linear extrapolation toward zero indicates an ultimate of approximately 200 billion barrels. Hubbert was right on the peak in 1970 when he used an ultimate of 200 billion barrels (a rounded-up value because he knew that the accuracy was low). In Fig. 2, the linear trend

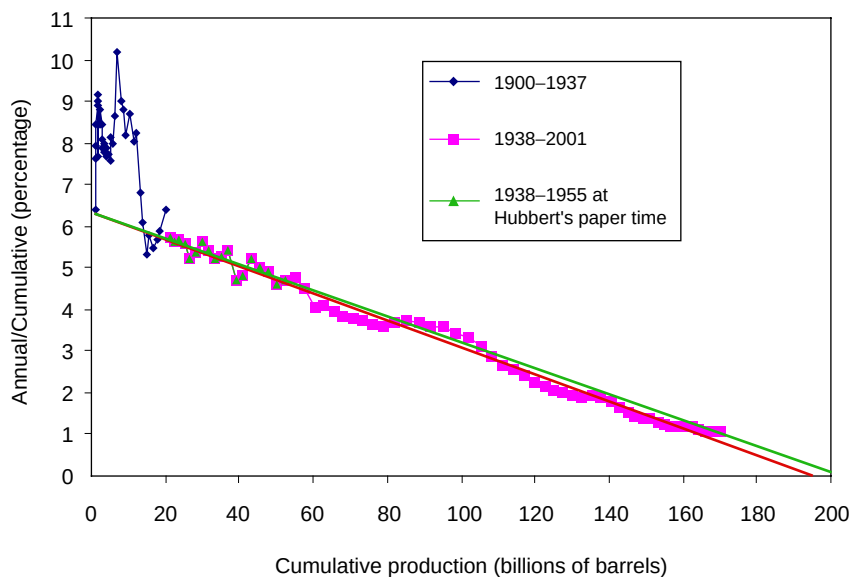


FIGURE 2 Continental U.S. (48 states) oil production: Annual/Cumulative versus cumulative.

from 1938 to 1955 (at the time of the Hubbert forecast) already indicates this 200-billion barrel ultimate. Hubbert was lucky that the real value was close to a rounded number.

Plotting the annual mean discovery, it is easy to draw a Hubbert curve that fits the discovery data (smoothed over a 5-year period) and has an ultimate of 200 billion barrels (Fig. 3). U.S. discovery in the 48 mainland states peaked around 1935 (the largest oilfield, in East Texas, was discovered in 1930) at a level of 3.2 billion barrels/year. This Hubbert discovery curve, shifted by 30 years, fits the production curve perfectly. Why is oil production in the continental United States as symmetrical in the rise as in the decline?

Plotting cumulative production versus time shows a perfect fit using one logistic curve with an ultimate of 200 billion barrels, whereas cumulative discovery needs two logistic curves: the first with an ultimate of 150 billion barrels and the second with an ultimate of 50 billion barrels.

Oil production in the mainland United States comes from more than 22 000 producers (a very large number), and randomness has to be taken into account, as in the air where there is a very large number of molecules with a random Brownian move that gives a perfect law among pressure, volume, and temperature. According to the central limit theorem (CLT), in probability, the addition of a large number of independent asymmetrical distributions gives a normal (symmetrical) distribution. The large number of U.S. independent producers leads to random behavior, and the aggregation of the very large

number of fields is normal. In terms of the previous modeling of production by a bell-shaped curve, it occurs that the model seems wrong around 1930 (Great Depression), 1950 (pro-rationing); and 1980 (oil prices) as well as when political or economic events obliged all operators to act in the same direction so that randomness cannot apply.

It is easier to work on cumulative discovery (CD) and production (CP) as small details are smoothed. In the continental United States, the discovery can be well modeled with two logistic curves for an ultimate of 200 billion barrels when only one logistic curve fits the production well (Fig. 4).

2.1.1 Multiple Cycles

But there are few places with such a large number of producers and continuous activity for more than a century as in the continental United States. Most other countries display several cycles of exploration activity and then of production. One good example is the nearly depleted oil production in France. A graph of the 1995 “World Oil Supply 1930–2050” for France was illustrated in a 1999 OGJ article by the author titled “World Oil Supply: What Goes Up Must Come Down, but When Will It Peak?” France’s oil production was modeled with two cycles fitting the past production (up to 1994) and forecasting approximately 10 million barrels/year in 2000. Figure 5 is an update with the real 2000 value of exactly 10 million barrels. France’s oil production, if a new cycle is not found, will cease in 2010. The modeling of the production gives a good fit for an 800-million barrel ultimate, whereas the ultimate

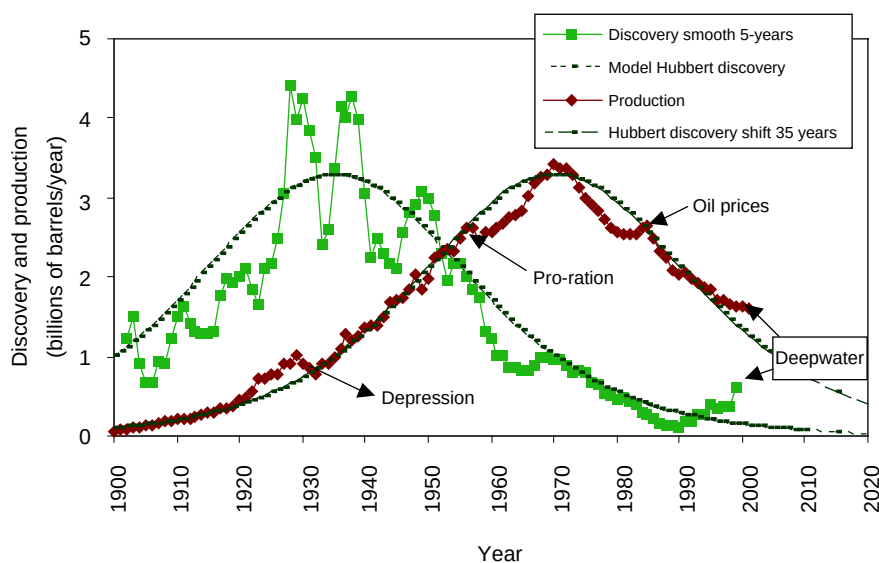


FIGURE 3 Continental U.S. (48 states): Annual oil discovery and production.

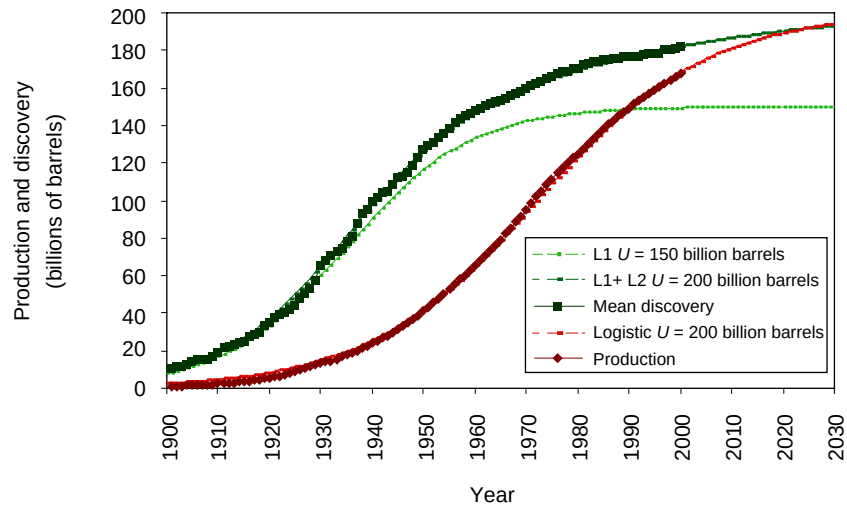


FIGURE 4 Continental United States (48 states): Cumulative oil discovery and production.

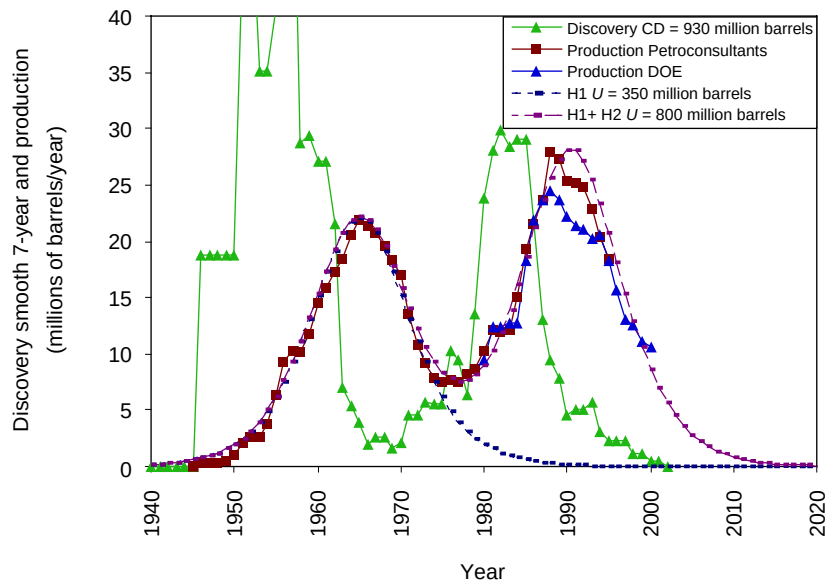


FIGURE 5 France oil discovery and production.

from discovery is higher at approximately 950 million barrels. Discoveries seems to be somewhat overestimated. But the fit between production data between Petroconsultants and DOE is poor for the first half of the 1990s, confirming the poor quality of published data.

2.2 Ultimate

Drawing a good model for oil production implies estimating the oil ultimate, for example, 200 billion barrels for the continental United States. The plot of

annual/cumulative percentage versus cumulative needs to be linear to give a reliable value, but that is not always the case. The best approach is the creaming curve. This curve was “invented” by Shell during the 1980s, displaying the cumulative discoveries versus the cumulative number of new field wildcats (NFWs) so as to eliminate the ups and downs of exploration when plotting versus time. When NFWs are not available, the plot versus number of discoveries (in time sequence) can give good results. Some economists call the creaming curve the curve of cumulative discoveries

versus the rank of the field by decreasing order (as in the fractal display). This is an inaccurate label because the curve varies with new discoveries as rank changes.

2.2.1 Creaming Curve

All of the creaming curves the author modeled were easily fitted with several hyperbolas, each representing a new cycle (as shown previously for the United States). The only problem is needing to guess whether a new cycle is possible, and only geologists can answer that by studying the geological potential. In the case of the Middle East, the last cycle is from 1974 to 2002 (28 years) corresponding to 2500 NFWs having an ultimate of 50 billion barrels with 830 discoveries, whereas earlier (70 years) the ultimate was 820 billion barrels corresponding to 1400 NFWs and 340 fields. This means that from 1974 to 2002, the success ratio was 33% but there was a potential of 120 million barrels per discovery, whereas earlier the success ratio was only 24% but the average field was approximately 600 million barrels. It is obvious that between 1974 and 2002, more fields were found but were much more smaller. It is inaccurate to say, as is often done, that the Middle East was poorly explored and has a huge potential (Fig. 6). The only poorly explored place is the Western Desert in Iraq, but the petroleum system is perfectly known and is mainly gas prone, the oil potential is small, and there is almost no deepwater.

2.2.2 Parabolic Fractal

Another way in which to assess the ultimate of a petroleum system (taking only a part of the system will disturb the natural behavior) is to plot in fractal display the field size versus the rank of the field by decreasing order in a log-log format. The linear fractal (power law), as developed by Benoit Mandelbrot, is only a theoretical interpretation given that every natural system displays a curved pattern. Natural distributions are fractal, and self-similarity is the rule of nature. On an illustration of a geological outcrop, a scale is needed. Fractal equations are used in movies to create “natural background.”

Every decade, the study of the evolution of discoveries leads to the drawing of the parabola representing the ultimate in the ground. The “yet to be found” can be deducted from the “yet in the ground” with some economic and time constraints. The example of the Niger delta petroleum system (covering three countries) shows that the 1960 pattern was quickly replaced by a pattern where the largest fields already found did not change much and only smaller fields were added (Fig. 7). The parabola represents the ultimate in the ground representing a large number of small fields that will be never discovered. The slope at the beginning (for the first 10 fields) characterizes the habitat, where it is a dispersed habitat (the slope is nearly flat), in contrast to a concentrated habitat (e.g., the Saharan Triassic, which shows a steep slope).

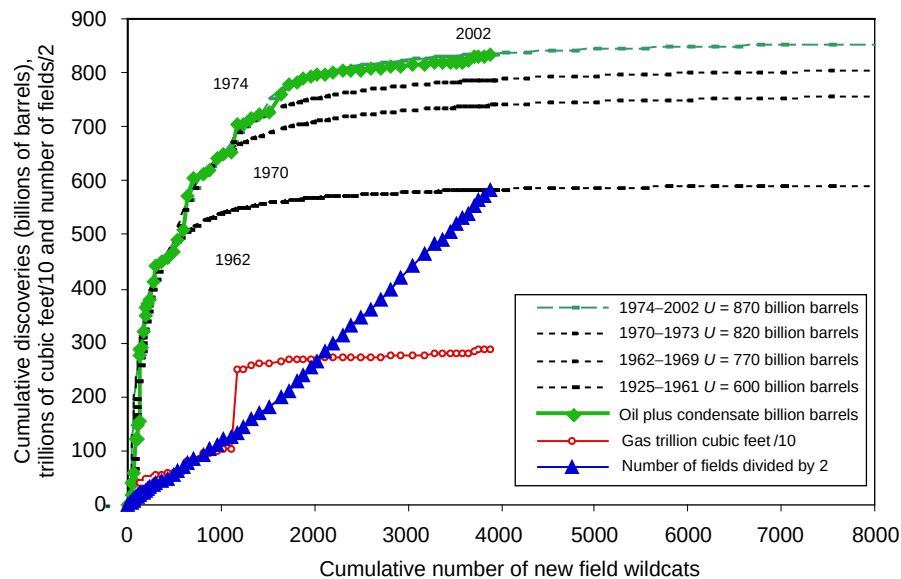


FIGURE 6 Middle East oil and gas creaming curve.

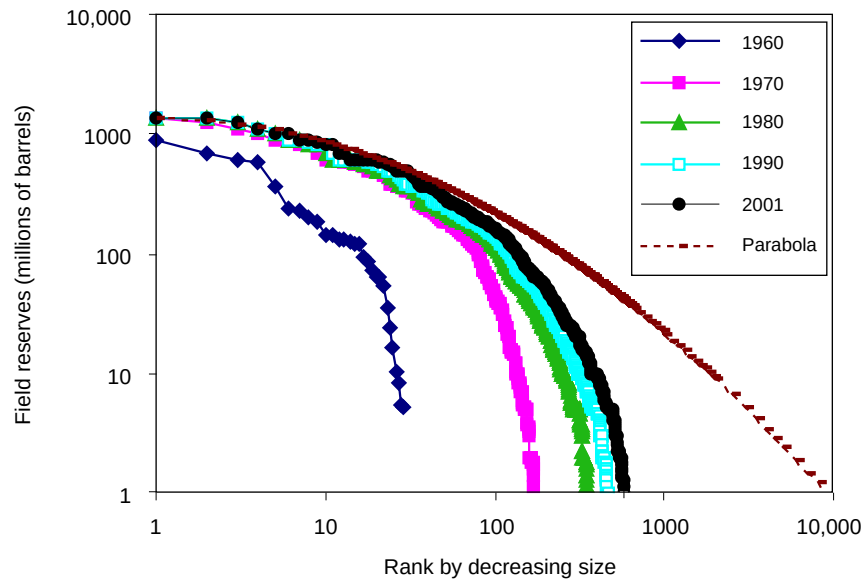


FIGURE 7 Niger delta oilfield size fractal display.

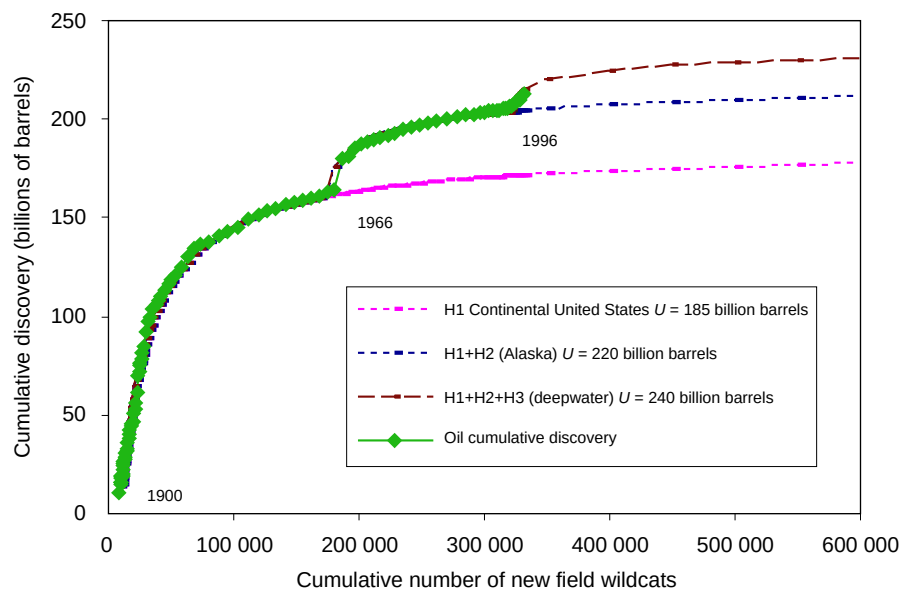


FIGURE 8 U.S. oil creaming curve.

3. CASES

3.1 United States

The creaming curve for the full United States (all 50 states) displays three cycles, well fitted to data with hyperbola curves. It is easy to recognize a first cycle for the continental United States (48 states), a second cycle for Alaska, and a third cycle for deepwater. Without a new cycle, the U.S. oil ultimate is approximately 225 billion barrels. The creaming

curve represents the well-known law of diminishing returns in mineral exploration. Big gold nuggets are found first, and fine particles are found at the end. The creaming curve indicates that finding an additional 10 billion barrels will require more than 50 000 NFWs (Fig. 8).

The annual/cumulative versus cumulative plot is nearly linear from 1940 to 2001, confirming an ultimate of approximately 220 billion barrels, very different from the ultimate estimated by the USGS 2000 report of an unrealistic 362 billion barrels (of

which 83 billion barrels were undiscovered and 76 billion barrels were of reserve growth). This annual/cumulative versus cumulative plot could be good in the case of one cycle, but it could prevent the new cycle from being seen coming or the deviation from the logistic function.

With a U.S. ultimate of 220 billion barrels coming from the production trend, it is easy to plot a logistic curve for the cumulative production. But it is interesting to compare it with the cumulative mean discovery. This mean discovery comes from the 1990 Energy Information Administration (EIA) report and the last annual reports that are grown to get the mean. This estimate is questionable, but it is striking to compare discovery shifted by 30 years with the production. Despite the jump due to Prudhoe Bay, the shifted discovery fits the production very well.

The published annual proven reserves plus cumulative production are plotted in Fig. 9 and in fact follow roughly in parallel the cumulative production, and a linear extrapolation for the last 30 years will go to 280 billion barrels in 2030. This is much lower than the estimates of the 2000 USGS with an ultimate in 2025 of 362 billion barrels (of which 76 billion barrels is for the so-called reserve growth), and even the ultimate without reserve growth at 296 billion barrels looks too high. The proven reserves are a poor estimation of the future and should be replaced by the practice of using mean reserves, as is done by the rest of the world.

3.2 World Oil

The world cumulative conventional oil discovery and production are fairly well modeled, with one logistic curve having an ultimate of 2 trillion barrels but with different slopes (Fig. 10). Again, the USGS ultimate of 3.012 trillion barrels looks unrealistic in light of the past mean discovery, but when the questionable reserve growth is removed, it reduces to a more acceptable 2.3 trillion barrels.

The cumulative discovery can be broken down by size of fields, separating the large fields of more than 2 billion barrels (which make approximately half of the total volume), the giants (500–2000 million barrels), the majors (100–499 million barrels), the rest (0–99 million barrels) (Fig. 11). All curves can be modeled well with one logistic, with the largest having reached close to the ultimate first. It is the principle of the creaming curve or the law of diminishing returns. Global results give a simple model as it smoothes the discrepancies.

The logistic model can also model the total primary energy consumption. (Unfortunately, these data omit the muscular energy from humans and animals—energy that built many empires—with only two cycles with an ultimate, in fact a peak, of 12 billion tonnes.) The second cycle is associated with the demographic explosion (Fig. 12).

When the ultimate has been assessed, for the world liquids, as 3 trillion barrels, the best way in which to plot the possible future production is as

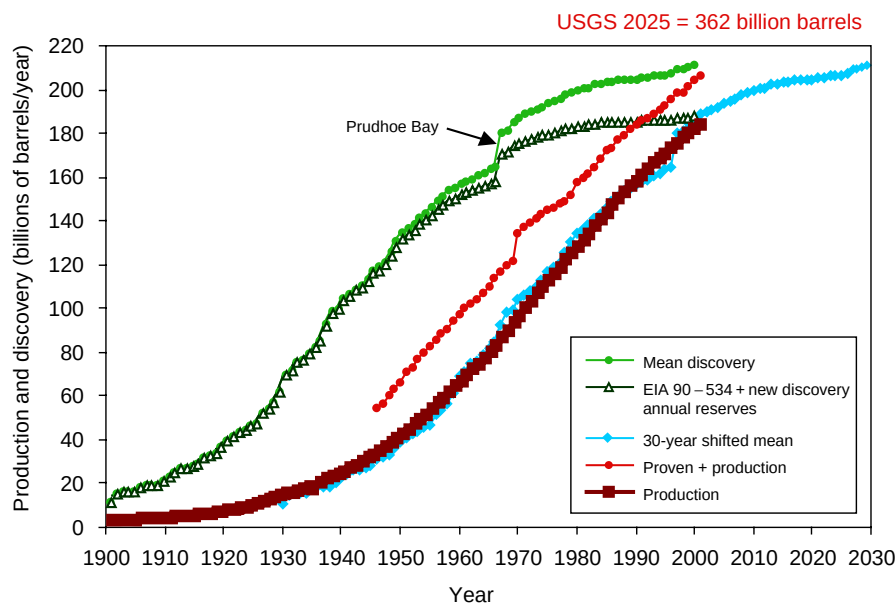


FIGURE 9 U.S. cumulative oil production and shifted mean discovery.

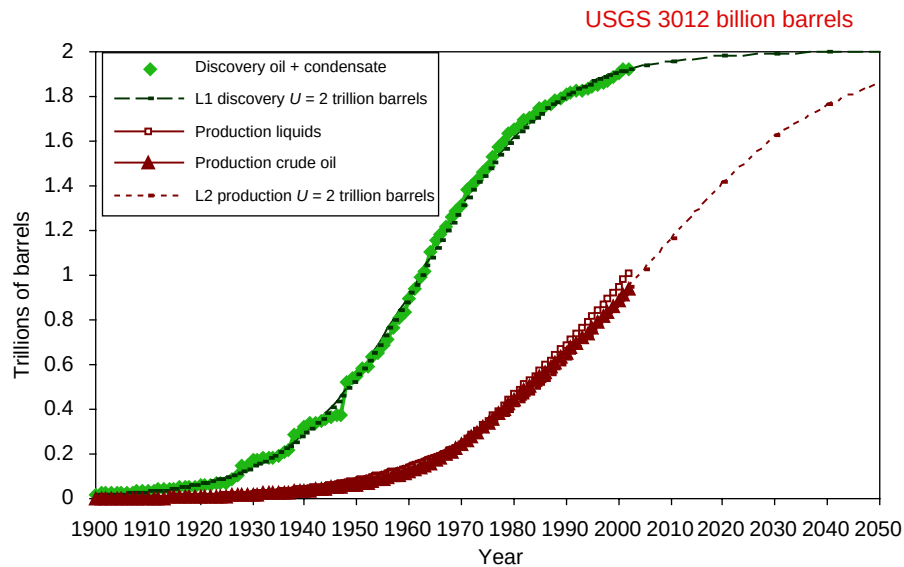


FIGURE 10 World cumulative conventional oil discovery and production.

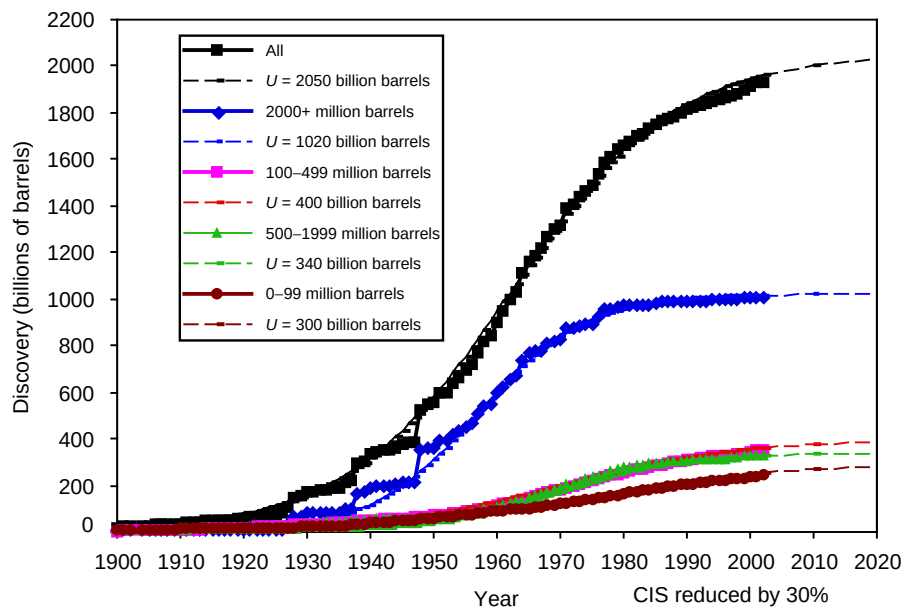


FIGURE 11 World oil cumulative discovery by field size.

being an addition to the OPEC Hubbert curve (ultimate 1.3 trillion barrels) and the non-OPEC curve (ultimate 1.6 trillion barrels), with the assumption that the production would be constrained not by the demand but rather by the supply. A smooth curve is better than some scenarios, as Wood noted from the USGS 2000 report, where many scenarios are drawn with a constant growth up to a peak that is followed by a constant decline. Such an angular pattern looks unnatural. The peak

could be approximately 90 million barrels/day by around 2015. It is likely that the demand will put down the supply curve given that the official scenarios of oil production are based on constant growth corresponding to economic growth of more than 3% per year for the next 20 years or so, leading to a DOE forecast of 120 million barrels/day in 2025 (without any sign of peaking). This level appears to be nearly impossible to reach from the standpoint of technical data.

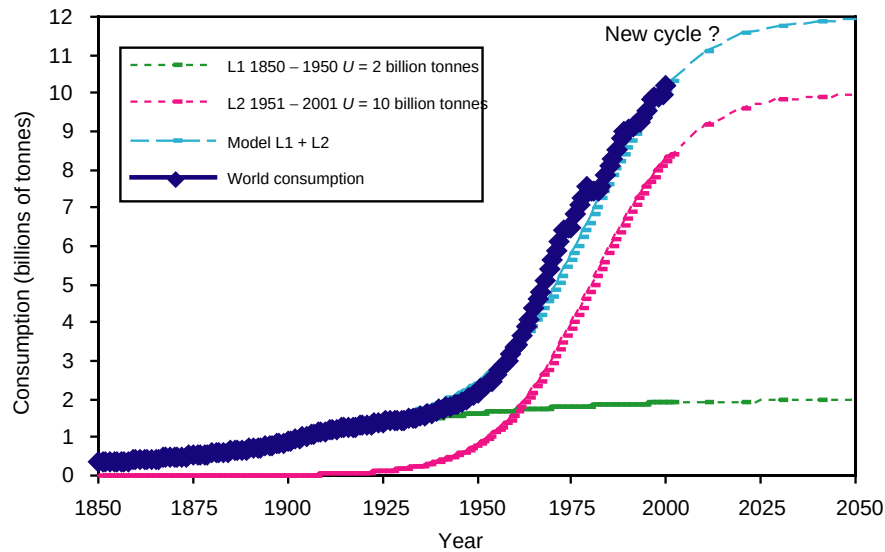


FIGURE 12 World primary energy consumption.

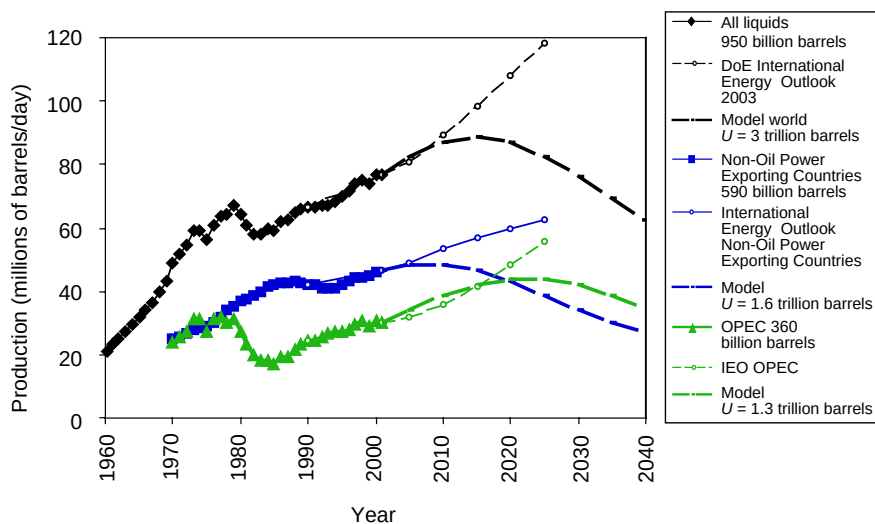


FIGURE 13 World liquids production and forecasts.

If the demand stays at a level of approximately 80 million barrels/day (a bumpy plateau), the real decline will be after 2020 (Fig. 13). The supply could constrain the demand only after 2020. If the supply is high enough to satisfy demand until that time, oil prices could be chaotic unless OPEC is able to keep its price mechanism working. Low prices will lead to more demand and then to a lack of supply.

For natural gas, the ultimate is 10 pounds/cubic foot for conventional gas and 12 pounds/cubic foot for all gas, giving a peak around 2030 of roughly 130 trillion cubic feet/year (Fig. 14). DOE's 2003 International Energy Outlook forecasts 180 trillion

cubic feet/year in 2025 with no sign of slowing down. An ultimate with 9 pounds/cubic foot will give a peak of approximately 110 trillion cubic feet/year around 2020.

3.3 Hubbert Peak, Midpoint, and Constraints for the World

In the media today, the Hubbert symmetrical curve (and peak) is often found to describe the coming oil decline, and many believe that the Hubbert peak arriving at the midpoint (when half of the ultimate is

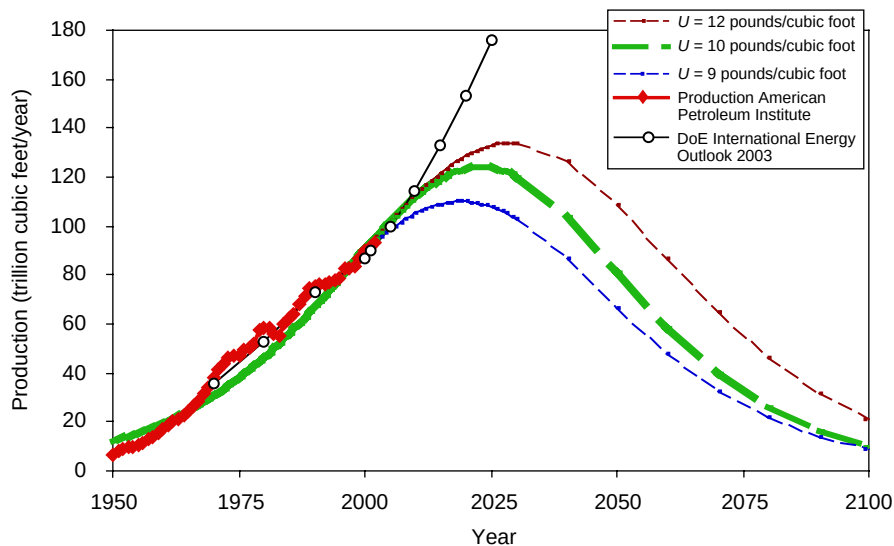


FIGURE 14 World natural gas production and forecasts.

produced) is the rule. A single symmetrical curve occurs only in places where exploration and production are active, without any interruption or any constraints, as in the continental United States and (probably) Norway. But even if so far there has been no constraint from the oil supply (except for a short time given that transport needs some time), there are many cases of constraint from the demand. Production was reduced in 1930 in Texas (with the creation of the Texas Railroad Commission, which was the model for the Organization of Petroleum Exporting Countries [OPEC] in 1967) as the discovery of the East Texas oilfield led to the fall of the oil price from \$1/barrel to \$0.1/barrel and in 1979 as the demand fell due to energy savings in light of the high oil prices. The world oil production peaked in 1979, and it took 15 years to get back to that same level. Figure 13 shows the forecast if there is no demand constraint, but despite the official forecast of 90 million barrels/day in 2010 and 100 million barrels/day in 2015, the supply could provide only 90 million barrels/day in 2010 and 2015 due to the anticipated problems in the world economy for the next decade. It is likely that the demand will constrain the supply given that the potential for oil savings is huge, particularly in the United States, where energy consumption is double that in Europe for a similar lifestyle. The supply will not really constrain the demand until after 2020.

Hubbert curves for a world liquid ultimate of 3 trillion barrels are plotted first in H1 as fitting the quick rise from 1950 to 1979 and giving a peak around 2000 at 140 million barrels/day, second in H2

as fitting the past 20 years or so and giving a peak in 2012 at 80 million barrels/day (but the model is too high before 1950), and in two cycles in H3 + H4 as fitting the data since 1950 with a peak around 2020 at 90 million barrels/day. The forecast in Fig. 15 (coming from a long study and adding several Hubbert curves) gives a peak at 90 million barrels/day around 2015, but the midpoint is not reached until 2019 because the global curve is not symmetrical.

3.4 North America Gas

Whereas there is only one oil market because transport is cheap, there are three gas markets because transport is 6 to 10 times more expensive. The North American gas market has been supplied only locally. The mean gas discovery (2P = proven + provable) of the United States, Canada, and Mexico is shifted by 20 years to fit the production that is plotted as raw, dry, and dry minus unconventional (coalbed methane and tight gas). The graph in Fig. 16 was done in 2000 when production was still rising. The author's forecast at that time was for a soon-to-come sharp decline.

The combined U.S. and Canada natural gas discovery is a perfect logistic curve for an ultimate of 1700 trillion cubic feet (Fig. 17), but the USGS 2000 report gave an ultimate of 2118 trillion cubic feet because USGS assumed that the reserve growth in the United States would bring 355 trillion cubic feet without saying how much in Canada. Removing this reserve growth, the ultimate is then 1763 trillion cubic feet, in close agreement with the estimate. The

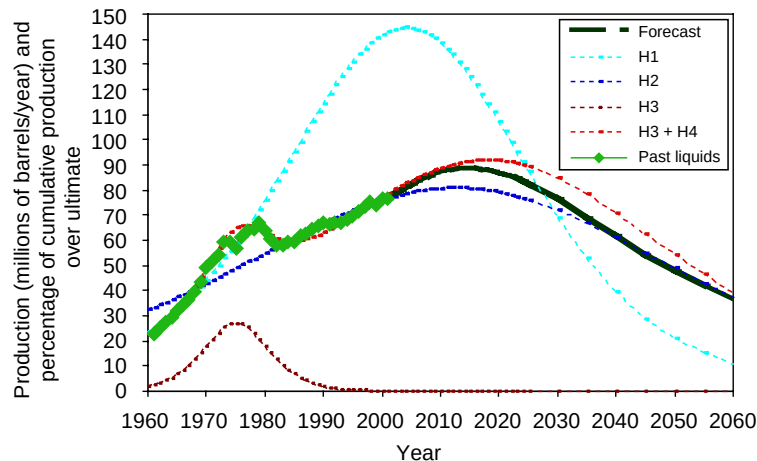


FIGURE 15 World liquids production with forecast and Hubbert curves for 3 trillion barrels.

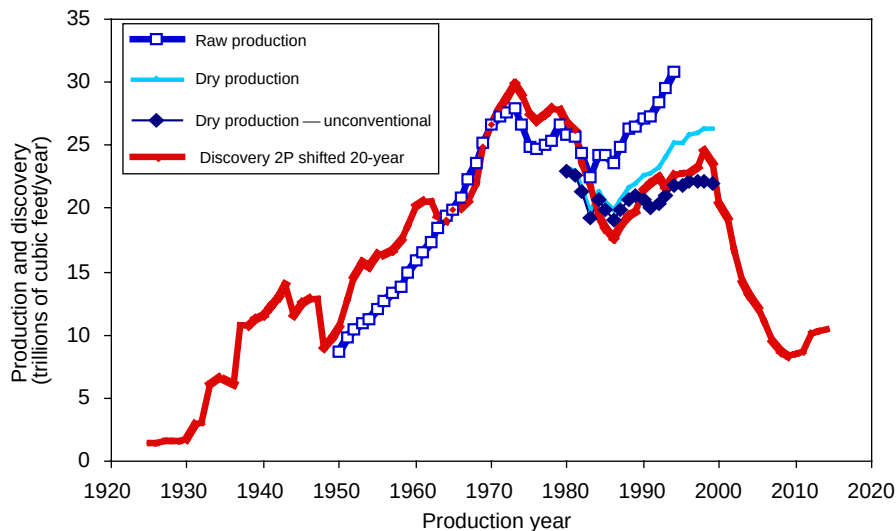


FIGURE 16 U.S., Canada, and Mexico natural gas: production and shifted discovery.

production is following the same logistic curve with a shift of 33 years. It means that the future gas production seems to be fairly settled for the next 30 years or so.

This confirms that the best modeling is not a model but rather a correlation of the production with a shifted discovery. Without any assessment of the ultimate, the part of the shifted discovery beyond current year is the best forecast.

4. OTHER MODELS AND ULTIMATE ASSESSMENTS

Albert Bartlett has written several articles on the Hubbert peak, but he uses a normal curve for annual

production instead of a derivative of logistic, with the latter being better in theory assuming the influence of randomness in countries where there is a large number of producers. But there are many countries with few producers. The cumulative production is then modeled by an integral of the normal curve, and this is more complex to handle than the logistic function.

Richard Duncan has developed some software (a kind of “black box”) modeling the coming peak, but because he has no access to the confidential discovery values, his model relies only on past production. He received some help from Walter Youngquist in getting some geology into his modeling. Duncan compared three different forecasts by Campbell, Duncan, and Smith, but the values of the

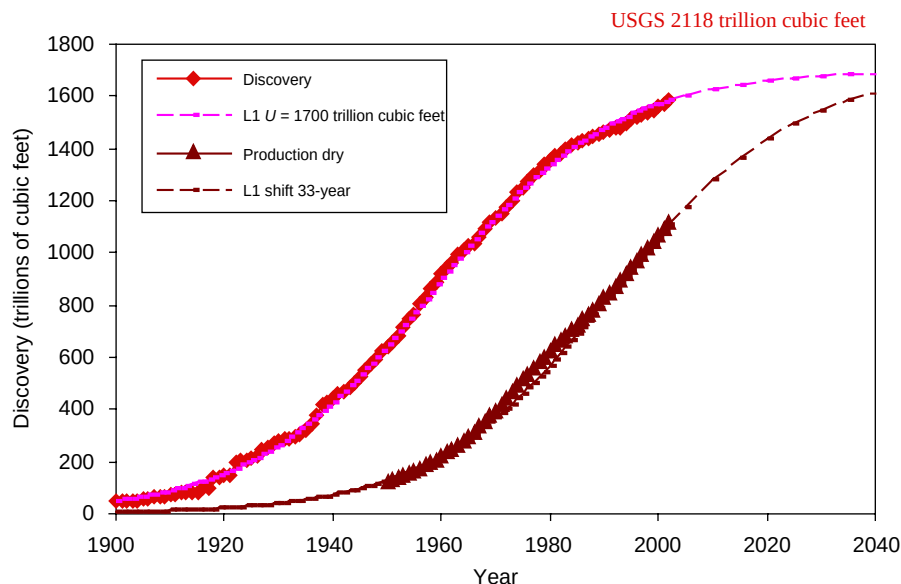


FIGURE 17 U.S. and Canada natural gas cumulative discovery and production.

three forecasts starting in 2000 are significantly different, meaning that they do not deal with the same product and so lead to differences due not only to the models.

Michael Smith published a report titled “Analysis of Global Oil Supply to 2050,” which provides forecasts up to 2100 for every producing country based on an ultimate and a constant decline from the peak. The peak of conventional and unconventional oil is forecasted at approximately 90 million barrels/day in 2012. It is a very thorough work, updating a similar report written in 1994 by Campbell.

Ken Deffeyes wrote a good book, *Hubbert’s Peak*, in which he stated that, despite having discussions and eating lunch many times with Hubbert, Deffeyes did not dare to ask him about the earliest roots of his prediction. Deffeyes uses the annual/cumulative versus cumulative plot, and Hubbert could have done so as well in 1956. Hubbert was clever enough to use rounded figures because he was aware of the poor accuracy of any oil data. But he apparently used his knowledge about discovery more than he used a plot on production.

The base of a production forecast is the ultimate (often called the estimated ultimate recovery [EUR]). Geologists have published their estimates for the world since 1942, and for the past 30 years or so the estimates have ranged from 1600 to 3000 billion barrels (according to various definitions). In 1977, the French Petroleum Institute (IFP) used a Delphi

approach by asking a large number of experts first with an individual inquiry and second with an inquiry after giving the results of the first inquiry. USGS used the Delphi process for its assessment of undiscovered reserves for their world petroleum estimates of 1984, 1987, 1992, and 1994.

The USGS 2000 report did a very good job of defining and describing the petroleum systems of the world with the help of the oil industry. But this industry help stopped when evaluating the potential of undiscovered reserves (due to confidentiality). Instead of using a Delphi process in light of the lack of experts, for each area, only one geologist was asked to give on a single sheet of paper (called the seventh approximation) six values of the minimum, median, and maximum numbers of the undiscovered fields as well as of the size. Then, Monte Carlo simulations were used to give a full distribution of the undiscovered reserves. It was the way in which East Greenland was provided with 47 billion barrels of mean undiscovered oil.

A kind of poor model used by the industry is the R/P, which is the ratio of the remaining reserves on the date versus the annual production of the current year. It is given as 40 years for oil, but this ratio is poor for two reasons: (1) the reserves are the political data rather than the technical data and (2) the model assumes that the production will stay flat for 40 years and drop suddenly to zero during the 41st year. This model is far from the Hubbert model and far from production reality.

5. CONCLUSIONS

There is no theory to justify any model for oil or gas production. The best approach is to use the best data (for reserves, it has to be the backdated mean and never the proven value) for the longest possible period on the largest possible area and to try the simplest model of second degree. Having modeled thousands of curves, it is amazing to the author to find that any event, particularly annual discovery and production, can be modeled with several symmetrical bell-shaped curves (as a Fourier analysis for a permanent wave) and that oil and gas cumulative discovery and production can be modeled fairly with several logistic curves versus time and with several hyperbolas versus the cumulative number of NFWs.

But the best tool is to find the best fit between the shifted (by a certain number of years varying between 5 and 35) annual discovery and the annual production, with this shift allowing the future production to be predicted by following the part of the discovery beyond today.

The Reserve (future production) estimate depends mainly on the reservoir characteristics that are mainly geological, and it is why the recovery factor ranges from 3% (fractured tight reservoir) to 85% (very porous reef). But economy and technology cannot change geology. Do not underestimate the geology.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Resource Assessment • Markets for Natural Gas • Markets for Petroleum • Natural Gas Resources, Global Distribution of • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil and Natural Gas Resource Assessment: Classifications and Terminology • Oil and Natural Gas Resource Assessment: Geological Methods • Petroleum Property Valuation

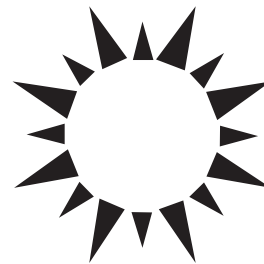
Further Reading

- Bartlett, A. A. (1999). An analysis of U.S. and world oil production patterns using Hubbert-style curves. *Math. Geol.* 32, 1–17.
- Campbell, C. J., and Laherrère, J. H. (1995). “The World’s Oil Supply, 1930–2050.” Petroconsultants, Geneva, Switzerland.
- Duncan, R. C. (2000). The peak of world oil production and the road to the Olduvai Gorge. Paper presented at Geological Society of America, “Summit 2000,” Reno, NV.
- Duncan, R. C. (2003). Three world oil forecasts predict peak oil production. *Oil Gas J.*, May 26, pp. 18–21.
- Hubbert, M. K. (1956). Nuclear energy and fossil fuels. In “Drilling and Production Practice,” American Petroleum Institute, proceedings of spring meeting, San Antonio, TX, pp. 7–25.
- Ivanhoe, L. J. (1995). Future world oil supplies: There is a finite limit. *World Oil*, October, pp. 77–88.
- Laherrère, J. H. (1996). “Distributions de type fractal parabolique dans la Nature.” www.oilcrisis.com/laherrere/fractal.htm.
- Laherrère, J. H. (1999). World oil supply: What goes up must come down—When will it peak? *Oil Gas J.*, Feb. 1, pp. 57–64. <http://dieoff.com/page177.htm>.
- Laherrère, J. H. (2000a). Distribution of field sizes in a petroleum system: log-normal, parabolic fractal, or stretched exponential? *Mar. Petrol. Geol.* 17, 539–546. www.elsevier.nl/cgi-bin/cas/tree/store/jmpg/cas_free/browse/browse.cgi?year=2000&volume=17&issue=4.
- Laherrère, J. H. (2000b). Learn strengths, weaknesses to understand Hubbert curve. *Oil Gas J.* April 17, <http://dieoff.com/page191.htm>.
- Laherrère, J. H. (2001). Forecasting future production with past discoveries. Paper presented at OPEC seminar, “OPEC and the Global Energy Balance: Towards a Sustainable Energy Future,” Vienna, Austria. www.hubbertpeak.com/laherrere/opec2001.pdf.
- Laherrère, J. H. (2003). “Future of Oil Supplies.” Swiss Institute of Technology, Zurich, Switzerland. www.hubbertpeak.com/laherrere/zurich/pdf.
- Laherrère, J. H., and Sornette, D. (1998). Stretched exponential distributions in nature and economy: Fat tails with characteristic scales. *Eur. Phys. J. B* 2, 525–539.
- Longwell, H. J. (2002). The future of the oil and gas industry: Past approaches, new challenges. *World Energy* 5(3). www.worldenergyresource.com/articles/pdf/longwell_we_v5n3.pdf.
- Masters, C.D., Attanasi, E.D., Root, D. (1994). World petroleum assessment and analysis. In “Proceedings of the 14th World Petroleum Congress,” vol. 2, pp. 529–542. Wiley, Stavanger Norway.
- Perrodon, A., Laherrère, J. H., and Campbell, C. J. (1998). “The World’s Non-conventional Oil and Gas.” Petroleum Economist, London.
- Smith, M. (2002). Analysts claim early peak in world oil production. *Oil Gas J.*, August 26, pp. 33–36.
- U.S. Department of Energy/Energy Information Administration. (1990). “US Oil and Gas Reserves by Year of Field Discovery.” DOE/EIA, Washington, DC.
- U.S. Department of Energy/Energy Information Administration. (2003). “IEO 2003.” www.eia.doe.gov/oiaf/ieo.
- U.S. Geological Survey. (2000). “World Petroleum Assessment.” <http://greenwood.cr.usgs.gov/energy/worldenergy/dds-60/espt4.html>.
- Wood, J. (2000). “Long Term World Oil Supply (a resource base/production path analysis).” www.eia.doe.gov/pub/oil-gas/petroleum/presentations/2000/long-term-supply/longtermoil-supplypresentation.ppt.
- Youngquist, W. (1997). “Geodestines: The Inevitable Control of Earth Resources over Nations and Individuals.” National Book, Portland, OR.

Oil Crises, Historical Perspective

MAMDOUH G. SALAMEH

Oil Market Consultancy Service
Haslemere, Surrey, United Kingdom



-
1. Introduction
 2. The Road to the First Oil Crisis
 3. The Oil Weapon
 4. The Embargo
 5. The Second Oil Crisis
 6. Anatomy of Oil Crises
 7. Third Oil Crisis?
 8. Implications for the Global Economy

Glossary

aramco The Arabian American oil company that obtained the concession for oil exploration in the Eastern Province of Saudi Arabia in 1933. It was originally made up of four American oil companies: Standard Oil of California (Socal), Texaco, Standard Oil of New Jersey (then called Esso and later changed to Exxon Mobil), and Socony-Vacuum (Mobil). In 1976, it was purchased by the Saudi government and renamed "Saudi Aramco."

conventional crude oil The oil produced from an underground reservoir, after being freed from any gas that may have dissolved in it under reservoir conditions, but before any other operation has been performed on it. In the oil industry, simply termed "crude."

proven reserves The quantities that geological and engineering information indicates with reasonable certainty can be recovered from known reserves under existing economic and operating conditions.

renewable energy An energy source that does not depend on finite reserves of fossil or nuclear fuels, such as solar energy, wind energy, biomass, hydroelectric power, and hydrogen. All of these renewable sources involve the generation of electricity.

shale oil A distillate obtained when oil shale (a rock of sedimentary origin) is heated in retorts.

ultimate global reserves This is the amount of oil reserves that would have been produced when production eventually ceases.

unconventional oil Oil that has been extracted from tar sands, oil shale, extra heavy oil, and the conversion of

natural gas to liquids, known collectively as synthetic fuels or "synfuels."

The 20th century was truly the century of oil. Though the modern history of oil begins in the latter half of the 19th century, it is the 20th century that has been completely transformed by the advent of oil. Oil has a unique position in the global economic system. One could not imagine modern societies existing without oil. Modern societies' transportation, industry, electricity, and agriculture are virtually dependent on oil. Oil makes the difference between war and peace. The importance of oil cannot be compared with that of any other commodity or raw material because of its versatility and dimensions, namely, economic, military, social, and political. The free enterprise system, which is the core of the capitalist thinking, and modern business owe their rise and development to the discovery of oil. Oil is the world's largest and most pervasive business. It is a business that touches every corner of the globe and every person on earth. The financial resources and the level of activity involved in exploring, refining, and marketing oil vastly exceed those of any other industry. Of the top 20 companies in the world, 7 are oil companies. Human beings are so dependent on oil, and oil is so embedded in daily life, that individuals hardly stop to comprehend its pervasive significance. Developing nations give no indication that they want to deny themselves the benefits of an oil-powered economy, whatever the environmental questions. In addition, any notion of scaling back the world's consumption of oil will be influenced by future population growth.

1. INTRODUCTION

No other commodity has been so intimately intertwined with national strategies and global politics

and power as oil. In World War I, the Allies floated to victory on a wave of oil. Oil was central to the course and outcome of World War II in both the Far East and Europe. One of the Allied powers' strategic advantages in World War II was that they controlled 86% of the world's oil reserves. The Japanese attack on Pearl Harbor was about oil security. Among Hitler's most important strategic objectives in the invasion of the Soviet Union was the capture of the oilfields in the Caucasus. In the Cold War years, the battle for the control of oil resources between international oil companies and developing countries was a major incentive and inspiration behind the great drama of decolonization and emergent nationalism.

During the 20th century, oil emerged as an effective instrument of power. The emergence of the United States as the world's leading power in the 20th century coincided with the discovery of oil in America and the replacement of coal by oil as the main energy source. As the age of coal gave way to oil, Great Britain, the world's first coal superpower, gave way to the United States, the world's first oil superpower.

Yet oil has also proved that it can be a blessing for some and a curse for others. Since its discovery, it has bedeviled the Middle East with conflict and wars. Oil was at the heart of the first post-Cold War crisis of the 1990s—the Gulf War. The Soviet Union—the world's second largest oil exporter—squandered its enormous oil earnings in the 1970s and 1980s in a futile military buildup. And the United States, once the largest oil producer and still its largest consumer, must import almost 60% of its oil needs, weakening its overall strategic position and adding greatly to an already burdensome trade deficit—a precarious position for the only superpower in the world.

The world could face an energy gap probably during the first decade of the 21st century once global conventional oil production has peaked. This gap will have to be filled with unconventional and renewable energy sources. A transition from fossil fuels to renewable energy sources is, therefore, inevitable if humans are to bridge the energy gap and create a sustainable future energy supply. Sometime in the 21st century, nuclear, solar, geothermal, wind, and hydrogen energy sources may be sufficiently developed to meet a larger share of the world's energy needs. But for now humans will continue to live in an age of oil. Oil will, therefore, still be supplying a major share of the global energy needs for most, perhaps all, of the 21st century and will continue to have far-reaching effects on the global economy.

2. THE ROAD TO THE FIRST OIL CRISIS

One distinctive feature dominated the global economic scene in the decades following World War II. It was the rising consumption of oil. Total world energy consumption more than tripled between 1949 and 1972. Yet that growth paled in comparison to the rise in oil demand, which during the same period increased more than $5\frac{1}{2}$ times over. Everywhere, growth in the demand for oil was strong. Between 1948 and 1972, consumption tripled in the United States, from 5.8 to 16.4 million barrels/day—unprecedented except when measured against what was happening elsewhere. In the same years, demand for oil in Western Europe increased 15 times over, from 970,000 barrels/day to 14.1 million barrels/day. In Japan, the change was nothing less than spectacular; consumption increased 137 times over, from 32,000 barrels/day to 4.4 million barrels/day.

The main drivers of this global surge in oil use were the rapid economic growth and the cheap price of oil. During the 1950s and 1960s, the price of oil fell until it became very cheap, which also contributed mightily to the swelling of consumption. Many governments encouraged its use to power economic growth and industrial modernization. There was one final reason that the market for oil grew so rapidly. Each oil-exporting country wanted higher volumes of its oil sold in order to gain larger revenues.

In the buoyant decades following World War II, oil overtook coal as the main fuel for economic growth. Huge volumes of oil surged out of Venezuela and the Middle East and flowed around the world. Oil was abundant. It was environmentally more attractive and was easier and more convenient to handle. And oil became cheaper than coal, which proved the most desirable and decisive characteristic of all. Its use provided a competitive advantage for energy-intensive industries. It also gave a competitive edge to countries that shifted to it. And yet, there was a haunting question: How reliable was the flow of oil on which modern societies had come to depend? What were the risks?

Among the Arabs, there had been talk for more than a decade about wielding the "oil weapon." This was their chance. On June 6, 1967, the day after the start of the Six-Day War, Arab oil ministers, members of Organization of Arab Petroleum Exporting Countries, formally called for an oil embargo against countries friendly to Israel. Saudi Arabia, Kuwait, Iraq, Libya, and Algeria thereupon banned shipments to the United States and Great Britain.

By June 8, the flow of Arab oil had been reduced by 60%. The overall initial loss of Middle Eastern oil was 6 million barrels/day. Moreover, logistics were in total chaos not only because of the interruptions but also because, as in 1956, the Suez Canal and the pipelines from Iraq and Saudi Arabia to the Mediterranean were closed. The situation grew more threatening in late June and early July when, coincidentally, civil war broke out in Nigeria, depriving the world oil market of 500,000 barrels/day at a critical moment.

However, by July 1967, a mere month after the Six-Day War, it became clear that the selective Arab oil embargo was a failure; supplies were being redistributed to where they were needed. And by the beginning of September, the embargo had been lifted.

The 1970s saw a dramatic shift in world oil. Demand was catching up with available supply and the 20-year surplus was over. As a result, the world was rapidly becoming more dependent on the Middle East and North Africa for its oil.

Oil consumption surged beyond expectation around the world, as ever-greater amounts of petroleum products were burned in factories, power plants, homes, and cars. The cheap price of oil in the 1960s and early 1970s meant that there was no incentive for fuel-efficient automobiles. The late 1960s and early 1970s were also the watershed years for the domestic U.S. oil industry. The United States ran out of surplus capacity. In the period 1957 to 1963, surplus capacity in the United States had totaled approximately 4 million barrels/day. By 1970, it had declined to only 1 million barrels/day. That was the year, too, that American oil production peaked at 10 million barrels/day. From then on, it began its decline, never to rise again. With consumption continuing to rise, the United States had to turn to the world oil market to satisfy its needs. Net imports tripled from 2.2 million barrels/day in 1967 to 6 million barrels/day by 1973. Imports as a share of total oil consumption over the same years rose from 19 to 36%.

The disappearance of surplus capacity in the United States would have major implications, for it meant that the “security margin” on which the Western world depended was gone. For the United States, it marked a shift from: (1) oil self-sufficiency to reliance on Middle East oil; (2) being a major exporter to becoming a major importer; (3) loss of the ability to control the global oil markets at a time when Middle East oil producers (Arab Gulf producers) began to assert themselves on the global

markets; and (4) inability to provide stand-by supply to its allies in an emergency. This meant a major shift from energy security to vulnerability and dependency. Indeed, the razor’s edge was the ever-increasing reliance on the oil of the Middle East. New production had come from Indonesia and Nigeria (in the latter case, after the end of its civil war in early 1970), but that output was dwarfed by the growth in Middle Eastern production. Between 1960 and 1970, Western world oil demand had grown by 21 million barrels/day. During that same period, production in the Middle East (including North Africa) had grown by 13 million barrels/day. In other words, two-thirds of the huge increase in oil consumption was being satisfied by Middle East oil.

In a wider sense, the disappearance of surplus capacity caused an economic and geopolitical transformation of the global oil market. In an economic sense, the “center of gravity” of oil production, energy security, and control of global oil supplies had shifted from the United States to the Middle East. In a geopolitical sense, the oil revenue and the global dependence on Middle East oil provided the Arab Gulf producers with unprecedented political influence. This they channeled into support of their political aspirations as they did during the 1973 war and the resulting oil embargo.

Another disturbing development was that the relationship between the oil companies and the producing nations was beginning to unravel. In gaining greater control over the oil companies, whether by participation or outright nationalization, the exporting countries also gained greater control over prices. The result was the new system that was forged in Tehran and Tripoli, under which prices were the subject of negotiation between companies and countries, with the producing countries taking the lead in pushing up the posted price.

However, the supply-demand balance that emerged at the beginning of the 1970s was sending a most important message: Cheap oil had been a tremendous boon to economic growth, but it could not be sustained. Demand could not continue growing at the rate it was; new supplies needed to be developed. That was what the disappearance of spare capacity meant. Something had to give, and that something was price.

By 1972, many experts reckoned that the world was heading for an acute oil shortage in a few years. The signs of a shortage were visible everywhere. The demand for oil in the summer of 1973 was going above the wildest predictions—in Europe, in Japan, and most of all in the United States. Imports from the

Middle East to the United States were still racing up: production inside the United States was still falling. In April, U.S. President Nixon lifted restrictions on imports of oil, so the Middle East oil flowed in still faster. There was a new danger sign when Kuwait decided in 1972 to conserve its resources and to keep its production below 3 million barrels/day.

As late as 1970, there were still approximately 3 million barrels/day of spare capacity in the world outside the United States, with most of it concentrated in the Middle East. By the second half of 1973, the spare capacity had shrunk to only 500,000 barrels/day. That was just 1% of world consumption. With a 99% capacity utilization and a 1% security margin, the oil supply-demand balance was indeed extremely precarious.

In June 1973, as prices were zooming up, the Organization of Petroleum Exporting Countries (OPEC) summoned another meeting in Geneva to insist on another price increase because of the further devaluation of the U.S. dollar. The radicals—Algeria, Libya and Iraq—were pressing for unilateral control of price, but eventually OPEC agreed on a new formula that increased prices by another 12%. By September 1973, for the first time since the founding of OPEC, the market price of oil had risen above the posted price. It was a sure sign that OPEC was in a very strong bargaining position. Armed with this knowledge, OPEC invited the oil companies to meet them in Vienna on October 8 to discuss substantial increases in the price of oil.

In this atmosphere of crisis, the oil company delegates prepared to confront OPEC in Vienna on October 8. And then, just as they were leaving for Vienna, Egypt and Syria invaded Israeli-occupied territories. There was war.

While the shortage loomed, the Arabs were at last achieving closer unity. They were determined to use oil as a weapon against Israel and by 1973 the militants were being joined by Saudi Arabia. The very fact that Saudi Arabia had become the largest oil exporter made King Feisal more vulnerable in the face of his Arab colleagues and the danger of an embargo more likely, for he could not afford to be seen as a blackleg.

The international oil order had been irrevocably changed. However, it was not only a question of price, but of power. The extent of dependence by the industrial countries on oil as a source of energy had been exposed and the practicality of controlling supply as means of exerting pressure for raising the price of oil had been dramatically demonstrated. Although the oil weapon had not worked in 1967,

the “rationale of those who called for its use as a weapon in the Middle East conflict has been strengthened in current circumstances.”

3. THE OIL WEAPON

Contrary to popular belief, the Americans, not the Arabs, were the first to wield the oil weapon. They used it against Japan when on July 25, 1941, the United States announced a freeze on Japanese funds necessary for Japan to buy American oil, which, in practice, meant an embargo on oil. The embargo was the result of Japanese military aggression in Asia.

Increasingly worried about a cut-off of oil supplies from the United States, Tokyo instituted a policy to establish self-sufficiency and to try to eliminate dependence on U.S. oil supplies. In 1940–1941, it was energy security that led Japan to occupy the Dutch East Indies and take control of its oilfields. Indeed, the U.S. oil embargo was an important factor leading Japan to attack Pearl Harbor, bringing the United States into World War II. Oil had been central to Japan’s decision to go to war.

Ever since the 1950s, the Arab world had been talking about using the oil weapon to force Israel to give up occupied Arab territories. Yet the weapon had always been deflected by the fact that Arab oil, though it seemed endlessly abundant, was not the supply of last resort.

In June 1967, 2 days into the Six-Day War, the Arabs wielded the oil weapon when they imposed an oil embargo against the United States and Great Britain. However, by July 1967, it became clear that the Arab embargo had failed. The Arabs would have to wait for another chance to wield the oil weapon again. That chance came their way when just moments before 2:00 P.M. on October 6, 1973, Egyptian and Syrian armies launched an attack on Israeli-held positions in the Sinai and the Golan Heights. Thus began the October War or, what became known as the Yom Kippur War, the fourth of the Arab–Israeli wars—the most destructive and intense of all of them and the one with the most far-reaching consequences. One of the most potent weapons used in this war was the oil weapon, wielded in the form of an embargo—production cutbacks and restrictions on exports—that, in the words of Henry Kissinger, “altered irrevocably the world as it had grown up in the postwar period.”

The embargo, like the war itself, came as a surprise and a shock. Yet the pathway to both in retrospect seemed in some ways unmistakable. By

1973, oil had become the lifeblood of the world's industrial economies and it was being pumped and circulated with very little to spare. Never before in the entire postwar period had the supply-demand equation been so tight, while the relationships between the oil-exporting countries and the oil companies continued to unravel. It was a situation in which any additional pressure could precipitate a crisis—in this case, one of global proportions.

With supply problems becoming chronic in the early 1970s, talk about an energy crisis began to circulate in the United States. There was agreement, in limited circles, that the United States faced a major problem. Price controls on oil, imposed by Nixon in 1971 as part of his overall anti-inflation program, were discouraging domestic oil production while stimulating consumption. The artificially low prices provided little incentive either for new exploration or for conservation. By the summer of 1973, United States oil imports were 6.2 million barrels/day, compared to 4.5 million barrels/day in 1972 and 3.2 million barrels/day in 1970. The oil trade journal *Petroleum Intelligence Weekly* reported in August 1973 that “near-panic buying by the U.S., the Europeans, and the Japanese was sending oil prices sky-rocketing.”

As global demand continued to rise against the limit of available supply, market prices exceeded the official posted prices. It was a decisive change, truly underlining the end of surplus. For so long, reflecting the chronic condition of oversupply, market prices had been below posted prices, irritating relations between companies and governments. But the situation had reversed and the exporting countries did not want to see the growing gap between the posted price and the market price go to the companies. Wasting little time, the exporters sought to revise their participation and buy-back arrangements so that they would be able to obtain a larger share of the rising prices.

One of the principal characteristics governing the operations of the oil industry is that it generates an important “economic rent” or “oil surplus,” the appropriation of which involves three players: the exporting countries, the consuming countries, and the multinational oil companies. Both the exporting countries and the consuming countries are effectively staking a claim to the significant element of economic rent built into the price of oil. For the exporters, such as the Arab Gulf producers, oil remains the single most important source of income, generating approximately 85 to 90% of their revenues. Significantly, consumer countries have

always looked on oil as an important source of taxation since demand for it is relatively inelastic; that is to say, it varies little as the price changes. These countries have more maneuverability when the price of oil is low. This was amply demonstrated when, as a result of the oil price collapse in 1986, many of them took the opportunity to raise tax rates on petroleum products. This practice was, to a lesser extent, in operation in the 1970s but has accelerated since 1986 in Europe, with tax levels on petroleum products reaching between 80 and 87%. In other words, the sharp increases in taxes by the consuming countries were intended to cream off more of the rent. Is it any wonder that the United Kingdom has been for years earning far more revenue from taxes on petroleum products than from its North Sea oil exports?

On September 1, 1973, the fourth anniversary of Muammar Qaddafi's coup, Libya nationalized 51% of those company operations it had not already taken over. The radicals in OPEC—Iran, Algeria, and Libya—began pushing for a revision in the Tehran and Tripoli agreements. By the late summer of 1973, the other exporters, observing the upward trend of prices on the open market, came around to that same point of view. They cited rising inflation, the dollar's devaluation, and also the rising price of oil. By September 1973, the Saudi oil minister Sheikh Ahmed Zaki Yamani was able to announce that the Tehran Agreement was dead. Even as the economics of oil were changing, so were the politics that surrounded it—and dramatically so.

By April of 1973, President Anwar Sadat of Egypt had begun formulating with Syria's Hafez Al-Asad strategic plans for a joint Egyptian-Syrian attack against Israel. Sadat's secret was tightly kept. One of the few people outside the high commands of Egypt and Syria with whom he shared it was King Feisal. And that meant oil would be central to the coming conflict.

In the early 1970s, as the market tightened, various elements in the Arab world became more vocal in calling for use of the oil weapon to achieve their economic and political objectives. King Feisal was not among them. He had gone out of his way to reject the use of the oil weapon. “It was not only useless,” he said, “but dangerous even to think of that. Politics and oil should not be mixed.” Yet, by early 1973, Feisal was changing his mind. Why?

Part of the answer lay in the marketplace. Much sooner than expected, Middle Eastern oil had become the supply of last resort. In particular, Saudi Arabia had become the marginal supplier for everybody,

including the United States; American dependence on the Gulf had come not by the widely predicted 1985, but by 1973. The United States would no longer be able to increase production to supply its allies in the event of a crisis and the United States itself was now, finally, vulnerable. The supply–demand balance was working to make Saudi Arabia even more powerful. Its share of world exports had risen rapidly from 16% in 1970 to 25% in 1973 and was continuing to rise.

In addition, there was a growing view within Saudi Arabia that it was earning revenues in excess of what it could spend. Two devaluations of the U.S. dollar had abruptly cut the worth of the financial holdings of countries with large dollar reserves, including Saudi Arabia.

The changing conditions in the marketplace, which with each passing day made the Arab oil weapon more potent, coincided with significant political developments. By the spring of 1973, Sadat was strongly pressing Feisal to consider using the oil weapon to support Egypt in a confrontation with Israel. King Feisal also felt growing pressure from many elements within his kingdom and throughout the Arab world. Thus, politics and economics had come together to change Feisal's mind. Thereupon the Saudis began a campaign to make their views known, warning that they would not increase their oil production capacity to meet rising demand and that the Arab oil weapon would be used, in some fashion, unless the United States moved closer to the Arab viewpoint and away from Israel.

On August 23, 1973, Sadat made an unannounced trip to Riyadh to see King Feisal. He told the king that he was considering going to war against Israel. It would begin with a surprise attack and he wanted Saudi Arabia's support and cooperation. He got it.

On October 17, 1973, 11 days into the war, Arab oil ministers meeting in Kuwait agreed to institute a total oil embargo against the United States and other countries friendly to Israel. They decided to cut production 5% from the September level and to keep cutting by 5% in each succeeding month until their objectives were met. Oil supplies at previous levels would be maintained to "friendly states." One clear objective of the plan was to split the industrial countries right from the start.

On October 19, Nixon publicly proposed a \$2.2 billion military aid package for Israel. In retaliation for the Israeli aid proposal, Saudi Arabia had gone beyond the rolling cutbacks; it would now cut off all shipments of oil, every last barrel, to the United States. The oil weapon was now fully in battle—a weapon, in Kissinger's words, of political blackmail. To counter

that blackmail, Kissinger called for the industrialized nations to meet in Washington, DC at the earliest possible moment. He wanted the oil-consuming West to make a united stand against the Arabs.

4. THE EMBARGO

The embargo came as an almost complete surprise despite the evidence at hand: almost two decades of discussion in the Arab world about the oil weapon, the failed embargo in 1967, Sadat's public discussion of the "oil option" in early 1973, and the exceedingly tight oil market of 1973. What transformed the situation and galvanized the production cuts and the embargo against the United States was the very public nature of the resupply of ammunitions and armaments to Israel and then the \$2.2 billion aid package.

On October 21, Sheikh Yamani met with the president of Aramco, Frank Junkers. Using computer data about exports and destinations that the Saudis had requested from Aramco a few days earlier, Yamani laid out the ground rules for the cutbacks and the embargo the Saudis were about to impose. He told Junkers that any deviations from the ground rules would be harshly dealt with.

At the time of the embargo, the management of the Saudi oil industry was in the hands of Aramco (the Arabian–American Oil Company), the joint venture between Standard Oil of California (Socal), Texaco, Standard Oil of New Jersey (then called Esso and later changed to ExxonMobil), and Socony–Vacuum (Mobil).

In 1948, U.S. imports of crude oil and products together exceeded exports for the first time. No longer could the United States continue its historical role as supplier to the rest of the world. That shift added a new dimension to the vexing question of energy security. The lessons of World War II, the growing economic significance of oil, and the magnitude of Middle Eastern oil reserves all served, in the context of the developing Cold War with the Soviet Union, to define access to that oil as a prime element in Western security. Oil provided the point at which foreign policy, international economic considerations, national security, and corporate interests would all converge. The Middle East would be the focus. There the oil companies were already building up production and making new arrangements to secure their positions.

In Saudi Arabia, development was in the hands of Aramco. The company understood from the time it obtained the Saudi oil concession in 1933 that the

concession would always be in jeopardy if it could not satisfy the expectations and demands of King Abdul Aziz Ibn Saud, the founder of Saudi Arabia, and the royal family. Since then, it has worked tirelessly to enhance Saudi oil reserves and production and build terminals and pipelines for exporting the oil worldwide.

In October 1972, Sheikh Yamani negotiated a participation agreement between Saudi Arabia and Aramco. It provided for an immediate 25% participation share, rising to 51% by 1983. Aramco had finally agreed to participation with Saudi Arabia because the alternative was worse—outright nationalization.

In June 1974, Saudi Arabia, operating on Yamani's principle of participation, took a 60% share in Aramco. By the end of the year, the Saudis told Aramco that 60% was simply not enough. They wanted 100%. An agreement to that effect was eventually reached in 1976 between Aramco and Saudi Arabia, almost 43 years after the concession was granted. By then, the proven reserves of Saudi Arabia were estimated at 149 billion barrels—more than one-quarter of the world's total reserves.

But the agreement did not by any means provide for a severing of links. Thus, under the new arrangement, Saudi Arabia would take over ownership of all Aramco's assets and rights within the country. Aramco could continue to be the operator and provide services to Saudi Arabia, for which it would receive 21 cents per barrel. In return, it would market 80% of Saudi production. In 1980, Saudi Arabia finally paid compensation, based on net book value, for all Aramco's holdings within the kingdom. With that, the sun finally set on the great concessions.

The Saudis had already worked out the embargo in some detail. They insisted that on top of the 10% cutback, Aramco must subtract all shipments to the United States, including the military. The Saudis asked Aramco for details of all crude oil used to supply American military forces throughout the world. The details were provided and the Saudis duly instructed Aramco to stop the supplies to the U.S. military. The situation was serious enough for Washington to ask whether British Petroleum (BP) could supply the U.S. Sixth Fleet in the Mediterranean.

At the beginning of November 1973, only 2 weeks after the initial decision to use the oil weapon, the Arab oil ministers decided to increase the size of the across-the-board cuts. This resulted in a gross loss of 5 million barrels/day of supply from the market. This time, however, there was no spare capacity in the United States. Without it, the United States had lost its critical ability to influence the world oil market.

And with the price of a barrel of oil skyrocketing, the oil exporters could cut back on volumes and still increase their total income.

The panic and shortage of oil supplies caused by the embargo led to a quadrupling of crude oil price and precipitated a severe recession, which adversely affected the economies of the industrialized nations. Panic buying meant extra demand in the market. The bidding propelled prices even further upward. The posted price for Iranian oil, in accordance with the October 16 agreement, was \$4.50 per barrel. In December, it sold for \$22.60.

The oil crisis had far-reaching political and economic effects. The quadrupling of prices by the Arab oil embargo and the exporters' assumption of complete control in setting those prices brought massive changes to the world economy. The combined oil earnings of the oil exporters rose from \$23 billion in 1972 to \$140 billion by 1977. For the industrial countries, the sudden hike in oil prices brought profound dislocations. The oil rents flooding into the treasuries of the exporters added to a huge withdrawal of their purchasing power and sent them into deep recession. The U.S. gross national product (GNP) plunged 6% between 1973 and 1975 and unemployment doubled to 9%. Japan's GNP declined in 1974 for the first time since the end of World War II. At the same time, the price increases delivered a powerful inflationary shock to economies in which inflationary forces had already taken hold. President Nixon later commented that: "The oil embargo made it transparently clear that the economies of Western Europe and Japan could be devastated almost as completely by an oil cutoff as they could be by a nuclear attack."

On March 18, 1974, the Arab oil ministers agreed to end the embargo after the United States warned that peace efforts between the Arabs and Israel could not proceed without the lifting of the embargo. After two decades of talk and several failed attempts, the oil weapon had finally been successfully used, with an impact not merely convincing, but overwhelming, and far greater than even its proponents have dared to expect. It had transformed world oil and the relations between producers and consumers and it had remade the international economy. Now it could be resheathed. But the threat would remain.

5. THE SECOND OIL CRISIS

The Iranian revolution was at the heart of the second oil crisis. The astronomical oil price rises of 1979

and the emergence of the Rotterdam “Spot Market” were a direct consequence of the Iranian revolution. In January 7, 1978, a Tehran newspaper published a savage attack on an implacable opponent of the Shah, an elderly Shiite cleric named Ayatollah Ruhollah Khomeini, who was then living in exile in Iraq. This journalistic assault on Khomeini set off riots in the holy city of Qom, which remained his spiritual home. Troops were called in and demonstrators were killed. The disturbance in Qom ignited riots and demonstrations across the country, with further dramatic clashes and more people killed. Strikes immobilized the economy and the government and demonstrations and riots went on unchecked.

All through 1978, the internal political strife against the Shah’s regime and the political drama that was unfolding simultaneously in Paris and Tehran were pushing Iran toward an explosion.

It had become evident in the mid-1970s that Iran simply could not absorb the vast increase in oil revenues that was flooding the country. The petrodollars, misspent on grandiose modernization programs or lost to waste and corruption, were generating economic chaos and social and political unrest throughout the nation. Iranians from every sector of national life were losing patience with the Shah’s regime and the rush to modernization. Grasping for some certitude in the melee, they increasingly heeded the call of traditional Islam and of an ever more fervent fundamentalism. The beneficiary was Ayatollah Khomeini, whose religious rectitude and unyielding resistance made him the embodiment of opposition to the Shah’s regime. The Iranian oil industry was in a state of escalating chaos. The impact of the strikes was felt immediately. On October 13, 1978, workers at the world’s largest oil refinery, in Abadan, suddenly went on strike at the instigation of the exiled Ayatollah Khomeini, who was inciting them from Paris in recorded speeches on cassettes smuggled to Iran. Within a week, the strike had spread throughout most of the oil installations and Iran was, for all intents and purposes, out of the oil business.

Iran was the second-largest exporter of oil after Saudi Arabia. Of the 5.7 million barrels/day produced in Iran, approximately 4.5 million barrels/day were exported. By early November 1978, production decreased from 5.7 million barrels/day to 700,000 barrels/day. And by December 25, Iranian oil exports ceased altogether. That would prove to be a pivotal event in the world oil market. Spot prices in Europe surged 10 to 20% above official prices. The ceasing of Iranian exports came at a time when, in the

international market, the winter demand surge was beginning. Oil companies, responding to the earlier general softness in the market, had been letting their inventories fall.

On December 26, Khomeini declared, “as long as the Shah has not left the country there will be no oil exports.” On January 16, 1979, the Shah left the country. And on February 1, Khomeini landed in Tehran.

By the time the Khomeini regime decided to resume pumping, they could not restore production to prerevolutionary levels because the Ayatollah kicked out of Iran all the Western companies that operated the Iranian oilfields. During the production disruption, these fields lost gas pressure and the lack of maintenance made matters worse. The Iranians apparently did not have enough technical know-how to maintain or operate the fields. In effect, the Iranians pulled the rug from under the feet of the oil market, the world panicked, and the prices started to hit the roof.

Up until September 1977, there was actually a glut of oil in the market, which meant that the continuous rise in oil prices had come to a halt. The oil surplus was due to a number of factors. North Sea oil production increased much faster than expected, Mexico became a large oil exporter, and in July 1977 oil from Alaska started its long-awaited flow to the United States. At the same time, Saudi Arabia, the world’s largest oil exporter, was pumping 8.5 million barrels/day in the first half of 1977 in an attempt to frustrate demands for high price increases by some members of OPEC. Another reason is that demand for crude oil in Europe barely increased from the 1976 levels, as the economies there recovered more slowly than expected from the worldwide recession that followed the Arab oil embargo 3 years earlier. Also, the demand for gasoline in the United States, expected to rise 3% in 1977, grew only 2% due to the introduction of new fuel-efficient American cars.

However, between 1974 and 1978, the world oil supplies seemed to have reached some kind of uneasy stability. Before Iranian oil production ceased, the world oil output approximately matched demand at 63 million barrels/day. By December 26, 1978, when all oil exports from Iran ceased, the world oil market had a shortfall of 5 million barrels/day. Part of the shortfall had been made up by other OPEC members. This left a shortfall of 2 million barrels/day.

Yet when measured against world demand of 63 million barrels/day, the shortage was no more than 3%. Why should a 3% loss of supplies have resulted in a 150% increase in the price? The answer was

panic. The rush to build inventories by oil companies resulted in an additional 3 million barrels/day of demand above actual consumption. When added to the 2 million barrels/day of net lost supplies, the outcome was a total shortfall of 5 million barrels/day, which was equivalent to approximately 8% of global consumption. The panic buying more than doubled the actual shortage and further fueled the panic. That drove the price from \$13/barrel to \$34/barrel.

As the world oil shortage became more serious, there was frenzied activity on the Rotterdam Spot Market, which rapidly became the barometer of the crisis and also an indicator of the extreme price levels, and this market became the new frontier of the oil trade. Many factors conspired in early 1979 to make the Spot Market more excitable. The New Iranian regime decided to sell more and more oil on a day-to-day basis instead of on term contracts, so that in effect became part of the Spot Market. The Japanese had been dependent on Iranian oil and as the crisis deepened, they came unstuck. It was Japan that led the panic buying and by May 1979 spot crude leapt up to \$34.5/barrel. The Spot Market became still more excited as the Europeans tried to replenish their declining oil stocks. American oil companies with refined oil in the Caribbean began shipping it to Rotterdam instead of the east coast of the United States.

Throughout the cold spring of 1979, the oil shortage was being felt across the industrial world and by the week beginning May 14, the Spot Market began to go crazy. Oil deals by the major oil companies at the Spot Market helped push the price further up. It became evident that the major oil companies previously sold cargos of their cheap oil bought out at the long-term contract prices, making huge profits. They were now rushing in to buy oil for storage because they were not sure how long the Iranian oil exports cutoff would last. The next big move in the Spot Market came in October 1979, when the spot prices hit \$38/barrel. Then, in early November, 90 people, including 63 Americans, were taken hostage at the U.S. Embassy in Tehran. An international crisis with grave military overtones suddenly erupted onto the world scene. Spot crude hit \$40/barrel.

However, the panic buying of 1979–1980 would become the glut of 1980–1986. Prices would eventually tumble. By the summer of 1980, oil inventories were very high; a pronounced economic recession was already emerging; in the consuming countries, both product prices and demand were even falling; and the inventory surplus continuing to

swell. Now it was the buyers' turn to walk away from contracts and the demand for OPEC oil was decreasing. Indeed, in mid-September, a number of OPEC countries agreed to voluntarily cut back production by 10% in an effort to firm prices.

But as the world oil market was becoming calmer, on September 22, 1980 Iraqi troops began an attack on Iran. The outbreak of war threw the oil supply system into jeopardy, threatening a third oil crisis. In its initial stages, the Iran–Iraq War abruptly removed almost 4 million barrels/day of oil from the world market—6% of world demand. Spot prices jumped up again. Arab light crude oil reached its highest price ever: \$42/barrel. Panic was once again driving the market. However, supply from other sources was making up for the lost output from Iran and Iraq. Within days of the war, OPEC producers raised their production. At the same time, production in Mexico, the North Sea, Alaska, and other non-OPEC countries was continuing to increase as well. Non-OPEC producers, anxious to increase market share, were making significant cuts in their official prices. As a result, OPEC's output in 1981 was 26% lower than the 1979 output and in fact was the lowest it had been since 1970.

In retrospect, one important question presents itself. Could the astronomical rise in the oil prices have been held in check? Sheikh Yamani seems to think so. He expressed the view that if the United States government and the governments of other major consuming nations intervened at the time and forbade the oil companies from trading in the Spot Market, the prices could have been checked and the panic would have ended.

6. ANATOMY OF OIL CRISES

The world witnessed two oil crises during the last 30 years of the 20th century. The topic will return to the front pages soon, however, because the world may be confronted by a third oil crisis. This third crisis promises to be similar to but have a more modest economic impact than the two previous crises, unless the price of oil hits the \$50/barrel mark.

6.1 Definition of Oil Crises

For economic purposes, an oil crisis is defined here as an increase in oil prices large enough to cause a worldwide recession or a significant reduction in global real gross domestic product (GDP) below projected rates by two to three percentage points.

The 1973 and 1979 episodes both qualify as oil crises by this definition. The 1973 oil crisis caused a decline in GDP of 4.7% in the United States, 2.5% in Europe, and 7% in Japan. According to the U.S. government, the 1979 increase in oil prices caused world GDP to drop by 3% from the trend.

The price increase following the first oil crisis raised consumer payments for oil by approximately \$473 billion (in real 1999 U.S. dollars), whereas the second oil crisis increased consumer expenditure by \$1048 billion. By contrast, the oil price hikes during 1999–2001 raised consumer expenditure by \$480 billion.

6.2 Characteristics of Oil Crises

The 1973 and 1979 crises shared four characteristics. First, the disruption in oil supplies occurred at a time of rapid expansion in the global economy. The rapid economic growth fueled greater consumption of oil. In the 5 years that preceded the 1973 crisis, global oil consumption had grown from 38.3 million barrels/day in 1968 to 52.7 million barrels/day in 1972, an average annual increase of 7.5%. Similarly, in the 5 years preceding the 1979 crisis, global consumption had risen from 53 million barrels/day in 1974 to 63 million barrels/day in 1978, an average annual increase of 3.8%.

Second, both disruptions occurred at a time when the world crude oil production was operating at virtually full capacity. Global capacity utilization reached 99% in 1973, with OPEC accounting for 56% of total production. The second oil crisis had seen a deficit of 5 million barrels/day resulting from the disruption in Iranian oil production. Third, each crisis took place at a time when global investment in oil exploration had been declining, making it impossible to achieve a speedy increase in non-OPEC production. In both 1973 and 1979–1980, the global oil industry was at the end, rather than the start, of a new surge in non-OPEC output. Fourth, in both crises, OPEC members had made a deliberate decision to reduce oil production in order to achieve political ends.

6.3 Underlying Factors

In a tight oil market, any of the following underlying factors could, individually or collectively, trigger a price escalation reminiscent of the spot market prices of 1979 and precipitate an oil crisis.

6.3.1 The Global Oil Demand

Economic growth and population growth are the most important drivers behind increasing global

energy demand. The global economy is projected to grow by 3.2% per annum, on average, to 2025. Global GDP is projected to rise from \$49 trillion in 2000 (year 2000 dollars purchasing power parity) to \$108 trillion in 2025 and \$196 trillion in 2050.

World population is expected to grow from 6 billion in 2000 to 8 billion in 2020. The population growth among the 4.8 billion people living in developing countries is estimated at 1.7% per annum. This compares with an average 0.3% per annum in the developed countries. Expanding industrialization and improving standards of living will contribute significantly to the growing energy demand.

The developed countries produce approximately one-third of global oil but consume two-thirds, whereas the developing countries produce two-thirds but consume only one-third (see Table I). Annual per capita consumption in the developing countries is 2 barrels/year. This compares with 14.2 barrels/year in the developed countries and 25 barrels/year in the United States.

The International Energy Agency (IEA) and the U.S. Department of Energy forecast that the world oil demand will grow from 76.1 million barrels/day in 2001 to 95.8 million barrels/day in 2010 and 115 million barrels/day in 2020, with Middle East producers having to meet the major part of the additional demand. However, this will depend on sufficient investment to expand production capacity (see Table II).

6.3.2 The Global Sustainable Productive Capacity

Sustainable productive capacity is here defined as “being attainable within thirty days and sustainable

TABLE I

World Crude Oil Production versus Demand in 2001

Region	Production (billion barrels)	Share (%)	Demand (billion barrels)	Share (%)
Developed countries	8	30	18	64
Developing countries	19	70	10	36
World	27	100	28	100
United States	2	7	7	25

Sources: British Petroleum Statistical Review of World Energy, June 2002; and U.S. Energy Information Administration, June 2001.

TABLE II

World Oil Demand and Supply (Million Barrels/Day), 2000–2020

	2000	2001	2005	2010	2020
World demand	76.2	76.1	83.5	95.8	114.7
World supply					
Non-OPEC	45.2	46.1	44.7	43.6	49.6
OPEC	29.3	30.0	36.0	45.9	51.1
Stock change	–1.7	—	—	—	—
Synfuels	1.2 ^a	1.3 ^a	1.8 ^a	2.7 ^a	4.2 ^a
Total supply	76.2	76.1	80.7	89.5	100.7
Global oil deficit	—	—	–2.8	–6.3	–14.0

Sources: U.S. Department of Energy; British Petroleum Statistical Review of World Energy, July 2002; and International Energy Agency.

^aSynfuel oil production is already included in non-OPEC supply figures.

for three months.” OPEC’s sustainable productive capacity stood at 29.7 million barrels/day in January 2001 with a 99% capacity utilization (see Table III). The organization’s immediately available spare capacity stood then at 390,000 barrels/day. However, with the production cutbacks since March 2001, spare capacity has risen to 4 million barrels/day.

The capital costs of maintaining and expanding OPEC’s capacity over a 5-year period are estimated at \$112 billion, money that members do not have. These projected costs are based on the member countries’ planned capacity increase of 7 to 36.7 million barrels/day.

There is no non-OPEC spare capacity. The financial incentive of high oil prices and firm demand mean that every non-OPEC oilfield is being exploited and any capacity brought onstream is being utilized as quickly as possible.

6.3.3 The Ultimate Global Proven Reserves

World ultimate conventional oil reserves are estimated at 2000 billion barrels. This is the amount of production that would have been produced when production eventually ceases. Different countries are at different stages of their reserve depletion curves. Some, such as the United States, are past their midpoint and in terminal decline, whereas others are close to midpoint, such as the United Kingdom and Norway. The U.K. sector of the North Sea is currently at peak production and is set to decline at approximately 6% per year. However, the five major Gulf producers—Saudi Arabia, Iraq, Iran, Kuwait, and

TABLE III

OPEC Sustainable Capacity and Capacity Utilization (Million Barrels/Day) in January 2001

Country	Capacity	Production	Capacity utilization
Algeria	0.880	0.860	98%
Indonesia	1.300	1.300	100%
Iran	3.500	3.500	100%
Iraq	3.000	2.900	97%
Kuwait	2.200	2.200	100%
Libya	1.450	1.450	100%
Nigeria	2.100	2.100	100%
Qatar	0.720	0.700	97%
Saudi Arabia	9.250	9.000	97%
United Arab Emirates	2.400	2.400	100%
Venezuela	2.900	2.900	100%
Total	29.700	29.310	99%

Sources: Energy Intelligence Group’s “Oil Market Intelligence”; Petroleum Review, April 2000; International Energy Agency; and Author’s projections.

United Arab Emirates—are at an early stage of depletion and can exert a “swing” role, making up the difference between world demand and what others can supply. They can do this only until they themselves reach their midpoint of depletion, probably by 2013.

The expert consensus is that the world’s midpoint of reserve depletion will be reached when 1000 billion barrels of oil have been produced—that is to say, half the ultimate reserves of 2000 billion barrels. With 935 billion barrels already produced, this will occur probably between 2004 and 2005. The yet-to-find (YTF) oil reserves are estimated at 280 billion barrels (see Table IV). As the world production peak approaches, the oil price will soar.

However, if the potential of unconventional oil, such as tar sand oil and extra heavy oil, is included, amounting to 572 billion barrels, then the midpoint of depletion could be delayed for a few more years—but not beyond 2010.

In 1956, the geologist M. King Hubbert predicted that U.S. oil production would peak in the early 1970s. Almost everyone, inside and outside the oil industry, rejected Hubbert’s analysis. The controversy raged until 1970, when U.S. production of crude oil started to fall. U.S. production peaked at 9.64 million barrels/day in 1970 and has been falling since then, reaching 5.77 million barrels/day by 2001. Hubbert was proven right and his bell-shaped curve became a useful tool of oil production analysis.

TABLE IV

Ultimate Global Reserves of Conventional Oil and Depletion Rate (End of 2001)

	Volume or rate	Description
Ultimate reserves (billion barrels)	2000	Amount of production when production ceases
Produced so far (billion barrels)	935	Until the end of 2001
Yet-to-produce (billion barrels)	1065	Ultimate reserves less produced
Discovered so far (billion barrels)	1720	Produced plus remaining reserves
Yet-to-find (billion barrels/year)	280	Ultimate reserves less discovered
Discovery rate (billion barrels/year)	7	Annual additions from new fields
Depletion rate (%)	3	Annual production a percentage of yet-to-produce

Sources. United States Geological Survey, and British petroleum Statistical Review of World Energy, June 2002.

Around 1995, several analysts began applying Hubbert's method to world oil production. Based on Hubbert's pioneering work and an estimated 1.8–2.1 trillion barrels of ultimate reserves, they established that the peak production year will be sometime between 2004 and 2009. If the predictions are correct, there will be a huge impact on the global economy, with the industrialized nations bidding against one another for the dwindling oil supply. One promising oil province that remains unexplored is the Spratly Islands in the South China Sea, where exploration has been delayed by conflicting claims to the islands by six different countries. Potential reserves in the disputed territories are estimated at multibillion barrels of oil and gas. But even if the South China Sea oil reserves are proven, they could hardly quench China's thirst for oil. By 2010, China is projected to overtake Japan to become the world's second largest oil importer after the United States.

Another promising province is the Caspian Basin. Estimates of 40 to 60 billion barrels as the ultimate reserve base of the Caspian region are judged to be reasonable by most geologists familiar with the region. Apart from the limited size of the reserves, the area's oil is very costly to find, develop, produce, and transport to world markets. Projected Caspian Sea oil production of 2–3 million barrels/day by 2010 can be achieved only when prices exceed \$20/barrel (in real terms). Oil prices will be the key factor in the expansion of Caspian Sea oil.

6.3.4 New Oil Discovery Rates

The widely held view that improved seismic surveying and interpretation have improved drilling success rates is not borne out by the level of discoveries during the period 1992–2001 (see Table V).

TABLE V

Global Crude Oil Reserves Additions^a (Billion Barrels) 1992–2001

Year	Added in year	Annual production	As % of annual production
1992	7.80	23.98	33
1993	4.00	24.09	17
1994	6.95	24.42	28
1995	5.62	24.77	23
1996	5.24	25.42	21
1997	5.92	26.22	23
1998	7.60	26.75	28
1999	13.00	26.22	50
2000	12.60	27.19	46
2001	8.90	27.81	32
1992–2001	77.63	256.87	30
Average	7.76	25.83	30

Sources. IHS Energy Group's 2002 World Petroleum Trends Report, and British Petroleum Statistical Review of World Energy, 1993–2002.

^aExcluding the United States and Canada.

The race for reserves is, therefore, on. With the demand for oil envisaged for the next 10 years, the world will consume an average 30 billion barrels per year over that period. If the global oil industry wants to replace this consumption with new reserves without diluting the world's existing proven reserves of some 1 trillion barrels, it must find an additional 300 billion barrels of new oil in the next decade—a daunting challenge indeed.

6.3.5 Global Reserve Depletion Rate

Globally the reserve depletion rate is generally calculated at 3%. This means that to sustain the

world's current 76 million barrels/day consumption at that level, approximately 4 million barrels/day of new capacity is needed every year. Against this, the world production capacity has remained static or even declined while consumption has been increasing.

This means that the Middle East producers, with 65% of the world's proven reserves and just one-third of global production, will assume clear-cut leadership of the supply side of the oil market.

7. *THIRD OIL CRISIS?*

The parallels between current conditions and the early 1970s are unnerving. The conditions that made the 1973 and 1979 oil crises possible exist in the early 2000s. First, world oil consumption is growing rapidly and is projected to continue expanding. World oil demand grew by 2.4% in 2000, 2.2% in 2001 with a projected additional annual increase of 1.8% until 2005.

Second, every indicator points to the fact that production from OPEC will not rise substantially. OPEC's maximum production capacity is estimated at 29.7 million barrels/day. The organization must produce approximately 30 million barrels/day to maintain a supply-demand balance under the current level of stocks. Yet, an output level of 30 million barrels/day may be beyond the organization's reach or, if it is feasible, achievable only with production from Iraq.

Third, the pattern of low-level industry investment in 2000–2001 is also very similar to that observed during the first and second oil crises. Oil companies' capital investment in exploration in 2000/2001 was slow in the aftermath of low oil prices during 1998–1999. Capital spending on exploration and production by the supergiants—Exxon Mobil, BP Amoco, and Shell—fell 20% to \$6.91 billion in the first half of 2000 from a year earlier. Most companies decided to wait before boosting investment, preferring instead to buy back stocks. This has reduced the capital investment available for production and production capacity expansion.

Fourth, as in 1973 and 1979, the causes of this crisis will be a reduction in OPEC production and stagnation in non-OPEC's taking place at a time of rapid economic growth.

Fifth, as in 1973 and 1979, it will be a major political event that could send oil prices rocketing and thus precipitate a third oil crisis. The 1973 Arab–Israeli war was behind the first oil crisis, whereas the Iranian revolution was at the heart of

the second. Similarly, a war against Iraq could precipitate a third oil crisis.

Such conditions make for a truly tight market with the potential for developing into a genuine crisis. But even if the world manages to escape a third oil crisis, oil-market turmoil can be expected to last until 2004.

However, a growing number of opinions among energy experts suggest that global conventional oil production will probably peak sometime during this decade, between 2004 and 2010. Declining oil production will cause a global energy gap, which will have to be filled by unconventional and renewable energy sources.

Various projections of global ultimate conventional oil reserves and peak years have been suggested by energy experts and researchers between 1969 and 2002. The extreme end of opinion is represented by the United States Geological Survey (USGS), IEA, and the U.S. Energy Information Administration (EIA) (see Table VI).

The estimate by EIA is so implausibly high that it can be ignored, whereas the USGS estimate includes 724 billion barrels of YTF reserves. Such huge YTF reserves require discovering an additional amount of oil equivalent to the entire Middle East. But since 90% of global conventional oil has already been found, trying to find 724 billion barrels of new oil is not only an exceptionally daunting task, but virtually impossible.

However, such estimates are of only limited relevance. What is important when attempting to identify future supplies are two key factors: the discovery rate and the development rate, and their relationship to the production rate.

The technology for extracting oil from tar sands, oil shale, and extra heavy oil, known collectively as synfuels, exists but extraction costs are high. Synfuel oil is usually 3 times as labor-to-energy intensive and 10 times as capital-to-energy intensive as conventional oil.

Whereas some—and possibly a great deal—of unconventional oil (synfuels) will eventually be available, there will not be enough to replace the shortfalls in conventional oil. Synfuels will be hard-pressed to meet 3% of the global oil demand in 2010 and 4% in 2020, because of the slow extraction rate and the huge investments needed.

In 2002, only 35,000 barrels/day of natural gas to liquid oil is produced worldwide. This is projected to rise to 685,000 barrels/day by 2010, equivalent to 0.7% of global demand. The constraint, however, might be the very large capital commitment. For a

TABLE VI

Various Projections of Global Ultimate Conventional Oil Reserves and Peak Year (Billion Barrels)

Author	Affiliation	Year	Estimated ultimate reserves	Peak year
Hubert	Shell	1969	2100	2000
Bookout	Shell	1989	2000	2010
Mackenzie	Researcher	1996	2600	2007–2019
Appleby	BP	1996		2010
Invanhoe	Consultant	1996		2010
Edwards	University of Colorado	1997	2836	2020 ^a
Campbell	Consultant	1997	1800–2000	2010
Bernaby	ENI	1998		2005
Schollenberger	Amoco	1998		2015–2035 ^a
IEA	OECD	1998	2800	2010–2020 ^a
EIA	DOE	1998	4700	2030 ^a
Laherrere	Consultant	1999	2700	2010 ^a
USGS	International Department	2000	3270	—
Salameh	Consultant	2000	2000	2004–2005
Deffeyes	Princeton University	2001	1800–2100	2004

Sources. Various.

Note. BP, British Petroleum; DOE, U.S. Department of Energy; EIA, U.S. Energy Information Administration; IEA, International Energy Agency; OECD, Organisation for Economic Cooperation and Development; USGS, U.S. Geological Survey; ENI, Ente Nazionale Idrocarburi.

^aThese ultimate reserve estimates include extra heavy crude, tar sands, oil shale, and also projected production of gas-to-liquid oil.

production of 400,000 barrels/day, this would amount to \$8 billion.

In 2000, renewable energy sources contributed 2% to the global primary energy demand. However, by 2025 they are projected to contribute 4%, rising to 5% by 2050 (see Table VII).

Fuel-cell motor technology will eventually have a great impact on the global consumption of gasoline and diesel. But it could take years before hydrogen-powered cars dominate the highways and certainly not before they are able to compete with today's cars in terms of range, convenience, and affordability.

Fossil fuels, with a growing contribution from nuclear energy, will, therefore, still be supplying the main share of the global energy needs for most—perhaps all, of the 21st century.

8. IMPLICATIONS FOR THE GLOBAL ECONOMY

Twenty years ago, oil crises wreaked havoc on the world economy. Today, although oil remains important, a new crisis will have a much more modest impact because of the diminished role of oil in the global economy, unless the price of oil rises to \$50/barrel.

TABLE VII

World Primary Energy Consumption^a (mtoe), 2000–2050

	2000	2025	2050
Primary energy	9631	16,618	19,760
Oil	3835	6429	7344
Natural gas	2190	4760	6207
Coal	2136	3283	3037
Nuclear	636	733	825
Hydro	617	818	1440
Renewables ^b	217	595	907
Renewables: % of total	2	4	5

Sources. Shell International, Scenarios of 2050; Organization of Petroleum Exporting Countries Statistical Review of World Energy, June 2002; and U.S. Energy Information Administration International Energy Outlook 2002.

^amtoe = Million tonnes oil equivalent.

^bExcluding hydro (i.e., hydroelectric power).

One reason is that oil represents a much smaller share of global GDP today than it did during the two previous crises. Another reason is that the economy is more open, which makes it harder for companies to pass on cost increases to customers. A third reason is that customers can hedge against price increases.

Finally, economies around the world have become more adaptable in their use of every resource, including oil. Consequently, increases in oil prices trigger substitutions of natural gas or any other energy sources in manufacturing. For instance, if one takes the case of the United States, which accounts for 25% of the global oil consumption, it is found that a rise in oil prices today is much less important to the U.S. economy than it has been in the past. The decline in importance of oil prices can be seen from Table VIII, which shows the share of crude oil in nominal GDP.

Table VIII shows that crude oil accounted for 4% of GDP at the time of the first oil crisis, 6% at the time of the second oil crisis, and 2.4% at the time of the Gulf War. In 2001, it accounted for 2%. However, if the price of oil hits the \$50/barrel mark in 2002, then crude oil will have the same share in the U.S. economy as in 1974, with adverse economic implications for the U.S. and the global economy.

Despite the above, higher oil prices still are very important to the world economy. A new report by the International Monetary Fund (IMF) notes that a war on Iraq could be the final straw for an international economy still struggling with the aftermath of September 11, the 20% drop in world share prices this year, and the implosion of the Argentine and Brazilian financial services. The IMF argues that the risks are “primarily on the downside” even before the possibility of a conflagration in Iraq and the Middle East is taken into account.

The IMF warns of the consequences for the oil markets of a war in the Gulf region. It says that a

\$15/barrel increase in the price of oil would be a severe blow to the global economy, knocking at least 1 percentage point off global GDP and sending the world economy spiraling toward recession.

Certainly, all the evidence from past conflicts in the region, including the 1973 Arab–Israeli War, Iranian revolution, and the Gulf War of 1990–1991, suggests that each of these conflicts has been followed by recession and unemployment, particularly in the West.

This time around, the scenario is even more worrying. Financial markets are already paralyzed as a result of the burst of the technology bubble and the mammoth corporate accounting frauds in the United States.

Before Washington’s decision to up the ante with Iraq, there was some hope at the IMF that there would be some kind of bounce-back for the global economy in 2003. The world economy was projected to pick up speed, growing by 3.7% in 2003, after slipping this year to 2.4%. The United States was forecast to pick up from a 1.4% expansion this year to 2.3% in 2003.

But higher oil prices could make all the difference. The IMF notes that oil prices began to surge in the second half of August 2002 as the U.S. administration began to talk up the war with Iraq. The price of crude oil on both sides of the Atlantic is already above the \$22 to \$28 per barrel price range built into most major economic projections. For oil-importing countries, higher oil prices act as an immediate tax on consumption and business. An oil price hike makes it less likely that consumers will spend and business will invest.

With the global economy struggling to pick up momentum, a prolonged period of high oil prices would, without doubt, delay an upturn. It would almost certainly give another leg to the downturn in share prices—in 2003, down by 40% since the peak in March 2000—hitting confidence hard.

Although the loss of Iraqi oil supplies over the short term might make little significant difference, it is the collateral damage that could deliver the most serious blow. There is the possibility that Iraq might seek to disable the critical Saudi and Kuwaiti oilfields, as in the Gulf War, or that the Arab countries—outraged by America’s attack on one of their own—might seek to impose an oil embargo or refuse to increase supplies to make up for the loss of Iraqi output.

The result would be a hike in prices on world oil markets and disruption to an already misfiring global economy. However, once the war is over, there may well be cheaper and more plentiful oil as Iraq’s

TABLE VIII

Oil in the U.S. Economy: Consumption, Prices, and Percentage of GDP

	Oil consumption (million barrels/day)	Oil prices (\$/barrel)	Nominal GDP (\$ billion)	Oil consumption as percentage of GDP
1974	16.65	9.07	1382	4.0
1980	17.06	28.07	2784	6.3
1990	16.99	22.22	5744	2.4
1999	19.36	17.60	8857	1.4
2000	19.63	27.72	9224	2.2
2001	19.90	25.93	9462	2.0
2002 ^a	20.18	30.00	9611	2.3
2002 ^a	20.18	50.00	9611	4.0

Sources. Courtesy of PKVerleger LLC; British Petroleum Statistical Review of World Energy, June 2002; U.S. Energy Information Administration Annual Energy Outlook 2002.

^a Estimates.

production is restored. But by then the damage to a rickety world economy will have been inflicted.

SEE ALSO THE FOLLOWING ARTICLES

Geopolitics of Energy • Markets for Petroleum • Nationalism and Oil • National Security and Energy • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Industry, History of • Oil-Led Development: Social, Political, and Economic Consequences • Oil Price Volatility • OPEC, History of • OPEC Market Behavior, 1973–2003 • Strategic Petroleum Reserves • War and Energy

Further Reading

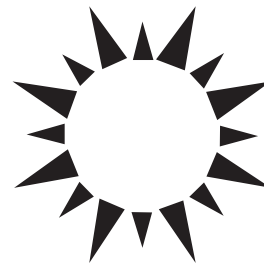
Deffeyes, K. S. (2001). "Hubbert's Peak: The Impending World Oil Shortage." Princeton University Press, Princeton, NJ.

- Nixon, R. (1980). "The Real War." Sidgwick & Jackson, London.
- Robinson, J. (1988). "Yamani: the inside story." Simon & Schuster, London.
- Salameh, M. G. (1990). "Is a Third Oil Crisis Inevitable." Biddles, Guildford, UK.
- Salameh, M. G. (1999). Technology, oil reserve depletion and the myth of reserve-to-production (R/P) ratio. *OPEC Review*, June 1999, pp. 113–124.
- Salameh, M. G. (2001). Anatomy of an impending third oil crisis. *In* Proceedings of the 24th IAEE International Conference, April 25–27, 2001, Houston, TX, pp. 1–11.
- Salameh, M. G. (2001). "The Quest for Middle East Oil: the U S Versus the Asia-Pacific Region." International Asian Energy Conference, August, Hong Kong.
- Sampson, A. (1980). "The Seven Sisters." Hodder & Stoughton, London.
- Verleger, P. K. (2000). "Third Oil Shock: Real or Imaginary?" Consequences and Policy Alternatives. International Economics Policy Briefs, No. 00-4, Institute for International Economics, Washington, DC.
- Yergin, D. (1991). "The Prize: The Epic Quest for Oil, Money and Power." Simon & Schuster, New York.

Oil Industry, History of

AUGUST W. GIEBELHAUS

Georgia Institute of Technology
Atlanta, Georgia, United States



1. Petroleum Prior to the Modern Era
2. Oil's Age of Illumination
3. Competition and Monopoly
4. Technological Change and Energy Transition
5. Abundance, Scarcity, and Conservation
6. The Globalization of the Industry
7. International Competition and War
8. The Middle East, OPEC, and the Energy Crisis
9. Oil in the Modern Economy

Glossary

barrel The international standard of measure for crude oil and oil products equivalent to 42 U.S. gallons and used since first defined in the U.S. oil fields in the 1860s.

cracking Various refining processes using heat, pressure, and catalysts to change less volatile and heavier petroleum fractions into compounds with lower boiling points.

crude oil Unrefined petroleum as it comes from the well, consisting almost entirely of carbon and hydrogen compounds and varying widely in appearance, color, odor, and the presence of sulfur, nitrogen, oxygen compounds, and ash.

illuminating oil Common 19th-century name for kerosene, the fraction of crude oil between gas oil and gasoline on the refiner's scale, usually ranging between 105 and 300°F; represented the first major commercial petroleum product and defined the early decades of the industry.

law of capture Legal principle emanating from the courts in the 19th century that encouraged the rapid drilling and pumping of oil from underground pools; a major cause of wasteful practice in the early oil industry.

Organization of Petroleum Exporting Countries (OPEC) Organized in 1960 to establish a united front against the power of the world's multinational oil companies, it became a force to be reckoned with in the 1970s and remains a relevant player in the economics of world oil.

refining Processes employed to transform crude oil into useful commercial products; includes fractional distillation, cracking, purifying, and treating.

rotary drilling Method for drilling deep holes in search of petroleum or gas in which a drill bit is attached to a revolving drill pipe; used first in the early 20th century.

Spindletop Enormous oil strike or "gusher" on the Texas Gulf Coast near Beaumont in 1901 that changed the face of the oil industry. Economic forces unleashed brought a market erosion of Standard Oil dominance and the birth of oil's new "age of energy" with the increased production of fuel oil and gasoline.

Standard Oil Parent firm of the John D. Rockefeller oil interests, initially capitalized as an Ohio corporation, reorganized as a New Jersey holding company in 1899, and dissolved by the U.S. Supreme Court in a famous antitrust decision in 1911.

Most global citizens today view the international oil industry as an enormous industrial power whose financial profits have arisen from the utilization of one of nature's gifts, often through shrewdness and technological innovation but sometimes also as a result of monopolistic practices and exploitation of Third World resources. During the time of the price dislocations of 1973 and 1979 that defined the "energy crisis" of the decade, many pointed to domination of an oligopoly of multinational firms defined by British journalist Anthony Sampson as the "seven sisters." An aura of monopolistic competition and "big business" has surrounded this industry from its beginnings, as the Rockefeller Standard Oil empire remains imprinted as a past symbol of rampant corporate power. However, the history of this industry has been much more complex than a narrative of wealth accumulation by a few major companies. It represents a story of vision, risk, business success, and often disaster played out in an increasingly multinational arena. Oil has emerged as the dominant fuel consumed in the United States, the country in which the industry took life in the modern era, but domestic crude oil production peaked in 1970 and has declined ever since. Once the "Saudi Arabia of the world," the United States

became a net importer of petroleum at the historical moment that an oil regime had come to dominate world energy. The energy crises of 1973 and 1979, the Persian Gulf War of 1991, and the invasion of Iraq in 2003 have all highlighted the geopolitical importance of Middle Eastern oil, and the major vertically integrated, multinational companies have become almost equal players with governments in this arena.

1. PETROLEUM PRIOR TO THE MODERN ERA

Energy historians have emphasized the concept of transition from one energy regime to another as an analytical tool. One model, for example, suggests that early mankind entered the first transition from hunting and gathering to the harvesting of biomass energy through agriculture and silviculture. This relatively straightforward and simple energy form dominated in antiquity, the medieval period, and well into the early modern era. The era of the industrial revolution saw the transition from biomass harvesting to the exploitation of fossil energy in the form of coal. The shift from a predominant reliance on coal to an overwhelmingly significant role for oil and natural gas has defined the last transition. Two aspects derive from this analysis. The first is to understand how very brief has been the “oil age” within the context of human history (approximately 140 years, from 1859 to the present). The other is the speculative argument that since all fossil fuels (coal, oil, and natural gas) will eventually be depleted, we must plan as a global society for the next transition to a more sustainable energy source. These arguments have most often focused on one type of “renewable” fuel or another, whether it be, in the words of energy analyst Amory Lovins, hard or soft. For example, many advocates of nuclear energy still argue that fission breeder reactors promise a potential renewable energy source, but one that many social scientists would criticize as undesirable or hard because of the great potential for environmental degradation. For others, more environmentally benign, soft solutions, such as solar, wind, tidal, or geothermal energy or renewable biomass fuels such as ethanol, provide the answer. These arguments must be left for another discussion, but all experts agree that the age of overabundant, cheap petroleum will end at some point; the problem is to determine when this will occur.

Prior to the industrial era, petroleum had a very limited utility. Records show human uses for oil in the form of natural seepage of asphaltic bitumen as far back as ancient Mesopotamia prior to 3000 BC. Employed as a mastic in construction, a water-proofing for ships, and a medicinal poultice, this oil apparently did have a small market. Later, Near Eastern cultures successfully distilled natural crude oil to obtain lamp fuel and the basic ingredient for “Greek fire,” a military incendiary introduced with great effect by the Byzantine Empire against the rigging of attacking ships. In the West, limited supplies of crude oil seepage were confined mostly to medicinal use. Native Americans had also discovered the medical benefits of crude oil, and when Edwin L. Drake and associates decided to drill for oil in northwestern Pennsylvania in 1859, they were aware that local tribes had been using oil for a very long time.

A market for medicinal oil developed by the early 19th century in Pennsylvania and a number of entrepreneurs entered the business. Samuel M. Kier of Pittsburgh, the most successful of these businessmen, established a successful medicinal market for petroleum or rock oil before the Civil War. These original “snake oil salesmen” claimed a host of cures for rock oil to be used both externally and internally. In a parallel development, a small Russian industry developed around the presence of natural seepage in the Baku region, initially also for medicinal use. However, it took a 19th-century shortage of lamp oil, precipitated by a whale oil crisis and relative scarcity of coal oil, to spark the modern petroleum industry.

2. OIL'S AGE OF ILLUMINATION

In the 1850s, experiments with refined petroleum demonstrated that it could serve as a satisfactory lamp oil (in kerosene form). The problem was that it was not available in suitable quantities. The drilling of water wells in northwest Pennsylvania had uncovered the presence of oil along with brine in a number of instances. Why couldn't one drill for oil? A group of investors hired “Colonel” Edwin L. Drake to obtain leases and attempt to drill for oil using the standard percussion techniques then used for drilling for water. On August 27, 1859, Drake's rig struck oil at 69 ft near Oil Creek, a short distance from the town of Titusville, Pennsylvania. The oil age was born. The market for illuminating oil was

insatiable and the growth of the industry was sudden and dramatic. Crude oil production increased from only 2000 barrels in 1859 to 4.8 million barrels a decade later and 5,350,000 barrels in 1871. From the beginning, the demand for illuminating oil reflected both the burgeoning domestic population and economy and an expanding export business. Between the 1860s and 1900, crude oil and oil products exported ranged from one-third to three-fourths of total U.S. production.

This new industry initially operated in a classically competitive fashion. Ease of entry defined all aspects of the business, and capital investment was minimal. Refining technology was primitive, consisting of heating crude in a still in order to obtain the desired kerosene fraction; other constituent fractions such as gasoline were often run into streams or onto the ground in ditches. The evolution of the legal doctrine of law of capture by the Pennsylvania courts also contributed to wasteful practices in the oil fields. Several different landowners or leaseholders might sit astride an underground oil pool. Drawing an analogy with game being legally captured if it were lured onto another's land, and citing precedent in English common law, the courts upheld the right of the property owner to "capture" oil that had migrated under the owner's land. Limited understanding of oil geology at that time mistakenly led to the belief that oil flowed in underground rivers. Each owner or leaseholder rushed to drill and pump before their neighbors depleted the oil pool. Oil markets became characterized by alternate periods of oil glut followed quickly by scarcity. This wasteful practice continued as the oil frontier moved westward from Pennsylvania, Ohio, and Indiana in the latter 19th century to Texas, California, and the midcontinent in the 20th century.

3. COMPETITION AND MONOPOLY

The young John D. Rockefeller entered this rough-and-tumble business in 1863. He had cut his teeth as a partner in the wholesale grocery business in Cleveland, doing well as purveyor of supplies to Ohio military units. However, Cleveland possessed several locational advantages that encouraged its development as an important refining center. Located on Lake Erie, close to the Pennsylvania oil fields, it had railroad trunkline connections to major eastern markets, a readily available workforce, and sources of financial capital. The industry, then as today,

consisted of four main functional sectors: production, refining, transportation, and marketing. Production entails all the activities involved in "getting it out of the ground." This includes exploration, drilling, leasing, pumping, primary and secondary recovery techniques, and all other associated activities. Refining entails all relevant processes developed to obtain usable products from crude oil in its natural state. Initially, this included only fractional distillation, the heating of crude to boil off its constituent parts, but has evolved today to include a host of sophisticated specialty practices. Transportation, the third element of the industry, encompasses the movement of crude oil from the oil field to the refinery and the shipment of refined product to market. In the beginning of the industry in the 1860s and 1870s, this meant by barge, railroad car, or horse-drawn wagon. Oil transport evolved to include petroleum pipelines, tanker trucks on the highway, and fleets of tankers on the high seas. Marketing envelops the distribution and sale of petroleum products to the consumer. 1870s marketing might have consisted of a Standard Oil horse-drawn tank wagon selling branded "illuminating oil" or kerosene in neighborhoods; today, we think of well-advertised gasoline stations and slick television commercials designed to get us to use one firm's brand over another's. Although we think of the industry today as consisting of a few vertically integrated firms (those that are engaged in all four major functions from production through marketing), the oil fraternity has always contained a strong element of independents. These individual producers, refiners, pipeline operators, and marketers often viewed themselves as opponents of the big companies such as Standard Oil and they still represent a distinct voice in the industry.

Rockefeller's early endeavors were concentrated in the refining sector of the industry, achieving growth through horizontal combination; he did not adopt a strategy of vertical integration until later. First as a partnership, then organized as a corporation in 1870, Rockefeller and associates' Standard Oil Company proceeded to achieve what was later termed the conquest of Cleveland. Using tactics that one could define as either shrewd or unethical, Rockefeller succeeded in dominating the Cleveland refining business. Concentrating on transportation as the key to control, Standard used its size to obtain preferential shipping arrangements with the railroads through the use of rebates and then sought to eliminate competition with a series of pooling and cartel agreements. Rockefeller's attempts to control

the refining business with the South Improvement Company and later the National Refiners Association remain textbook examples of business strategy in the late 19th century. When Rockefeller found informal agreements inadequate to achieve tight industry control, he changed his approach to one of merger and acquisition.

By 1878, Standard controlled more than 90% of total U.S. refining capacity. To formalize this economic empire, Rockefeller and associates created the Standard Oil Trust Agreement in 1882, later reorganized as the Standard Oil Company (New Jersey) as a legal holding company under New Jersey law in 1899. Rockefeller integrated backward in the 1880s by acquiring producing properties in Ohio, into transportation with his own pipelines, and forward into marketing with his branded cans of kerosene for domestic sale and export. The Standard near monopoly became the hated enemy of independent oil producers, who found their prices driven down by Rockefeller buying power, and by smaller operators, who had been forced to sell out in the face of price-cutting attacks. As a public outcry against the power of large Gilded Age corporations emerged, "muckraking" exposés by Henry Demerest Lloyd and Ida M. Tarbell, a daughter of an independent oilman, painted Standard Oil as a hated symbol of monopoly in America.

Standard had been under legal attack in state courts by its enemies and competitors for years, but the passage of the Sherman Antitrust Act in 1890 had laid the groundwork for federal challenge. With a change in political climate during the presidency of Theodore Roosevelt toward progressivism, federal investigations went forward, culminating in the forced dissolution of Standard Oil into several constituent companies in 1911. This did not hurt Rockefeller personally too much because he remained a stockholder in these many companies, and historians have argued that the former monopoly had simply transformed into an oligopoly of large vertically integrated companies. For example, Standard Oil (New Jersey) suffered for a time because it was a large refiner and marketer but lacked integrated production facilities. Soon, however, it rectified this situation with the acquisition of oil production in the Oklahoma and Texas fields. Similar strategies had enabled most of the spun-off firms to emerge as strong vertically integrated units by the early 1920s. However, market forces other than court decisions had already begun to change the U.S. oil industry in significant ways.

4. TECHNOLOGICAL CHANGE AND ENERGY TRANSITION

Economists Harold F. Williamson and Ralph Andreado have argued that by the eve of the 1911 dissolution decree the competitive structure of oil already had been altered, the key factor being the discovery of vast supplies of oil on the Texas Gulf Coast at Spindletop in 1901. If the Drake well in 1859 had trumpeted the birth of the illuminating oil industry, Spindletop marked the birth of its new age of energy. Spindletop and subsequent other new western fields provided vast oil for the growing economy but also enabled firms such as Gulf, Texas, and Sun Oil, independent of the Standard interests, to gain a substantial foothold. The gulf region was underdeveloped and attractive to young companies, and the state of Texas was hostile toward the Standard Oil monopoly. Asphaltic-based Spindletop crude made inferior grades of kerosene and lubricants but yielded a satisfactory fuel oil. New firms, such as Gulf and Texaco, easily integrated forward into the refining, transportation, and marketing of this fuel oil in the coal-starved Southwest. Located near tidewater on the Gulf Coast, small investments in pipelines also enabled operators to get their crude to seagoing tank ships, which could then carry it to other markets. The Sun Oil Company of Pennsylvania, a long-time competitor of Standard, built a robust business by shipping Texas crude to its refinery outside Philadelphia.

In 1900, Standard Oil (New Jersey) controlled approximately 86% of all crude oil supplies, 82% of refining capacity, and 85% of all kerosene and gasoline sold in the United States. On the eve of the 1911 court decree, Standard's control of crude production had declined to approximately 60–65% and refining capacity to 64%. Moreover, Standard's competitors now supplied approximately 70% of the fuel oil, 45% of the lubricants, 33% of gasoline and waxes, and 25% of the kerosene in the domestic market. Newly formed post-Spindletop companies such as Gulf and Texaco, along with invigorated older independent firms such as Sun and Pure, had captured significant market share. Meanwhile, Standard, heavily invested in the traditional kerosene business, was slow to move into the production and marketing of gasoline.

At approximately the same time, the automobile, which had appeared in the 1890s as a novelty, had begun to shed its elitist image. The introduction of mass-produced and relatively inexpensive vehicles, led by the 1908 Ford Model T, very quickly

influenced developments in the oil business. In 1908, the total output of the U.S. auto industry was 65,000 vehicles. In less than a decade, Ford alone sold more than 500,000 units annually. Within that same decade, both the volume of production and the total value of gasoline passed those of kerosene. By the turn of the century, the increasing electrification of urban America and widespread use of the Edison incandescent lamp were already a worry to oil companies whose major product was kerosene. Now the oil industry was becoming increasingly concerned about how it could boost production of gasoline to meet demand.

As the oil industry faced real or apparent shortages of supply, one response was technological. New production methods such as the introduction of rotary drilling early in the century increased crude supplies, and refinery innovations enabled crude stocks to be further extended. However, as the demand for gasoline increased, new oil discoveries in Texas, California, and Oklahoma in the early 1900s proved insufficient. Once discarded as relatively useless, this lighter fraction typically constituted 10–15% of a barrel of crude oil. Refiners stretched the gasoline fraction by including more of the heavier kerosene, but this resulted in an inferior motor fuel. One could enrich the blend of gasoline with the addition of what would later be termed higher octane product obtained from selected premium crudes (e.g., from California) or “natural” or “casinghead” gasoline yielded from highly saturated natural gas. The most important breakthrough, however, occurred with the introduction of thermal cracking technology in 1913 by a Ph.D. chemist employed by the Standard Oil Company (Indiana), William Burton. The industry had employed light cracking, the application of heat to distillation to literally rearrange hydrocarbon molecules, since the 1860s to obtain higher yields of kerosene from feed stock. By dramatically increasing the temperature and pressure of his cracking stills, Burton discovered that he could double the output of gasoline obtained over previous fractional distillation and cracking methods. An additional advantage was that this gasoline was of generally superior quality. Although the problem of “knocking” in the internal combustion engine was not yet understood fully, it would become a major technological challenge in the 1920s as higher compression, higher performance engines appeared.

Oil’s growth as an energy source between 1900 and 1920 coincided with an enormous increase in total U.S. energy consumption. Samuel J. Schurr and

Bruce Netschert calculated that total energy use increased by 123% during this span of two decades. Although oil companies in the Southwest had begun to charge that “King Coal’s” reign was over, aggregate statistics show that the industry’s share of total energy consumption had only risen from 2.4% of the total in 1900 to 12.3% in 1920. Coal remained dominant, with its share of total energy consumption actually slightly increasing from 71.4 to 72.5%. The actual transition of fossil fuel dominance from coal to oil would come later in the century, but the foundation had been set.

5. ABUNDANCE, SCARCITY, AND CONSERVATION

From its earliest beginnings in the 19th century, the oil industry had to deal with the recurrent feast or famine that accompanied the alternate discovery of new fields followed by their depletion. In “Special Report on the Petroleum of Pennsylvania,” compiled in 1874 by geologist Henry Wrigley, it was argued that “we have reaped this fine harvest of mineral wealth in a most reckless and wasteful manner.” John F. Carl of the Pennsylvania Geologic Survey concluded in 1886 that “the great Pennsylvania oil fields, which have supplied the world for years, are being exhausted, and cannot respond to the heavy drafts made upon them many years longer, unless reinforced by new deposits from deeper horizons.” History soon demonstrated that many of these concerns were unfounded, as the discovery of new producing fields once again replaced scarcity and high crude prices with glut and price depression. However, the projections of future supply in the United States published by the U.S. Geologic Survey in 1908 predicted total depletion of U.S. reserves by 1927, based on then current levels of consumption, and the secretary of the interior warned President Taft in 1909 of an impending oil shortage. Driven by the rush to get oil out of the ground by the law of capture, the industry’s production sector remained chaotic and highly unpredictable as the “forest of derricks” that defined each field moved ever westward. Within the context of this volatile business one can read the arguments of the Standard Oil (New Jersey) attorneys defending the company from antitrust attack in a different light. They maintained that Rockefeller’s strategy of consolidation, denounced as evil monopoly, in fact represented a rational attempt to impose order and stability on an unstable industry.

A new era of conservation as well as the first significant enforcement of antitrust law had arrived on the political scene at the beginning of the 20th century with the presidential administration of Theodore Roosevelt. It is important to differentiate the conservation approach, best defined as wise utilization of natural resources, from the preservationist approach of Roosevelt contemporary John Muir, founder of the Sierra Club. Roosevelt and his key advisers, such as chief forester of the United States Gifford Pinchot, did not want to lock away America's resources but did favor planning and responsible exploitation. One should interpret the efforts of the U.S. Geological Survey and the Bureau of Mines (created in 1910) in this light. These two institutions provided some attention to petroleum-related issues, but the creation of the Bureau of Mines' petroleum division in 1914 signaled a heightened federal government interest in oil conservation. Work centered at the division's Bartlesville, Oklahoma Petroleum Experiment Station focused on reservoir behavior, efficient drilling practices, well spacing, and secondary recovery techniques and disseminated knowledge to industry through influential Bureau of Mines publications.

In the 1920s, petroleum conservationists began to encourage unitization or the unit management of oil pools as a central alternative to the wasteful law of capture. This approach encouraged each pool to be operated as a cooperative unit, with the individual leaseholder's percentage share defining the amount of oil that could be pumped out. Operators would drill fewer wells, produce oil at a controlled rate, and either use or return to the producing zone under pressure all natural gas obtained along with the oil rather than the standard practice of flaring it or venting it into the atmosphere. Unitization was supported by the scientific principle that all recoverable oil in its undisturbed state contains gas in solution and that this solution has a lower viscosity, lower specific gravity, and lower surface tension than gas-free oil. It had long been known that natural gas had the same expulsive function in the well that carbon dioxide has in a shaken soda bottle, but reduced viscosity of the oil meant more fluidity and ultimately enhanced recovery. There were champions of this approach in the early 1920s within the industry and the government, among them Henry L. Doherty of the Cities Service Company and George Otis Smith, Director of the U.S. Geologic Survey. However, Doherty failed in efforts to get the American Petroleum Institute, the powerful trade association of the industry founded after World War I, to

support the practice. The industry only began to come around after discoveries of huge amounts of oil in the midcontinental United States in the mid- to late 1920s. Doherty's plea that wise management led to the elimination of economic as well as physical waste began to receive attention.

The problem in the United States was that many operators were involved in the production of any given pool. This was free enterprise at work accompanied by individualistic opposition to government mandate to enforce unit operation. Because of the differing circumstances in most foreign fields, however, a different story unfolded. The Anglo-Persian Oil Company (APOC), like other European firms, had lagged behind its U.S. counterparts in a number of technical areas, including the adoption of thermal cracking and gas absorption plants to obtain rich "natural gasoline." However, because APOC had a monopolistic position in the huge Iranian fields it was able to unitize its operations. By carefully regulating the number of wells drilled, APOC could conserve the volume of natural gas in solution with the petroleum and operate the pool in the most efficient manner to sustain long-term recoverability.

The U.S. oil industry was still operating within the schizophrenic world of alternate scarcity and overabundance. A perceived scarcity of petroleum reserves at the beginning of the 1920s, coupled with increased demand, stimulated other technological developments. The Bureau of Mines conducted extensive research into enhanced and secondary recovery techniques such as the waterflooding of older fields to boost production. Integrated firms developed their own thermal cracking technologies in efforts to circumvent the Burton patents held by Standard of Indiana and increase their own gasoline output. There was also brief flirtation with alternative liquid fuels in the early part of the decade. Standard Oil (New Jersey) marketed a 25% ethanol-gasoline blend in 1922–1923 and later obtained the basic German patents for coal liquefaction in a licensing with the I. G. Farben interests. There was also a boom in Western shale oil in the 1920s, which has left an interesting if eccentric history. All these liquid fuel alternatives would soon prove unnecessary as new strikes of oil again dampened anxiety about shortages.

By 1930, oil's share of total U.S. energy consumption had increased to an impressive 23.8%, the most dramatic shift being in gasoline consumption. In 1920, oil-based motor fuel represented only 2.2% of total oil consumption compared with 48.1% for fuel oil. By 1930, gasoline's percentage

had increased to 42.6% of the total and fuel oil's relative figure had shrunk to 43.5%. Kerosene production and sale had decreased sharply. However, traditional problems still plagued the industry. The opening of the East Texas field in 1930 again demonstrated that abundance was a double-edged sword. As unrestricted flow accelerated, the price of crude that had sold for \$3 a barrel in 1919 had decreased to 10¢ in the summer of 1931. Voluntary or even compulsory unitization of pools was powerless in the face of such a major problem of overproduction. A new emergency approach developed in the states to combat this economic disaster—regulated prorationing, or the enforced limiting of production at each wellhead. Initially introduced under martial law by the Texas and Oklahoma governors in 1931, legislation in 1932 ceded the power to regulate production to the Texas Railroad Commission and the Oklahoma Corporation Commission. These actions did provide some stability: The price of crude increased to 85¢ a barrel by 1934. Federal New Deal legislation supported and augmented state practice as Franklin Roosevelt's National Recovery Administration (NRA) supported "fair competition" under the NRA Oil Code and policed the interstate shipment of "hot oil," oil produced illegally above state-mandated prorationing quotas. When the Supreme Court struck down the NRA in 1935, senators and congressmen from the oil-producing states swiftly moved to fill in the gap of cooperative regulation with passage of the Connolly Hot Oil Act and the Interstate Oil Compact Commission in 1935. These laws would serve to regulate production and combat "economic waste" (e.g., low prices) to the present day.

6. THE GLOBALIZATION OF THE INDUSTRY

At the end of World War I, British Foreign Secretary George Curzon stated that "the allies had floated to victory on a sea of oil." The major world navies had converted their coal-fired boilers to fuel oil in the critical years before the war and the noninterrupted flow of fuel for these vessels did indeed prove crucial. Winston Churchill, who had pushed for the conversion of the Royal Navy to fuel oil, was also the major player in the British government decision to take a controlling 51% financial position in the Anglo-Persian Oil Company (BP) in 1914. Unlike much of the activity of the former European colonial

powers, the history of U.S. operations abroad has centered on the efforts of private firms obtaining concessions and the government has never owned or directly participated financially in actual operations. The U.S. State Department and other agencies have at various times, however, given strong support to these private efforts. Initial U.S. exploratory ventures occurred in Mexico early in the 20th century and in Venezuela in the 1920s. The most important producing area in the world today, the Middle East, had remained a British sphere of influence for a very long time. U.S. firms first obtained concessions in Iraq (Mesopotamia) in 1928 and in 1934 Standard of New Jersey (Exxon) and Standard of New York (Mobil) acquired one-fourth of the Iraqi Petroleum Company. Standard Oil of California (Chevron) gained leases in Bahrain Island in 1927 and Saudi Arabia in 1933, selling half of its rights to Texaco in 1936 when the two companies jointly formed the Arabian-American Oil Company (ARAMCO). The large vertically integrated major firms had control of large domestic supplies, but they viewed the obtaining of foreign crude as a rational economic strategy for their foreign marketing activities and as a hedge against continuing predictions of long-term depletion of U.S. supplies.

Oil imports into the United States prior to 1945 remained low, rarely exceeding 5% of total consumption. However, from 1945 to 1959 the total demand for petroleum products increased by approximately 80%, the single largest share being for gasoline as the automobile culture emerged triumphant in the postwar era. By the 1950s, cheaper, imported oil accounted for approximately 12% of needs and domestic U.S. producers were becoming alarmed by the challenge. Moreover, the Internal Revenue Service had ruled that U.S. companies operating abroad could deduct all foreign oil royalties and taxes from their corporate income tax bill, thus providing a huge financial incentive to pump Middle Eastern crude. President Dwight Eisenhower responded to political pressure from independent oilmen, most of them in western states, by asking the majors to implement voluntary import curbs. In 1959, acknowledging the failure of such voluntary measures, Eisenhower imposed mandatory import quotas.

Despite these quotas, as demand continued to expand in the gas-guzzling United States, imports became seen as the only way to meet needs. When the Nixon administration lifted quotas in 1973, oil imports as a percentage of total U.S. consumption dramatically increased, reaching 38% in 1974. At

the same time, a regulated price structure for domestically produced oil, a carryover from Nixon's 1971 anti-inflation pricing policies, discouraged exploration and drilling at home. Imports increased to approximately one-half of total U.S. consumption by 1979 and have continued at these levels or above.

7. INTERNATIONAL COMPETITION AND WAR

World War II further crystallized the strategic importance of oil to international security, and those nations with ready access to supply were greatly advantaged. When Hitler invaded Poland in 1939, fully one-third of the total oil capacity with which Nazi Germany went to war was represented by synthetic oil derived from brown coal (approximately 19 million barrels per year). Germany's need for oil reserves is evident in Hitler's decision to invade Soviet Russia in June 1941. Most are aware of the German thrust toward Moscow that was only thwarted with Hitler's legions in view of the capital. An equally and perhaps more important southern line of attack aimed at the rich oil-producing properties in Baku in the southern Caucasus is less known. The United States' decision to embargo shipments of high-octane aviation fuel to Japan in 1941 was a significant factor in the Imperial government's decision to launch the war at Pearl Harbor and to attack Dutch interests in the East Indies in search of oil to fuel its military machine.

World War II also represented an important watershed in the history of the U.S. petroleum industry. As the nation mobilized for conflict, there was a validation of the New Deal consensus that a cooperative relationship between business and government represented the best way to achieve and maintain a stable industry in the wake of the inherently uncertain realm of unpredictable resources and rising demand. Much of the suspicion toward New Deal regulation was superseded by a new spirit, imbued with patriotism, but also pleased that production curtailment was being replaced by dramatic efforts aimed at increasing production. The wartime Petroleum Administration for War (PAW) fostered cooperation in pipeline construction, the patent pooling of technical processes, and the development of new catalytic cracking plants needed to produce huge quantities of 100 octane aviation gasoline demanded by new high-performance aircraft. If the allies had indeed floated to victory on a

sea of oil during the 1914–1918 conflict, one can make a good case that they flew to victory in the 1939–1945 war.

A council of oil executives, the Petroleum Industry War Service Committee, was formed to advise the PAW, headed by Secretary of the Interior and New Deal zealot Harold Ickes. The oil czar's new image of working with industry to boost oil production contrasted with his previous reputation among oilmen as Horrible Harold. In 1946, the success of this industry advisory group was so valued by government and industry that President Harry Truman continued it in peacetime with the creation of the National Petroleum Council (NPC). The NPC has continued to function as a consulting and advisory body to government and as a generator of reports and studies on all aspects of the oil industry. It has primarily worked in coordination with the Department of the Interior and, after 1977, the Department of Energy.

Another major outcome of the war was an increased awakening of U.S. interest in foreign oil resources and the laying of the foundation for the structure of oil importation that characterizes the industry today. During World War II, the protection of Middle East oil reserves became an essential aim of U.S. foreign policy.

After the war, the exploitation of oil abroad continued to expand at the same time that conservation in the form of market-demand prorationing continued at home. When Chevron and Texaco invited Mobil and Exxon into partnership with ARAMCO in 1946, the U.S. government response was to grant clearance to the companies from antitrust prosecution, citing the value to the national interest. However, U.S. oil policy was ambivalent during the period of the early Cold War. The State Department continued to support aggressive U.S. firms operating abroad, but the industry was still split between independent companies, which were primarily engaged in the domestic industry, and the larger major multinational firms, which were seeking to expand their interests overseas.

8. THE MIDDLE EAST, OPEC, AND THE ENERGY CRISIS

In the 1950s and 1960s, an increasing number of former European colonies in Asia and Africa were gaining independence. In other areas where colonial powers had not maintained political control in the

strictest sense for some time, such as the Middle East, multinational oil companies effectively controlled the destinies of these nations through their economic power. The formation of the Arab League in 1945 and the first Arab–Israeli war in 1948–1949, which erupted in response to the birth of modern Israel, also pointed to a growing Arab nationalism. With Israel’s victory, the Palestinians were now without a homeland and this issue has dominated Middle Eastern politics ever since. The first direct threat to Western oil interests, however, came not from an Arab state but from a strategically situated nation whose interests have often coincided with her Arab neighbors, Iran.

Iranian Prime Minister Mohammed Mossadegh’s nationalization of British Petroleum’s properties in 1951 led to the first real postwar oil crisis. This nationalist challenge to Western interests, coupled with the proximity of Iran to Soviet Russia, prompted the United States to act. The Central Intelligence Agency took a direct hand in a political counterrevolution that overthrew Mossadegh and installed the U.S.-backed Shah, Reza Pahlavi, to the throne in 1953. With U.S. State Department support, U.S. companies now obtained 40% of a new foreign consortium established under the Shah to exploit Iranian oil. The Shah remained America’s man in Iran until his overthrow during the Iranian revolution of 1979. During that time, private oil interests were served well with access to Iranian crude, and the State Department enjoyed Persian Gulf stability through military and economic support of the Shah.

A new breed of Arab leaders, such as Egypt’s Gamal Abdel Nasser, also symbolized this new nationalistic and independent attitude. Nasser’s seizure of the Suez Canal in 1956 was the precipitating factor in the second Arab–Israeli war. Libya’s Colonel Muammar Qaddafi’s ascension to power in 1969 represented an even more aggressive leadership and an assertion of Muslim values. Superimposed over all these events was the playing out of Cold War tensions as both the Soviet Union and the United States sought to assert their influence in this vital area. In the midst of this changing political landscape, representatives of the major petroleum-producing countries meeting in Baghdad, Iraq, in 1960 formed the Organization of Petroleum Exporting Countries (OPEC). This new cartel of countries, not companies, viewed itself as a counterweight to the international cartel of multinational oil companies later christened the Seven Sisters (British Petroleum, Royal Dutch/Shell, Exxon, Chevron, Mobil, Texaco, and Gulf). Arab spokesmen among OPEC began to

talk about “using the oil weapon” but little happened until the outbreak of the third Arab–Israeli conflict in 1967. On June 6, 1967, the day after hostilities began, Arab oil ministers called for an embargo on oil shipped to nations friendly to Israel. The flow of Arab oil was reduced by 60% by June 8 and a crisis threatened. Although there were examples of property destruction and temporary economic dislocation, the embargo was a failure in the wake of the brief 6-day war. The oil companies were able to weather the storm and the major losers were the Muslim nations, which lost tremendous oil revenue during the embargo. The situation would not be the same when similar problems occurred in 1973.

Israel had defeated the Arabs with relative ease during the three previous Middle Eastern conflicts since World War II. Despite their huge oil reserves, it appeared that the Arab nations were no military match for a modern technologically equipped military power with American and European support. The Yom Kippur War that broke out in the fall of 1973 had a different outcome. Catching Israel by surprise on Judaism’s high holy day, Egypt and Syria inflicted heavy casualties on Israel in 3 weeks of bloody fighting. At the time the war broke out in October 1973, OPEC was meeting in Vienna. OPEC took two strong measures: They voted to raise the posted price of oil from \$3 to \$5 a barrel, and they announced an embargo against countries supporting Israel. The Nixon administration had been caught off guard at a point when the sitting president was fighting to keep his job in the midst of the Watergate scandal. Secretary of State Henry Kissinger performed brilliantly in negotiating a cease-fire through rounds of shuttle diplomacy that saw him flying from capital to capital. Although military operations halted at precisely the time when Israel appeared to be gaining the upper hand in the conflict, Kissinger’s immediate goal was to cease hostilities. Further diplomacy brought a peace settlement, but the world would experience the aftermath of the embargo and price shocks of October 1973 for the rest of the decade.

High oil prices fueled the increasing problem of stagflation (inflation coupled with stagnating economic growth) that haunted the remaining months of the Nixon presidency until his resignation over Watergate-related matters in August 1974. A similar set of problems limited the effectiveness of the brief presidency of Gerald R. Ford. The presidential election of 1976 brought a Washington outsider, James Earl “Jimmy” Carter, to office and energy policy emerged as a central issue of his

administration. The multinational oil companies had come increasingly under attack in the 1970s on charges of complicity in the OPEC embargo, the taking of “obscene” profits when the market price of crude tripled, and for opposing alternative forms of energy out of selfish economic motive. There was a growing literature on multinational firms generally that asked the question, To whom do these firms owe loyalty? Examples of the fuzzy nature of this question existed particularly with regard to instances of the behavior of oil companies in the international arena. For example, there were many complaints that ARAMCO, ostensibly a U.S. company, had fully cooperated with Saudi Arabia and OPEC during the embargo. In another case, there was evidence that Gulf Oil had supported one faction in the Angola civil war that controlled its area of business operation, whereas the U.S. government backed a competing rebel group. The political economy of world oil had become incredibly more complicated, and the aims of the multinational oil industry had become ever more intertwined with international diplomacy.

9. OIL IN THE MODERN ECONOMY

President Carter urged the American people to view the 1970s energy crisis as the moral equivalent of war. The problem was that the idea never really took hold. To be sure, there had been frightening events in 1973–1974, such as citizens freezing to death for want of fuel oil in New England and a man shot to death in California when he cut into the line at a service station. However, by 1977 oil prices had begun to stabilize as price incentives for new exploration in Alaska, Mexico, and the North Sea along with completion of the Alaska pipeline kicked in. President Carter’s and Energy Secretary Schlesinger’s appeal for energy conservation had also begun to pay off. Perhaps the most significant legacy of Carter’s policies will be his deregulation of oil and natural gas that reintroduced a program of more realistic market pricing. Strong political support for investment in alternative energy technologies, such as solar power, synthetic fuel, wind energy, and biomass conversion, attracted much publicity and enthusiasm but faded amid cynical charges that the energy crisis had been contrived by the oil companies and with the return of relatively lower oil prices. Nuclear power advocates believing that Carter, a nuclear engineer, would give a boost to that industry had their hopes dashed with the Three Mile Island disaster of March 28, 1978—a blow from which the industry has never

fully recovered. Negotiations with the Saudi and Iranian governments in 1978 led to a realization that it was in the best economic and political interests of the oil producers to curtail prices and it appeared that oil prices had achieved stability. However, the international oil economy would have a more cruel blow in store for Jimmy Carter—one that probably cost him reelection in 1980.

President Carter visited Iran and congratulated the Shah as a force of stability in the Persian Gulf. Soon after, when the Iranian people overthrew the Shah in 1979, halting Iranian oil production in the process, another major spike in prices resulted. As had occurred in 1973, the consuming countries seemed unable to compensate for what was, on a global basis, only a small reduction in supplies. Jimmy Carter’s political fate became sealed when the United States allowed the Shah into the country for medical and humanitarian reasons and the U.S. embassy was stormed and 50 hostages were ultimately held captive. Candidate Ronald Reagan benefited from a number of Carter’s difficulties, including very high interest rates and an apparent weak policy abroad, but high oil prices coupled with the embarrassment of the hostage crisis comprised a fatal mix. The revolutionary government of Iran rubbed salt in the wound when it released the hostages on the day of Ronald Reagan’s inaugural as president. President Reagan also benefited from a weakening of OPEC solidarity as members sold oil at below posted prices in order to obtain badly needed money. In 1982, non-OPEC production exceeded OPEC production for the first time as production from the North Sea, Soviet Union, Mexico, Alaska, Egypt, Angola, and China added to world reserves. Although OPEC still remains a relevant player in the political economy of oil, its role is much diminished from its position in the early 1970s. The long-running Iran–Iraq war caused price dislocations when it first started in 1980, but it actually bolstered world supplies in the long term because both nations were forced to sell above quota for needed revenue.

The industry has also had to deal with an increasing number of environmental issues in recent decades. One can date offshore drilling technology, the placing of wells in submerged lands, to the early years of the 20th century, but the technique has only become of critical importance recently. Operators in Santa Barbara, California, for example, had drilled more than 200 wells from piers extending out over the water in 1906, and drilling had been attempted off the coast of Peru that same year. Further experiments went forward in the United States and

abroad during the 1920s and the 1930s, but it would not be until after World War II that offshore drilling began to take on the importance that it has today with deeper drilling techniques. Drilling off the Texas Gulf Coast, for example, became very extensive after the war. However, the events of January 1969 off the coast of Santa Barbara sent shocks throughout the industry. When the drilling of an offshore well in the Santa Barbara channel ran into difficulties that sent a slick of heavy crude onto 30 miles of beaches, there was a huge public outcry. It was also in 1969 that BP, Humble Oil (Exxon), and Atlantic Richfield announced plans to construct a pipeline across the Alaskan tundra aimed at transporting oil from Prudhoe Bay, north of the Arctic Circle, to the port of Valdez. The ensuing debate delayed the project, brought cost overruns, and pitted the environmental movement against the industry in a very public way before the project was finally completed in 1977. Things quieted down with the wide recognition of the part that Alaskan oil had played in dampening high oil prices. Then the Exxon Valdez incident occurred. Oil spills from seagoing tankers have always presented a risk, but the size of the new generation of supertankers exacerbated concerns. When the supertanker Exxon Valdez went aground in Alaska's Prince William Sound spilling 240,000 barrels of Prudhoe Bay oil in 1989, environmental critics of the industry received a further boost. Although it is environmentally risky, deep-water drilling as is occurring in the North Sea, wilderness pipelines, and the passage of supertankers on the oceans have become essential elements of commerce in world oil.

The industrialized nations have enjoyed a relatively more benign oil economy in the 1990s and into the early new century, but the historic pattern of dependency on oil controlled by the producing nations remains a reality. In 2003, the United States imported 55% of its oil; in the crisis year of 1973, imports amounted to only 35%. The Iraqi invasion of Kuwait on August 2, 1990, precipitated a series of problems in the Persian Gulf region. Future historians will interpret the defeat of Saddam Hussein in the 1991 Gulf War by the first Bush administration as but a chapter in the events that led to the U.S. invasion of Iraq in the spring of 2003 directed by his son. Neither in 1991 nor in 2003 was oil the sole factor in precipitating international crisis, but one cannot ignore the extremely significant role that it has played. Critics point to the very close ties between the administration of the second President Bush and the oil industry, and many have uttered the protest refrain of "No Blood for Oil" with

reference to the war in Iraq. However, in a figurative sense oil has become the blood of international commerce and industry more than ever as we embark on a new century.

One of the main business trends in the global oil industry in recent years exhibits a touch of irony for the United States. This is the movement toward megamergers initiated by the acquisition by Chevron (Standard Oil of California) of Gulf Oil in 1984 and continuing through 2000. The demise of Gulf occurred after swashbuckling financier T. Boone Pickens' small Mesa Petroleum Company shocked the industry with a takeover bid for the large multinational. Chevron's ultimate acquisition of Gulf was followed by a series of additional moves, including the Royal Dutch/Shell group acquiring that part of Shell USA that it did not already own; BP's merger with Sohio (Standard Oil of Ohio); the joining of Exxon (Standard Oil of New Jersey) and Mobil (Standard Oil of New York); Chevron's merger with Texaco; and the merger of BP, Amoco (Standard Oil of Indiana), and ARCO (Rockefeller's old Atlantic Refining Company). The notorious Seven Sisters of the 1970s had shrunk considerably by the turn of the century. Historical irony lies in this partial reassembling of the pieces of the old Rockefeller Standard Oil Company, which had been fragmented into 34 separate companies in 1911.

Why this rush toward consolidation? In general terms, this oil industry behavior relates to the broader trend of U.S. merger activity beginning in the 1980s and present today. The argument that firms required a leaner, more competitive profile to respond to the challenges of the new global economy became most prominent, and a more benign antitrust environment created by recent Republican administrations in Washington has abetted the trend. For the oil industry specifically, there are other reasons cited. One of these is the argument that the companies representing the consuming nations need to have a stronger bargaining position with OPEC and other producing groups. Others see consolidation as a key to mobilizing more effective resistance to world environmental movements, such as the Kyoto Protocol, which have called for the reduction of heat-trapping greenhouse gas emissions caused by the burning of coal, oil, and gasoline. Certainly, environmental groups have been among those most critical of the industry mergers. Some industry spokesmen have referred to declining profits in the wake of relatively lower world oil prices as the reason for moving toward the cost reductions resulting from merger. Within the context of the

history of the oil industry, one might ask simply “how big should big oil be?”

Oil supplies will eventually become exhausted; this is an accepted reality. However, there remain huge supplies of petroleum and a world economy that depends on access to it. There will be an historic transition to other energy sources in the future and the individual companies that comprise the industry may very well have diversified to the extent that they will become part of this new political economy. The history of the oil industry has been relatively brief in years, approximately a century and a half, but it has been a richly diverse and fascinating story that is still not concluded.

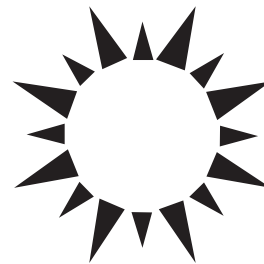
SEE ALSO THE FOLLOWING ARTICLES

Coal Industry, History of • Electricity Use, History of • Geopolitics of Energy • Hydrogen, History of • Manufactured Gas, History of • Nationalism and Oil • Natural Gas, History of • Nuclear Power, History of • Oil Crises, Historical Perspective • Oil-Led Development: Social, Political, and Economic Consequences • OPEC, History of • OPEC Market Behavior, 1973–2003 • War and Energy

Further Reading

- Brantly, J. E. (1971). “History of Oil Well Drilling.” Gulf, Houston, TX.
- Chernow, R. (1998). “Titan: The Life of John D. Rockefeller, Sr.” Random House, New York.
- Clark, J. G. (1987). “Energy and the Federal Government: Fossil Fuel Policies, 1900–1946.” Univ. of Illinois Press, Urbana.
- Clark, J. G. (1990). “The Political Economy of World Energy: A Twentieth Century Perspective.” Univ. of North Carolina Press, Chapel Hill.
- Gorman, H. S. (2001). “Redefining Efficiency: Pollution Concerns, Regulatory Mechanisms, and Technological Change in the U.S. Petroleum Industry.” Univ. of Akron Press, Akron, OH.
- Olien, R. M., and Olien, D. D. (2000). “Oil and Ideology: The Cultural Creation of the American Petroleum Industry.” Univ. of North Carolina Press, Chapel Hill.
- Painter, D. S. (1986). “Oil and the American Century: The Political Economy of U.S. Foreign Oil Policy, 1941–1954.” Johns Hopkins Univ. Press, Baltimore.
- Pratt, J. A., Becker, W. H., and McClenahan, W. Jr. (2002). “Voice of the Marketplace: A History of the National Petroleum Council.” Texas A&M Univ. Press, College Station.
- Schneider, S. A. (1983). “The Oil Price Revolution.” Johns Hopkins Univ. Press, Baltimore.
- Vietor, R. H. K. (1984). “Energy Policy in America Since 1945.” Cambridge Univ. Press, New York.
- Williamson, H. F., and Daum, A. (1959). “The American Petroleum Industry: The Age of Illumination, 1859–1899.” Northwestern Univ. Press, Evanston, IL.
- Williamson, H. F., Andreano, R. L., Daum, A., and Klose, G. C. (1963). “The American Petroleum Industry: The Age of Energy, 1899–1959.” Northwestern Univ. Press, Evanston, IL.
- Yergin, D. (1991). “The Prize: The Epic Quest for Oil, Money & Power.” Simon & Schuster, New York.

Oil-Led Development: Social, Political, and Economic Consequences



TERRY LYNN KARL

Stanford University

Stanford, California, United States

1. Introduction
2. Definitions: Oil Dependence, Resource Curse, Dutch Disease, and Rentier States
3. Poverty and Social Welfare Consequences of Oil-Led Development
4. Oil-Related Changes in Social Structure
5. The Rentier State
6. Oil, Political Stability, and State Capacity
7. Social and Environmental Impacts at Regional and Local Levels
8. Petroviolence and Civil War
9. Conclusion

Glossary

corruption Though often used interchangeably with rent seeking, corruption is more narrowly defined as the misuse of public power or resources for private gain, and it is generally illegal.

Dutch disease Named after the negative effects of the North Sea oil boom on industrial production in The Netherlands, this phenomenon occurs when resource booms cause real exchange rates to rise and labor and capital to migrate to the booming sector. This results in higher costs and reduced competitiveness for domestically produced goods and services, effectively “crowding out” previously productive sectors.

oil-led development An overwhelming dependence on revenues from the export (and not the internal consumption) of petroleum, as measured by the ratio of oil and gas to gross domestic product, total exports, and the contribution to central government revenues.

rent In Adam Smith’s classic definition, this is unearned income or profits “reaped by those who did not sow.” According to economists, rents are earnings in excess of all relevant costs, including the market rate of return on

invested assets. They are the equivalent of what most noneconomists consider to be monopoly profits.

rentier state A state that lives from externally generated rents rather than from the surplus production of the population. In oil-exporting states, this is measured by the percentage of natural resource rents in total government revenues.

rent seeking The efforts, both legal and illegal, to acquire access to or control over opportunities for earning rents. In oil-dependent countries, rent seeking refers to widespread behavior, in both the public and the private sectors, aimed at capturing oil money through unproductive means.

resource curse The negative growth and development outcomes associated with minerals and petroleum-led development. In its narrowest sense, it is the inverse relationship between high levels of natural resource dependence and growth rates.

Proponents of oil-led development believe that countries lucky enough to have “black gold” can base their development on this resource. They point to the potential benefits of enhanced economic growth and the creation of jobs, increased government revenues to finance poverty alleviation, the transfer of technology, the improvement of infrastructure, and the encouragement of related industries. But the experience of almost all oil-exporting countries to date illustrates few of these benefits. To the contrary, the consequences of oil-led development tend to be negative, including slower than expected growth, barriers to economic diversification, poor social welfare indicators, and high levels of poverty, inequality, and unemployment. Furthermore, countries dependent on oil as their major resource for development are characterized by corruption and

exceptionally poor governance, a culture of rent seeking, often devastating economic, health, and environmental consequences at the local level, and high incidences of conflict and war. In sum, countries that depend on oil for their livelihood eventually become among the most economically troubled, the most authoritarian, and the most conflict-ridden in the world.

1. INTRODUCTION

Oil is a commodity with special characteristics. These include its unique role as both the common natural heritage of a country and the motor of global industrialization, its depletable, its price volatility and consequent boom–bust cycles, its especially high capital intensity and technological sophistication, its enclave nature, and the exceptional generation of profits that accrue to the state and to private actors. The combination of these factors produces what has been called “the paradox of plenty” or the “resource curse.” This is not due to the resource per se, which is simply a black and viscous substance, and it is not inevitable. A resource boom can be beneficial or detrimental: Norway, an oil exporter, has used the benefits of North Sea petroleum to earn the highest place on the United Nations Development Program’s list of best development performance, whereas other exporters, such as Nigeria and Angola, are clustered near the bottom. Instead, what matters for the social consequences generated by petroleum dependence are, first, the type of preexisting political, social, and economic institutions available to manage oil wealth as it comes onstream and, second, the extent to which oil revenues subsequently transform these institutions in a rentier direction. Because almost all proved oil reserves exist in less developed countries, where administrative institutions tend to be weak (only 4% can be found in advanced industrialized democracies), the probability of the resource curse is exceptionally high.

2. DEFINITIONS: OIL DEPENDENCE, RESOURCE CURSE, DUTCH DISEASE, AND RENTIER STATES

Mineral (especially oil-led) development is often promoted as a key path for developing countries seeking sustained economic growth. But the oil-led

development model of today is significantly different from the role that energy played in the late 19th and early 20th centuries in the United States, Canada, and Australia. In those earlier and more successful experiences, mining and oil exploitation contributed only a very small percentage of total economic output, never dominated exports, and never came anywhere near the magnitude of dependence that characterizes contemporary oil-led development. While leaving a considerable regional impact, oil and minerals were never the motor of development.

Today, to the contrary, oil-led development means that countries are overwhelmingly dependent on revenues gleaned from the export of petroleum. This dependence generally is measured by the ratio of oil and gas exports to gross domestic product; in countries that are sustained by petroleum rents, this figure ranges from a low of 4.9% (in Cameroon, a dependent country that is running out of oil) to a high of 86% (in Equatorial Guinea, one of the newest oil producers). Dependence is also reflected in export profiles, with oil in dependent countries generally making up 60–95% of a country’s total exports. Oil-dependent countries can be found in all geographic regions of the world, although they are most commonly associated with the Middle East and, more recently, Africa.

Oil-dependent countries suffer from what economists call the “resource curse.” In its simplest form, this refers to the inverse association between growth and natural resource abundance, especially minerals and oil. This association repeatedly has been observed across time and in countries that vary by population size and composition, income level, and type of government; it is so persistent that has been called a “constant motif” of economic history. Specifically, countries that are resource poor (without petroleum) grew four times more rapidly than resource-rich (with petroleum) countries between 1970 and 1993, despite the fact that they had half the savings. Similar findings have been replicated through a study of the members of the Organization of Petroleum Exporting Countries (OPEC), using a different and longer time period, from 1965 to 1998. OPEC members experienced an average decrease in their per capita gross national product (GNP) of 1.3% per year during this period, whereas lower and middle-income developing countries as a whole grew by an average rate of 2.2% per year over the same time. Moreover, studies show that the greater the dependence on oil and mineral resources, the worse the growth performance. Finally, countries dependent on the export of oil have not only performed

worse than their resource-poor counterparts; they have also performed far worse than they should have given their revenue streams.

The causes of this resource curse are a matter of debate, but the negative association between growth and oil and mineral wealth is not attributed to the mere existence of the natural resource. Oil in itself cannot encourage or hinder growth. Instead, this association is less direct and, though the weight of various specific causal mechanisms is still debated, it is generally attributed to some combination of the factors. First, oil windfalls can hurt other sectors of the economy by pushing up the real exchange rate of a country's currency and thus rendering most other exports noncompetitive, a phenomenon called the "Dutch disease." The reduced competitiveness in agricultural and manufacturing exports then "crowds out" other productive sectors and makes the diversification of the economy particularly difficult. This in turn reinforces the dependence on oil and, over time, it can result in a permanent loss of competitiveness.

Second, the long-term price deflation and price volatility of the international primary commodities market hinders economic development. Since 1970, this volatility has grown worse, and oil prices are twice as variable as those of other commodities. This means that oil economies are more likely to face more frequent economic shocks, with their attendant problems, and they are especially susceptible to acute boom-bust cycles. This oil price volatility exerts a strong negative influence on budgetary discipline and the control of public finances as well as on state planning, which subsequently means that economic performance deviates from planned targets by as much as 30%. Volatility also exerts a negative influence on investment, income distribution, and poverty alleviation.

Third, the enclave nature of the industry combined with its capital intensity fosters especially weak linkages to the broader economy and does little to create employment. Because oil is the world's most capital-intensive industry, the sector creates few jobs per unit of capital invested, and the skills required by these jobs often do not fit the profile of the unemployed. If growth in the oil sector had a significant multiplier effect, this would not be such a great problem, but the productive linkages between this sector and the rest of the economy tend to be weak. Furthermore, the opportunities for technology diffusion are very limited, and so is infrastructure development. Downstream processing industries have typically not emerged, and when they do, they are often at a competitive disadvantage.

Perhaps most important, petroleum may be one of the hardest resources to utilize well; countries dependent on oil exports seem particularly susceptible to policy failure. The reason lies in the weakness of preexisting institutions in places where oil for export is found, their frequently authoritarian character, and the ease with which they can be transformed by an overwhelmingly powerful export sector. Generally, oil rents produce a rentier state—one that lives from the profits of oil rather than from the extraction of a surplus from its own population. In rentier states, economic influence and political power are especially concentrated, the lines between public and private are very blurred, and rent seeking as a wealth creation strategy is rampant. Rentier states are notoriously inefficient because productive activity suffers and self-reinforcing "vicious" development cycles can set in. Together, all of these factors slow growth, raise powerful barriers to the diversification away from petroleum dependence, and produce the skewed development patterns described by the resource curse.

3. POVERTY AND SOCIAL WELFARE CONSEQUENCES OF OIL-LED DEVELOPMENT

One of the most important social consequences of the resource curse is that oil-exporting countries have unusually high poverty rates, poor health care, high rates of child mortality, and poor educational performance given their revenues—outcomes that contradict the beliefs about what should happen within oil-exporting countries. Though it is true that most forms of primary commodity dependence are associated with poverty, not all commodities are equally culpable. Countries dependent on agricultural commodities tend to perform better with respect to poverty, minerals in general are linked to high levels of poverty, and oil dependence in particular is correlated with low life expectancy and high malnutrition rates.

Oil dependence has an ambiguous relationship with poverty alleviation, and this is related to the boom-bust cycles accompanying dependence on the resource. At the beginning of oil exploitation for export, per capita income rises during the "euphoric," or "boom," period. Especially in the initial stages of production for export, petroleum revenues initially transform a society, often suddenly and dramatically. Employment increases, infrastructure is

improved, and per capita income grows rapidly. Thus, for example, per capita oil exports for North Africa and the Middle East soared from \$270 in 1970 to \$2042 in 1980, and this fueled accelerated economic activity. But the failure to diversify from oil dependence into other self-sustaining economic activities, especially agriculture and labor-intensive industry, becomes a significant obstacle to pro-poor development. Over time, as booms peter out, oil exporters are plagued by (often sudden) declines in per capita income. In North Africa and the Middle East, for example, per capita oil exports plunged from the 1980 high of \$2042 in 1980 to \$407 by 1992 as oil prices dropped and population increased. Paradoxically, in what seems to be the midst of plenty, a high percentage of people living in oil-exporting countries tend to remain poor or suffer from dramatic shifts in their welfare that ultimately leave them in poverty. Thus, despite significant rises in per capita income, over the past several decades, all oil-dependent countries have seen the living standards of their populations drop, and sometimes drop very dramatically. This boom-bust cycle affects even the world's richest oil exporters. In Saudi Arabia, for example, where proved reserves are the greatest in the world, per capita income has plunged from \$28,600 in 1981 to \$6800 in 2001.

For many countries, including Algeria, Angola, Congo, Ecuador, Gabon, Iran, Iraq, Kuwait, Libya, Qatar, Saudi Arabia, and Trinidad Tobago, this plunge has been very severe, moving real per capita incomes back to the 1970s and 1980s. For a few countries, most notably Nigeria and Venezuela, the growth of poverty has been catastrophic; in these cases, real per capita income has plummeted to 1960 levels. It is almost as if 40 years of development had not taken place. In Nigeria, the disparity between oil wealth and poverty is especially notable. Despite the fact that over \$300 billion in oil profits has been generated over the past 25 years, the proportion of households living below the United Nation's absolute poverty line of \$1/day grew from 27% in 1980 to 66% by 1996. Income disparities are shocking: the richest 10% controls 40% of the country's wealth and its poorest 20% has a share of just 4.4%. But oil dependence is associated with more than sudden shifts in levels of poverty and exceptionally low living standards for much of the population in petrostates. It is also linked to unusually high rates of child mortality, child malnutrition, low life expectancy, poor health care, and reduced expenditures on education. In countries dependent on oil and/or minerals, both infant

mortality and life expectancy at birth are worse than in non-oil and non-mineral countries at the same income levels. Simply put, when taken as a group, the greater the dependency of a country on oil, the greater the likelihood that children born there will die at birth, will have poorer health care, nutrition, and education than their resource-poor counterparts, and will die sooner, if they survive at birth.

The statistics are startling. For each 5-point increase in oil dependence, the under-5-years-old mortality rate rises by 3.8 per 1000. This may be due to the fact that oil dependence is also negatively correlated with health care expenditures. Paradoxically, the more countries are dependent on oil, the less they spend on health as a percentage of gross domestic product (GDP). In Nigeria, for example, the government spends about \$2 per person per year on health care, which is far less than the \$34 per year recommended for developing countries by the World Health Organization. But poor child welfare performance is also due to higher malnutrition rates that exist in oil-dependent countries. Indeed, once the effects of per capita income are taken into account, for every 5-point rise in oil dependence, there is a corresponding 1% rise in the percentage of children under 5 years old who are malnourished. Compare, for example, the global average of 26.5 malnourished children per 1000 to the 37.7 per 1000 rate in oil-rich Nigeria.

Given the available resources, education also performs worse than expected, affecting future prospects for growth. Countries that are dependent on natural resources, inadvertently or deliberately, neglect the development of their human resources by devoting inadequate attention and expenditure to education. Thus, school enrollments tend to be lower than in their non-resource-rich counterparts. In the OPEC countries, for example, 57% of all children go to secondary school compared with 64% for the world as a whole; OPEC spends less than 4% of the GNP on education compared with almost 5% for the world as a whole (in 1997 figures). The explanation for poor educational performance in oil-exporting countries is not clear. Perhaps because the high skill level needed by oil-rich countries in their leading sector can be bought or imported, their governments do not face the same urgent educational imperatives and may underrate the need for strong educational policies. Flooded with easy money, they may perceive more urgent needs than the long-term investments in education that results in long-term development benefits.

4. OIL-RELATED CHANGES IN SOCIAL STRUCTURE

Dependence on petroleum skews the social structure of countries. Because of the enormous capital and technological resources necessary to exploit this resource, foreigners (principally oil companies) become a dominant, if not *the* dominant internal social force, especially at the beginning stages of development. This has important implication for the creation of a domestic entrepreneurial class. Though foreign companies may form partnerships with domestic elites, their overwhelming economic presence and capital and technological advantages mean that domestic entrepreneurs have less opportunity to develop on their own. To be successful, the domestic businesses must forge close ties either to the state or to foreign capital, or they may be marginalized, e.g., merchants in Middle East oil-exporting countries. This pattern exists for other types of primary commodity exporters, but it is more exaggerated in oil-exporting countries because domestic capitalist economic groups, notoriously concentrated in monopolies or oligopolies, are dependent on oil rents and the political power arrangements that choose their distribution through patronage. Thus, instead of a capitalist class, the *nouveau riche*, who are fabulously and ostentatiously rich and owe their success to the state, characterize oil states. But because this wealth is the result of a windfall and privileged links to the state and because it may be largely independent of merit-based efforts made by citizens, this pattern of wealth creation encourages rent seeking as well as a tendency to live beyond one's means.

Middle and professional classes are also shaped by dependence on oil exports as the engine of the economy. Labor markets in oil-exporting countries tend to offer only three major types of jobs—oil related, public sector, and private services—and this retards the growth of a large middle and professional class. When these groups do appear, they differ from other middle and professional classes because their job prospects and standard of living are directly linked to the fortunes of the major export industry, petroleum, thus they are exceptionally vulnerable. During boom periods (for example, the 1970s and early 1980s), jobs and wealth are readily available for the educated, but during bust cycles, middle and professional classes may be educated but have few job opportunities and little prospect for wealth. The outcome is often intense social and generational tension, especially in urban areas, as the population and number of educated grow and employment

shrinks. This is most notable in the Middle East, where the young generation of urban middle sectors has seen their parents' situation deteriorate and has arrived too late for economic benefits or adequate social services in stagnant economies.

At the same time, the formation of a broad-based urban working class is compromised. Because oil employs relatively few workers and their skill level must be especially high, and because the rest of the labor market is skewed, dependence on oil fosters a type of labor aristocracy that is separate from most of the workforce. Though this separation, delineated by educational and skill level, can be found in most developing countries, it is especially notable in oil-exporting countries because they have among the fastest rural-to-urban migration rates in the world. Rural poor not only experience the normal economic pulls to cities, but oil rents also create exaggerated expectations of new opportunities—even as the Dutch disease begins to bias against agriculture and agrarian interests. So rapid is the outflow from the countryside that some landlords, most notably those in Iran, have been compelled to import foreign workers to till their lands. This especially rapid rural-to-urban migration means that cities are filled with a relatively small middle and professional class when compared to the vast majority of underskilled and underemployed workers.

Finally, exceptional levels of in-migration characterize oil states. This is encouraged by the structure of the labor market as well as by the pull of oil wealth. In some cases, migration has dramatically altered the profile of petrostates. Most of the oil-exporting countries in the Gulf region, for example, have more foreign than citizen residents! Somewhere between 50 and 90% of private sector workers in the Gulf are foreigners. There are 6 million foreigners among Saudi Arabia's 18 million residents, and foreigners are 98% of manufacturing workers, 97% of construction workers, and 93% of the service sector. This extensive in-migration further distorts levels of inequality because immigrants are generally paid less, compared to nationals. Saudi youth, for example, often expect to earn at least SR2500 (\$670) a month for unskilled work, whereas foreigners filling these jobs typically earn SR500–SR1000 a month.

This peculiar social structure is linked to a specific "culture of oil" that permeates all social classes and groups as well as the state. As novelists, journalists, and other observers have repeatedly illustrated, petroleum creates a world of illusion because some people become wealthy without effort. This means

that work ethics are undermined and negative attitudes toward certain forms of work, especially manual labor, can prevail in many oil-exporting countries. This in turn can translate into lower levels of productivity than those found in comparable resource-poor states. As novelists, journalists, and other observers of oil countries have repeatedly noted, states and people that experience a sudden influx of income they did not work hard to get have not usually developed the fiscal and financial discipline or work habits normally required to get and keep such windfalls. They tend to become accustomed to relatively high salaries and little work. For this reason, employers in oil-exporting countries report that they prefer foreign workers who will work harder for less money, seemingly grateful that they may be earning five times the salary possible in their country of origin.

Embedded in this social structure are two fundamental cleavages. First, because the West dominates the oil industry through both production and consumption, a Western definition of modernity based on the market rather than indigenous cultures is transmitted; indeed, oil exporters may experience the most accelerated Westernization and may be the most exposed to Western influence when compared to non-oil countries. At least part of the country will be linked together through this form of modernization, e.g., technocrats, public sector employees, and educated elites. But precisely because oil development accelerates the rate of change, because oil countries are so exposed to the West, and because the discontinuities provoked by petroleum wealth are so great, the failure of the promise of an apparently easy modernization may give rise to decidedly conservative anti-Western movements based on different principles for the organization of economic life, as in Algeria and Iran, or distinctive traditional notions about the depletion of finite resources, as among the U'wa indigenous people in Colombia.

Second, though the inequalities created by oil-led development appear to be at about the same levels as in non-oil states with similar incomes, people in oil-exporting countries may experience these inequalities very differently because they occur in what is widely perceived to be a rich country. The visibility of oil wealth may compound the problem. Where traditional practices are essentially conservative and egalitarian, as in some Latin American indigenous groups, or where religious practices emphasize the need to redistribute income fairly, avoid earning interest, take care of the poor, and prohibit waste and idleness, as in the Islamic world, the cultural shock

can be especially powerful. When rulers appear to be wasteful, despotic, and dominated by foreigners, this can produce an especially potent political mix.

5. THE RENTIER STATE

Ineffective and inefficient governance, perhaps more than any other factor, may explain the extent of poverty in oil-dependent countries, but this, too, is related to the presence of oil. Because the revenue base of the state *is* the state, oil rents affect state capacity. Oil dependence skews the institutional development of the state because oil rents weaken agencies of restraint. In resource-poor countries, intense population pressure on scarce resources reduces the tolerance for inefficiency and predation, and the economy cannot support extensive protection or an over-expanded bureaucracy. But in oil states, the brake of scarcity does not exist. Instead, oil dependence encourages the expansion of states into new arenas while weakening opportunities to strengthen administrative capacities, especially non-oil-based tax systems, merit-based civil services, and the rule of law—fundamental elements for creating efficient states.

The impact of oil rents on effective governance has a pernicious effect on the quality of administrative institutions in less developed countries, regardless of whether they are democratic or authoritarian. First, because oil states do not have to extract the majority of their resources from their own populations, they do not have to build the institutional capacities that have historically been required by such extraction. This means that they are denied the information that is generated by a robust tax bureaucracy, and they are also denied the incentives for innovation within a civil service that stems from scarcity. Even where state capacity embedded in tax authorities may have previously existed, oil rents tend to be undermining. With the discovery of oil, these tax authorities are often disbanded because they appear to be no longer necessary. Second, because windfall gains that arise from petroleum encourage rent-seeking behavior, the state becomes a type of “honey pot” in which competing interests try to capture a significant portion of resource rents by capturing portions of the state. A vicious cycle results in which all actors try to gain parts of the bureaucracy while governments, in turn, reward their supporters by funneling favors their way. But this means that the public sector tends to lack the corporate cohesiveness and authority necessary to exercise effective public policy.

Finally, if the state is especially weak and the target of capture, it is also especially overloaded. Oil revenues are the catalyst for a chronic tendency of the state to become overextended, over-centralized, and captured by special interests. This can be seen through the accelerated and especially large growth of the public sector, the overextension of public expenditure, and the unusually extended periods of protection for import-competing sectors. Yet, without the institutional and administrative capacity to cope with this enhanced state role, this overextension is a formula for ineffectiveness.

The most telling indicator of declining state capacity is the loss of fiscal control, measured by overspending and soaring debt as well as by the inability of oil states to reform themselves. This is because oil states degenerate into sophisticated mechanisms to transfer resources from the primary sector to politically influential urban groups, especially as windfall gains provoke a type of “feeding frenzy” to capture petrodollars. This does not occur to the same extent when labor-intensive activities drive economic growth, centering on food and agricultural products, for example, in part because they tend to generate fewer rents. The political competitions for resource rents (when combined with the often nontransparent mechanisms for distributing them) have important efficiency costs. For example, they make it more difficult for governments to moderate spending in response to the price volatility of petroleum, thereby further distorting the economy. In general, oil rents permit incapable state institutions to endure, and ineffective policies to persist, considerably longer than they can in less resource-rich countries. To avoid unpopular reforms, governments use their oil as collateral for borrowing abroad or intensify the squeeze on the export sector. Petrodollars simply permit more scope for cumulative policy error.

States that have the greatest resource endowments, and especially oil-exporting countries, also have the extraordinarily high levels of corruption—a reality confirmed by stunning quantitative evidence and numerous case studies. With incomes of the order \$35 billion/year for Mexico; \$30 billion/year for Venezuela; \$22 billion/year for Nigeria, the temptations for abuse are immense, and with weak state capacity and rule of law in place, there is little institutional restraint. “People rob,” one finance minister of an oil-exporting country remarked, “because there is no reason not to.” Oil rents and institutional weakness form a vicious cycle. Quantitative evidence suggests that the extent of corruption

is higher in countries in which civil service recruitment and promotion procedures rely less on merit-based considerations; where this is the case, efforts to reform the civil service are blocked in order to sustain patterns of corruption. At its worst, this can degenerate into a “corruption trap,” whereby payoffs at the top of political and business institutions encourage the corruption of others until a large percentage of public and private sector figures are involved, as the case of Nigeria demonstrates.

Corruption takes place not only at the production and exports stage through secret signature bonuses and opaque financial arrangements, but also as a result of extremely high-level and difficult-to-absorb investments at the “upstream” stage as well as at the trading, or “downstream,” stage, where massive resources tend to disappear through price transfers that are difficult to track. Though transactions are obviously clandestine, evidence of oil-related corruption abounds in both the private sector and the state. The former president of the French state oil company, Elf Aquitaine, is charged with presiding over the commission payments on oil deals with African countries. Mobil oil executives are charged with illegal payments in Kazakhstan. In Angola, more than \$1 billion/year of oil revenues disappeared between 1996 and 2001, a full one-sixth of the national income, in a country where more than 70% of the population lives on less than \$1/day.

Corruption contributes to the resource curse. Rulers will support policies that produce personalized rents even if these policies result in lower overall social welfare; because they need to share these rents with supporters and subordinates, the level of distortion can be very great. Policy choices are deformed in a number of ways. First, when huge oil rents are present, officials tend to favor larger public sectors with overly excessive regulatory interventions that enhance opportunities for rent seeking. Second, policy choices are distorted toward the financing of megaprojects in which payoffs can be more easily hidden and the collection of bribes is easier, whereas productive long-term investment remains undersupplied. Highly capital-intensive and specialized one-of-a-kind designs may be favored so that there are no reliable cost benchmarks; for example, an aluminum smelter was built for \$2.4 billion in Nigeria even though it served no valid development objective and its cost was between 60 and 100% higher than the cost of similar plants elsewhere. Infrastructure and defense projects are also favored over health and education expenditures, thereby reducing the quality of public services as well

as lowering the quality of public infrastructure. Most important, corruption affects both economic growth and income levels. Economists estimate, for example, that Venezuela's average GDP growth rate would be raised by some 1.4% annually had it reduced its corruption to the level of corruption in Chile.

6. OIL, POLITICAL STABILITY, AND STATE CAPACITY

Oil and centralized rule seem to go together; and oil and democracy do not generally mix. Political scientists have repeatedly documented this relationship through case studies, and they have found a robust and statistically significant association between oil dependence and authoritarian governments. Oil appears to impede the appearance of democracy in most cases, especially in the Middle East and North Africa, though it facilitated democratization in Venezuela.

The hindering of democratization seems to occur primarily through different, though related, mechanisms. The first is based on how rentier states collect revenues. Because these states live from oil rents rather than from direct taxation, they are likely to tax their populations lightly or not at all. Thus, they are unusually detached from and unaccountable to the general population, and their populations, in turn, are less likely to demand accountability from and representation in government. In effect, the vital link between taxation and representation is broken. Studies have shown, for example, that the governments of Kuwait and Qatar became less accountable to the traditional merchant class in this way. Even in Venezuela, where some type of democracy exists, the lack of taxation has made both representation and state accountability less than expected.

A second causal mechanism depends on how regimes spend state revenues. Oil wealth produces greater spending on patronage, which, in turn, weakens existing pressures for representation and accountability. In effect, popular acquiescence is achieved through the political distribution of rents. Oil states can buy political consensus, and their access to rents facilitates the cooptation of potential opponents or dissident voices. With basic needs met by an often generous welfare state, with the absence of taxation, and with little more than demands for quiet and loyalty in return, populations tend to be politically inactive, relatively obedient and loyal, and levels of protest remain low, at least as long as the oil

state can deliver. Thus, for long periods, an unusual combination of dependence, passivity, and entitlement marks the political culture of petroleum exporters. This is especially the case in smaller exporting states such as the Gulf monarchies, where oil reserves per capita are 43 times those of large exporting states such as Algeria, Indonesia, Nigeria, Venezuela, and Iran, and where such costly distributive policies can be sustained for a longer time. Regimes have even used their largess to prevent the formation of social groups, independent from the state, that might someday prove to be political challengers or to rid themselves of already existing challengers—a phenomenon that has been documented (during various historical periods) in Venezuela, Algeria, Iraq, Iran, Kuwait, and Qatar. In the latter two countries, for example, the political distribution of oil rents eliminated the influence of the merchant class in decision making, leaving the rulers with no real political opponents that could base themselves in a social class. In Iran, under the Shah, the agricultural class was simply transformed into urban commercial (and dependent) interests through the politically judicious use of oil rents. But the spending of oil rents supports repression as well as cooptation to keep authoritarian rulers in power. Not surprisingly, then, oil dependence is closely associated with military spending and the creation of vast repressive apparatuses. This is in part due to the fact that superpowers are wary of letting oil reserves fall out of the control of their allies and into the hands of possible opposition groups. As a group, oil exporters spend much more money and a greater percentage of their revenues on their military and security forces than do non-mineral-dependent countries. For example, whereas the average developing country spends about 12.5% of its budget on the military, Ecuador, in contrast, spends 20.3% and Saudi Arabia spends a whopping 35.8%. The extent of militarization is stunning. In the decade from 1984 to 1994, for example, the share of annual military expenditures as a percentage of total central government expenditures in OPEC countries was three times that of the developed countries, and two to ten times that of the non-oil-dependent developing countries.

For these reasons, oil revenues tend to lend support, at least in the short to medium term, to whatever type of regime is in place, whether it is the occasional democracy or the more likely authoritarian ruler. Though all states may use their fiscal powers to reduce dissent through coercion or cooptation, oil wealth provides states with exceptional possibilities to do so—a phenomenon that has

been observed throughout the Middle East and in Mexico and Venezuela. Thus oil wealth is robustly associated with more durable regimes, and oil dependence is a positive predictor of durability. Even though authoritarian regimes in general are more likely to fall during economic crises, oil-based authoritarian regimes have some cushion from this general rule. Regimes such as Suharto's in Indonesia, Saddam Hussein's in Iraq, Gomez's in Venezuela, and the long-lived monarchies of the Persian Gulf (all of which lasted at least three decades) are representative of this unusual durability. Even if power shifts from one type of authoritarian rule to another (or to some form of elite democracy), political elites inherit the power that comes from control over the process of rent distribution because they control the state during windfalls, and they can consolidate this form of control through their allocative power. Thus, oil rents initially help regimes to consolidate, then enable them to endure for unusually long periods, and even enable them to persist during periods of bust.

The norm of regime stability is only part of the story. Richly detailed case studies of Nigeria, Venezuela, and Iran show that oil can help to undermine political stability over time, especially in authoritarian regimes. Virtually all oil-rich states tend to face significantly higher levels of social protest when oil revenues fall, and some of these regimes collapse. Where regimes have developed mechanisms of social control, permit rotation in power, or have sources of legitimacy that are not based on oil rents, they are more likely to endure through boom-bust cycles. But where initial oil exploitation coincides with regime and state building, non-oil-based interests do not form and patronage rents may be the main glue of the polity. Under these circumstances, these regimes are especially fragile and vulnerable during oil busts.

7. SOCIAL AND ENVIRONMENTAL IMPACTS AT REGIONAL AND LOCAL LEVELS

The exploitation of oil has a profound regional and local impact, and from the standpoint of the majority of the local population, this impact is alarming. Rather than bring prosperity to a region, as is often the claim, the boom-bust cycle associated with petroleum dependence is magnified. Over time, localities where oil is actually located, compared to

the rest of the country, tend to suffer from lower economic growth, lower per capita incomes, greater dislocations, higher environmental and health hazards, and higher levels of conflict. Economically, petroleum fails to offer long-term sustainable employment alternatives at the local level, but it can seriously disrupt preexisting patterns of production. The promise of new jobs that new oil exploitation seems to offer typically attracts large numbers of migrants to an exploitation area. The rapid influx of people and the higher relative salaries of oil project workers inflate the local prices of key goods and services, bringing about a significant increase in the cost of living, even for those who do not share in the benefits of an oil project. For example, the municipality of Yopal, in the state of Casanare, Colombia, abruptly filled with migrants hoping to find employment at salaries three to four times the minimum wage, even before nearby massive oil fields at Cusiana-Cupiagua came onstream. Rents and prices increased 300%, virtually overnight. But because most jobs created by the petroleum industry are temporary or seasonal in nature, and because the growth in jobs generally occurs only during the exploration phase (land needs to be cleared or equipment transported), the industry actually offers comparatively few jobs over time. Thus, though discoveries trigger massive changes, beginning with the influx of workers seeking employment on the construction of roads, pipelines, and other infrastructure, these increased employment opportunities do not last; employment levels tend to decline dramatically when infrastructure construction is complete. These problems are compounded by the expropriation of arable land for resource extraction activity and by environmental damage, which promote a shift away from subsistence agriculture. The resulting instability in employment and income and food instability stress the local economy.

The social fabric of oil localities also changes, as disparities in income emerge and migrants pour in, often from other countries, ethnic groups, or religions. After the construction phase has been completed, the most likely local result of an oil boom (along with higher than average local inflation, increased migration, chronic underemployment, and food shortages) is increased prostitution, autoimmune disease syndrome (AIDS), and crime. Original residents who may not have been able to share in oil benefits increasingly clash with "newcomers," as they see their own ways of life greatly disrupted. This is the case of the Bakola/Bagyeli pygmies, for example, an ethnic minority in the region around Kribi,

Cameroon, who depend on forest products and hunting for their subsistence. They claim that the Chad–Cameroon pipeline construction has destroyed their medicinal plants and fishing and game areas, benefiting foreign workers and the Bantu population without providing meaningful compensation to them. The adverse impact on public health near oil localities is especially great. The migration of workers and the conditions of their housing lead to an increase in the incidences of communicable diseases, such as AIDS, other sexually transmitted diseases, tuberculosis, and cholera. Along the Chad–Cameroon pipeline, for example, temporary encampments have led to the rise of prostitution and, consequently, the appearance of human immune deficiency virus (HIV)/AIDS.

The environmental dimension of oil exploration is a chief cause of social dislocation. Hazardous wastes, site contamination, and the lack of sufficient protection of surface and subsurface waters, biodiversity, and air quality (both in the immediate vicinity of the oil project and in relation to global concerns such as ozone-depleting substances and greenhouse gases) have endangered the health of local populations near oil installations and pipelines; local livelihoods such as farming and fishing have been destroyed. Local communities have reported a sharp rise in infantile leukemia near oil facilities. This disruption is most profound among ethnic minorities and indigenous peoples who live off the land and whose customs and traditions may also be threatened. In Ecuador, the Cofan Indian Tribe has reported the contamination of its drinking supply. In Colombia, where at least 2.1 million barrels of petroleum have been spilled since 1987 (approximately 11 times as much oil as was spilled in the Exxon Valdez disaster of 1989), severe damage to this tropical ecosystem includes air pollution, land clearings, water contamination, soil erosion, sedimentation, and the disturbance of wildlife habitats. Petroleum wastes wash directly into local waterways, and Colombia's Institute of Natural Resources (INDERENA) has repeatedly condemned the presence of high concentrations of heavy metals and toxic polycyclic aromatic hydrocarbons, which are 300 times higher than drinking water standards allow in northern Colombia and 50% higher than international standards allow for oil discharges to surface waters.

The fate of the Niger Delta region, where exploration began in 1958, is the best known example of the local impact of oil exploration. Although 2 million barrels are pumped out of the Niger Delta mangrove swamps every day, providing

Nigeria with a large share of its GDP, over 90% of its export earnings, and almost all its tax revenues, the people in the region have barely benefited. Despite producing energy for the country and the world, many of the Nigerians do not even have electricity. Though compensation has paid for land acquisition and oil spillages have aided some individuals from the Ogoni minority, whose land is affected, the local economy and the environment have been devastated. Gas flaring has permanently scorched the earth, destroying food crops and rendering farmlands barren. Some scientists believe that the incomplete combustion of the flares has resulted in acid rain that, in turn, has damaged crops and drinking water. Oil spillages (an average of three per month) and ruptured pipelines (either from improper maintenance or sabotage) have destroyed streams, farmlands, and aquatic life. Thousands of villagers have been killed in pipeline explosions resulting from leaks, including over 700 people in one leak alone in October 1998. This has made unlivable the Alwa Ibom community in Iko, a once-thriving economically stable and self-supporting community. By most calculations, the region remains one of the most economically backward and politically marginalized in the country. As popular protest against the activities of oil companies rises and security forces are increasingly called on for protection of facilities, it is also one of the most conflict-ridden and politically explosive regions.

8. PETROVIOLENCE AND CIVIL WAR

Natural resources and war are linked, but oil plays a special role in this relationship. Economists have found that high levels of primary commodity export dependence are associated with civil war, but petroleum dependence is even more likely to be associated with conflict compared to any other commodity. Countries dependent on oil are more likely to have civil wars than are their resource-poor counterparts; these wars are more likely to be secessionist and they are likely to be of even greater duration and intensity compared to wars where oil is not present. Evidence of this relationship is both statistical and case study based.

Because oil produces such high rents, it can be the main impetus for going to war, either directly or indirectly. Oil revenues may be the catalyst for a conflict that might not otherwise have happened. In

the Congo Republic, for example, an opposition group received \$150 million funding from the French oil company, Elf Aquitaine, to support a takeover of the government so that the company could receive more favorable treatment under the new regime (which it subsequently did). The payment financed a 4-month war that resulted in 10,000 dead and the destruction of parts of Brazzaville. More frequently, the impact of oil on the outbreak of civil conflict is more indirect, i.e., the result of long-standing grievances over land expropriation, environmental damage, corruption, or the maldistribution of resources. This is especially true during bust cycles, as economic opportunities dry up. Recent civil wars and violent conflicts in oil-exporting countries have occurred in Algeria (since 1991), Angola (1975, 2002), Indonesia/Aceh (since 1986), Yemen (1990–1994), the Sudan (since 1983), Nigeria (1980–1984), Iraq (1985–1992), and the Republic of Congo (1997, 1999). Though cross border wars, e.g., the Iraqi invasion of Kuwait, have also occurred, the powerful association is with civil wars.

Secessionist wars are statistically more frequent in oil-exporting than in non-oil-exporting countries. When secessionist movements are present, the likelihood of conflict is especially high because the promise of oil wealth appears to make viable a secession that might not seem possible in poorly endowed areas. Not surprisingly, when oil is regionally concentrated and when benefits accrue to the nation while most adverse effects are local, secessionist wars are more likely. Examples abound. In the Sudan, war was triggered by President Numeiry's decision to place newly discovered oil fields, in the country's Christian south, under the control of the Muslim north. In Indonesia, the Aceh Freedom Movement has denounced the government for stealing Aceh's oil and natural gas resources as a main reason for its separatist struggle, and it has used the analogy of Brunei to convince its followers that Aceh could be equally as rich. In Nigeria, Biafra's move to secede occurred only after the government had made fiscal decisions that treated oil as a centralized, rather than a regional, asset. In this way, fights over oil revenues may become the reason for ratcheting up levels of preexisting conflict in a society.

Oil dependence is associated with particularly intense conflict. Because petroleum is so technologically sophisticated and requires so much capital, it is not easily extracted and transported; it is not "lootable" like drugs or gems. This means that it is difficult for rebels or generally unskilled groups to exploit, but governments can use this wealth to

attempt preemptive repression. This is the case in Sudan, for example, where the government tried to forcibly clear entire populations away from the oil fields and the pipelines. Oil rents paid for the destruction of crops and razing of houses as well as for widespread terror against the local population. Oil's nonlootability also means that separatist conflicts (like that of the Sudan) may be especially bloody and intractable where petroleum deposits coincide with the presence of minority groups. When more straightforward fights over the distribution of oil rents between groups can be resolved by a new pattern of distribution, this is often not the case in separatist wars. But when oil is involved, such struggles are generally resolved only by the seizure of control of oil fields by rebels and a subsequent declaration of autonomy, or by the government's total defeat of the minority located near the oil fields.

Oil dependence, like dependence on other mineral resources, is associated with civil wars of long duration. Wars are expensive to pursue, and both governments and rebels can use oil rents to finance their armies. Because petroleum is transported generally through pipelines, it can be easily disrupted, and pipelines are an invitation to extortion. In Colombia, for example, oil revenues support the government's battle against rebel movements, but because petroleum must be transported to the coast through two pipelines that are both over 400 miles long, there are almost unlimited opportunities for rebels to extract "protection rests" and other forms of resources from oil companies. In 2000 alone, the pipelines were bombed 98 times and kidnappings for ransom were frequent; according to one estimate, rebel groups have managed to earn an estimated windfall of \$140 million annually.

9. CONCLUSION

More than any other group of countries, oil-dependent countries demonstrate perverse linkages between economic performance, poverty, bad governance, injustice, and conflict. This is not due to the resource per se, but to the structures and incentives that oil dependence creates. Various proposals exist to mitigate this "paradox of plenty," including demands for revenue transparency by oil companies and exporting governments, revenue management schemes, stabilization funds to mitigate price shocks, reforms of taxation and civil service, and the democratization and deconcentration of both the industry and the exporting countries. Without the

implementation of reforms, the consequences of oil dependence will continue to be adverse.

SEE ALSO THE FOLLOWING ARTICLES

Development and Energy, Overview • Geopolitics of Energy • Markets for Petroleum • Nationalism and Oil • National Security and Energy • Oil and Natural Gas Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Industry, History of • Oil Price Volatility • Resource Curse and Investment in Energy Industries • Sustainable Development: Basic Concepts and Application to Energy

Further Reading

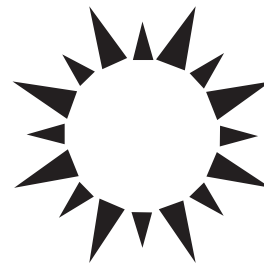
- Auty, R. (1993). "Sustaining Development in the Mineral Economies: The Resource Curse Thesis." Routledge, London.
- Auty, R. (ed.). (2001). "Resource Abundance and Economic Development." Oxford Univ. Press, Oxford.
- Beblawi, H., and Luciani, G. (eds.). (1987). "The Rentier State." Croom Helm, New York.
- Chaudry, K. (1997). "The Price of Wealth: Economies and Institutions in the Middle East." Cornell Univ. Press, Ithaca, New York.
- Bannon, I., and Collier, P. (eds.). (2003). "Natural Resources and Violent Conflict: Actions and Options." World Bank, Washington, D.C.
- Collier, P., and Hoeffler, A. (1998). On economic causes of civil war. *Oxford Econ. Pap.* 50, 563–573.
- Gary, I., and Karl, T. (2003). "Bottom of the Barrel: Africa's Oil Boom and the Poor." Catholic Relief Services, Baltimore, Maryland.
- Gedicks, A. (2001). "Resource Rebels: Native Challenges to Mining and Oil Corporations." South End Press, Cambridge, Massachusetts.
- Gelb, A. (1988). "Oil Windfalls: Blessing or Curse?" Oxford Univ. Press, Oxford, United Kingdom.
- Glyfason, T. (2001). Natural resources, education and economic development. *Eur. Econ. Rev.* 45(4–6), 847–859.
- Human Rights Watch. (1999). "The Price of Oil: Corporate Responsibility and Human Rights Violations in Nigeria's Oil Producing Communities." Available on the Internet at www.hrw.org.
- Karl, T. (1997). "The Paradox of Plenty: Oil Booms and Petro-States." University of California Press, Berkeley, California.
- Karl, T. (1999). "The perils of petroleum: Reflections on the paradox of plenty. In "Fueling the 21st Century: The New Political Economy of Energy." [Special edition of *J. Int. Affairs* 53 (No. 1).]
- LeBillon, P. (2001). The political ecology of war: Natural resources and armed conflict. *Pol. Geogr.* 20, 561–584.
- Leite, C., and Weidmann, J. (1999). "Does Mother Nature Corrupt? Natural Resources, Corruption, and Economic Growth." IMF Working Paper WP/99/85. International Monetary Fund Publ., Washington, D.C.
- Mahdavy, H. (1970). Patterns and problems of economic development in rentier states: The case of Iran. In "Studies in the Economic History of the Middle East (M. A. Cook, ed.). Oxford Univ. Press, Oxford, United Kingdom.
- Ross, M. (2001). "Extractive Sectors and the Poor." Oxfam America Report. Oxfam, London.
- Ross, M. (2001). Does oil hinder democracy? *World Pol.* 53
- Sachs, J., and Warner, A. (1999). The big push: Natural resource booms and growth. *J. Dev. Econ.* 59, 1.
- Smith, B. (2004). Oil wealth and regime survival in the developing world, 1960–1999. *Am. J. Pol. Sci.* (in press).

Oil Pipelines

PRASANTA KUMAR DEY

University of the West Indies

Bridgetown, Barbados



-
1. Introduction
 2. Pipeline System Design Issues
 3. Pipe Coating System Selection
 4. Pipeline Route Selection
 5. Construction Management Issues
 6. Risk-Based Inspection and Maintenance

Glossary

corrosion One of the major causes of pipeline failure is corrosion, an electrochemical process that changes metal back to ore. Corrosion generally takes place when there is a difference of potential between two areas having a path for the flow of current. Due to this flow, one of the areas loses metal.

corrosion protection system The combination of coating and cathodic protection.

Environmental Impact Assessment (EIA) A planning aid concerned with identifying, predicting, and assessing impacts arising from proposed activities such as policies, programs, plans, and development projects that may affect the environment.

pipe coating system To protect pipelines from corrosion, pipes are coated with chemical compounds, which work as insulator. The coating system commonly uses coal tar enamel, fusion bonded epoxy (FBE), and three-layer side extruded polyethylene (3LPE) as coating materials.

pipeline failure The use of the term failure in the context of pipelines varies from country to country and across organizations. For example, in Western Europe, any loss of gas/oil is considered a failure, whereas in the United States, an incident is considered to be a failure only when it is associated with the loss of commodity and also involves a fatality or injury, or damage of more than \$50,000. In India, generally any loss of commodity, however small, is considered to be a failure. Hence, failure is defined as any unintended loss of commodity from a pipeline that is engaged in the transportation of that commodity.

pipeline system Consists of oil terminal (offshore/on-shore), tank farm, originating pumping station, inter-

mediate pump station, scrapper station, pipelines connecting terminal, tank farm, and stations.

risk-based maintenance A selecting and prioritizing method of inspection and maintenance practices (tools and techniques) for facilities, equipment, and machines on the basis of their failure characteristics (likelihood and severity of failure).

Socio-Economic Impact Assessment (SEIA) A study for assessing, in advance, the social and economic consequences of the initiation of any industrial activity to the human population. SEIA forms a part of the EIA, which is mandatory for certain industrial projects in many countries.

strategic sourcing An institutional process of developing partnership among all stakeholders across the supply chain.

Oil pipelines are the nervous system of the oil industry, as they transport crude oil from sources to refineries and petroleum products from refineries to demand points. Therefore, the efficient operation of these pipelines determines the effectiveness of the entire business.

1. INTRODUCTION

Cross-country oil pipelines are the most energy-efficient, safe, environment-friendly, and economic way to ship hydrocarbons (gas, crude oil, and finished products) over long distances, either within the geographical boundary of a country or beyond it. A significant portion of many nations' energy requirements is now transported through pipelines. The economies of many countries depend on the smooth and uninterrupted operation of these lines, so it is increasingly important to ensure the safe and failure-free operation of pipelines.

Pipelines are the irreplaceable core of the U.S. petroleum transportation system and hence the key to meeting petroleum demand. Without oil pipelines, petroleum products would not reach the millions of

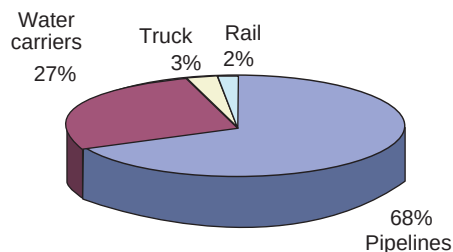


FIGURE 1 Domestic shipments of petroleum products in the United States.

consumers in all 50 states. Oil pipelines transport roughly two-thirds of the petroleum shipped in the United States; they deliver more than 14 billion barrels of petroleum per year. Figure 1 shows the domestic shipments of petroleum in the United States using various modes. Figure 2 shows both the crude and petroleum products pipelines network in the United States.

While pipelines are one of the safest modes of transporting bulk energy, and have failure rates much lower than the railroads or highway transportation,

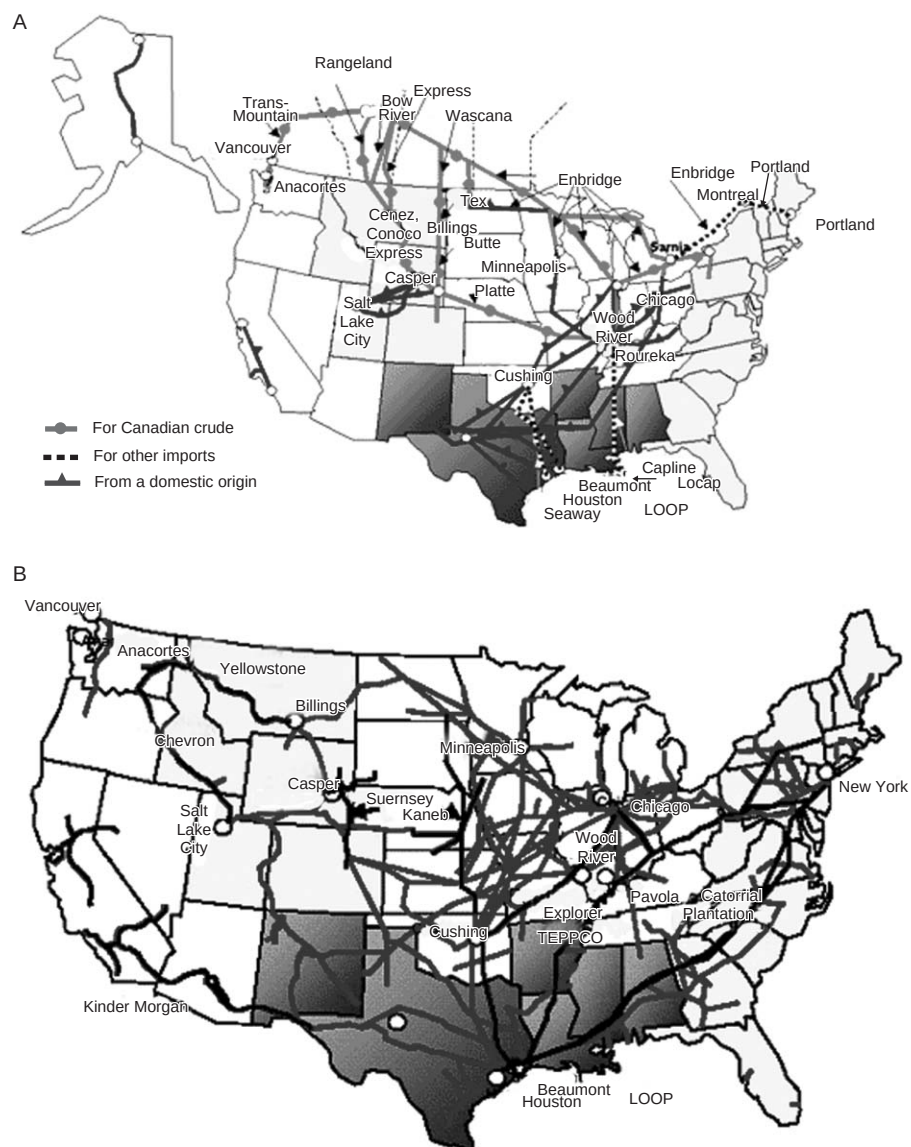


FIGURE 2 Crude and petroleum products pipelines network in the United States. The crude oil pipeline length is 55,000 miles, and the products oil pipeline length is 95,000 miles. (A) Selected crude oil truckline systems. From www.pipeline101.com. (B) Major refined products pipeline. State shading shows Petroleum Administration for Defense Districts (PADDs). From Allegro Energy Group (2001).

TABLE I

Length of Pipeline and Service (000s km) in Western Europe

	1971–1975	1976–1980	1981–1985	1986–1990	1991–1995	1996–2000
Crude	7.9	8.7	8.7	8.0	8.6	8.1
Product	7.4	8.6	9.3	15.6	20.8	21.7
Hot	0.7	0.7	0.6	0.5	0.5	0.5
Total	16.0	18.0	18.7	24.1	29.9	30.3

The spillage frequencies by cause and their proportions shown over three decades

	1971–1980		1981–1990		1991–2000	
Cold pipelines	Spillages per 1000 km	% of total	Spillages per 1000 km	% of total	Spillages per 1000 km	% of total
Third party	0.31	41%	0.19	38%	0.14	41%
Natural	0.04	5%	0.02	3%	0.01	2%
Corrosion	0.12	16%	0.12	24%	0.07	20%
Operational	0.06	7%	0.06	12%	0.03	8%
Mechanical	0.23	31%	0.11	22%	0.10	30%

Source. Concawe (2002).

failures do occur, and sometimes with catastrophic consequences. In 1993 in Venezuela, 51 people were burned to death when a gas pipeline failed and the escaping gas ignited. In 1994, a 36-in. (914 mm) pipeline in New Jersey failed, resulting in the death of one person and more than 50 injuries. Similar failures also have occurred in the United Kingdom, Russia, Canada, Pakistan, and India. While pipeline failure rarely cause fatalities, disruptions in operation lead to large business losses. Failures can be very expensive and cause considerable damage to the environment. The safety performance of the oil pipeline industry in United States has improved substantially. All oil pipelines together had 129 spills larger than 50 barrels (2100 gallons) in 2001. Deaths and injuries resulting from oil pipeline transportation are rare but do occur occasionally. Table I shows the statistics of spillage incidents of Western European cross-country oil pipelines over the past 30 years along with the information about the pipeline network over the past 30 years.

Achieving pipeline throughput depends on failure-free operations, which are feasible through efficient pipeline system design, appropriate coating system selection, best pipeline route selection, effective construction planning, and an optimum maintenance policy.

2. PIPELINE SYSTEM DESIGN ISSUES

Figure 3 shows a schematic diagram of a typical pipeline system. Crude oil is received from an offshore

terminal/oil field and transported to the tank farm and from the tank farm to the refinery through pipelines for refining. Depending on the distance between the tank farm and refinery, hydraulic gradient, and pipe size, the number of pump stations is designed to facilitate the transportation. Refinery produces petroleum products through distillation processes and they are stored into various product tanks. From these tanks, products are pumped via products pipelines to different destination as per demand. Intermediate pump stations are designed in accordance to length of the pipeline, hydraulic gradient, and pipe size.

For a given distance in the same terrain and constant flow rate, with the increase of pipe size (diameter), the power requirement reduces as pipe friction reduces with pipe diameter. This increases capital expenditure but reduces operating cost. Therefore, pipeline operators require determining the size (diameter) of the pipeline with the consideration of life cycle costing (capital cost and operating cost) of the project. Figure 4 demonstrates the relationships between pipe diameter and the total cost of the pipeline project. Accordingly, the number of pump stations and pumping units and their capacities are derived. However, with the increase in oil demand, pipeline throughput (flow rate) must be enhanced, which can be done in one of the following ways:

- Augmenting capacity of the pumping unit (pump and prime mover), either by replacing or modifying configuration
- Constructing additional intermediate pump stations

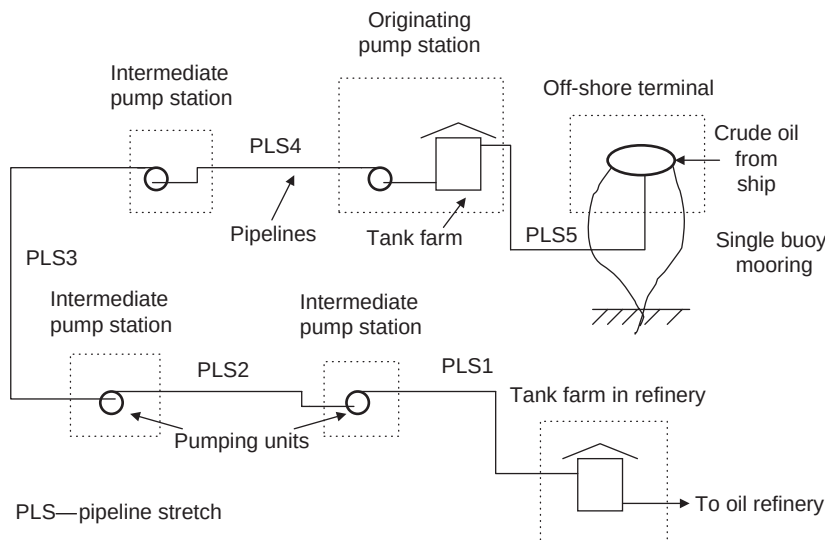


FIGURE 3 Schematic of cross-country crude oil pipelines.

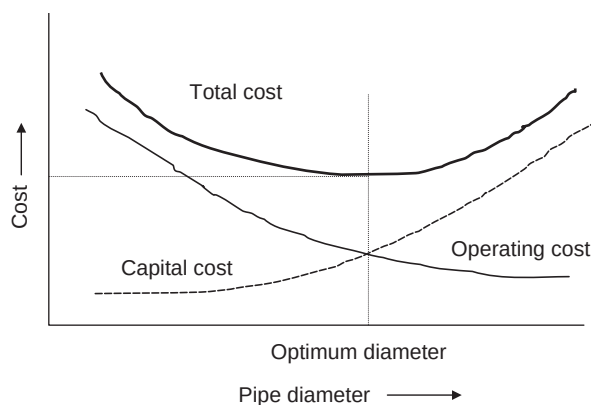


FIGURE 4 Relationship between pipeline diameter and total cost (capital and operating) of pipeline.

- Replacing the existing pipeline with a bigger pipeline
- Laying twin line/loop lines

Each method has its own pros and cons. Therefore, consideration of projected oil demand during the feasibility study leads to the design of a better pipeline system by making a rationale decision on pipe size (diameter and thickness), number of pump stations, and number and capacity of pumping units.

3. PIPE COATING SYSTEM SELECTION

Pipeline coatings are the primary means for protecting pipelines against external corrosion. In fact, the

life of pipelines depends mostly on coating materials and their applications. These not only reduce the chance of pipeline failure from external corrosion considerably, but also reduce inspection and maintenance cost throughout their lives. Hence, selection of the right type of coating materials for pipelines improves the overall productivity of the pipeline organizations. Wide varieties of coating materials are available for cross-country petroleum pipeline application. The coal tar enamel (CTE), three-layer side extruded polyethylene (3LPE) coating, and fusion bonded epoxy (FBE) coatings are commonly used by pipeline operators. They are characterized by varied physical and chemical properties, and each has its own varied application procedure. Although materials and application cost are the major factors for coating selection, factors like soil characteristics, environment concern, maintainability, and safety are also getting increased attention from pipeline operators. All these factors again behave in a particular manner when cathodic protection is applied on a coated pipeline. Coating behavior will to a large extent also depend on cathodic protection parameters. Therefore, to achieve corrosion protection for the design life of a pipeline, the corrosion protection system—the combination of coatings and cathodic protection—has to be designed, specified, applied, constructed, maintained, and managed correctly.

Study shows that FBE coating is mostly employed by the pipeline operators compared to CTE and 3LPE tape coatings, primarily due to its superior performance with respect to soil stress resistance,

adhesion, and cathodic protection shielding, although FBE is cost intensive for both materials and application. It has excellent compatibility with cathodic protection and stands up well against high temperature. However, it is weak from the point of view of impact damage and moisture absorption. There are instances of poor performance of the coating system using FBE, mainly due to faulty application. In the early days of using FBE, a number of problems were found and recognized. Some of these are as follows:

- Poor adhesion on one side of the longitudinal seam weld. Caused by inadequate surface preparation and/or low temperature of application.
- Pinholes in or roughening of the coating. Caused by inadequate surface preparation (salts).
- Craters/blistering of the coating. Caused by extraneous material (e.g., rubber) trapped in the coating as it gels, or mill scale trapped in the steel and giving off gasses while coating is curing.
- Foaming. Can be caused by high levels of moisture in the powder or too high an application temperature.
- Strain lines/cracks in the coating. Caused by hydrostatic testing at very low temperatures.
- Blisters in the parent coating when induction heating field joints. Can be caused by moisture uptake, trapped moisture, contamination under the coating, or incorrect temperature profile during coating (cold ends).
- Loss of adhesion at the cutback. Can be caused by moisture uptake or contamination under the coating.
- Blistering of the coating soon after application over the circumferential weld—joint coating. Following cellulosic welding techniques, hydrogen gas can be evolved.
- Leopard spotting in the ground. Caused by the formation of corrosion products underneath coating that has lost adhesion to the substrate, disbonded, or delaminated. This usually occurs at pinholes.

CTE coatings have a long history of pipeline protection. CTE has good performance record across the world. It has excellent electrical insulating property and low permeation properties that resist bacterial attack and resist solvent action of petroleum oils. It is less expensive compared to other types of coating materials. Pipelines with CTE coating can be easily maintained as field repair of external coating can be easily organized. A temporary

operations facility can be built near the site, as manual and semiautomatic application of coating is feasible. This reduces the logistic expenditure considerably. The performance of CTE depends mostly on its application. Hence, there are instances of pipeline failures due to poor application CTE coating. Handling during storage, transportation to the site, and handling during construction also contribute pipeline failure. CTE suffers from drawbacks such as the negative environmental impact as it produces fumes, which is injurious to health during application. CTE coatings are subjected to soil stress, which can cause coating to wrinkle, crack, disband, and leave exposed steel surfaces. Additionally, if water penetrates under disbonded coatings, cathodic protection requirement generally increases as the pipelines with CTE coatings age. Operating temperature of pipelines with CTE coating should not go beyond 65°C.

The 3LPE coatings (3-layer, low density, polyethylene) called for 50 to 75 microns of fusion bonded epoxy powder, ~150 microns of a polyethylene co-polymer adhesive and an overall total film thickness of 2.3 mm using a low-density polyethylene (PE). As low-density polyethylene (LDPE) has a higher rate of moisture vapor transmission, it has more chance of failure. Adhesion is the major problem of this type of coating. Coating performance mostly depends on surface preparation, application, and storage facilities of coated pipes.

Selection of a specific coating for a pipeline is governed by the factors like coating characteristics, application complexity, environment and safety aspects, the physical facility requirement for coating application, and the cost of materials and application.

The coating is primarily characterized by its soil stress, adhesion, and cathodic protection shielding properties. Soil stress is the amount of pressure or pull applied to the coating surface after being buried. This can cause the coating to wrinkle, sag, or crack. Different coating materials have different soil stress properties. Coating that is least affected by soil stress performs better than coating that is more affected by soil stress. Adhesion is the measure of bond between the coating materials and pipe materials. Hence, coating with a better bond is treated as superior. If the coating loses adhesion to the pipe materials, corrosion enters into the annular space between the pipe surface and the disbond coating. In this situation, corrosion takes place if the disbonded coating does not permit entry of the cathodic protection current. This phenomenon is called

cathodic protection shielding and depends on the type of coating. A coating that has lesser tendency to shield cathodic protection current is preferred.

Different coating materials have different application procedures and requirements. Pipe surface preparation is an important activity for improving adhesion. FBE and CTE require more stringent surface preparation than other types of coatings. Some coating types are vulnerable from handling and construction damage and call for advanced logistic facilities for transporting to the site for construction. Pipe-to-pipe welding onsite requires field joint coating for all types of coating. Hence, preference is given to the type coating that has better compatibility to a wide range of field coating systems. Depending on the application complexity, the coating types can be classified and prioritized. None of the coating material has infinite life and is liable to get damaged for various reasons. Hence, repair and maintenance should be taken up for effective operations of the pipeline system. The coating that is compatible with a wide range of repair and maintenance practices is preferred over others.

The physical facility requirement for the coating application is an important consideration for selecting a coating type. This depends on plant, logistics, and warehouse requirements. The selection of technology (automatic, semiautomatic, and manual) has a strong bearing on it. Pipeline operators often choose to construct a temporary coating plant near to the construction site so as to get the advantage in expensive logistics and warehousing. However, this reduces automation and, in turn, productivity. Some

coating applications are not at all technically and economically feasible through temporary operations facilities. Although through superior technology, good quality coating can be produced in a competitive price through scale economy, it increases logistics and warehouse cost considerably.

Environmental friendliness and the tendency to be less hazardous to health are other factors to be considered when selecting the right coating, as all pipe coating production technologies affect the environment negatively and present considerable health hazards to the operators. Many countries have their own standard environmental regulations, which restrict the use of some coating materials. Manual application always results in more hazard to operators than automatic and semiautomatic application. Hence, even in a labor-intensive economy, preference should be given for selecting a coating that calls for less human intervention during application. Additionally, adherence to a stringent inspection requirement is another factor that can attract considerable attention to a specific coating system.

Along with these factors, materials cost and application cost are to be considered while selecting the coating type for any pipeline system. Figure 5 shows the selection model in an integrated framework.

The following steps will ensure that the pipeline coating system is effective and improves pipeline performance:

- Select the best coating system for specific project.
- Develop specifications for application.

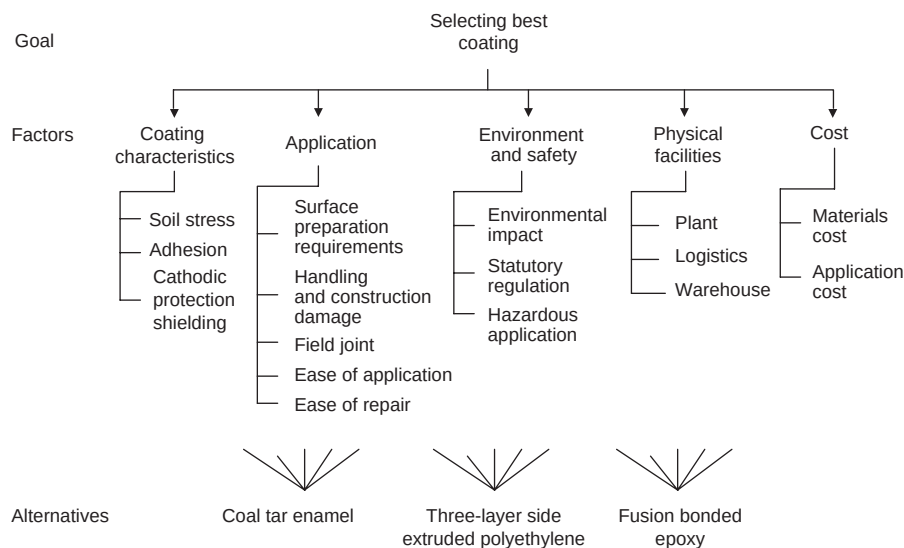


FIGURE 5 Coating system selection model.

- Customize production facilities and develop inspection procedures (project specific).
- Provide storage and logistics for coated pipes.
- Organize field joint coating and subsequent maintenance activities.

4. PIPELINE ROUTE SELECTION

Pipeline route selection plays a major role when designing an effective pipeline system, as not only the health of the pipeline depends on its terrain, but also the productivity of pipeline business depends on length, operability, constructability, and maintainability.

Cross-country petroleum pipeline route selection is governed by the following goals:

- Establish the shortest possible route connecting the originating, intermediate, and terminal locations.
- Ensure, as far as practicable, accessibility during operation and maintenance stages.
- Preserve the ecological balance and avoid/minimize environmental damage; the route should be kept clear of forests as much as possible.
- Avoid populated areas for public safety reasons.
- Keep rail, road, river, and canal crossings to the bare minimum.
- Avoid hilly or rocky terrain.
- Avoid a route running parallel to high-voltage transmission lines or D.C. circuits.
- Use an existing right of way, if possible.
- Avoid other obstacles such as wells, houses, orchards, lakes, or ponds.

These characteristics must be determined by a reconnaissance survey and the goal of finding the shortest possible route is always important. There are many examples of loss of productivity due to the wrong pipeline route selection.

A specific stretch of pipeline in eastern India (102 km, 12.75 in./324 mm in diameter, 0.25 in./6.35 mm wall thickness, API 5L X-46 pipe grade) was commissioned in September 1977. The pipeline is protected against external corrosion by an impressed current cathodic protection (CP) system, over and above the coal tar enamel external coating. The right of way includes a waterlogged of 30 km area that remains under water for 6 to 7 months in a year, with a water depth varying between 3 and four 4 m. In this 30-km stretch, extensive external corrosion has occurred. Between pipeline chainage (distance from a reference point) of 11 km to 41 km,

two leaks were encountered, and at 203 locations, sleeving (reinforcement cladding) has been carried out due to heavy external corrosion. The pipeline operates at a lower throughput than its designed capacity; it is estimated that this pipeline incurs a loss of US\$ 2.5 million per year on average (US\$ 2.3 million in lost production due to reduced capacity operation plus US\$ 0.2 million for additional maintenance cost). If the pipeline had been rerouted during the initial design phase, the capital cost would have increased by US\$ 2 million (US\$ 1 million for additional mainline cost, plus US\$ 1 million for an additional station). The benefit of an alternative route is clear.

Another example of poor route selection is a 40-km stretch of coal belt along the route of a petroleum pipeline in eastern India. This stretch is vulnerable to failure due to third-party activities (coal mining) in the surrounding area. A risk analysis study determined that the probability and severity of a failure here were very high, with an expected cost of US\$ 20 million. Rerouting the pipeline during the design stage would have cost US\$ 1.25 million. Avoiding this area would have been cost-effective.

Petroleum pipelines are designed to carry a specific capacity for a specific life period and rely on a forecast of a supply-demand scenario of petroleum products. It is quite likely that during a pipeline's life span, the capacity of the line may need to be increased to maximize returns commensurate with petroleum product demand. However, if the pipeline route does not provide adequate room for augmentation, this cannot be accomplished.

Pipeline route selection should be governed not only by factors like the shortest total distance, approachability, and constructability; factors such as operability, augmentation capability, and maintainability also need to be considered. The following steps can be taken to select the best pipeline route:

- Identify a few feasible alternate pipeline routes using a geographic information system (GIS).
- Develop a database for all alternate feasible pipeline routes for analysis. Table II shows a typical database for four alternate pipeline routes.
- Identify factors that lead to selecting the best pipeline route among the alternatives. Table III shows a list of factors and subfactors that are identified by the experienced pipeline operators for a typical pipeline route selection problem.
- Select the optimum route using the preceding information with active involvement of the concerned stakeholders.

TABLE II
Pipeline Database

Description	Route I	Route II	Route III	Route IV
Throughput (MMTPA) ^a	3	3	3	3
Length (km)	780	1000	750	800
Number of stations ^b	3	3	3	3
Terrain detail (km)	430	785	570	770
Normal terrain	2	5	45	15
Slushy terrain	—	1	3	2
Rocky terrain	—	5	7	2
Forest terrain	3	4	5	1
River crossing	330	200	120	10
Populated area	15	—	—	—
Coal belt area				
Soil conditions	Less corrosive soil	Less corrosive soil	Corrosive soil for slushy terrain	Less corrosive soil
Third-party activity	More due to coal belt and populated area	More due to populated area	More due to populated area	—
Chances of pilferage	Higher due to populated area	Higher due to populated area	Higher due to populated area	—

^a MMTPA = million metric tons per annum.

^b 1-originating pumping station, 1-intermediate pump station, 1-terminal delivery station.

TABLE III
Factors and Subfactors

1 Pipeline length
2 Operability
a Hydraulic gradient
b Expansion possibility
3 Maintainability
a Corrosion
b Chances of pilferage
c Third party activities
4 Approachability
a Proximity to railway and highway
b Terrain characteristics
5 Constructability
a Statutory clearances from various authorities
b Mobilization
c Availability of power/water
d Ease of construction
6 Environmental friendliness

- Carry out a reconnaissance survey.
- Finalize the route through a detailed survey.

The following paragraphs describe each factor and subfactor.

Pipeline length governs the capacity requirement of almost all the equipment for the entire pipeline project, as pipeline head loss is directly proportional to the length of the pipeline. Hence, the shorter the length of a pipeline, the lower the capital cost of the project and vice versa.

The hydraulic gradient is a major factor in selecting prime mover power for pipeline operations, as a negative hydraulic gradient demands higher prime mover power. Similarly, more route diversion causes more friction loss, resulting in higher prime mover power for the same throughput. These result in more capital investment.

A pipeline is designed for specific throughput in line with demand; a pipeline may need to be augmented in the future to cope with the demand for maximizing profit. Therefore, expansion/augmentation capability is one attribute of a properly designed pipeline. In addition to improving the existing prime mover capacity, a pipeline can also be augmented by installing more pumping stations along the route and laying loop lines/parallel lines. Therefore, the pipeline route should be chosen with a view to its augmentation capability.

Although pipelines are designed with adequate safety factors, they are subject to failure due to various reasons. Pipeline corrosion, pilferage, and third-party activities are factors that may create

quantum throughput loss along with chances of disaster; therefore, these factors should be carefully considered during the feasibility study. In a decision model, these factors may influence the selection of a specific route.

One of the major causes of pipeline failure is corrosion, an electrochemical process that changes metal back to ore. Corrosion generally takes place when there is a difference of potential between two areas having a path for the flow of current. Due to this flow, one of the areas loses metal.

External interference is another leading cause of pipeline failure. It can be malicious (sabotage or pilferage) or be caused by other agencies sharing the same utility corridor. The latter is known as third-party activity. In both cases, a pipeline can be damaged severely. External interference with malicious intent is more common in socioeconomically backward areas, while in regions with more industrial activity, third-party damage is common.

All activities, industrial or otherwise, are prone to natural calamities, but pipelines are especially vulnerable. A pipeline passes through all types of terrain, including geologically sensitive areas. Earthquakes, landslides, floods, and other natural disasters are common reasons for pipeline failures.

Although a cross-country petroleum pipeline is buried underground, the right of way should allow uninterrupted construction activities as well as operation, inspection, and maintenance. The ideal pipeline route should be along a railway track or a major highway. This is not always possible due to the long length of pipelines, which may require river crossings and passage through forests, deserts, and so on. Therefore, a pipeline route with better approachability gains an edge over other routes.

Laying a pipeline across states/province or national boundaries requires permission from statutory governmental authorities. Stringent safety and environmental stipulations are sometimes hindrances to project activities.

Mobilization is a major construction activity. One factor in pipeline routing is the provision for effective mobilization by the contractor. Distance to market, the availability of power and water, and the number of skilled and unskilled workers are typical requirements for starting effective construction activities.

Pipeline construction methods vary greatly with terrain conditions. For example, laying a pipeline across a river requires horizontal direction drilling (HDD), while laying pipe across a rocky area requires rock-trenching techniques. Therefore, location characteristics are a major cost component of

pipeline construction. Inappropriate route selection can cause major time and cost overruns.

Poor construction, combined with inadequate inspections and low-quality materials, also contributes to pipeline failures. Other reasons include human and operational error and equipment malfunctions. Computerized control systems considerably reduce the chance of failure from these factors.

Pipelines handle hazardous petroleum products. Although pipelines are designed with safety features, failure is not uncommon. Sometimes failures result in a release of large quantities of petroleum products into the environment. If this should happen, a pipeline located in a remote area is less of a safety concern.

Figure 6 depicts a pipeline route selection model.

5. CONSTRUCTION MANAGEMENT ISSUES

Estimated miles of natural gas, crude oil, and refined-products pipelines under way or planned for construction outside the United States and Canada total 75,995 miles (122,276 km) (Pipeline and Gas Industry, 2001). Table IV shows the regional breakup of pipelines under construction or planned.

A 600-mile (1072 km) Chad–Cameroon pipeline has been planned to commission by 2004. The pipeline would increase government revenues by 45 to 50% per year and allow it to use those resources for important investments in health, education, environment, infrastructure, and rural development, necessary to reduce poverty. However, the plan suffers from tremendous negative impacts like social and ecological risk as the pipeline passes through environment-sensitive zones, loss of livelihood and resettlement, and loss of biodiversity. Moreover, the selection of a pipeline route had political intent. Although the status of the project is quite impressive, the project is vulnerable with respect to financial and political risk due to the involvement of various international organizations including the World Bank, ExxonMobil of the United States, Petronas of Malaysia, and ChevronTexaco of the United States.

The government of Georgia has authorized the construction of the Baku–Tbilisi–Ceyhan oil pipeline. The document was signed by environment minister Nino Chkhobadze in the presence of the country's president and minister of state. Authorization was given provided a number of ecological conditions are complied with; otherwise the authorization may be automatically annulled, the environment ministry

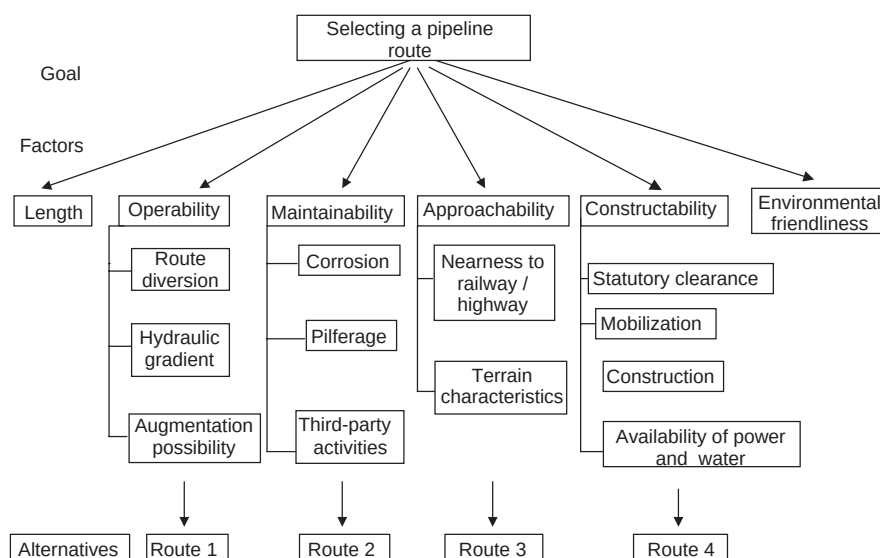


FIGURE 6 Model for selecting oil pipeline route.

TABLE IV

Outside United States/Canada – 75,995 Miles of Pipelines Under Way/Planned/Proposed

Country	Gas		Crude		Products		Other		Totals		
	Current	Future	Current	Future	Current	Future	Current	Future	Current	Future	Totals
Europe	2704	14,990	568	1830	40	9			3312	16,829	20,141
Middle East	878	9786		2548		284			878	12,618	13,496
Africa	1278	3442	725	511	484	500			2487	4453	6940
South Pacific	1017	6816		120		101			1107	7037	8144
Far East	649	9384	206	2341		565			855	12,290	13,145
Mexico/Central America	101	2717		41		346			101	3104	3205
South America	745	6889	477	746	590	1113	364		2176	8748	10,924
Totals	7372	54,024	1976	8137	1204	2918	364		10,916	65,079	75,995

Source. Pipeline and Gas Industry (2001).

reported. Thus, the pipeline will not be laid via sanitary zones of water bodies, including the Tsalkskoye water reservoir and Lake Tabatskuri. The same rule will be applied to the Borjomi Gorge, where popular mineral water is produced. It has not been disclosed how much longer the pipeline will become as a result. Trying to allay the public anxiety because of the pipeline being laid in the Borjomi Gorge, where famous mineral springs are situated, the president of Georgia, Eduard Shevardnadze, said in an interview on Georgian national radio that the project's investors had proposed almost "unprecedented" measures for environmental protection. This clearly indicates construction management issues of mega pipeline projects.

A cross-country petroleum pipeline construction project is characterized by the complexity of its execution with respect to lack of experience in relation to certain design conditions being exceeded (water depth, ground condition, pipeline size, etc.), the influence of external factors that are beyond human control, external causes that limit resource availability (of techniques and technology), various environment impacts, government laws and regulations, and changes in the economic and political environment. Cost and time overruns and the unsatisfactory quality of a project are the general sources of disappointment to the management of a pipeline organization.

When it comes to projects that involve worldwide resources and stakeholders, project managers know

the drill: plan early, rally support, and appease critics. Without a proactive strategy, ventures could languish and excessive costs could accrue. This is specially the case in pipeline projects. Stakeholders must be wooed before the engineering and logistics of pipelines can be considered. If authorized, the project is studied and designed before managers are in a position to recruit well-trained workers, synchronize management philosophies, and work to ensure the timely arrival of materials, parts, and equipment.

Oil pipelines are environmentally sensitive because they traverse through varied terrain covering crop fields, forests, rivers, populated areas, desert, hills, and offshore. Any malfunction of these pipelines may cause a devastating effect on the environment. Hence, the pipeline operators plan and design pipeline projects with sufficient consideration of environmental and social aspects along with the technological alternatives. Traditionally, in project appraisal, the optimum technical alternative is selected using financial analysis. Impact assessments (IA) are then carried out to justify the selection and subsequent statutory approval. However, IAs often suggest alternative sites or alternate technology and implementation methodology, resulting in a revision of the entire technical and financial analysis. This problem can be resolved by carrying out technical analysis, social assessment, and environmental impact assessment together in an integrated framework. This will help planners to select the best project among a few feasible projects. Subsequent financial analysis then justifies the selection.

Once a market need for a new pipeline is established, potential routes are determined by looking at the predicted flow of crude oil from sources to refinery, and petroleum products from refinery to demand points. Potential routes are then chosen based on pipes and construction costs, projected residents, and probable returns. Certainly, selecting the shortest distance between two points and routing along a railway or major highway is preferable, but it is not always possible because of wildlife and aquatic concerns, as well as issues involving right of way. Project managers must therefore approach the regulators with jurisdiction over such proposals and learn the potential pitfalls. Field surveys and environmental impact assessments must be performed. While the degree of permissiveness varies in different region of the world, overall there must not be any significant effect on either the natural habitat or the landowners who lease their right of way. Typically after months—if not years—of wrangling, the common solution is to reroute the pipeline to accommodate all concerns.

The success of any project is determined by whether it is completed on time, stays within budget, and achieves the requisite performance (technical requirement). The main barriers for achievement are changes in the project environment. The problem multiplies with the size of the project as uncertainties in project outcome increases with size. Large-scale construction projects are exposed to uncertain environments because of such factors as planning and design complexity, the presence of various interest groups (project owner, owner's project group, consultants, contractors, vendors, etc.), resources (materials, equipment, funds, etc.) availability, the climatic environment, the economic and political environment, and statutory regulations.

Although risk and uncertainty affect all projects, size can be a major cause of risk. Other risk factors include the complexity of the project, the speed of its construction, the location of the project, and its degree of unfamiliarity.

As oil pipeline construction projects are characterized by all of these factors, a conventional approach to project management is not sufficient, as it does not enable the project management team to establish an adequate relationship among all phases of project, to forecast project achievement for building confidence of project team, to make decisions objectively with the help of an available database, to provide adequate information for effective project management, and to establish close cooperation among project team members. Moreover, oil pipeline construction should be completed on a fast track, not only to achieve faster return on investment but also to reduce the negative impact of the project, as considerable negative impacts can occur during the construction of a pipeline, especially in highly populated areas and in crop fields. To achieve success, therefore, the pipeline operators must increasingly adopt and manage construction risks effectively and manage procurement through strategic sourcing.

5.1 Construction Risk Management

Figure 7 shows a flowchart for construction risk management, and Fig. 8 shows the work breakdown structure of an oil pipeline project. The risk factors and subfactors are identified with the involvement of executives working in projects through brainstorming sessions. In the brainstorming session, executives were given a checklist of risk, which was used initially to identify risk factors and subfactors; then, through group consensus, the executives developed a

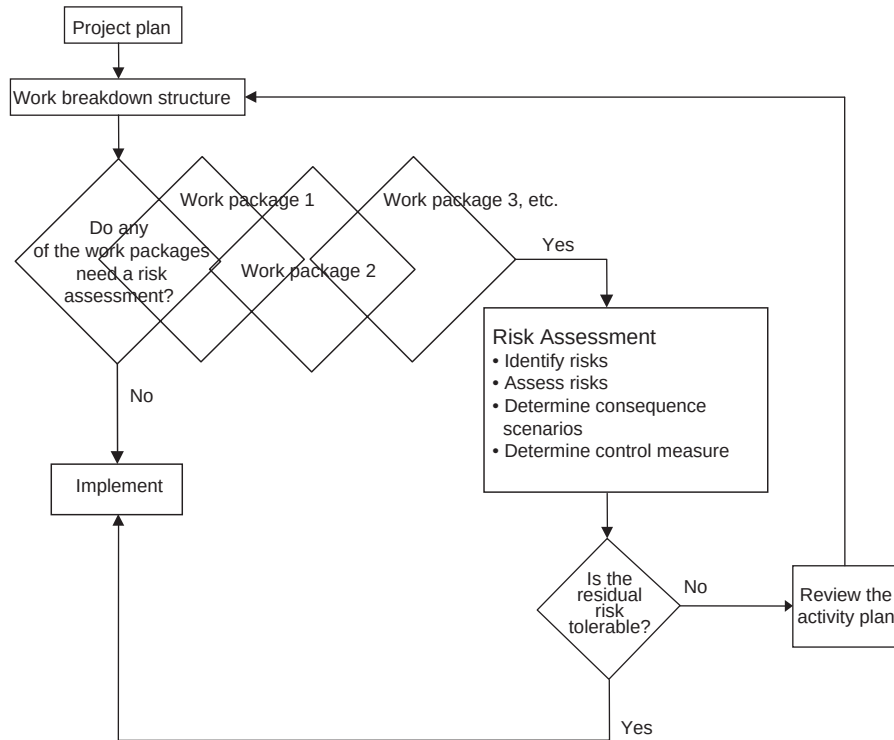


FIGURE 7 Construction risk management flowchart.

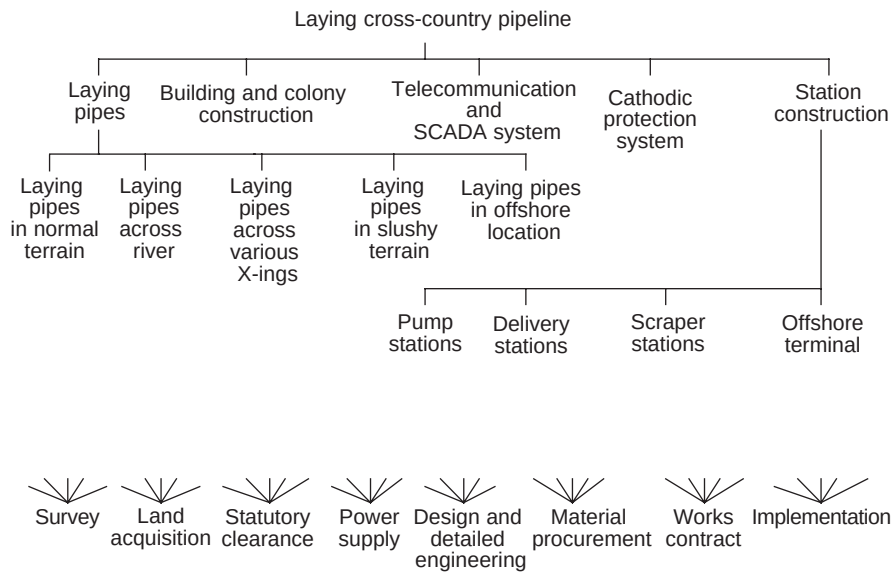


FIGURE 8 Work breakdown structure.

risk structure. The following are the risk factors and subfactors of a typical pipeline construction project:

1. Technical risk
 - a. Scope change
 - b. Technology selection
2. Acts of God
 - a. Natural calamities normal
- c. Implementation methodology selection
- d. Equipment risk
- e. Materials risk
- f. Engineering and design change

- b. Natural calamities abnormal
- 3. Financial, economical, and political risk
 - a. Inflation risk
 - b. Fund risk
 - c. Changes of local law
 - d. Changes in government policy
 - e. Improper estimation
- 4. Organizational risk
 - a. Capability risk of owner's project group
 - b. Contractor's failure
 - c. Vendor's failure
 - d. Consultant's failure
- 5. Statutory clearance risk
 - a. Environmental clearance
 - b. Land acquisition
 - c. Clearance from chief controller of explosive (CCE)
 - d. Other clearance from government authorities

A model for determining the risk level of a pipeline project is shown in Fig. 9.

Risk analysis consists of determining the likelihood of occurrence and severity of risk factors. Experienced pipeline executives take part in analysis, and an experienced-based method is often adopted to determine both the probability and the severity of each risk factor. The risk analysis results clearly derive responses, which are to be implemented to achieve project success.

5.2 Strategic Sourcing

Construction procurement is another issue that requires attention to achieve success in pipeline construction, as various specialized services and unique items are required from wide varieties of suppliers, contractors, and consultants across the world.

Table V demonstrates how strategic sourcing provides a paradigm shift from the typical L1 (offering work to lowest quoted bidder) environment.

In the pipeline industry, the following typical opportunities exist:

- Procurement of pipes, valves, station equipment, and construction services across a geographically diverse pipeline network not fully centralized
- Traditional approach to purchasing geared toward issuing quotes, buying from the lowest bidder, and negotiating contracts frequently
- Insufficient coordination between purchasing and engineering to determine the most cost-optimal solution
- Lack of long-term relationships with most suppliers
- Suboptimal utilization of procurement scale
- Globalization of pipeline industry

Strategic sourcing enables companies to enjoy significant cost benefits without painful headcount reductions. It views the project procurement function

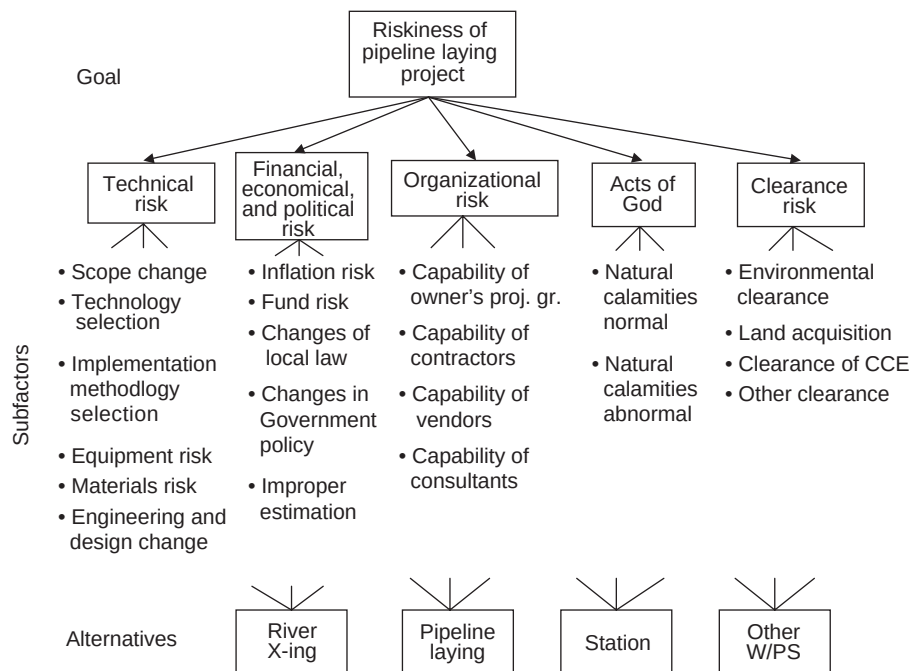


FIGURE 9 Model for determining riskiness of pipeline project.

from transactional to strategic. In the transactional perspective, procurement is viewed as being made up of people who place and transact orders, whereas in the strategic perspective, procurement is viewed as a knowledge-based buying organization. It focuses on total cost of ownership as opposed to price paid per commodity. It develops an optimal mix of supply relationships to provide competitive advantage.

A 15 to 30% cost reduction in procurement is feasible using strategic sourcing concepts. Reduction of the vendor base and consolidating volumes can yield a 5 to 15% cost improvement, and reduction of acquisition costs through structural change in the vendor relationship yields a 10 to 20% cost savings. Additionally, there are the following benefits:

- Enhancing product/service quality, reliability, and deliverability
- Leveraging supplier's technology and innovation capabilities
- Building internal sourcing capabilities that can be transferred throughout the organization.

The formation of winning alliances among client, consultants, contractors, and suppliers requires a total mindset and understanding of the changing role. Table VI demonstrates the transformation of the consultant's/contractor's/supplier's role for forming winning alliance. The winning alliances are made of partners with proven expertise, common experience, compatible business culture, and objectives.

Both global and local components are equally important for the client to bring international resources to be transformed at a local scene, and the output of the process has to serve local business needs or local market demand. Good global market coverage enables a client to establish links with potential consultants, contractors, and suppliers to procure globally. The development in technology and communication enables the client to exploit these opportunities and build global networks to gain a competitive edge.

Well-established local links provide the client with an understanding of the socioeconomic environment within which it has to operate. Alliances with local

TABLE V

Comparison of Strategic Sourcing Environment with Typical "L1" Environment

Typical "L1" environment	Strategic sourcing environment
Clearly defined requirements and specifications	Develop deep understanding of requirements — value engineer to identify optimal value/cost trade-offs
Open bid process with little/no ability for suppliers to offer alternative designs/specifications – purchase price focus	Develop deep understanding of supply industry product and service offerings, and performance drives of key suppliers
Elaborate internal controls and rigid processes to prevent changes/modifications to stated specifications	Develop robust total cost of ownership model, make buying decisions based on total cost over relevant time horizon
Typically, short-term contracts (1 year or less) that invite frequent competition and limit long-term supplier investment	True apples-to-apples comparison

TABLE VI

Transformation of the Consultant's/Contractor's/Supplier's Role

	Delivering value to a client	Coproducing value with a client	Delivering value to the consumers
Scope	Product delivery	Product and service delivery	Business function delivery
Horizon	Construction project	Product life cycle	Business project life-cycle consortia
Risk allocation	Client	Shared	Consortia
Competence	Understanding client's need	Understanding client's business	Understanding socioeconomic need
Dominant culture	Bidding culture	Cooperation	Partnering
Performance criteria	Design, construction, and product performance	Operational performance	Business and social performance
Alliances	Operational/tactical	Tactical/strategic	Strategic
Benefit of partnering	Cost savings	Improved performance	Competitive advantage
Extent of partnering	Within construction supply chain	Includes client/operator/lender	Includes stakeholders

Source. Mitrovic (1999).

consultants, contractors, and supplier organizations may also help market penetration by better addressing a customer's requirement with cost-effective design and delivery.

Close links with local and global consultants, contractors, and suppliers, preferably developed through long-term partnership relations, provide a client with a better chance to understand the business goals of other stakeholders within a dynamic business environment.

6. RISK-BASED INSPECTION AND MAINTENANCE

Traditionally, most pipeline operators ensure that during the design stage, safety provisions are created to provide a theoretical minimum failure rate for the life of the pipeline. Safety provisions are considered when selecting pipes and other fittings. To prevent corrosion, a pipeline is electrically isolated by providing high-resistance external coating materials. As a secondary protective measure, a low-voltage direct current is impressed in the pipe at a precalculated distance to transfer any corrosion that occurs due to breaks in the coating caused by a heap of buried iron junk, rails, and so on. This is called impressed current cathodic protection. The quality of the commodity that is being transported through the line is also ensured, and sometimes corrosion-preventing chemicals (corrosion inhibitors) are mixed with the commodity. To avoid deliberate damage of the pipeline in isolated locations, regular patrolling of the right of way from the air as well as on foot is carried out, and all third-party activities near the route are monitored.

Various techniques are routinely used to monitor the status of a pipeline. Any deterioration in the line may cause a leak or rupture. Modern methodologies can ensure the structural integrity of an operating pipeline without taking it out of service.

The existing inspection and maintenance practices commonly followed by most pipeline operators are formulated mainly on the basis of experience. However, operators are developing an organized maintenance policy based on data analysis and other in-house studies to replace rule-of-thumb-based policies. The primary reasons for this are stringent environmental protection laws, scarce resources, and excessive inspection costs. Existing policies are not sharply focused from the point of view of the greatest damage/defect risk to a pipeline. The basis for

selecting health monitoring and inspection techniques is not clear to many operators. In many cases, a survey is conducted over an entire pipeline or on a particular segment, when another segment needs it more. Avoidable expenditures are thus incurred.

A strong motivation exists, therefore, to derive a technique that will help pipeline operators select the right type of inspection/monitoring technique for segments that need it. A more clearly focused inspection and maintenance policy that has a low investment-to-benefit ratio should be formulated.

Figure 10 demonstrates an inspection and maintenance system of oil pipelines and its relationship with pipeline operations. The following steps demonstrate a risk-based inspection and maintenance method for oil pipelines.

- Step 1: Entire oil pipeline is classified into a few stretches.
- Step 2: All information related to the pipeline, including the terrain detail under study, is prepared and documented section wise.
- Step 3: The risk factors that can cause failures are identified. Generally, pipelines fail because of one of these reasons: corrosion (internal and external), free span, external interference, construction and materials defects, acts of God, or human and operational error.

One of the major causes of pipeline failure is corrosion, an electrochemical process that changes metal back to ore. Corrosion generally takes place when there is a difference of potential between two areas having a path for the flow of current. Due to this flow, one of the areas loses metal.

Internal corrosion is normally caused by chemical reaction between the pipeline material and the fluid or gas, such as CO_2 , H_2S , and O_2 . Corrosion can take the form of general or localized metal loss from pipe, and it may also give rise to cracking in pipeline material. The rate of corrosion depends on pipe materials, type of products being transported through pipelines, and corrosion inhibitor. The four major reasons for internal corrosion are as follows:

- Sweet corrosion is caused by the presence of carbon dioxide dissolved in the fluids, also called carbonic acid corrosion. This form of corrosion is typically slow, localized, and has pitting attack on pipeline metal.
- Sour corrosion is caused by hydrogen sulfide in the fluids. This form of corrosion causes rapid failure by cracking of the pipe metal.

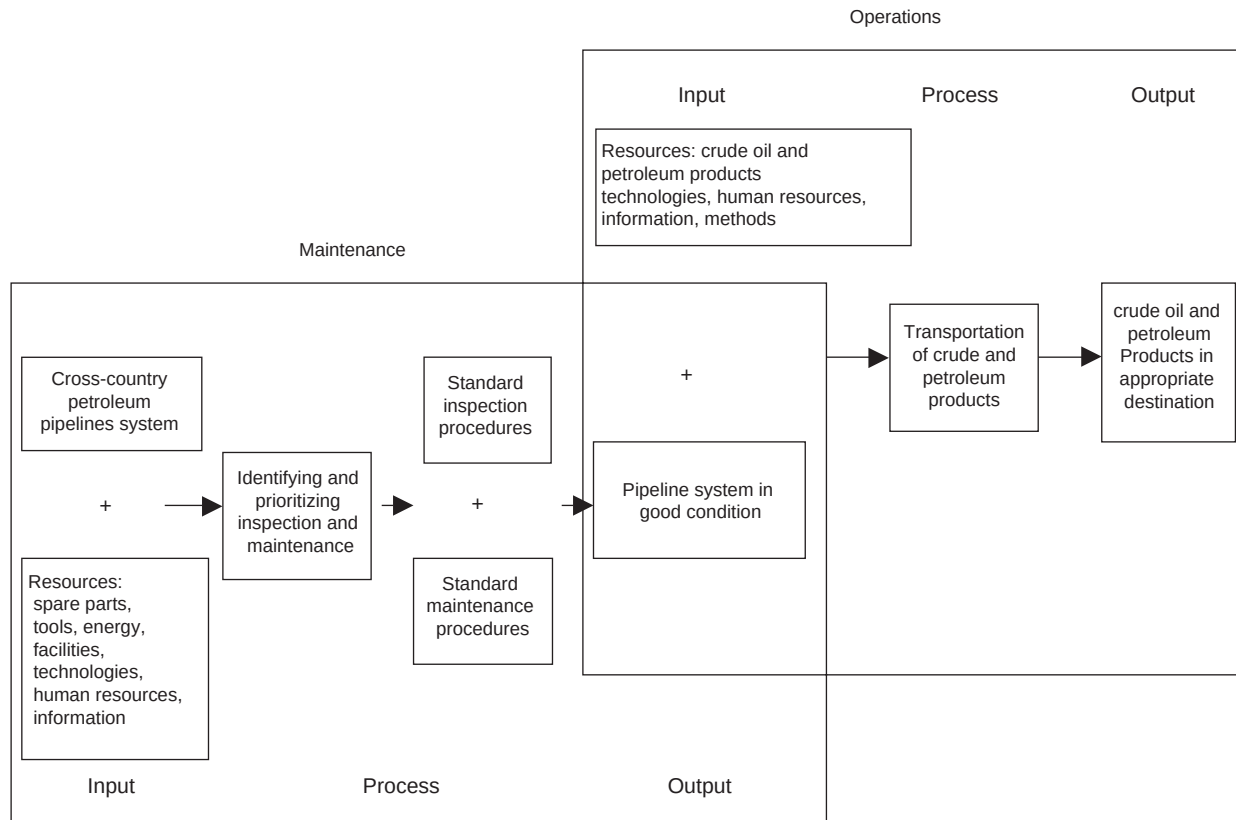


FIGURE 10 Relationship between operations and maintenance of cross-country petroleum pipeline system.

- Corrosion mainly results from the presence of oxygen, carbon dioxide, and hydrogen sulfide in the water, which comes from some oil fields.
- Microbiological corrosion results from the growth of bacteria in the pipeline.

Chemical reaction between pipe metal and seawater causes an external corrosion. The rate of external corrosion depends on the condition of the coating and cathodic protection. The external corrosion also includes external erosion, which is caused by solid substances in the seawater when they come in contact with pipelines.

A free span is another criteria that can cause pipeline failure. A free span is an area where the pipe does not have support because of subsidence. If the free span is long, pipe self-weight and earth load for onshore pipeline, and pipe self-weight and ocean currents might damage the pipeline.

External interference is another leading cause of pipeline failure. It can be malicious (sabotage or pilferage) or may be caused by other agencies sharing the same utility corridor. The latter is known as third-party activity. In both cases, a pipeline can be damaged severely. External interference with mal-

icious intent is more common in socioeconomically backward areas, while in regions with more industrial activity, third-party damage is common.

All activities, industrial or otherwise, are prone to natural calamities, but pipelines are especially vulnerable. A pipeline passes through all types of terrain, including geologically sensitive areas. Earthquakes, landslides, floods, and other natural disasters are common causes of pipeline failures.

Poor construction, combined with inadequate inspections and low-quality materials, also contributes to pipeline failures. Other reasons include human and operational error and equipment malfunctions. Computerized control systems considerably reduce the chance of failure from these factors.

Human and operational errors are another sources of pipeline failure. Inadequate instrumentation, a foolproof operating system, lack of standardized operating procedures, untrained operators, and so on are the common causes of pipeline failure due to human and operational errors.

Step 4: The next step is to form of a risk structure (Fig. 11) based on the identified risk factors.

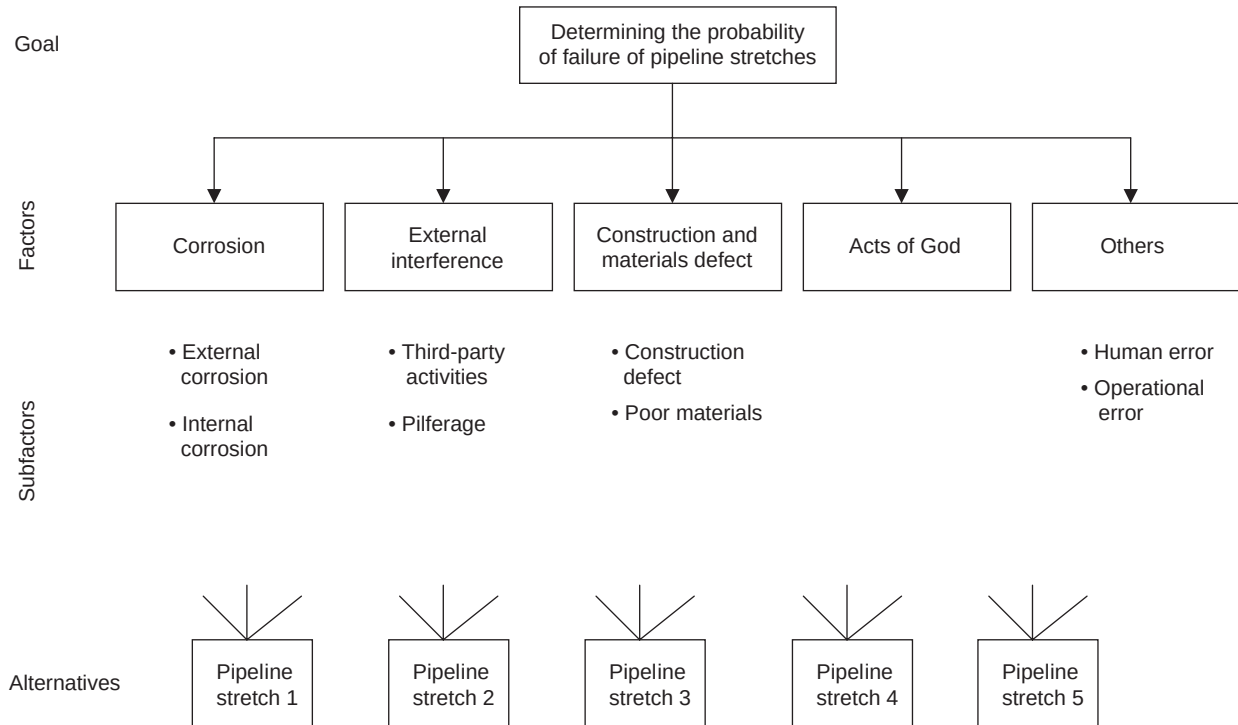


FIGURE 11 Risk structure to determine failure probability of various stretches.

Step 5: In this step, the likelihood of pipeline failure due to each factor is determined. The alternative pipeline stretches are compared with respect to each risk factor to determine the likelihood of failure for each pipeline stretch.

Step 6: In this step, specific inspection/maintenance requirements are determined for specific segments of pipelines from the likelihood of failure data; this is done to mitigate risk.

Step 7: The last step demonstrates the cost-benefit analysis of a suggested inspection and maintenance strategy along with a cost-effective insurance plan for the pipeline. The basis of the insurance premium depends on the likelihood of its failure, the expected failure cost in a given period, the risk perception of the management/organization, and the inspection/maintenance programs undertaken.

Advantages of this risk-based inspection and maintenance model are as follows:

- Reducing subjectivity in the decision making process when selecting an inspection technique
- Identifying the right pipeline or segment for inspection and maintenance
- Formulating an inspection and maintenance policy

- Deriving the budget allocation for inspection and maintenance
- Providing guidance to deploy the right mix of labor in inspection and maintenance
- Enhancing emergency preparations
- Assessing risk and fixing an insurance premium
- Forming a basis for demonstrating the risk level to governments and other regulatory agencies

If a pipeline system is designed, constructed, and operated ideally, many inspection and maintenance problems will not crop up. The overall performance of pipeline operations and maintenance would be improved through the following actions:

- Pipeline routes are to be decided on the basis of the life cycle costing approach, not on the basis of the shortest route.
- The maintenance characteristics of the pipeline are to be considered along with pressure and temperature parameters while designing pipe thickness for various stretches of pipeline.
- Pipeline coating shall be selected on the basis of the terrain condition, the environmental policy of the organization, the cost of coating materials, the construction methodology, inspection, and the maintenance philosophy.

- The construction methodology of a pipeline in a critical section is to be formulated during the feasibility stage of the project, and this shall commensurate with the design and operational philosophy of the pipeline as a whole. Factors such as availability of technology, availability of consultants, contractors and vendors, experience of owner project group, government regulations, and environmental requirements throughout the life of pipeline are to be rationally considered when selecting the best construction methodology.
- Networking in pipeline operations demands a foolproof mechanism in the system for minimizing operational and human errors. Improved instrumentation shall be designed that is commensurate with the design philosophy of the entire pipeline system.
- All pipeline operators are to be suitably trained in pipeline operation before taking charge of specific pipelines. Pipeline simulation training may be one of these areas. Criticality of pipelines and the expertise of personnel are to be considered when finding staff for pipeline operations.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Releases to the Environment: Natural Fate and Remediation Options • Crude Oil Spills, Environmental Impact of • Occupational Health Risks in Crude Oil and Natural Gas Extraction • Oil and Natural Gas Drilling • Oil and Natural Gas: Offshore Operations • Oil Industry, History of • Oil Recovery • Oil Refining and Products • Petroleum System: Nature's Distribution System for Oil and Gas • Public Reaction to Offshore Oil

Further Reading

- Awakul, P., and Ogunlana, S. O. (2002). The effect of attitudinal differences on interface conflicts in large scale construction projects: A case study. *Construction Manage. Econ.* 20, 365–377.
- Calvin, H., and Dey, P. K. (2002). Social impact assessment of a sewerage project in Barbados. *Impact Assess. Proj. Appraisal* 20(2).
- Dey, P. K., and Gupta, S. S. (2000). Decision support system yields better pipeline route. *Oil Gas J., PennWell*, 98(22), 68–73.
- Dey, P. K., and Gupta, S. S. (2001). Risk-based model aids selection of pipeline inspection, maintenance strategies. *Oil Gas J., PennWell*, 99(28), 39–67.
- Dey, P. K. (2003). Analytic hierarchy process analyses risk of operating cross-country petroleum pipelines in India. *Natural Hazard Rev., ASCE*, 4(4), 213–221.
- Dey, P. K. (2002). An integrated assessment model for cross-country pipelines. *Environmental Impact Assess. Rev.* 22(6), 703–721.
- Dey, P. K. (2001). Decision support system for project risk management: A case study. *Manage. Decision*, MCB University Press 39(8), 634–649.
- Dey, P. K., and Gupta, S. S. (2001). Feasibility analysis of cross-country pipeline projects: A quantitative approach. *Project Manage. J.* 32(4), 50–58.
- Dey, P. K., Tabucanon, M. T., and Ogunlana, S. O. (1994). Planning for project control through risk analysis: A case of petroleum pipeline laying. *Int. J. Project Manage.* 12(1), 23–33.
- Dey, P. K., Tabucanon, M. T., and Ogunlana, S. O. (1996). Petroleum pipeline construction planning: A conceptual framework. *Int. J. Project Manage.* 14(4), 231–240.
- Dey, P. K. (2001). Integrated approach to project feasibility analysis. *Impact Assess. Proj. Appraisal* 19(3), 235–245.
- Dey, P. K. (February, 2004). Decision support system for inspection and maintenance of cross-country petroleum pipeline. *IEEE Transactions Engineering Manage.*, in press.
- Lyons, D. (February, 2002). Western European cross-country oil pipelines 30-year performance statistics, Report no. 1/02, CONCAWE, Brussels.
- Mitrovic, D. (1999). Winning alliances for large scale construction projects on the world market. In “Profitable Partnering in Construction Procurement” (S. O. Ogunlana, Ed.), pp. 189–199. E & FN Spon. London and New York.
- Montemurro, D., and Barnett, S. (1998). GIS-based process helps Trans Canada select best route for expansion line. *Oil Gas J.*, 22.
- Ogunlana, S. O., Yisa, S., and Yotsinsak, T. (2001). An assessment of people's satisfaction with the hearing on the Yadana gas pipeline project. *J. Environ. Monitoring Assess.* 72, 207–225.
- Pipeline and Gas Industry (2001). “Worldwide Pipeline Construction Report,” Vol. 84, no. 11, November–December.
- Ramanathan, R., and Geetha, S. (1998). Socio-economic impact assessment of industrial projects in India. *Impact Assess. Proj. Appraisal* 16(1), 27–32.
- U.S. Department of Transportation (1995). *Pipeline Safety Regulation*, October 1.

Oil Price Volatility

M. J. HWANG

West Virginia University
Morgantown, West Virginia, United States

C. W. YANG

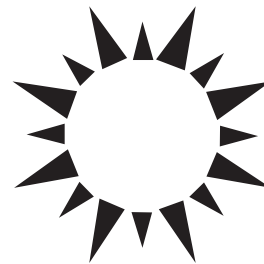
Clarion University of Pennsylvania
Clarion, Pennsylvania, United States

B. N. HUANG

Providence University
Shalu, Taiwan

H. OHTA

Aoyama Gakuin University
Tokyo, Japan



-
1. Introduction
 2. Microeconomic Foundation of the Price Elasticity of Demand and Its Estimation
 3. Volatility of an Oil Price Change and Macroeconomic Activity
 4. Conclusion

Glossary

crude oil price Crude oil is sold through many contract arrangements, in spot transactions, and on futures markets. In 1971, the power to control crude oil prices shifted from the United States to OPEC when the Texas Railroad Commission set prorating at 100% for the first time.

demand elasticity The relative responsiveness of quantity demanded to changes in price. The price elasticity of demand is important in affecting the pricing behavior of OPEC.

economic impact The study of oil price shocks and their effects on economic activities.

oil price volatility The volatility of crude oil prices is influenced more by both demand structure and shifting demand and supply conditions and less by the cost of producing crude oil. Oil price volatility creates uncertainty and therefore an unstable economy.

Organization of Petroleum Exporting Countries (OPEC) Formed in 1960 with five founding members: Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela. By the end of 1971, six other countries had joined the group: Qatar, Indonesia, Libya, United Arab Emirates, Algeria, and

Nigeria. OPEC effectively controlled oil prices independent from the United States during the Arab oil embargo.

The world oil price has been extremely volatile in the past three decades. The cartel pricing is largely affected by the aggregate demand the cartel faces and related elasticity. The stable and unstable cases of price elasticity of demand are investigated in this article to shed light on the seemingly mysterious Organization of Petroleum Exporting Countries pricing behavior. We estimate the elasticity of oil demands in the U.S. market (the world's largest energy consumer) and use this information to probe and predict movements in the market price of crude oil. The volatility of crude oil prices creates uncertainty and therefore an unstable economy. Employing recent data, our empirical results indicate that a higher oil price seems to have a greater impact on the stock market than on the output market.

1. INTRODUCTION

The world oil price has been extremely volatile in the past three decades. It was as low as \$2.17 per barrel in 1971 but spiked to \$34 in 1981. It soared to approximately \$40 per barrel toward the end of February 2003. The Organization of Petroleum Exporting Countries (OPEC) price increases of the

1970s drove Western industrialized economies into recession. During the 1950s and 1960s, many competitive independent producers characterized the crude oil industry. The demand facing the individual producer is relatively elastic, even though the demand curve for the entire industry is rather inelastic. The price of crude oil was close to the marginal cost during that period of time. OPEC emerged as an effective cartel in approximately 1971 when it successfully raised the pattern of world prices with the Tehran and Tripoli agreements. With the Arab-Israeli war of 1973, its consequent oil embargo, and the nationalization of oil production in member countries, the structure of the cartel provided the means to raise prices substantially from \$4.10 per barrel in 1973 to \$11.11 in 1974 to reap a monopoly profit.

With the beginning of political problems in Iran, another large increase in oil prices occurred in 1978 when Iranian production declined from a peak of 6 million barrels per day to 0.5 million barrels. Even though half of the reduction in Iranian oil was offset by expanded production by other OPEC producers, the effect was immediate and substantial, causing the price to increase to \$13.49 in 1978, because the elasticity of demand was rather inelastic in the short term. Because of the war between Iraq and Iran, Iraq's crude oil production declined by 2.7 million barrels per day and Iran's production declined by 600,000 barrels per day. As a result, the price of crude oil more than doubled from \$13.49 per barrel in 1978 to \$34 in 1981.

Higher oil prices create inflationary pressure and slow worldwide economic activities. The recession of the early 1980s reduced demand for oil. From 1982 to 1985, OPEC tried to stabilize the world oil price with low production quotas. These attempts were unsuccessful because various members of OPEC produced beyond their quotas, causing crude oil prices to decrease below \$10 per barrel by mid-1986. In particular, Saudi Arabia's increases in oil production frequently depressed the oil price. The price of crude oil remained weak until the start of the Gulf War in 1990, when it eclipsed \$40 per barrel. Because of the uncertainty associated with the invasion of Kuwait by Iraq and the ensuing Gulf War, the oil price spiked again to \$34 in late 1990. After the war and the recession of the early 1990s, the crude oil price began a steady decline. The economy, however, started to turn around in 1994. With a strong economy in the United States and a booming economy in Asia, increased demand led to a steady price recovery well into 1997. The financial crisis and subsequent economic setbacks in Asia

started in 1997 and oil prices plummeted to approximately \$10 in late 1998 and early 1999. With the recovery in Asia and the decrease in oil quotas by OPEC, the price has recovered to approximately \$30 a barrel in recent years. In 2002, for the first time, Russia's oil production surpassed that of Saudi Arabia, signaling a more complicated pricing scheme for oil.

The literature on the volatility of crude oil prices relates oil price changes either to the effect of the price elasticity of demand or to the instability of the market structures. It is apparent that the stable price is established through the equilibrium of total world demand and supply, including OPEC and non-OPEC production. In the short term, the price change is largely impacted by the immediate substantial increase or decrease in oil production from OPEC members and political events. The cartel's pricing policy is largely affected by the aggregate demand it faces and related elasticity.

The volatility of crude oil prices in turn creates uncertainty and therefore an unstable economy. In his pioneering work, Hamilton indicated that oil price increases have partially accounted for every U.S. depression since World War II. Many researchers, using different estimation procedures and data, have tested the relationships between the oil price increases and many different macroeconomic variables. Using a multiequation statistical approach incorporating the U.S. interest rate, oil price, industrial production, and real stock returns with daily data from 1947 to 1996, Sadorsky found that oil price volatility does have an impact on real stock returns. Recently, emphasis has shifted to the asymmetry of the impact of oil price shocks on economic activities and on stock markets. Mork was the first to provide the asymmetry of oil price shocks or its volatility on economic activities. Using data from industrial nations, Mork and Olson verified that there is a negative relationship between an oil price increase and national output, whereas no statistical significance can be attributed to them when the oil price declines. Lee *et al.* estimated oil price shocks from a generalized econometric model and investigated the impacts of positive and negative oil price shocks on economic activities. They came to the same conclusion that positive shocks have a statistically significant impact on economic activities, whereas negative shocks have no significant impact.

This article examines the volatility of crude oil prices by first determining the potential maximal price that OPEC can extract based on the microeconomic foundation of the elasticity theory proposed by

Greenhut *et al.* The market structure of OPEC, the stable and unstable demand structure, and related elasticity of demand are discussed. In particular, the theory of unstable price elasticity of demand helps explain some of the pricing behavior of OPEC. The price elasticity of demand is then estimated to shed light on the volatility of oil prices. This article further investigates the significance of changing oil prices on the economy by examining the relationship between oil price volatility and industrial production and/or the stock market.

2. MICROECONOMIC FOUNDATION OF THE PRICE ELASTICITY OF DEMAND AND ITS ESTIMATION

Consider a cartel (e.g., OPEC) whose objective is to maximize joint profit:

$$\pi = PQ - \sum_{i=1}^n TC_i(q_i) = PQ - TC(Q), \quad (1)$$

where P is a uniform cartel price with $P' < 0$, $Q = \sum q_i$ is the sum of outputs of n cartel members (q_i), and TC_i is the total cost function of cartel member i . Note that OPEC behaves much like a monopoly or an effective cartel despite occasional squabbles over output quotas. Under recent OPEC arrangements, if the price becomes too low, OPEC would reduce output by half a million barrels per day ($\Delta Q = 500,000$) at one time. A 500,000-barrel decrease (increase) in Q is approximately 2% of the total cartel output, and each member must accept a 2% decrease (increase) in its quota in the case of a tightly controlled cartel. Within this framework, it is the total cartel output Q , instead of q_i , that plays a crucial role in the pricing decision.

The first-order condition requires

$$\pi' = QP' + P - MC = 0, \quad (2)$$

where $\pi' = d\pi/dQ$ is the marginal profit.

Equation (2) states that the marginal revenue (MR) equals the common cartel marginal cost (MC, or horizontal summation of marginal cost curves for all cartel members) in equilibrium. Note that if some MC_i 's exceed the going market price, these members have no role to play in the pricing decision. However, this is not likely the case for OPEC because production costs fall far short of the market price.

Substituting $MR = P(1 - 1/e)$ into Eq. (2) yields

$$P(1 - 1/e) = MC$$

or

$$P(1 - 1/e) = P/k, \quad (3)$$

where $k = P/MC$. Solving for the profit maximizing price elasticity leads to

$$e^* = k/(k - 1). \quad (4)$$

The second-order condition plays a critical role in describing the switching demand conditions. Here, we classify two major demand cases via expanding the elasticity theory formulated by Greenhut *et al.* in terms of the second-order condition.

By using the chain rule, we have

$$\begin{aligned} \frac{d^2\pi}{dQ^2} &= \frac{d\pi'}{dQ} \\ &= \frac{dP}{dQ} \left(\frac{\eta - (1 - e)}{e} - \frac{-e}{\beta k} \right) < 0, \end{aligned} \quad (5)$$

where $\beta = (dQ/dMC)(MC/Q)$ is the output elasticity on the cartel marginal cost curve, and the price elasticity of elasticity $\eta = (de/dP)(P/e)$ measures the percentage change in price elasticity from the percentage price change. Alternatively, by taking the derivative of the marginal profit π' with respect to price P , the second-order condition in Eq. (5) appears as

$$\frac{d\pi'}{dP} = \frac{\eta - (1 - e)}{e} + \frac{e}{\beta k} > 0. \quad (6)$$

For a given output elasticity β , if the marginal cost is an insignificant portion of price (i.e., a large k) or the marginal cost is constant ($\beta \rightarrow \infty$, as is true in a large-scale crude oil production with a sizable fixed cost), the second term on the right side of Eq. (6) plays a trivial role and can be ignored. It follows that $d\pi'/dP > 0$ for $d\pi'/dQ < 0$, and it is sufficient that relation A holds:

$$\frac{d\pi'}{dP} > 0 \quad \text{if } \eta > 1 - e, \eta = 1, \text{ or } \eta > 1. \quad (A)$$

The residual term $e/\beta k$ would reinforce the relation, but its magnitude may be insignificant, with k and/or β being large enough.

On the other hand, the unstable relation B would follow if the second-order condition is not satisfied:

$$\frac{d\pi'}{dP} < 0 \quad \text{if } \eta < 1 - e. \quad (B)$$

Relation A indicates that no matter what the value of e is, the marginal profit π' will increase with price less than, proportionately equal to, or more rapidly than price if and only if $1 - e < \eta < 1$, $\eta = 1$, or $\eta > 1$, respectively. If prior market conditions establish a price at a level at which the elasticity of demand is less than unity and MR is below MC (i.e., the marginal profit being negative), an OPEC price hike

could be favorable. Under relation A, an increasingly elastic demand at higher prices (e.g., $\eta > 0$) would generally create conditions in which alternative fuels become more competitive following the price hike. The desired result under profit-maximizing principles is to have the marginal profit increase to 0 and elasticity close to $e^* = k/(k-1)$ depending on the ratio k ($k = P/MC$).

Since the marginal cost for OPEC countries is very low relative to price, and k is accordingly large, e^* should be very close to 1. Under the condition $\eta = 1$ or $\eta > 1$, the increase in elasticity at a higher price is proportionately equal to or greater than the increase in price, and we reach the point at which negative marginal profit increases to zero more rapidly than under $1 - e < \eta < 1$. On the other hand, if the price elasticity of demand is greater than unity and MR is greater than MC, it becomes advantageous for OPEC to lower its price. In general, the elasticity of demand would converge toward $e^* = k/(k-1)$ and the marginal profit to zero. That is, the market system would adjust itself under stable relation A.

Relation B indicates that if $\eta - (1 - e) < 0$, the marginal profit will decline (increase) as the price of crude oil is raised (lowered). This relation is an unusual and basically unstable case. Relation B also implies that the second-order condition in Eqs. (5) or (6) is not satisfied. The elasticity will therefore diverge from $e^* = k/(k-1)$. A cartel can benefit from an increase in price if demand is inelastic. Under the unstable relation B, the elasticity of demand decreases markedly as the price is raised. Inasmuch as marginal profit decreases and is negative, output will be curtailed and the price will be raised further. This can occur only in unusual circumstances (e.g., the energy crisis of 1973–1974). It is not a mathematical curiosity to have relation B—that is, an unusually strong convex demand curve. This could be a direct result of a structural break in the world oil market from a preembargo competitive oil market to an after-embargo cartel market. For example, the market demand changed from an elastic competitive market demand (before the 1973 embargo) to an inelastic cartel demand (after the embargo, from point F to C in Fig. 1).

The regime shift occurred when OPEC emerged as an effective cartel in 1971 after the Tehran Agreement. The ensuing Arab–Israeli War of 1973, subsequent oil embargo, and the nationalization of oil production in member countries enabled the cartel to raise the price of crude oil from \$2 per barrel in 1971 to \$34 in 1981. The price skyrocketed again to \$37 in 2000 and \$39.99 in late February

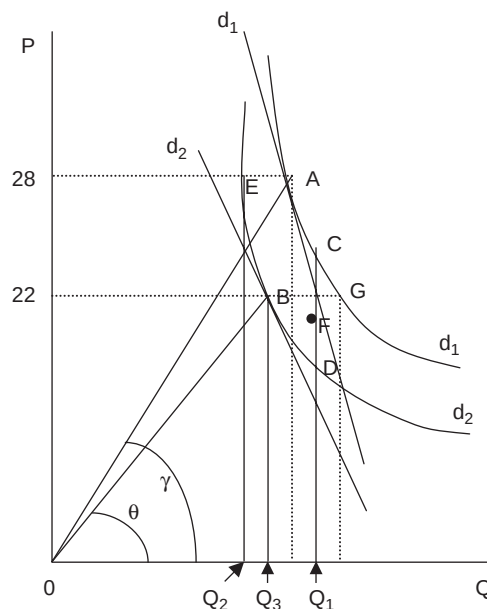


FIGURE 1 Switch in demand structure for oil. F, preembargo point; C, after-embargo point; A and E, upper limit of the price band; B and G, lower limit of the price band; A to B, market positions due to recession; C to D, market positions due to recession without a production cut.

2003. The market changed from a competitive one to one in which OPEC was able to exercise a considerable degree of monopoly power. Demand, in turn, changed from a relatively elastic individual market demand curve to a relatively inelastic industry cartel demand curve. The OPEC countries could therefore raise the oil price even further under this structural break in demand. This was the direct result of the unstable relation B.

In order to estimate the demand relations, we use data from “The Annual Energy Review” and “The Economic Report of the President.” The price of coal (PC) is measured in cents per million Btu (cost, insurance, and freight price to electric utility power plants), as are prices of petroleum (P) and natural gas (PN). The quantity of petroleum consumption (Q) is in quadrillion Btu (10^{15}) and real gross domestic product is used to reflect income (Y). The sample period extends from 1949 to 1998. As with other economic variables, the unit root (unstable variables or not stationary) property needs to be examined before the estimations. Note that we fail to reject the null hypothesis of a unit root for all the variables even after the logarithmic transformation. That is, these variables are not stationary for the ordinary least squares technique: A technique is needed that relates a set of independent variables and dependent variables in terms of a linear equation. Since all the

variables in our model are not stationary, we apply the two-step technique by Engle-Granger to explore the cointegration relation (comovements of unstable variables). The cointegration model, used in the presence of unstable variables, suggests the use of the error correction model (ECM) in estimating the demand relation. That is, demand for petroleum can be appropriately formulated as

$$dQ_t = \text{ECM}_{t-1} + dP_t + dPC_t + dPN_t + dY_t + v_t, \quad (7)$$

where the prefix d denotes first difference of the variables (in logarithm); Q_t denotes consumption of crude oil; ECM_{t-1} is representing past disequilibrium; P_t denotes the price of crude oil; Y_t is income; PC_t is the price of coal; and PN_t represents the price of natural gas. Despite its complexity, the model is more general because it allows past disequilibrium or perturbations to be included in explaining oil consumption. The estimated results for the entire sample period are shown as follows:

$$\begin{aligned} dQ_t = & -0.1645 \text{ECM}_{t-1} - 0.0502 dP_t \\ & \quad (-1.56) \quad (-3.22) \\ & + 0.0275 dPN_t - 0.0210 dPC_t \\ & \quad (1.02) \quad (-0.58) \\ & + 0.7664 dY_t + 0.7462 \text{AR}(1) \\ & \quad (6.89) \quad (5.33) \\ & - 0.0054 + e_{1t} \\ & \quad (-0.45) \end{aligned} \quad (8)$$

$$\bar{R}^2 = 0.71,$$

where $\bar{R}^2$ is the adjusted R square, which reflects general fitness of the model; t statistics in parentheses indicate the significance of estimated coefficients. Within the framework of the partial adjustment model (or model with lagged dependent variable) with ECM, the price elasticity tends to converge approximately to $-0.198 = [-0.0502/(1-0.7426)]$ when estimated in the level of first difference. Similarly, we report the estimation results for the two-subsample periods (1949–1975 and 1976–1998):

$$\begin{aligned} dQ_t = & -0.1182 \text{ECM}_{t-1} - 0.1778 dP_t \\ & \quad (-2.72) \quad (-2.52) \\ & - 0.060 dPN_t + 0.1150 dPC_t \\ & \quad (-1.22) \quad (1.92) \\ & + 0.6097 dY_t + 0.0230 + e_{2t} \\ & \quad (4.26) \quad (2.89) \end{aligned} \quad (9)$$

$$\bar{R}^2 = 0.70$$

and

$$\begin{aligned} dQ_t = & -0.1643 \text{ECM}_{t-1} - 0.0625 dP_t \\ & \quad (-2.83) \quad (-3.06) \\ & + 0.0673 dPN_t + 0.0403 dPC_t \\ & \quad (2.23) \quad (0.26) \\ & + 1.3127 dY_t - 0.0366 + e_{3t} \\ & \quad (6.61) \quad (-5.11) \end{aligned} \quad (10)$$

$$\bar{R}^2 = 0.76,$$

The price elasticity of demand for petroleum within the entire sample period tends to approach -0.198 $[-0.0502/(1-0.7462)]$ in the long term, as indicated by Eq. (8). Thus, there appears to be enough room for further price hikes because the estimated price elasticity is not close to $k/(k-1)$ or approximately -1.05 in the empirical estimation as predicted by our theory.

The short-term price elasticities before and after the structural break can thus be estimated as -0.1778 and -0.0625 , respectively. The empirical results indicate that (i) the elasticity after the energy crisis decreased in absolute value from 0.1778 to 0.0625 (a 65% decrease), clearly an unstable case of $\eta < 0$ (i.e., the price elasticity decreases as price increases), and (ii) there seems to be considerable room for a price hike because the long-term elasticity from Eq. (8) of -0.198 falls far short of $k/(k-1)$.

Neither the long-term price elasticity for the entire sample period (-0.198) nor the short-term price elasticity after the structural break (-0.0625) are close to the theoretical limit: $k/(k-1) \approx 1.05$ as is expected by a profit-maximizing cartel. The discrepancy may be explained by the significant income elasticity of 1.3127 in Eq. (10). Since the business cycle is both inevitable and unpredictable, a recession could certainly shift the demand curve to the left (Fig. 1). Continuous and gradual price hikes without disruptions in demand would render price elasticity toward $k/(k-1)$ in the long term. A recession would generally depress both price and quantity, as shown in Fig. 1 from point A to B. Given that the price elasticity can be geometrically measured as $e_A = \tan \gamma / \text{slope}(d_1)$ at point A, the size of the price elasticity does not appear to have changed appreciably. That is, $e_A = \tan \gamma / \text{slope}(d_1)$ and $e_B = \tan \theta / \text{slope}(d_2)$ are similar because both $\tan \theta$ and the slope of the line d_2 have decreased. Unless the U.S. economy is recession-proof in the long term, it does not seem possible that the long-term price elasticity would approach $k/(k-1)$ as implied by the theory. The wild swings in oil price after 1973 speak to the fact that demand for and supply of crude oil are not independent of political events. The

significant change in the oil price was and can be detrimental to the suppliers with limited capacity and relatively higher marginal cost. In contrast, producers with lower marginal cost and greater capacity (e.g., Saudi Arabia) would benefit from greater output quota. To prevent such violent price changes, it is advantageous to OPEC to have a price band in which the price elasticity is not too low. However, can $k/(k-1) \approx 1.05$, a theoretical limit developed previously, ever be reached? There are at least three reasons in favor of this argument. First, switching to alternative fuels becomes more plausible at high oil prices and thereby tends to push long-term price elasticity close to $k/(k-1)$. Second, noncartel output (e.g., 5–7 million barrels per day from Russia) in the long term can present a credible threat when the price is high enough. This renders OPEC's residual demand curve flatter for every high price. Third, significant *ad valorem* tariffs in G8 countries would effectively pivot the demand curve downward.

Substantial price fluctuations have been witnessed for the past three decades, especially since the 1973 energy crisis. Based on our estimate, the average price elasticity was very low, -0.1778 before 1975, but the change in market structure has given rise to the unstable demand case. It has been borne out with the short-term elasticity of -0.0625 after 1975. Notice that the negative and significant ECM coefficient (-0.1643) in Eq. (10) suggests a convergent trend toward equilibrium values after the structural break. In the absence of major war, it does not seem plausible that the price of oil will significantly exceed \$40 due to the income effect in the demand structure. The price elasticity, hovering at -0.0625 after the structural break, suggests that the oil price would more likely approach the upper limit of the price band (\$28) than the lower limit (\$22) if OPEC members strictly adhered to the production cut (e.g., from Q_1 to between Q_2 and Q_3 ; Fig. 1) in the presence of a recession. The result is borne out because oil prices lingered around \$28 at the beginning of May 2001, 1 month after the production cut. As the recession deepened after the September 11, 2001, terrorist attacks, the price of crude oil dropped to \$20 in November 2001 when the OPEC cartel was reluctant to further reduce production. Failing to do so would result in a drastic price reduction (from point C to point D in Fig. 1) as occurred before. On the other hand, the upward pressure on price could be too irresistible because -0.0625 is far below the theoretical limit of -1.05 , as inelastic demand promotes price increase.

3. VOLATILITY OF AN OIL PRICE CHANGE AND MACROECONOMIC ACTIVITY

Given that oil is of great importance in the production process, its impact on industrial output cannot be ignored. Furthermore, the growing magnitude of oil price changes reflects the major role of uncertainty in statistical modeling. For instance, Lee *et al.* found that uncertainty in oil price (conditional variance of an oil price change) could significantly impact economic growth. In particular, one unit of positive normal shock (a large price increase) gives rise to a decreased output, whereas one unit of negative shock does not necessarily lead to an increased output. Similarly, the impact of an oil price change on the financial market has received increased attention. Sadorsky was among the first to apply a formal statistical model to investigate such an impact. To capture the uncertainty inherent in the macroeconomic foundation, we need to first define the real oil price after adjusting for the consumer price index or roilp_t . To ensure stationarity property, let Δroilp_t represent the first difference of the logarithmic transformation on roilp_t . One of the common ways to measure the volatility of oil price is to apply an econometrical model called autoregressive and conditional heteroskedastic (ARCH) by Engle or its generalized version, the GARCH model, by Bollerslev. The GARCH model shown below is fairly general in describing the behavior of a volatile time series such as oil price in macroeconomics:

$$\begin{aligned}\Delta \text{roilp}_t &= a_0 + \sum_{i=1}^p \phi_i \Delta \text{roilp}_{t-i} + \varepsilon_t - \sum_{j=1}^k \theta_j \varepsilon_{t-j} \\ \varepsilon_t &= z_t \sqrt{h_t} \quad z_t \sim N(0, 1) \\ h_t &= \alpha_0 + \beta_1 h_{t-1} + \alpha_1 \varepsilon_{t-1}^2,\end{aligned}\quad (11)$$

where h_t is the conditional variance often used to represent volatility, and $z_t = \varepsilon_t / \sqrt{h_t}$ denotes a standardized disturbance term. Note that optimal lags of the GARCH (1,1) – ARMA(p, q) are selected based on the criterion that a series correlation (correlation via residuals) is not found in z_t and z_t^2 . Finally, z_t is normally distributed with zero mean and unit variance or $N(0,1)$. Even though h_t is frequently chosen to measure the volatility, the major problem is that variance masks the direction of changes to which shocks are administered. The asymmetric impact—only positive shocks (or price increases) slow economic activity and/or the stock market—begs the use of z_t instead of h_t as a proxy for

volatility. As such, we use z_t as the proxy in our research model.

Monthly data of industrial production (ip_t), real stock return (Δr_{stk}_t), consumer price index (cpi_t), exchange rate ($exch_t$), interest rate (r_t), and stock indices from July 1984 to March 2002 were obtained from the International Financial Statistics (IFS) data bank. Data for oil prices are from "Energy Prices and Taxes" published by the International Energy Agency. Note that we include exchange rates because oil is exported to many industrialized economies. In addition, the interest rate is incorporated to account for monetary policy. Major industrialized countries included in the study are the United States, Canada, Italy, Japan, Germany, and the United Kingdom.

As in many other time series analyses, we need to examine the existence of a unit root (unstable) on the variables of lip_t , lr_{stk}_t , lr_t , and $lexch_t$. The prefix I indicates the logarithmic transformation of the original variables, used to damp unstable variables. If these variables are $I(0)$ or integrated of order zero indicating stationarity or stability of the variable, we could add z_t to form a five-variable VAR model. However, if these variables are $I(1)$ or integrated of order 1 (not stationary), we need to first examine the potential existence of any cointegration relation (or comovements of unstable variables). A five-variable vector error correction model can be formulated in the presence of such a cointegration relation. Otherwise, the four-variable (first difference) model along with z_t would suffice to analyze the direction of causality among the variables.

The techniques for examining variable stationarity by Phillips and Perron and by Kwiatkowski *et al.* are applied to the four variables for the six countries. The result is unanimous in that all the variables are of $I(1)$ or not stationary. Consequently, we examine the potential existence of cointegration relations via the trace method developed by Johansen. An

examination of the trace statistics suggests that no cointegration relation exists for the five countries except Japan, and two sets of cointegration relations exist for Japan. That is, we can include the past disturbance term ECM_{t-1} in the model to explain industrial production for Japan as follows:

$$\begin{aligned} \Delta lip_t = & a_4 + r_4 ECM_{t-1} + \sum_{i=1}^k b_{4i} \Delta lr_{t-i} \\ & + \sum_{i=1}^k c_{4i} \Delta lexch_{t-i} + \sum_{i=1}^k d_{4i} z_{t-i} \\ & + \sum_{i=1}^k f_{4i} \Delta lip_{t-i} + \sum_{i=1}^k g_{4i} \Delta lr_{stk}_{t-1} + e_{4t}, \end{aligned} \quad (12)$$

where ECM_{t-1} is the error correction term representing the long-term cointegration. On the other hand, the five-variable vector autoregression model is applied to the other five countries as follows (only the stock return equation is shown):

$$\begin{aligned} \Delta lr_{stk}_t = & a_5 + \sum_{i=1}^k b_{5i} \Delta lr_{t-i} + \sum_{i=1}^k c_{5i} \Delta lexch_{t-i} \\ & + \sum_{i=1}^k d_{5i} z_{t-i} \\ & + \sum_{i=1}^k f_{5i} \Delta lip_{t-i} + \sum_{i=1}^k g_{5i} \Delta lr_{stk}_{t-i} + e_{5t}. \end{aligned} \quad (13)$$

The optimal lag lengths are determined based on the Akaike's Information Criterion, and directions of the Granger causality can be analyzed by the impulse response function.

To purge serial correlation in z_t of Eq. (11), for more accurate estimation, a lag of 5 on autoregression and a lag of 1 on moving average are needed.

TABLE I
Results of the Granger Causality^a

H_0	Canada	Germany	Italy	Japan	United Kingdom	United States
z_t not $\rightarrow \Delta lr_t$	Cannot reject	Cannot reject	(+) Reject *	Cannot reject	Cannot reject	Cannot reject
z_t not $\rightarrow \Delta lexch_t$	Cannot reject	Cannot reject	Cannot reject	Cannot reject	Cannot reject	Cannot reject
z_t not $\rightarrow \Delta lip_t$	Cannot reject	Cannot reject	(-/+) Reject **	Cannot reject	(+) Reject ***	Cannot reject
z_t not $\rightarrow \Delta lr_{stk}_t$	Cannot reject	(-) Reject **	(-) Reject *	Cannot reject	Cannot reject	(-) Reject ***

^a "not $\rightarrow$ " means "does not Granger cause." The Granger causality carries the information that variable x causes variable y , whereas the ordinary regression model provides x associates y . A rejection of H_0 implies the existence of a Granger causality. *, **, and *** denote 1, 5, and 10% significance level, respectively. The signs in parentheses denote the direction of the causality from the impulse response analysis. -/+ indicates the impact is first negative and then turns positive.

The result indicates that the volatile behaviors of oil prices are best described by an ARCH model for Canada due to its insignificant β_1 . However, the GARCH model is preferred for the remaining countries, of which Japan has the strongest persistence in oil price volatility (i.e., $\alpha_1 + \beta_1 > 1$). The United States has the smallest persistence, with $\alpha_1 + \beta_1 = 0.7866$. Except for Japan, in which $\alpha_1 > \beta_1$ (unexpected shock impacts current conditional variance more than past conditional variance), the reverse is true for other countries. Note that the GARCH model is selected so that z_t and z_t^2 are free of serial correlation, an undesirable property in a regression analysis.

It is not surprising that the greatest z_t (greatest change in oil price) value occurred during the Gulf War in 1990. Table I reports the causality results from the statistical models. The lengths of impacts can be determined from the analysis. In the case of Germany, the volatility of a real oil price change exerts a negative impact on stock returns for three periods before it tapers off to zero. For Italy, the volatility of a real oil price change leads the monthly interest rate to change positively for 7 months, the monthly stock return to change negatively for approximately 4 months, and the monthly industrial production to change negatively in the first month but positively during the next 3 months.

Surprisingly, the same volatility exerts positive impacts on industrial production (3 months) for the United Kingdom before leveling off to zero. The favorable impact of higher oil prices on industrial output can be attributed to the fact that Britain is one of the non-OPEC producers: A higher oil price commands the greater revenue ($\eta < 0$) from export, which can in turn be reinvested. For the United States, the volatility of a real oil price change leads to negative stock returns for 3 months before diminishing gradually to zero. The same volatility has no appreciable impacts on either industrial production or stock returns for Canada and Japan. It seems that the volatility of oil price changes leads to negative stock returns in three of the six countries. It affects industrial production in only two countries, including a positive impact in the United Kingdom.

4. CONCLUSION

Recent studies have highlighted the role of oil price uncertainty. The results seem to favor the long-held conjecture that higher oil prices have a greater impact on the stock market than on the output

market. This is supported in our model for the United States. Higher oil prices seem to have a lower impact on major economies such as those of the United States and Japan due to the presence of strategic oil reserve. A strategic oil reserve is essential to absorb the shocks from an excessive price hike. At the heart of the volatility analysis is the concept of oscillating price elasticity at the microeconomic foundation developed by Greenhut *et al.* The existence of unstable price elasticity indeed sows the seed of inherent volatile oil price changes, which can readily inflict shocks to the economy, especially a relatively small economy whose demand for oil depends solely on import. Our volatility model, employing the most recent data, supports this conclusion. The impacts on the interest rate and other monetary variables remain a challenging research avenue for the future.

SEE ALSO THE FOLLOWING ARTICLES

Business Cycles and Energy Prices • Energy Futures and Options • Inflation and Energy Prices • Markets for Petroleum • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Leasing • Oil Crises, Historical Perspective • Oil Industry, History of • Oil-Led Development: Social, Political, and Economic Consequences

Further Reading

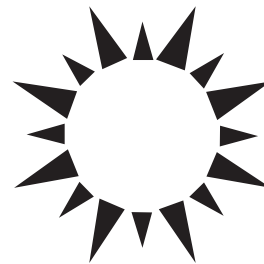
- Alhajji, A. F., and Huettner, D. (2000). OPEC and world crude oil market from 1973 to 1994: Cartel, oligopoly, or competitive? *Energy J.* 21(3), 31–58.
- Bollerslev, T. (1986). Generalized autoregressive heteroskedasticity model. *J. Econometrics* 31, 307–327.
- Burbridge, J., and Harrison, A. (1984). Testing for the effect of oil price rises using vector autogressions. *Int. Econ. Rev.* 25(1), 459–484.
- Dargay, J., and Gately, D. (1995). The response of world energy and oil demand to income growth and changes in oil prices. *Annu. Rev. Energy Environ.* 20, 145–178.
- Department of Energy, Energy Information Administration (1998). *Annu. Energy Rev.*
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of variance of United Kingdom inflation. *Econometrica* 50, 987–1007.
- Engle, R. F., and Granger, C. W. J. (1987). Cointegration and error correction: Representation, estimation, and testing. *Econometrica* 55, 251–276.
- Ferderer, J. (1996). Oil price volatility and the macroeconomy. *J. Macroecon.* 18(1), 1–26.
- Gately, D. (1993). The imperfect price—Reversibility of world oil demand. *Energy J.* 14(4), 163–182.

- Greene, D. L. (1991). A note on OPEC market power and oil prices. *Energy Econ.* 13, 123–129.
- Greene, D. L., Jones, D. W., and Leiby, P. N. (1998). The outlook for U.S. oil dependence. *Energy Policy* 26(1), 55–69.
- Greenhut, M. L., Hwang, M. J., and Ohta, H. (1974). Price discrimination by regulated motor carriers: Comment. *Am. Econ. Rev.* 64(4), 780–784.
- Griffin, J. M. (1985). OPEC behavior: A test of alternative hypotheses. *Am. Econ. Rev.* 75(5), 954–963.
- Griffin, J. M., and Teece, D. J. (1982). “OPEC Behavior and World Oil Prices.” Allen & Unwin, London.
- Hamilton, J. D. (1983). Oil and the macroeconomy since World War II. *J. Political Econ.* 99(2), 228–248.
- Huang, R. D., Masulis, R. W., and Stoll, H. R. (1996). Energy shocks and financial markets. *J. Future Market* 16(1), 1–27.
- Hwang, M. J. (1982). Crude oil pricing in the world market. *Atlantic Econ. J.* 10(2), 1–5.
- Johansen, S. (1988). Statistical and hypothesis testing of cointegration vectors. *J. Econ. Dynamic Control* 12, 231–254.
- Kwiatkowski, D., Phillips, C. B., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *J. Econometrics* 54, 159–178.
- Lee, K., Ni, S., and Ratti, R. A. (1995). Oil shocks and the macroeconomy: The role of price variability. *Energy J.* 16(4), 39–56.
- Mork, K. A., and Olson, M. H. T. (1994). Macroeconomic responses to oil price increases and decreases in seven OECD countries. *Energy J.* 15(4), 19–35.
- Papapetrou, E. (2001). Oil price shocks stock market: Economic activity and employment in Greece. *Energy Econ.* 23, 511–532.
- Phillips, P. C. B., and Perron, P. (1989). Testing for unit root in time series regression. *Biometrika* 75, 335–346.
- Sadorsky, P. (1999). Oil price shocks and stock market activity. *Energy Econ.* 21, 449–469.
- U.S. Government Printing Office (1998). “Economic Report of the President.” U.S. Government Printing Office, Washington, DC.
- Yang, C. W., Hwang, M. J., and Huang, B. N. (2002). An analysis of factors affecting price volatility of the U.S. oil market. *Energy Econ.* 24, 107–119.

Oil Recovery

RUSSELL T. JOHNS

University of Texas at Austin
Austin, Texas, United States



-
1. Importance of Oil and Gas Recovery
 2. Overview
 3. Petroleum Reservoir Fluids
 4. Thermodynamics and Modeling of Oil and Gas
 5. Primary Recovery Processes
 6. Secondary and Tertiary Recovery Processes

Glossary

enhanced oil recovery processes Those processes that stimulate oil recovery by the injection of materials not normally present within the reservoir; usually initiated as a secondary or tertiary recovery method.

miscibility Occurs between two fluids when no distinct interface exists when mixed in any proportion.

oil saturation The ratio of the oil volume that occupies the pores to the volume of the pores.

permeability The ability of fluids to flow through pores of the rock that is largely dependent on the size, shape, and connectivity of the pores.

porosity The ratio of the pore volume of the rock where fluids reside to the bulk volume of rock that includes rock grains and pores.

primary recovery processes Those processes that recover oil by relying on the natural energy available in the reservoir and adjacent aquifers.

recovery efficiency or factor The fraction of the original volume of oil in place that is recovered.

reservoir The portion of porous and permeable rock that contains oil, gas, or water that typically consists of multiple layers of sandstones, limestones, or dolomites from a few meters to nearly a thousand meters thick.

secondary and tertiary recovery processes Those processes that stimulate oil recovery by external injection of fluids to increase the energy available in the reservoir following primary recovery.

Oil recovery is about the efficient and cost-effective recovery of oil from subsurface reservoirs. The design of recovery methods to satisfy these objectives

is the responsibility of petroleum engineers. Petroleum engineers handle the analysis, design, and implementation of recovery projects, which include topics in reservoir, drilling, petrophysical, and production engineering.

1. IMPORTANCE OF OIL AND GAS RECOVERY

1.1 World Energy Consumption

Oil and gas recovery are, and will remain, critical to the energy needs of the world. Figure 1 shows the world's energy consumption for the five major energy sources. In 2002, nearly two-thirds of the world's energy supply came from oil and gas recovery, and oil was the single most important source of energy in the world.

Oil and gas are expected to remain important far into the future. World energy consumption is expected to increase by nearly 58% between 2001 and 2025. Over that period, oil and gas energy sources will remain at nearly two-thirds of the world's energy needs (Fig. 2). Renewables such as wind, solar, and geothermal energy currently represent approximately 2.5% of the world's energy consumption and are not expected to increase dramatically over this period.

1.2 Environmental Benefits

Petroleum engineers have also been pivotal in solving complex environmental problems by extending oil recovery technology to the effective cleanup of aquifers contaminated with toxic chemicals. An example of such an extension is the use of surfactants to clean up aquifers contaminated with nonaqueous phase liquids such as trichloroethane. Surfactant enhanced aquifer remediation, has been very effective in the cleanup of aquifers where conventional cleanup methods have failed.

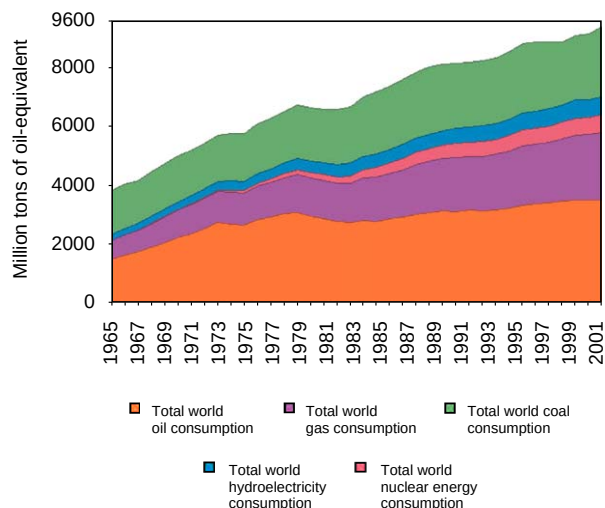


FIGURE 1 Historic world energy consumption of primary energy sources (in millions of tonnes oil-equivalent). From British Petroleum (2003).

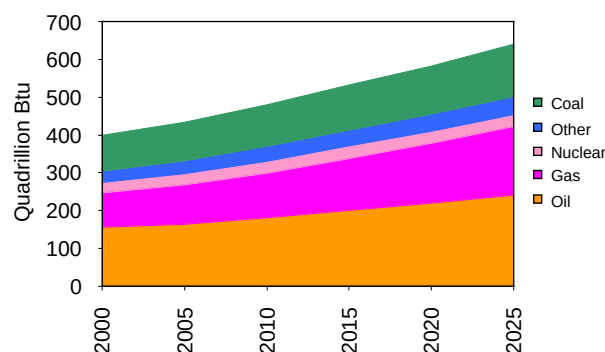


FIGURE 2 Projected world energy consumption by energy source (in quadrillions of Btu). Data from Energy Information Administration. (2003). "System for the Analysis of Global Energy Markets." EIA, Washington, DC.

Achieving high oil recovery by good engineering practices has additional environmental benefits. For example, the use of enhanced oil recovery (EOR) techniques extends the field life beyond its primary recovery period, which lessens the need for new field developments. In addition, EOR reduces greenhouse gas (GHG) emissions by sequestration of carbon dioxide, and injection of hydrocarbon gases that would otherwise be flared.

2. OVERVIEW

Subsurface oil reservoirs are found at depths at or near the Earth's surface to several kilometers below. Oil reservoirs are not pools or lakes of oil below the surface; the oil resides in pores within rock, and this

greatly complicates recovery. Most oil very near the Earth's surface is heavy oil because lighter components in the oil (e.g., methane, ethane) migrate into the atmosphere over time, leaving only the heaviest components behind. However, the processes involved in the migration and formation of oil reservoirs is only partially understood.

Recovery depends largely on the nature of the reservoir and the fluids. Reservoir heterogeneities, such as layering within the reservoir, play a significant role in how quickly the oil (or gas) is produced. For example, oil in low-permeability layers is difficult to recover due to its resistance to flow. Reservoirs may also contain isolated compartments or fractures of high permeability. Fractures, in particular, require good characterization to reduce uncertainty in the recovery predictions. The composition of the oil, the reservoir temperature and pressure, and the distribution of the oil, water, and gas within the reservoir are also important to the selected recovery process. Thus, the most important step in oil recovery design is to characterize accurately the reservoir formation and its fluid properties.

One of the most important outcomes of any characterization is to determine the oil target for a recovery process by calculation of the volume of the original oil in place (OOIP). Only with a reliable OOIP number can petroleum engineers evaluate the success of a particular recovery process.

Surface seismic borehole cores, pressure transient and tracer tests, and geophysical methods (e.g., well logging) are often used to make a three-dimensional characterization of the oil reservoir. Well logging is a geophysical method that introduces a device, called a sonde, down a borehole to measure, using electrical and nuclear methods, properties such as porosity and fluid saturations. However, the characterization of a reservoir is a daunting task given that reservoirs are highly heterogeneous, and these methods sample limited and varying volumes of reservoirs at differing degrees of resolution.

The number and placement of wells in the reservoir is also important to recovery. Wells can be perforated over a selected depth interval where it is believed that good oil recovery will occur. Advances in drilling engineering allow for wells to be drilled a great length horizontally through thin oil layers. Where multiple wells had to be drilled in the past, now sometimes only a few wells are required. Advances in offshore drilling have also opened up new possibilities for finding and producing oil reservoirs. Offshore wells can now be drilled in oceans that are more than a mile deep.

Once the reservoir and its fluids are characterized, a recovery process is selected based on economic and engineering constraints. Primary recovery methods are those that are initially used to recover as much of the oil as possible by drilling production wells (boreholes) in predetermined patterns. Primary recovery eventually becomes uneconomical as the reservoir fluids lose energy (or pressure) over time as the oil is produced. Some reservoirs require pumps or other methods to assist in recovery, especially at later stages of production. Secondary and tertiary recovery methods, such as water and gas injection, are typically used (if cost-effective) after oil recovery from primary methods has peaked and a significant decline is taking place. Either new injection wells are drilled or some of the production wells are converted into injection wells in a predetermined pattern partially dictated by the pattern of wells used for primary recovery.

An accurate estimation of the ultimate recovery efficiency is one of the primary goals of petroleum engineers. The recovery efficiency or factor, which is a fraction between 0 and 1, is defined as the ratio of the volume of oil produced (or recovered) to the volume of OOIP.

The most important aspects of oil recovery from an economic viewpoint are the rate and volume of oil recovered. A decision about which recovery process to use is not based solely on what could be recovered but rather on economic limits and rankings, risk factors, and available assets. Besides safety and environmental concerns, engineering designs are driven by the desire to rapidly produce the oil in the most cost-effective way.

3. PETROLEUM RESERVOIR FLUIDS

The composition of oil and gas in a reservoir (or the fluids injected) is important in the design of recovery processes. Several experimental methods based on true boiling point distillation and gas chromatography are available to determine the composition of the fluids. Current methods can report molar compositions for individual carbon numbers up to approximately C_{35} and can also give PNA (paraffins, naphthenes, and aromatics) distributions. Older reports generally report only paraffin hydrocarbon components up to carbon numbers of C_6 , and the rest are lumped into a C_{7+} fraction or pseudo-component. Ideally, compositional analyses should give the mole fractions, boiling points, specific gravity, critical properties, and molecular weights

of as many components as possible so that equations of state (EOSs), which are mathematical models of the oil and gas phase behavior, can be constructed with reasonable accuracy and reliability. This is a daunting task given that many isomers exist for carbon numbers greater than C_6 . For example, the paraffin hydrocarbon decane (C_{10}) has 75 different structural configurations, each one of which has different physical properties. Thus, components with a similar number of carbon atoms are often lumped into one pseudo-component with their average properties reported.

Petroleum reservoir fluids are classified as black oils, volatile or near critical oils, retrograde gas condensates, wet gases, and dry gases. The fluid composition determines the phase envelopes (the boundary of the two-phase region), which are composed of the dew point and bubble point curves. The reservoir fluid type is determined relative to the phase envelope (Fig. 3). A single phase, usually gas or liquid, exists outside the phase envelope. Inside the envelope, two phases coexist at equilibrium. The dew point and bubble point curves are separated by the critical point (labeled C in Fig. 3), where the bubble point curve is to the left of the critical point. The dew point curve gives the pressure at which the first small drop of liquid forms as the pressure is increased at constant temperature. The bubble point curve gives the pressure at which the first small bubble of gas forms as the pressure is decreased at constant temperature. The critical point is the temperature and pressure at which the two phases and their properties become identical. The critical point is very important in the determination of miscibility.

Four phase envelopes are shown in Fig. 3, each of which is for a given reservoir fluid composition. For a

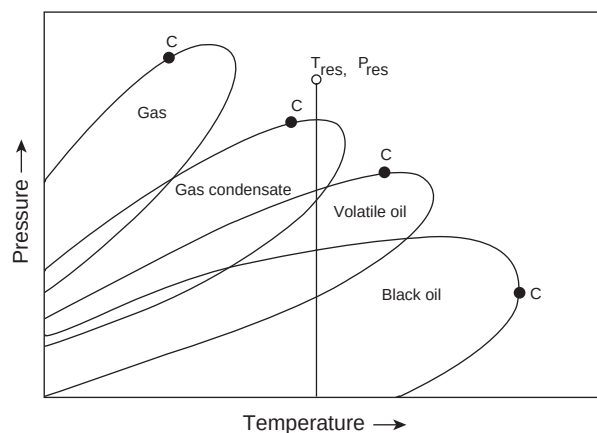


FIGURE 3 Reservoir fluid types and phase envelopes. From Pederson, Fredenslund, and Thomassen (1989).

fluid composition that gives the phase envelope at the upper left-hand side of the figure, the fluid is a gas at the reservoir temperature and pressure shown and remains a gas as the reservoir pressure is reduced. The reservoir pressure is typically reduced during production, especially for primary depletion drives. The reservoir temperature is essentially constant so long as fluids such as steam are not injected. For a fluid composition that gives the next phase envelope from the left in Fig. 3, the reservoir pressure intersects the dew point curve twice. Thus, as pressure is reduced during production, the liquid saturation will first increase and reach a maximum but will subsequently decrease to zero. Reservoir fluids that behave in this way are known as retrograde gas condensates. For the other two fluid compositions (and their corresponding phase envelopes), the fluids are single-phase oils, known as volatile or near critical oils and black oils. For both of these oils, the bubble point curve is first intersected as pressure is reduced at a pressure known as the bubble point pressure or saturation pressure, followed by the intersection of the dew point curve. Thus, the oil saturation continuously decreases, whereas the gas saturation continuously increases. Black oils contain a greater mole fraction of heavier components than do the other fluid types, and these heavier components will remain in the liquid phase even for very low pressures.

4. THERMODYNAMICS AND MODELING OF OIL AND GAS

Numerical simulation is often used to predict oil recovery from a reservoir, and the calculation of phase behavior is an integral part of those predictions. Phase behavior describes the complex interaction among physically distinct, separable portions of matter called phases that are in contact with each other. Typical phases are solids, liquids, and vapors. Phase behavior plays a vital role in many petroleum applications such as EOR, compositional simulation, geochemical behavior, wellbore stability, geothermal energy, environmental cleanup, multiphase flow in wellbores and pipes, and design of surface facilities.

Thermodynamics, which is central to understanding phase behavior, is the study of energy and its transformations. Thermodynamics began as the study of heat applied to steam power but was broadened substantially by J. W. Gibbs during the mid- to late 1800s. Thermodynamics describes the energy

changes that occur during phase transformations; it is used to predict the phase separations that occur. The most fundamental idea in thermodynamics is the conservation of total energy—the first law of thermodynamics. The first law is based on our everyday observation that for any change in thermodynamic properties, total energy, which includes internal, potential, kinetic, heat, and work energy, is conserved. The second fundamental idea in thermodynamics is the total entropy balance or the second law of thermodynamics. Entropy is a thermodynamic property that expresses the unidirectional nature of a process.

4.1 Phase Equilibrium Thermodynamics

Gibbs's most significant contribution was the development of phase equilibrium thermodynamics applied to multicomponent mixtures, particularly the concept of chemical potential. The concept of chemical potential leads to the simple result that, at equilibrium, the chemical potential of each component must be the same in all phases ($\mu_{iL} = \mu_{iV}$).

Phase equilibrium thermodynamics seeks to determine properties such as temperature, pressure, and phase compositions that establish themselves after all tendencies for further change have disappeared. Figure 4 is a schematic showing a closed container of liquid and vapor that could represent one grid block in a numerical simulation. Given constant and known temperature, pressure, and overall molar fractions or compositions (z_i where $i = 1, \dots, n_c$), the fundamental task in petroleum phase behavior calculations is to quantify the phase molar fractions (phase splits L and V) and the phase equilibrium molar fractions of the vapor (y_i where $i = 1, \dots, n_c$) and liquid phases (x_i where $i = 1, \dots, n_c$) that form at

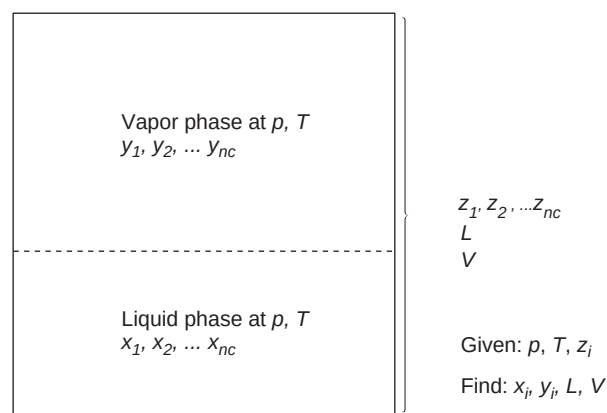


FIGURE 4 Vapor-liquid equilibrium at constant pressure, temperature, and overall molar composition. A dashed line shows a distinct interface between the two phases.

equilibrium. z_i is the ratio of the moles of component i in all phases to the total moles of all components in all phases, x_i is the ratio of the moles of component i in liquid to the total moles of liquid, y_i is the ratio of the moles of component i in vapor to the total moles of liquid, and n_c is the number of components. L is the ratio of the total moles of liquid to the total moles of all phases, and V is the ratio of the total moles of vapor to the total moles of all phases. The phases are assumed to be homogenous in that parameters such as pressure, temperature, density, viscosity, and phase compositions are uniform throughout the phases. By satisfying the equilibrium conditions that $\mu_{iL} = \mu_{iV}$, the vapor-liquid equilibrium properties can be calculated to determine the compositions of vapor and liquid in equilibrium along with the phase molar splits and volumes.

Oil recovery processes are never in equilibrium, and it would appear that equilibrium thermodynamics is not very useful. The concept of local equilibrium is used to overcome this apparent limitation of thermodynamics. Equilibrium at a point in a reservoir, termed local equilibrium, often applies when internal relaxation processes are rapid with respect to the rate at which changes are imposed on the system. That is, equilibrium thermodynamics can be applied locally over small volumes of the reservoir, even though pressure and other gradients remain across the reservoir. In reservoir simulation, the small volumes are taken to be grid blocks. The size of the grid blocks must be sufficiently small so that good accuracy is obtained.

4.2 Equations of State

An accurate characterization of phase behavior is critical to the prediction of oil recovery. Continuous experimental data covering the full range of pressures and compositions that may occur in a reservoir are never available. Thus, mathematical models that are “tuned” to the available experimental data are used to predict the phase behavior. EOS models are used for this purpose. The simplest and most fundamental EOS is the ideal gas equation, where the pressure, volume, and temperature of a fluid are related by $pV = RT$, where p is pressure, V is molar volume (inverse of molar density), R is gas constant, and T is temperature.

For compositional simulation purposes, more complex EOSs that accurately represent, in a continuous way, both the oil and gas phases are used. The two most widely used cubic EOSs in the petroleum industry are the Peng–Robinson EOS

(PREOS) and modified versions of the Redlich–Kwong EOS (RKEOS). These two cubic EOSs are sufficiently simple that they can be used in reservoir simulations with as many as hundreds of thousands of grid blocks so long as the number of components that represent the fluid is less than approximately 10.

An example of a cubic EOS is the PREOS, which is expressed for a pure fluid as

$$p = \frac{RT}{V - b} - \frac{a\alpha}{V(V + b) + b(V - b)}$$

where parameters a and b are functions of the component's critical temperature and pressure and α is a temperature-dependent function. This type of cubic EOS represents both the liquid and vapor in a continuous way. For example, when density decreases (or V is large), the PREOS approaches the correct limit of the ideal gas equation. For the other limit where V approaches b , the cubic EOS approaches an incompressible liquid in that the pressure increases rapidly with small changes in the density.

For fluid mixtures, the same form of the cubic EOS is used, but mixing rules determine the EOS parameters. Binary interaction parameters are used to account for nonideal interactions between different species. Volume shift parameters adjust liquid phase densities calculated from the EOS to experimentally measured densities. The phase viscosities are estimated from empirical relations and properties calculated from the EOS. Critical volumes of the components can be used to adjust the phase viscosities to match experimentally measured viscosities.

The accuracy of any EOS depends on its ability to model the attractions and repulsions among molecules over a wide range of temperatures and pressures. EOS models are empirical in that they do not attempt to model the detailed physics; rather, they attempt to model only the cumulative effect in terms of a small number of empirical parameters. In general, cubic EOS models of the Peng–Robinson type are more accurate for molecules with weak polarity or chemical interactions, and this explains why water, a polar substance, is difficult to model with a cubic EOS. Mixtures that contain alcohols, bases, organic or inorganic acids, and electrolytes are other examples of fluids that are not accurately modeled by cubic EOS models of the type shown here.

Fluid compositions may vary from one location in the reservoir to another. In such cases, multiple EOS characterizations may be required. Compositional variations can occur for a variety of reasons. For example, gravity can cause vertical compositional

gradients, where heavier components become more concentrated at greater depths. Composition variations in a reservoir can also be caused by thermal gradients.

4.3 Flash Calculation

The procedure for equilibrium calculation of a two-phase mixture shown in Fig. 4 is called a flash calculation. For an equilibrium flash calculation, the pressure, temperature, and overall mole fractions are specified. The general procedure for a flash calculation is as follows:

1. Make an initial guess of the K values, where $K_i = y_i/x_i$. Most EOS programs use empirical correlations to estimate the phase equilibrium molar fractions based on K values.
2. Calculate x_i and y_i using a material balance procedure such as the Rachford–Rice procedure. Typically, a nonlinear material balance equation is solved by Newton–Raphson iteration.
3. Calculate the cubic EOS parameters (e.g., a , b). This step is straightforward and depends on the selected EOS and its associated mixing rules. The critical temperatures, pressures, and acentric factors for each component, among others, are needed to calculate the EOS parameters.
4. Solve the cubic EOS for the molar volumes V of the liquid and vapor.
5. Calculate the chemical potentials (or, alternatively, the fugacities) of each component in each phase. The selected cubic EOS is used to determine an expression for the fugacity of a component in a phase.
6. Check whether equilibrium has been reached. The correct equilibrium solution is found when $\mu_{iL} = \mu_{iV}$ (or, alternatively, the component fugacities are equal) for each component relative to an acceptable tolerance.
7. If the tolerance has not been satisfied, update the K values and repeat steps 2 to 6. Successive substitution or Newton–Raphson iteration is often used to update the K values.

4.4 Simplified Fluid Characterization

Phase behavior calculations require that components and their properties be known. However, crude oils typically have hundreds of components, making the preceding flash calculation procedure computationally too intensive even for numerical simulation. Thus, components are often lumped into 10 pseudo-

components or less to characterize the *in situ* fluid. The characterization usually takes the following three steps:

1. Analyze the hydrocarbon components in the *in situ* fluid using analytical techniques such as chromatography and distillation.
2. Separate and lump the measured components into as few pseudo-components as is possible without sacrificing the required accuracy. The properties and selection of the pseudo-components are determined using a variety of methods and empirical correlations. The required pseudo-component properties are those needed for the cubic EOS calculations such as critical temperature, pressure, and acentric factor.
3. Adjust the pseudo-component properties so that the EOS will match all available pressure–volume–temperature (PVT) experiments. This process, which typically uses nonlinear regression, is known as EOS tuning. The objective of tuning is to match fluid properties and phase equilibria accurately over the range of pressures and compositions expected to occur in a reservoir. EOS tuning is nearly always needed due to the inherent uncertainty in the properties estimated from step 2, especially for the heavier components. Binary interaction parameters between the heaviest and lightest pseudo-components are typically the first parameters to be adjusted, although other parameters (e.g., the acentric factors) might need tuning.

The selection of pseudo-components and their properties is likely not unique, as is often the case when numerous model parameters are estimated by fitting measured data with nonlinear regression. Care should be taken to avoid estimates in the pseudo-component properties that are unphysical and to reduce the number of parameters fitted. Furthermore, the final EOS characterization is most accurate in the range of the experimental phase behavior data that have been measured and fitted. Phase behavior data that covers, to the extent possible, the conditions that occur in the reservoir should be collected.

4.5 Pressure–Volume–Temperature Experiments

Routine PVT experiments give volumetric information over a range of composition and pressure conditions expected for conventional methods such

as water flooding and pressure depletion. Typical routine experiments include differential liberation, constant composition, constant volume depletion, and separator tests. These data are used for tuning EOSs for black and volatile oils. Differential liberation simulates depletion of black oil below its bubble point pressure as gas is evolved. Constant composition expansion tests simulate black oil depletion from above to below the bubble point pressure. Constant volume depletion experiments simulate depletion of a volatile oil below its bubble point pressure or a retrograde gas condensate below its dew point. This test can provide an approximate estimate of oil recovery that is expected during depletion. Finally, separator experiments simulate single- or multiple-stage separation processes at the surface. This test can provide gas/oil ratios (GORs) and formation volume factors (B_o) and molecular weight of the oil that remains in the surface stock tank. The oil formation volume factor is the ratio of the volume of oil at reservoir conditions to the volume of oil in the stock tank. The stock tank conditions are typically at atmospheric pressure and 15°C, which are referred to as standard temperature and pressure (STP) conditions. Compositions of the gas and liquid phases that separate at each simulated stage are measured from separator tests.

When tertiary recovery methods (e.g., gas injection) are used, other PVT experiments are required for EOS tuning that may better represent the compositions expected in the reservoir during recovery. These tests include swelling tests and multiple-contact experiments. Swelling tests measure phase behavior data in a plot that is commonly called a pressure composition diagram in thermodynamics. The injection gas is mixed with reservoir oil along a dilution line to compositions beyond near critical conditions. Phase equilibrium compositions are measured for several overall compositions along with the liquid saturation. The swelling or increase in the formation volume factor as gas is added is also measured. Multiple-contact tests simulate the gas/oil contacting process that occurs in a reservoir. Forward experiments simulate multiple contacting in a vaporizing gas drive in that fresh gas repeatedly contacts equilibrium oil. A backward multicontact experiment simulates contacts that occur in a purely condensing gas drive, where equilibrium gas repeatedly contacts fresh oil. However, multiple-contact tests do not adequately simulate the combined condensing–vaporizing drives that commonly occur during enriched gas or carbon dioxide floods.

5. PRIMARY RECOVERY PROCESSES

Primary recovery processes are those that recover oil by relying on the natural energy available in the reservoir and adjacent aquifers. Primary recovery relies on the expansion of *in situ* fluids or rocks to produce oil (or gas); specific mechanisms include solution gas drive, gas cap drive, natural water drive, and rock compaction. A combination of these recovery mechanisms occurs in most reservoirs, but often one of them is predominant. Gravity can also assist in the recovery, depending on the fluids and the reservoir geometry. Detailed numerical simulations of these processes are often performed to predict oil recovery.

When oil and gas are recovered during primary production and depletion, oil, gas, and some water are expelled from the reservoir, thereby reducing the reservoir pressure and causing the remaining oil, gas, and water to expand so as to fill the pore spaces vacated by the fluids removed. When the reservoir is hydraulically connected to aquifers or water-bearing layers, water may flow or influx into the oil reservoir as the reservoir pressure drops during production. The recovery factor for all of these processes depends largely on the reservoir heterogeneities and properties, reservoir pressure and temperature, reservoir fluid composition, production well patterns, and economics.

The goal of any recovery process is to maximize the oil recovery at the stock tank. The formation volume factor is a convenient measure of the volume of oil that may exist at stock tank conditions compared with its volume *in situ*. For example, for an undersaturated black oil (i.e., an oil that exists at a reservoir pressure above its bubble point pressure), the formation volume factor is always greater than 1.0 because the oil contains solution gas. Typical formation volume factors for black oils in oilfield units are approximately 1.2 to 1.6 reservoir barrels (RB)/stock tank barrel (STB). Volatile oils have even higher formation volume factors than do black oils; thus, shrinkage is even more of a concern.

As oil is produced, the oil pressure is lowered below its bubble point pressure and solution gas evolves from the oil, shrinking the oil volume (see Fig. 3). The solution GOR is a measure of the volume of solution gas in standard cubic feet (SCF) per STB.

5.1 Solution Gas Drive

Many oil reservoirs are initially undersaturated with dissolved gas in that the reservoir pressure is above the bubble point pressure. For such reservoirs, if

there is no water influx from the periphery of the reservoirs, the oil and water compressibilities provide the primary energy for recovery. That is, the liquids expand as pressure is decreased, causing oil to flow to the production wells and to the surface. Because the fluid compressibilities are small, there is often a rapid decline in average reservoir pressure during the liquid expansion period until the bubble point pressure is reached. Often, little water is produced because the initial or connate water saturation is typically small, making the resistance to water flow large.

Once the average reservoir pressure drops below the bubble point pressure, solution gas evolves from the oil. The quantity of gas that evolves depends on the oil composition, that is, whether it is a black oil or a volatile oil. The solution gas that develops in the pores provides additional energy beyond liquid expansion with which to recover oil. Because gas is more compressible than liquids, the average reservoir pressure does not decrease as rapidly as it did when the pressure was above the bubble point pressure. When the reservoir pressure eventually becomes very low, the only remaining source of energy for recovery is often gravity drainage of oil to the wells. Gravity drainage occurs slowly and pumping is often required to flow oil to the surface.

Volumetric and material balances are often used to determine the initial oil in place and to give an approximate estimate of the potential oil recovery. For a reservoir under pressure depletion with little water influx, produced oil must be replaced by gas. Thus, the gas saturation increases while the oil saturation decreases; therefore, the volume of expansion is equal to the volume of production. At abandonment (i.e., when the oil is no longer recovered in economical quantities), the recovery in STB per acre-foot is given by the difference in the OOIP and the oil in place at abandonment,

$$7758\phi \left[\frac{(1 - S_w)}{B_{oi}} - \frac{(1 - S_w - S_g)}{B_o} \right],$$

where B_{oi} is the initial formation volume factor of the reservoir oil, S_w is the initial water saturation, ϕ is the porosity, B_o is the formation volume factor of the oil at abandonment conditions, and S_g is the gas saturation at abandonment conditions. Recoveries from solution gas drives are typically small and in the range of 10 to 30% of the OOIP.

An accurate determination of the recovery factors requires measuring the fluid saturations. Core experiments or core samples taken from the reservoir are often used to measure the water and gas

saturations that go into the volumetric balances. Other techniques for measuring saturations include tracer tests and (more commonly) well logging.

5.2 Gas Cap Drive

Some oil reservoirs occur at conditions below the bubble point pressure so that an initial gas cap has formed over geological time. In such cases, the oil is initially saturated with dissolved gas and the production is primarily controlled by the size and expansion of the gas cap (Fig. 5). Because gas compressibility is large compared with liquid compressibilities, a small gas cap can provide significant pressure maintenance and contribute to increased oil recovery. For example, as production proceeds and the reservoir pressure is slowly reduced, the gas cap expands, displacing oil downward toward the perforated intervals of the production wells. The recovery efficiency is particularly enhanced if the reservoir is tilted so that gravity maintains a stable gas/oil interface during production. The field life can be extended during the later stages of production by recompleting deeper or sealing the perforation interval of high-GOR wells. The ultimate recoveries by gas cap expansion can be in excess of 50% of the OOIP with gravity assistance but are typically in the range of 25 to 35%.

5.3 Natural Water Drive

For reservoirs that have natural water influx (water drive), the reservoir pressure during oil production can often be maintained, providing a significant increase in energy in which to recover oil. The rate of pressure decline will depend on the volume of water that flows inward to the reservoir from its boundaries compared with the volume of oil, water, and gas that is recovered or expelled via production wells. High recoveries are obtained especially for tilted reservoirs where the water can move vertically upward, thereby displacing the oil (Fig. 5).

Like gas displacing oil downward during a gas cap drive, water that flows into the reservoir displaces oil, but a significant amount of oil can be left behind when it is trapped by capillary forces in what is termed residual oil. Therefore, for natural water drive reservoirs, the recovery is limited by the residual oil saturation. Residual oil saturations depend on the geometry of the pores, among other factors, and are commonly in the range of 10 to 40%.

A simple volumetric balance can sometimes be used to estimate the maximum recovery for a water

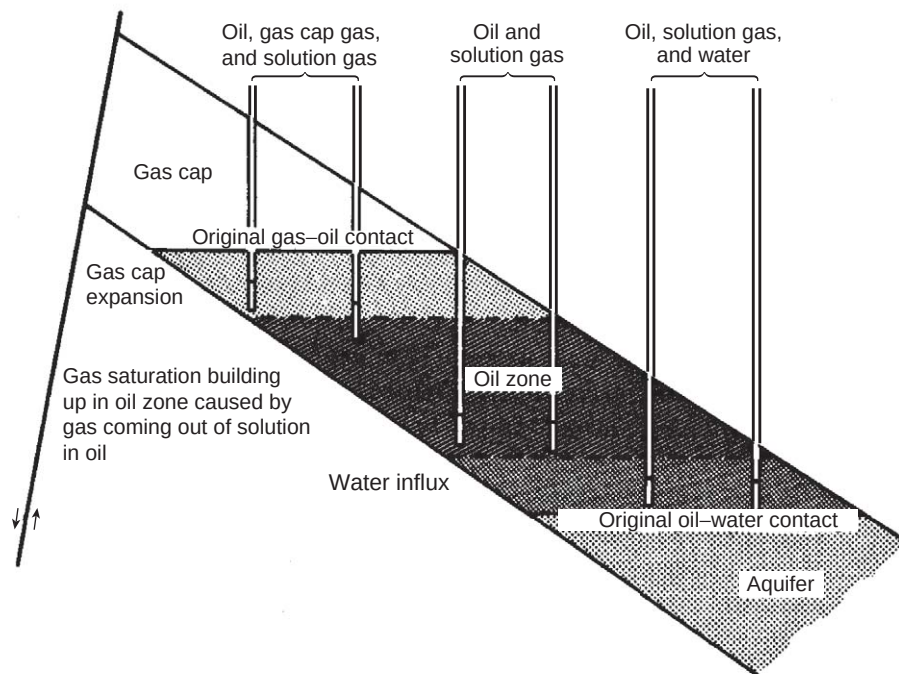


FIGURE 5 Primary recovery for a tilted reservoir under simultaneous gas cap drive and water drive. From Craft and Hawkins (1991).

drive. For example, when the reservoir pressure is maintained at high levels, no significant amount of gas evolves in the reservoir during production and the recovery in STB per acre-foot is given by

$$\frac{7758\phi(1 - S_w - S_{or})}{B_{oi}},$$

where S_{or} is the residual oil saturation. The recovery efficiency for natural water drive reservoirs is typically greater than that under solution gas drive and can be in excess of 50% of the OOIP.

5.4 Compaction Drive

The compressibility of the rock can also affect recovery, but typically this type of drive is secondary to other mechanisms. However, for some shallow reservoirs, compaction drives can provide a significant increase in primary recovery. Ultimate recovery factors of up to 50% for shallow reservoirs have been reported in a few cases.

The principles behind compaction drives are simple, but they are often difficult to model accurately. As the pressure is reduced, the pore volume of the rock may decrease (i.e., compact), and this can provide energy to squeeze the oil and displace it toward production wells. Compaction

can lead to subsidence at the surface, and this can be a major problem.

6. SECONDARY AND TERTIARY RECOVERY PROCESSES

Secondary and tertiary recovery processes are typically defined as those that add external energy to the reservoir to recover additional oil or gas. The most common type of secondary recovery process is that of water flooding, in which water is injected into the reservoir and displaces oil toward the production wells. As the name implies, water flooding as a secondary process typically occurs after primary production, although this is not necessarily so. Other secondary and tertiary processes that are commonly used include gas injection (both miscible and immiscible), microemulsion and polymer flooding, *in situ* combustion, and steam flooding.

Tertiary recovery processes often use EOR technology. EOR processes are those that stimulate oil recovery by the injection of materials not normally present within the reservoir. EOR techniques allow the economic value of existing fields to be maximized through increased oil recovery and field life extension. There are also environmental benefits associated with the use of EOR, as described previously.

6.1 Factors Controlling Recovery Efficiency

The recovery efficiency that results from secondary and tertiary recovery processes is the product of the displacement efficiency and the volumetric sweep efficiency (i.e., $E_R = E_D E_V$). The volumetric sweep efficiency E_V is defined as the ratio of the oil volume contacted by the recovery process to the volume of OOIP (Fig. 6). The fraction of oil that is contacted by the recovery process and ultimately recovered or displaced is defined as the displacement efficiency (E_D).

6.1.1 Reservoir Mixing

Mixing or dilution of injected fluids with reservoir fluids can cause the displacement efficiency E_D to decrease. Mixing in a reservoir can be caused by dispersion, diffusion, or cross-flow. Cross-flow is the process by which fluids move from one reservoir layer to another due to capillary, viscous, or gravity forces. Diffusion is the spreading of components in a single-phase fluid caused by random molecular movement. Dispersion is the spreading of components in a single-phase fluid caused by streamline splitting and dead-end pores during advection.

6.1.2 Fluid Mobility

The sweep efficiency E_V depends on the mobility of the injected fluids as well as the well pattern. The

mobility is defined as the ratio of the permeability to the fluid viscosity. Injected fluids of high mobility can flow through the high-permeability layers of the reservoir more easily. Conformance methods are often used to improve sweep efficiency by controlling where fluids are injected (or produced) or by injecting fluids that alter how the fluids move through the reservoir layers. Gels, polymers, and foams are examples of such fluids.

6.1.3 Fluid Density

The density of the injected fluids also affects the sweep efficiency. Light fluids can rapidly migrate upward in a gravity tongue, bypassing oil, while denser fluids can migrate downward. In such cases, heterogeneities may improve the sweep efficiency by not allowing fluids to migrate vertically upward (or downward). Water alternating gas injection (WAG), in which water and gas are injected in alternating cycles, is often used to improve the sweep efficiency by increasing the mixing of reservoir fluids with injected fluids. However, increased mixing can significantly reduce the displacement efficiency. Typically, there is a trade-off between the displacement and sweep efficiencies in maximizing oil recovery.

6.1.4 Multiphase Flow and Relative Permeability

Darcy's law governs flow of oil, water, and gas in porous media. Darcy's law says that the flow rate at any point in the reservoir is given by the fluid pressure gradient, the viscosity of the fluid, and its effective fluid permeability. That is, when gravity is neglected,

$$q = \frac{k}{\mu} \frac{dp}{dx},$$

where q is the volumetric flux of the fluid in meters per second, k is the effective permeability of the fluid, μ is the fluid viscosity, and dp/dx is the fluid pressure gradient. Therefore, for oil, the volumetric flux through the reservoir will be smaller for viscous or heavier oils than for lighter oils.

The effective permeability is the product of the absolute permeability of the rock and the relative permeability. The absolute permeability of the rock is typically measured with 100% water or gas occupying the pore space. However, when two or more phases are present in the pores, flow of each phase is restricted by the other phases. As the saturation of the water or gas phases is increased and the oil saturation decreases, the resistance for oil to flow increases until oil no longer flows at its residual oil saturation.

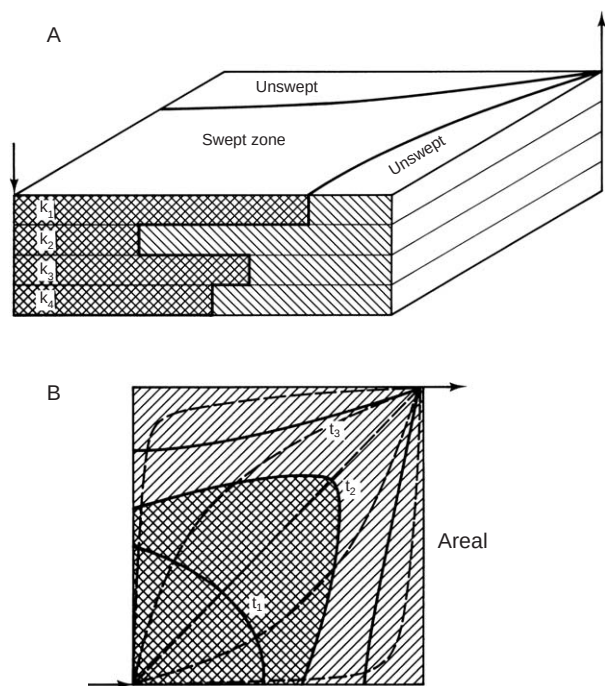


FIGURE 6 Schematic of sweep efficiency in a layered reservoir: (A) vertical and (B) areal. From Lake (1989).

Flow resistances also depend on the “wettability” of the rock. For water-wet rocks, the water resides in the smaller pores and on the surfaces of the rocks, whereas the oil is in the center of the larger pores. Fluids that are in large pores can flow easier than can those in smaller pores. Reservoirs can be water wet, oil wet, or mixed wet, although the wetting nature can vary significantly within the reservoirs.

A measure of the flow resistance is given by the phase relative permeability, which is usually a number between 0 and 1. Therefore, the effective permeability of a fluid is the absolute permeability multiplied by its relative permeability. The concept of effective permeability explains why oil recovery for a water flood or an immiscible gas flood is not 100%, even when sweep is nearly 100%.

6.2 Water Flooding

Water injection as a secondary method of recovery is very similar to the recovery process that occurs from natural water influx. The main advantage here is that the volume and placement of the water can be controlled so that sweep efficiency is improved and reservoir pressure is maintained. Polymers are sometimes injected with the water to increase the water's viscosity and to decrease its mobility. Sweep efficiency can be significantly improved with polymer injection.

6.3 Immiscible Gas Injection

Immiscible gas injection in the reservoir is similar to the recovery process that occurs during a gas cap drive. Like water flooding, the volume and placement of gas can be controlled to improve sweep efficiency and maintain reservoir energy or pressure. Typical gases for immiscible flooding are methane, nitrogen, carbon dioxide, and air. Many of these gases are not completely immiscible with the oil. For example, carbon dioxide nearly always has some limited miscibility with the oil and, therefore, can swell the oil and reduce its viscosity, both of which can improve recovery. However, carbon dioxide is relatively expensive to inject as an immiscible gas and is generally not used in this way today.

6.4 Miscible Gas Injection

A much more efficient and commonly used method today is to inject gas that is miscible with the oil at the reservoir temperature and pressure of interest. The oil recovery for miscible gas floods can increase

significantly beyond primary recovery processes because the interfacial tension between the residual oil and injected gas is reduced to zero. As the interfacial tension between the fluids is reduced, capillary pressure is decreased and the residual oil saturation decreases in many cases to values of less than 1 or 2%.

One miscible fluid used today is carbon dioxide, which is injected as a supercritical fluid at temperatures and pressures beyond its critical point (Fig. 7). Other gases that can be miscible with oil are methane and nitrogen, although these gases require much higher reservoir pressures to achieve miscibility. The type of gas used also depends on its cost and its availability at the field site.

There are two types of miscibility that can occur between resident oil and injected gas. Fluids can be multicontact miscible or first-contact miscible. First-contact miscibility occurs when the gas and oil are miscible when mixed in any proportions. This type of miscibility is ideal but is seldom achieved given the cost and availability of gas components in the field. Multicontact miscibility occurs more often and is the process by which gas and oil develop miscibility in the reservoir by repeated contact with each other. Historically, the multicontact miscible process was thought to occur either by vaporizing contacts or by condensing contacts. In vaporizing drives, fresh gas mixes with equilibrium oil and vaporizes the intermediate-weight components in the oil, whereas in condensing drives, fresh oil mixes with equilibrium gas and condenses intermediate-weight components in the gas into the oil. Vaporizing drives occur when components, such as methane and nitrogen, are injected at pressures above the minimum pressure for miscibility (or MMP). Condensing drives were thought to occur when gases are sufficiently enriched with hydrocarbon components, such as ethane and propane, to a composition called the minimum enrichment for miscibility (or MME).

The drive mechanisms for both enriched hydrocarbon gas floods and carbon dioxide floods have been shown recently to be combined condensing/vaporizing. In the combined condensing/vaporizing process, the gas at the leading edge of the gas front swells oil, while residual oil is vaporized by gas at the trailing edge. Miscibility is approached in the reservoir between the condensing and vaporizing regions.

Multicontact miscibility is only approached, and is never actually attained, in a reservoir. Dilution of the gas by mixing of resident oil with injected gas

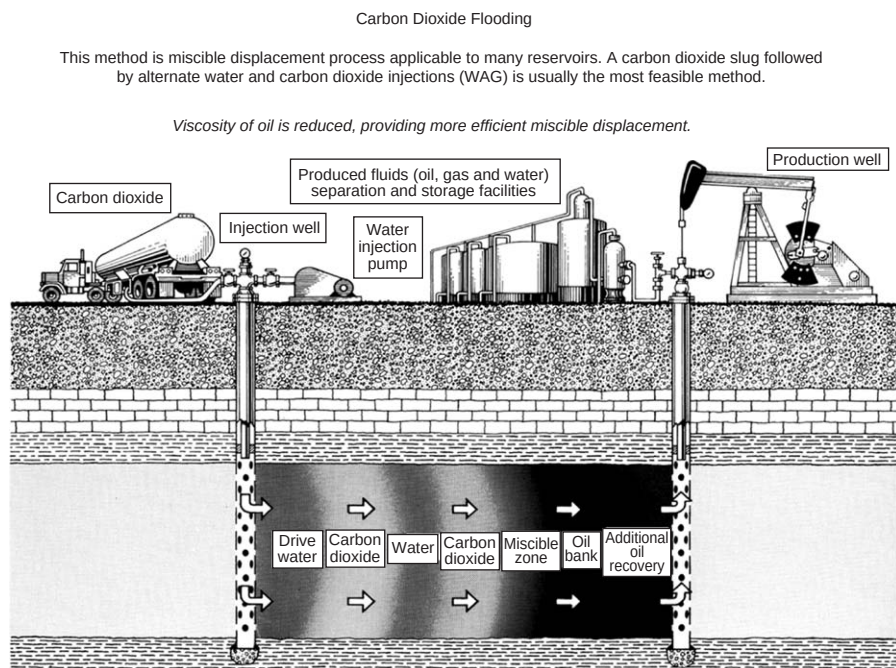


FIGURE 7 Schematic of a miscible gas flooding process. Drawing by Joe Lindley, U.S. Department of Energy, Bartlesville, OK. From Lake (1989).

always causes some two-phase flow or partial miscibility to occur. Fortunately, achieving near miscibility often is sufficient to greatly enhance recovery by a reduction in residual oil saturation. Floods are generally defined as miscible when the reservoir pressure is above the thermodynamic MMP or the composition of the injected gas is beyond the MME.

The estimation of the MMP and MME has been historically determined from slim-tube experiments, numerical simulations with fine grids, or correlations. In slim-tube experiments, oil recovery from the injection of gas into rock samples of thin diameter is plotted as a function of the pressure or enrichment. A bend in the recovery curve is thought to occur at the MMP or MME. However, both slim-tube experiments and numerical simulations can have a large uncertainty in these parameters due to the effect of dispersion on reducing the recovery. Recent methods have allowed for analytical calculation of these important parameters, but these methods rely on an accurate fluid characterization from an EOS.

6.5 Microemulsion Flooding

Microemulsion flooding is the injection of a microemulsion into the reservoir to reduce the interfacial tension that traps residual oil. Microemulsions are

surfactant stabilized dispersions of water and hydrocarbons. Micelles are formed at sufficiently high surfactant concentrations that are capable of solubilizing fluid within their core centers. One significant advantage of microemulsions is that they have viscosities significantly greater than water and, hence, can improve sweep efficiency.

Microemulsion flooding is more complex to perform than are other EOR methods and, thus, requires significant technical expertise. The injected slug of microemulsion depends on many parameters, including reservoir temperature, water salinity, and crude oil type. Because of its complexity, cost, and other concerns, microemulsion flooding is not used often for oil recovery. However, an extension of this method, called surfactant enhanced aquifer remediation (SEAR), is being used today to effectively clean up shallow contaminated aquifers that otherwise would remain contaminated.

6.6 *In situ* Combustion

In situ combustion is a process in which ignition of coke deposits left by the crude oil is initiated at an air injection well, causing a combustion front to propagate into the reservoir (Fig. 8). The coke region is the zone where carbonaceous material has been deposited as a result of thermal cracking of the crude

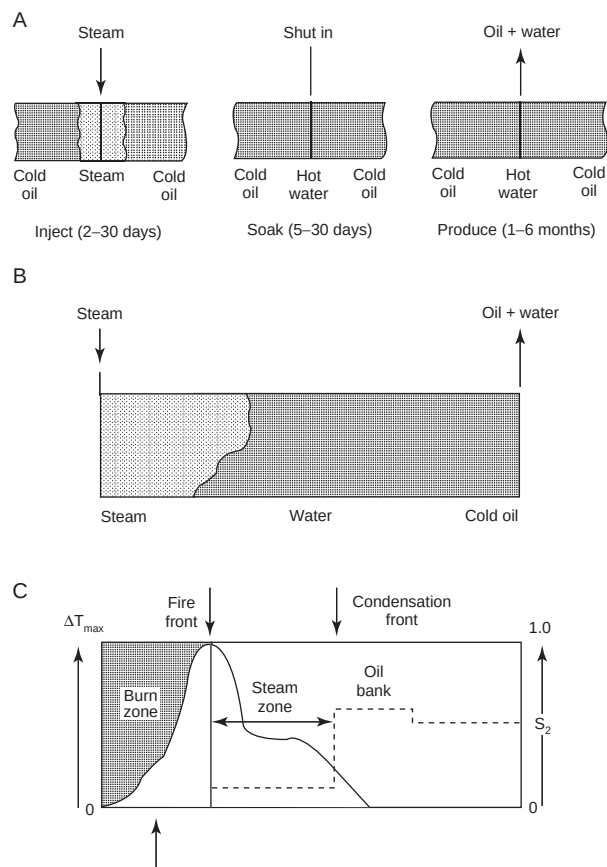


FIGURE 8 Thermal methods for improved recovery: (A) steam soak (or huff n' puff), (B) steam drive, and (C) *in situ* combustion. The oil saturation is denoted as S_2 . From Lake 1989, *Enhanced Oil Recovery*.

oil. In the zone immediately ahead of the hot combustion zone, the temperature is increased significantly and reaches a plateau based on the condensation of water at the reservoir pressure.

Recovery is increased primarily due to two main factors. A steam zone develops in front of the combustion zone, causing a steam flood of the oil. A small amount of carbon dioxide is also formed, increasing vaporization and swelling of the oil. *In situ* combustion is not widely used today primarily due to the economic risk involved in burning the oil and keeping the combustion front ignited. However, air injection as an immiscible gas flood has been reported to accidentally cause *in situ* combustion, and this process may occur more often than is realized.

6.7 Steam Flooding

In contrast to *in situ* combustion, steam flooding is a method that relies on the external injection of steam

or hot water into the reservoir. Two commonly used forms of steam injection are the steam drive, in which steam is continuously injected, and steam soak or cyclic steam injection, which involves repeated injection of steam and water (Fig. 8). Cyclic steam injection is also sometimes known as “huff n’ puff.” Steam aids oil recovery by reduction in the oil viscosity, reduction in the residual oil saturation, thermal expansion effects, and steam distillation. Steam flooding is very common for heavy oil reservoirs in which steam causes a significant reduction in the viscosity of the heavy oils. Steam flooding is currently the most used EOR technology and has enjoyed significant commercial success.

SEE ALSO THE FOLLOWING ARTICLES

Crude Oil Spills, Environmental Impact of • Oil and Natural Gas Drilling • Oil and Natural Gas Exploration • Oil and Natural Gas: Offshore Operations • Oil Pipelines • Oil Refining and Products • Petroleum System: Nature’s Distribution System for Oil and Gas

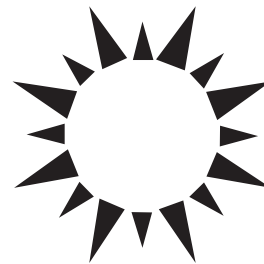
Further Reading

- Bradley, H. B. (ed.). (1987). “Petroleum Engineering Handbook.” Society of Petroleum Engineers, Richardson, TX.
- British Petroleum. (2003). “Statistical Review of World Energy.” BP, London.
- Craft, B. C., and Hawkins, M. F. (revised by R. E. Terry) (1991). “Applied Petroleum Reservoir Engineering,” 2nd ed. Prentice Hall, Englewood Cliffs, NJ.
- Craig, F. F., Jr. (1971). “The Reservoir Engineering Aspects of Waterflooding,” SPE Monograph Series, vol. 3. Society of Petroleum Engineers, Richardson, TX.
- Firoozabadi, A. (1999). “Thermodynamics of Hydrocarbon Reservoirs.” McGraw-Hill, New York.
- Jarrell, P. M., Fox, C. E., Stein, M. H., and Webb, S. L. (2002). “Practical Aspects of CO₂ Flooding,” SPE Monograph Series, vol. 22. Society of Petroleum Engineers, Richardson, TX.
- Lake, L. W. (1989). “Enhanced Oil Recovery.” Prentice Hall, Englewood Cliffs, NJ.
- McCain, W. D., Jr. (1990). “The Properties of Petroleum Fluids.” 2nd ed. PennWell Publishing, Tulsa, OK.
- Pederson, K. S., Fredenslund, A., and Thomassen, P. (1989). Properties of Oils and Natural Gases, Contributions in Petroleum Geology and Engineering, vol. 5. Gulf Publishing, Houston, TX.
- Petroleum Extension Service. (1983). “Oil and Gas: The Production Story.” University of Texas at Austin, Austin, TX.
- Petroleum Extension Service. (2001). “A Primer on Oil Well Drilling,” 6th ed. University of Texas at Austin, Austin, TX.
- Stalkup, F. I., Jr. (1984). “Miscible Displacement,” SPE Monograph Series. Society of Petroleum Engineers, Richardson, TX.

Oil Refining and Products

ABDULLAH M. AITANI

King Fahd University of Petroleum and Minerals
Dhahran, Saudi Arabia



1. Overview of Refinery Processing
2. Crude Oil and Products
3. Light Oil Processing
4. Heavy Distillate Processing
5. Residual Oil Processing
6. Treating Processes
7. Environmental and Future Issues

Glossary

alkylation A process using sulfuric acid or hydrofluoric acid as a catalyst to combine light olefins and isobutane to produce a high-octane product known as alkylate.

°API gravity A scale of liquid specific gravity (SG) that indicates the lightness or heaviness of hydrocarbons, defined by $[(141.5/SG) - 131.5]$.

catalytic cracking A process for the breaking-up of heavier hydrocarbons into lighter hydrocarbon fractions by the use of heat and catalysts.

cetane number A measure of ignition quality for kerosene, diesel, and heating oil, using a single-cylinder engine.

coking A process for thermally converting and upgrading heavy residues into lighter products and by-product petroleum coke.

crude oil A complex mixture of hydrocarbons containing low percentages of sulfur, nitrogen, and oxygen compounds and trace quantities of many other elements.

deasphalting A process for removing asphaltic materials from reduced crude, using liquid propane to dissolve nonasphaltic compounds.

hydrocracking A process used to convert heavier feedstock into lower boiling point, higher value products. The process employs high pressure, high temperature, a catalyst, and hydrogen.

hydrodesulfurization A catalytic process for the removal of sulfur compounds from hydrocarbons using hydrogen.

isomerization A catalytic process for the conversion and skeletal rearrangement of straight-chain hydrocarbons into branched-chain molecules of higher octane number.

methyl tertiary butyl ether (MTBE) An ether added to gasoline to raise octane number and enhance combustion.

octane number A measure of resistance to knocking of gasoline under laboratory conditions that simulate city driving conditions.

olefins Unsaturated hydrocarbons, such as ethylene and propylene, that have a double carbon bond, with the molecular formula C_nH_{2n} .

paraffins Saturated aliphatic hydrocarbons with the molecular formula C_nH_{2n+2} .

reforming A process for the transformation of naphtha into products with higher octane number. Reforming comprises isomerization, cracking, polymerization, and dehydrogenation.

visbreaking A low-temperature cracking process used to reduce the viscosity or pour point of straight-run residues.

This article discusses the various aspects of petroleum refining and oil products as a primary energy source and as a valuable feedstock for petrochemicals. The main objective of refining is to convert crude oils of various origins and different compositions into valuable products and fuels having the qualities and quantities demanded by the market. Various refining processes, such as separation, conversion, finishing, and environmental protection, are presented and briefly discussed. The ever-changing demand and quality of fuels, as well as environmental issues and the challenges facing the refining industry, are also highlighted. Environmental regulations have played a significant role in the progress of the refining industry and may even change the competition between petroleum and other alternative energy sources.

1. OVERVIEW OF REFINERY PROCESSING

1.1 Introduction

Refining is the processing of crude oil into a number of useful hydrocarbon products. Processing utilizes

chemicals, catalysts, heat, and pressure to separate and combine the basic types of hydrocarbon molecules naturally found in crude oil into groups of similar molecules. The refining process also rearranges their structures and bonding patterns into different hydrocarbon molecules and compounds. Therefore, it is the type of hydrocarbon (paraffinic, naphthenic, or aromatic) and its demand that configure the refining industry. Petroleum refining has evolved continuously in response to changing demands for better and different products. The trend in demand has also been accompanied by continuous improvement in product quality, such as octane number for gasoline and cetane number for diesel.

The original requirement was to produce kerosene for household use, followed by the development of the internal combustion engine and the production of transportation fuels (gasoline, diesel, and fuels). Refineries produce a variety of products including many required as feedstocks for the petrochemical industry. Early refining was the simple distillation (fractionation) of crude oil followed by the development in the 1920s of the thermal cracking processes, such as visbreaking and coking. The processes crack heavy fuels into more useful and desirable products by applying pressure and heat.

In the early 1940s, the catalytic processes were developed to meet the increasing demand of gasoline and higher octane numbers. Processes such as catalytic cracking, alkylation, isomerization, hydrocracking, and reforming were developed throughout the 1960s to increase gasoline yields and improve antiknock characteristics. Some of these processes also produced valuable feedstocks for the modern petrochemical industry. In the 21st century, the refinery uses an array of various catalytic and noncatalytic processes to meet new product specifications and to convert less desirable fractions into more valuable liquid fuels, petrochemical feedstocks, and electricity. The refinery has shifted from only physical separations to something close to a chemical plant.

1.2 Refinery Operations

Modern refineries incorporate fractionation, conversion, treatment, and blending operations and may also include petrochemical processing. Most light distillates are further converted into more usable products by changing the size and structure of the hydrocarbon molecules through cracking, reforming, and other conversion processes, as discussed further in this article. Figure 1 presents a typical scheme of a high-conversion refinery.

Various streams are subjected to various separation processes, such as extraction, hydrotreating, and sweetening, to remove undesirable constituents and improve product quality. In general, petroleum refining operations can be grouped as follows:

- Fractionation (distillation) is the separation of crude oil in atmospheric and vacuum distillation towers into groups of hydrocarbon compounds of differing boiling-point ranges called “fractions” or “cuts.”
- Light oil processing prepares light distillates through rearrangement of molecules using isomerization and catalytic reforming or combination processes such as alkylation and polymerization.
- Heavy oil processing changes the size and/or structure of hydrocarbon molecules through thermal or catalytic cracking processes.
- Treatment and environmental protection processes involve chemical or physical separation, such as dissolving, absorption, or precipitation, using a variety and combination of processes including drying, solvent refining, and sweetening.

1.3 World Refining

World refining capacity reached 81.9 million barrels/day in 2002 and is expected to increase by 4.3% per year to 115 million barrels/day by 2020. Table I presents a regional distribution capacity of major refining processes. There are 722 refineries in 116 countries, with more than 200 refineries in Asia-Pacific regions alone. The United States has maintained its leading position as the largest and most sophisticated oil-refining region. Approximately 25% of worldwide refining capacity is located in North America and another 25% is located in Asia followed by 17% in Western Europe. Remaining regions process approximately 33% of the world's crude oil in medium-type conversion refining schemes. World capacities of various refining processes are presented in Table II. Hydrotreating alone represents approximately 50% of total crude capacity, whereas all catalytic processes represent approximately 82% of crude capacity.

In general, the refining industry has always been characterized as a high-volume, low-profit-margin industry. World refining continues to be challenged by uncertainty of supply, difficult market conditions, government regulation, availability of capital, and slow growth. Although shipping of refined products has been increasing over the years, a close connection remains between domestic markets and domestic

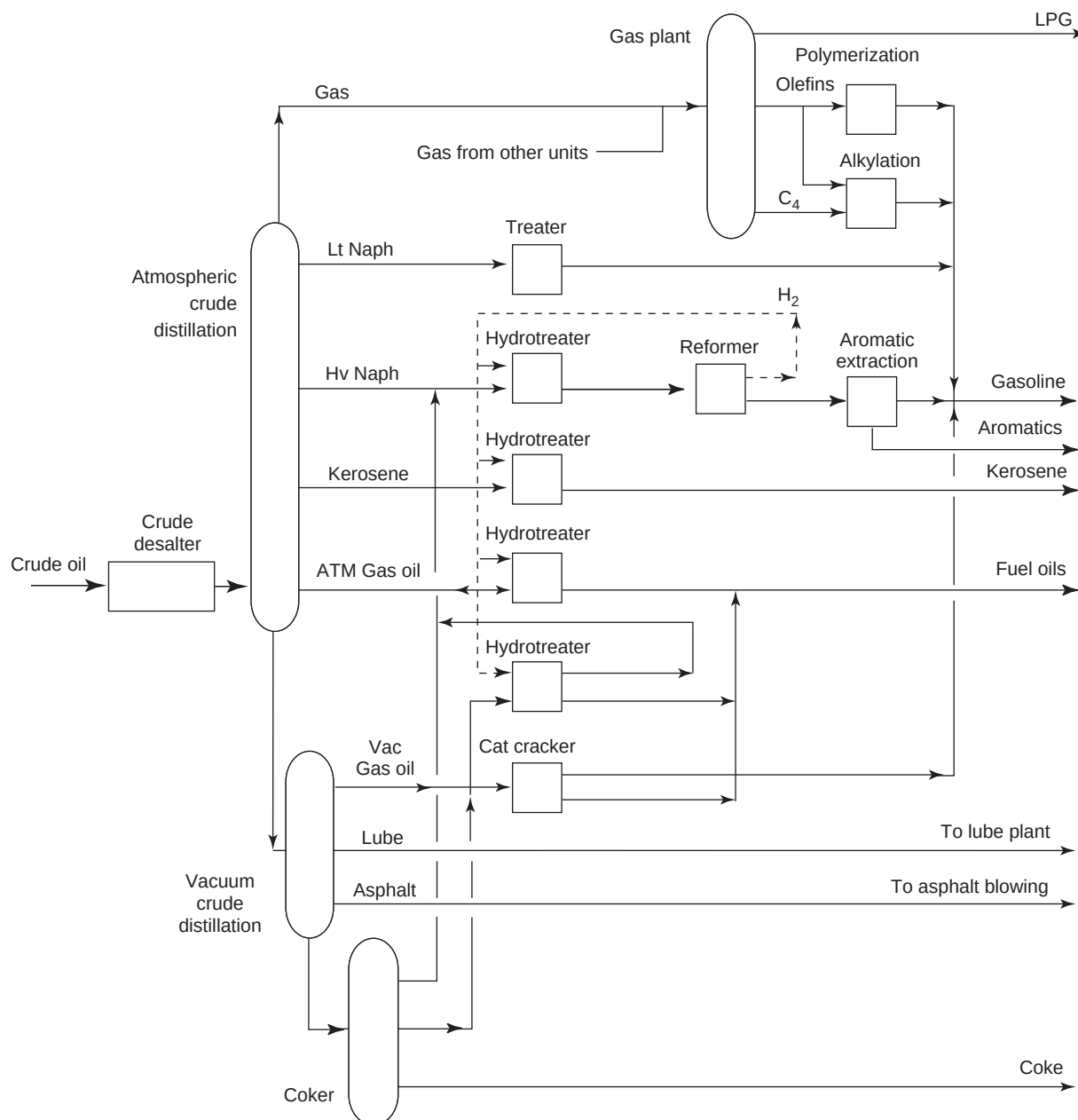


FIGURE 1 Schematic diagram of a high-conversion refinery. Cat cracker, catalytic cracker; Hv Naph, heavy naphtha; Lt Naph, light naphtha; LPG, liquefied petroleum gas; Vac gas oil, vacuum gas oil; ATM gas oil, atmospheric gas oil.

production. This explains the large differences in refinery schemes from one country to another and from one region to another.

2. CRUDE OIL AND PRODUCTS

2.1 Type and Composition of Crude Oils

Crude oil is a mixture of hydrocarbon compounds such as paraffins, naphthenes, and aromatics plus

small amounts of organic compounds of sulfur, oxygen, and nitrogen, in addition to small amounts of metallic compounds of vanadium, nickel, and sodium. Although the concentration of nonhydrocarbon compounds is very small, their influence on catalytic petroleum processing is large. There are a large number of individual components in a specific crude oil, reaching approximately 350 hydrocarbons and approximately 200 sulfur compounds. A specific crude oil can comprise a very large number of

TABLE I
Regional Distribution of Refining Operations Worldwide

Region	No. of refineries	Regional capacity	
		million barrels/day	%
North America	160	20.3	24.8
South America	69	6.7	8.2
Western Europe	105	14.6	17.8
Eastern Europe	95	10.6	12.9
Asia	202	20.2	24.7
Middle East	46	6.3	7.7
Africa	45	3.2	3.9
Total	722	81.9	100

TABLE II
World Refining Processes and Their Share of Crude Oil Capacity

Process	Capacity (million barrels/day)	% of crude capacity
Vacuum distillation	26.7	32.6
Coking	4.2	5.1
Visbreaking	3.7	4.5
Catalytic cracking	14.2	17.3
Naphtha reforming	11.2	13.7
Hydrocracking	4.4	5.4
Hydrotreating	38.4	46.9
Alkylation	1.9	2.3
Polymerization	0.2	0.2
Aromatics production	1.2	1.5
Isomerization	1.5	1.8
Oxygenates	0.3	0.4

compounds that are not easily identifiable or quantifiable. Most of these compounds have a carbon number less than 16 and these form a relatively high proportion of crude oil.

The elemental composition of crude oils depends on the type and origin of the crude; however, these elements vary within narrow limits. The proportions of these elements in a typical crude are 84.5% carbon, 13% hydrogen, 1–3% sulfur, and less than 1% each of nitrogen, oxygen, metals, and salts. The physical properties of crude oils vary within a wide range. Crude oils are defined in terms of API (American Petroleum Institute) gravity: the higher the API gravity, the lighter the crude. Crude oils with low carbon, high hydrogen, and high API gravity are usually rich in paraffins and tend to yield greater

proportions of gasoline and light petroleum products; those with high carbon, low hydrogen, and low API gravities are usually rich in aromatics.

Crude oils can be classified in many different ways, generally based on their density (API), sulfur content, and hydrocarbon composition. Condensate ranks highest, with densities reaching more than 50° API, whereas densities of heavy crudes may reach as low as 10° API. In general, refinery crude base stocks consist of mixtures of two or more different crude oils. There are more than 600 different commercial crude oils traded worldwide. In 2002, world oil production reached 66 million barrels/day, 40% of which is produced by members of the Organization of Petroleum Exporting Countries. Despite all energy alternatives, crude oil will remain the world's primary energy source, constituting approximately 40% of world energy up to the year 2020.

2.2 Crude Oil Processing

As a first step in the refining process, water, inorganic salts, suspended solids, and water-soluble trace metal contaminants are removed by desalting using chemical or electrostatic separation. This process is usually considered a part of the crude distillation unit. The desalted crude is continuously drawn from the top of the settling tanks and sent to the crude fractionation unit. Distillation of crude oil into straight-run cuts occurs in atmospheric and vacuum towers. The main fractions obtained have specific boiling-point ranges and can be classified in order of decreasing volatility into gases, light distillates, middle distillates, gas oils, and residue. The composition of the products is directly related to the characteristics of the crude processed. Desalted crude is processed in a vertical distillation column at pressures slightly above atmospheric and at temperatures ranging from 345 to 370°C (heating above these temperatures may cause undesirable thermal cracking). In order to further distill the residue from atmospheric distillation at higher temperatures, reduced pressure is required to prevent thermal cracking. Vacuum distillation resembles atmospheric distillation except that larger diameter columns are used to maintain comparable vapor velocities at the reduced pressures.

2.3 Transportation Fuels

Major oil products are mainly transportation fuels that represent approximately 52% of total worldwide oil consumption. Gasoline and diesel are the main concern, along with a large number of special

petroleum-based products. Crude oil type and refining configuration determine the quantity and quality of oil products. Tables III and IV present some typical data on the volume of atmospheric distillate cuts and refined products derived from the processing of Arabian light crude oil. For example, approximately 80–85 vol% of the refined products produced in a medium-type conversion refinery have a boiling temperature lower than 345°C compared to an amount of 55 vol% found in Arabian crude oil (see Tables III and IV). Almost half of the products are gasoline and lighter distillates. The demand for transportation fuels and petrochemical feedstocks has been increasing steadily compared to the decreasing demand for heating fuels and residual fuel oil, which are being replaced by natural gas.

2.3.1 Gasoline

Motor gasoline is the highest volume refinery product, having a mixture of hydrocarbons with boiling points ranging from ambient temperature to approximately 205°C. It flows easily, spreads quickly, and may evaporate completely in a few hours under temperate conditions. It is highly volatile and flammable. Gasoline is made up of different refinery streams, mainly straight-run naphtha, isomerized C₅/C₆ paraffins, reformat, hydrocracking, fluid catalytic cracking (FCC) gasoline, oligomerate, alkylate, and ethers. The most environmentally friendly gasoline comes from branched paraffins. The im-

TABLE III

Atmospheric Distillates Derived from Arabian Light Crude Oil

Product name	Boiling point (°C)	Volume (%)
Light naphtha	10–90	8
Heavy naphtha	90–200	21
Kerosene	200–260	11
Gas oil	260–345	15
Residue	345 +	45

TABLE IV

Typical Refined Products Derived from Arabian Light Crude Oil (Medium-Conversion Refinery)

Product name	Carbon no.	Boiling point (°C)	Volume (%)
Gases	1–4	–40–40	5
Gasoline	5–10	40–200	45
Kerosene/jet	10–16	150–260	5
Fuel oil	20–70	200–345	25
Residue	> 70	> 345	20

portant qualities for gasoline are octane number (antiknock), volatility (starting and vapor lock), vapor pressure, and sulfur content (environmental control). Additives are often used to enhance gasoline performance and to provide protection against oxidation and rust formation. Table V presents some typical data for current and future specifications for gasoline in Europe.

2.3.2 Diesel Fuel

Diesel fuel is usually second in volume next to gasoline. Diesel blend consists of cuts from atmospheric distillation, hydrocracking, FCC light cycle oil, and some products obtained from visbreaking and coking. The main property of diesel fuel for automotive engine combustion is cetane number, which is a measure of engine start-up and combustion. Diesel fuel and domestic heating oils have boiling point ranges of approximately 200–370°C. The desired properties of these distillate fuels include controlled flash and pour points, clean burning, and no deposit formation in storage tanks. Sulfur reduction and cetane improvement have been extensively investigated to produce ultralow-sulfur diesel (ULSD). Meeting future specifications for ULSD of 10–15 ppm sulfur will require significant hydrotreating investment. Table VI presents some typical data for current and future specifications for diesel fuel in Europe.

2.3.3 Jet Fuel (Kerosene)

Jet fuel is the third most important transportation fuel. It is a middle-distillate product that is used for jets (commercial and military) and is used around the world in cooking and heating (kerosene). When used as a jet fuel, some of the critical qualities are freeze

TABLE V

Current and Future Specifications for Gasoline

Specification ^a	Year	
	2000	2005
Sulfur (ppm)	150	30–50
Benzene content (vol%, maximum)	1	1
Aromatics (vol%, maximum)	42	35
Olefins (vol%, maximum)	18	15
Oxygen (wt%, maximum)	2.7	2.3
RVP (kPa)	60	50
RON/MON (minimum)	95/85	95/85

^aRVP, Reid vapor pressure; RON, research octane number; MON, motor octane number.

TABLE VI
Current and Future Specifications for Diesel Fuel

Specification	Year	
	2000	2005
Sulfur (ppm)	350	50
Specific gravity (maximum)	0.845	0.825
API (minimum)	36	40
Cetane number	51	54
Distillation T_{95} ($^{\circ}\text{C}$, maximum)	360	360
Polycyclic aromatic hydrocarbons (PAH wt%, maximum)	11	1

point, flash point, and smoke point. Commercial jet fuel has a boiling point range of approximately 190–275 $^{\circ}\text{C}$ and that of military jet fuel is 55–285 $^{\circ}\text{C}$. Kerosene, with less critical specifications, is used for lighting, heating, solvents, and blending into diesel fuel. *n*-Paraffins in the range C_{12} – C_{14} may be extracted from kerosene for use in the production of detergents.

2.4 Other Oil Products

2.4.1 Refinery Gases

Refinery gases are the lightest hydrocarbons containing a mixture of gases from methane to liquefied petroleum gas (LPG) and some pentanes. The gases are processed to separate LPG, which consists principally of propane and butane. Other refinery gases include lighter paraffins, unsaturates, and hydrogen sulfide. LPG is used as fuel and as an intermediate in the manufacture of olefins and selected petrochemical feedstocks. Butanes are also used in the manufacture of ethers and to adjust the vapor pressure of gasoline. LPG is also used in transportation and in domestic and household applications.

2.4.2 Petrochemical Feedstocks

Light olefins from FCC and benzene, toluene, xylenes (BTX) aromatics from naphtha reforming are the main petrochemical feedstocks derived from refineries. These products are the basis for integrating refining and petrochemical operations. Olefins include propylene and butanes, whereas benzene and xylenes are precursors for many valuable chemicals and intermediates, such as styrene and polyesters.

2.4.3 Residual Fuel Oil

Residual fuel oil is the least valuable of the refiner's products, selling at a price below that of crude oil.

Residual fuels are difficult to pump and may be heavier than water; they are also difficult to disperse and are likely to form tar balls, lumps, and emulsions. Many marine vessels, power plants, commercial buildings, and industrial facilities use residual fuels or combinations of residual and distillate fuels for heating and processing. The two most critical properties of residual fuels are viscosity and low sulfur content for environmental control.

2.4.4 Coke and Asphalt

Petroleum coke is produced in coking units and is almost pure carbon with a variety of uses from electrodes to charcoal briquets. Asphalt is a semisolid material produced from vacuum distillation and is classified into various commercial grades. It is used mainly for paving roads and roofing materials.

2.4.5 Lubricants

Vacuum distillation and special refining processes produce lube-oil base stocks. Additives such as antioxidants and viscosity enhancers are blended into the base stocks to provide the characteristics required for motor oils, industrial greases, lubricants, and cutting oils. The most critical quality is a high viscosity index, which provides for greater consistency under varying temperatures.

3. LIGHT OIL PROCESSING

3.1 Catalytic Naphtha Reforming

Catalytic naphtha reforming combines a catalyst, hardware, and processing to produce high-octane reformate for gasoline blending or BTX aromatics for petrochemical feedstocks. Reformers are also the source of much needed hydrogen for hydroprocessing operations. Reforming reactions comprise cracking, polymerization, dehydrogenation, and isomerization, which take place simultaneously. Universal Oil Products (UOP) and Axens-Institut Français du Pétrole (IFP) are the two major licensors and catalyst suppliers for catalytic naphtha reforming. Reforming processes differ in the mode of operation [semiregenerative or continuous catalyst regenerative (CCR)], catalyst type, and process engineering design. All licensors agree on the necessity of hydrotreating the feed to remove permanent reforming catalyst poisons and to reduce the temporary catalyst poisons to low levels. There are more than 700 reformers worldwide, with a total capacity of approximately 11.2 million barrels/day. Approximately 40% of this capacity is located in

North America, followed by 20% each in Western Europe and Asia-Pacific regions.

3.1.1 Reforming Processes

Reforming processes are generally classified into semiregenerative, cyclic, and CCR. This classification is based on the frequency and mode of regeneration. The semiregenerative process requires unit shutdown for catalyst regeneration, whereas the cyclic process utilizes a swing reactor for regeneration in addition to regular reactors. The continuous process permits catalyst replacement during normal operation. Globally, the semiregenerative scheme dominates reforming capacity at approximately 57%, followed by the continuous regenerative process at 27% and the cyclic process at 11%. Most grassroots reformers are designed to use continuous catalyst regeneration.

The semiregenerative process is a conventional reforming process that operates continuously over a period of up to 1 year. Conversion is kept more or less constant by raising the reactor temperatures as catalyst activity declines. The cyclic process typically uses five or six fixed catalyst beds, similar to the semiregenerative process, with one additional swing reactor, which is a spare reactor. CCR is characterized by high catalyst activity with reduced catalyst requirements, more uniform reformat of higher aromatic content, and high hydrogen purity. Figure 2 presents a schematic diagram of a CCR process. The continuous process represents a step change in reforming technology compared to semiregenerative and cyclic processes.

3.1.2 Reforming Catalysts

Since the 1950s, commercial reforming catalysts have been essentially heterogeneous monometallic

compounds and are composed of a base support material (usually chlorided alumina) on which platinum metal was placed. These catalysts were capable of producing high-octane products; however, because they quickly deactivated as a result of coke formation, they required high-pressure, low-octane operations. In the early 1970s, bimetallic catalysts were introduced to meet increasing severity requirements. Platinum and another metal (often rhenium, tin, or iridium) account for most commercial bimetallic reforming catalysts. The catalyst is most often presented as 1/16, 1/8, or 1/4 in. Al_2O_3 cylindrical extrudates or beads, into which platinum and other metal have been deposited. In commercial catalysts, platinum concentration ranges between 0.3 and 0.7 wt% and chloride is added (0.1–1.0 wt%) to the alumina support (η or γ) to provide acidity.

3.2 Isomerization

Isomerization is an intermediate, fed preparation-type process. There are more than 200 units worldwide, with a processing capacity of 1.5 million barrels/day of light paraffins. Two types of units exist: C_4 isomerization and C_5/C_6 isomerization. A C_4 unit will convert normal butane into isobutane, to provide additional feedstock for alkylation units, whereas a C_5/C_6 unit will isomerize mixtures of C_5/C_6 paraffins, saturate benzene, and remove naphthenes.

Isomerization is similar to catalytic reforming in that the hydrocarbon molecules are rearranged, but unlike catalytic reforming, isomerization just converts normal paraffins to isoparaffins. The greater value of branched paraffins over straight paraffins is a result of their higher octane contribution. The

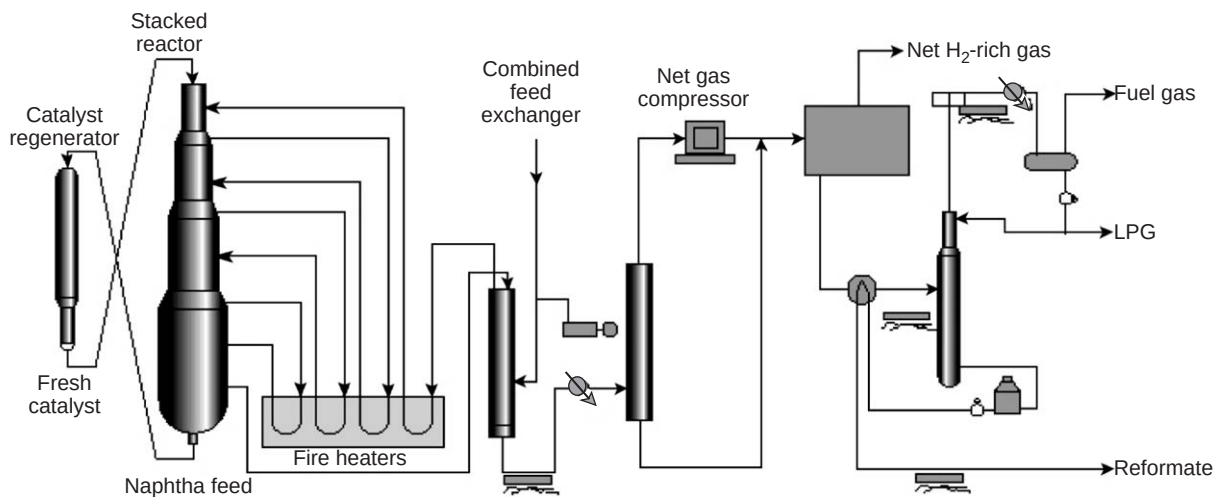


FIGURE 2 Flow scheme of a continuous catalyst regenerative naphtha reformer. LPG, liquefied petroleum gas.

formation of isobutane is a necessary step to produce alkylate gasoline or methyl tertiary butyl ether (MTBE). The extent of paraffin isomerization is limited by a temperature-dependent thermodynamic equilibrium. For these reactions, a more active catalyst permits a lower reaction temperature and that leads to higher equilibrium levels. Isomerization of paraffins takes place under medium pressure (typically 30 bar) in a hydrogen atmosphere.

C₄ isomerization produces isobutane feedstock for alkylation. Platinum or another metal catalyst, alumina chloride, is used for the higher temperature processes. In a typical low-temperature process where only aluminum chloride is used, the feed to the isomerization unit is *n*-butane or mixed butanes combined with hydrogen (to inhibit olefin formation). C₅/C₆ isomerization increases the octane number of the light gasoline components *n*-pentane and *n*-hexane, which are found in abundance in straight-run gasoline. The basic C₅/C₆ isomerization process is essentially the same as butane isomerization.

3.3 Alkylation

Alkylation is the process that produces gasoline-range compounds from the combination of light C₃–C₅ olefins (mainly a mixture of propylene and butylene) with isobutene. The highly exothermic reaction is carried out in the presence of a strong acid catalyst, either sulfuric acid (H₂SO₄) or hydrofluoric acid (HF). The world alkylation capacity stands at 1.9 million barrels/day and new grassroots units have been constructed in many refineries worldwide, especially those with FCC units. The alkylate product is composed of a mixture of high-octane, branched-chain paraffinic hydrocarbons. Alkylate is a premium clean gasoline blend, with octane number depending on the type of feedstocks and operating conditions. Research efforts are directed toward the development of environmentally acceptable solid superacids capable of replacing HF and H₂SO₄. Much of the work is concentrated on sulfonated zirconia catalysts.

3.3.1 Sulfuric Acid Process

In H₂SO₄-based alkylation units, the feedstock (propylene, butylene, amylene, and fresh isobutane) enters the reactor and contacts the concentrated sulfuric acid catalyst (in concentrations of 85–95%). The reactor effluent is separated into hydrocarbon and acid phases in a settler and the acid is returned to the reactor. The hydrocarbon phase is hot water washed with a caustic compound for pH control

before being successively depropanized, deisobutanized, and debutanized. The alkylate obtained from the deisobutanizer can then go directly to motor gasoline blending.

3.3.2 Hydrofluoric Acid Process

In a typical HF process, olefin and isobutane feedstock are dried and fed to a combination reactor/settler system. The process is operated at temperatures attainable by cooling water and at higher pressures to keep fluid in the liquid state. The reactor effluent flows to a separating vessel, where the acid separates from the hydrocarbons. The acid layer at the bottom of the separating vessel is recycled. Propane with a trace amount of HF goes to an HF stripper for HF removal and is then defluorinated, treated, and sent to storage. Isobutane is recycled to the reactor/settler and the alkylate is sent to gasoline blending.

3.4 Etherification

Etherification results from the selective reaction of methanol or ethanol to isobutene. The ether products, such as MTBE or other oxygenates, are used as components in gasoline because of their high octane blending value. The refinery capacity of oxygenate units is approximately 266,000 barrels/day, with almost all units associated with alkylation processes. The exothermic reaction is conducted in liquid phase at 85–90°C over a highly acidic ion-exchange polystyrene resin catalyst. The reaction is very rapid and equilibrium is limited under typical reaction conditions. A combination of catalytic distillation is applied to remove the product as vapor, thereby driving the reaction to almost 100% conversion. The etherification process is needed to supply oxygenates to meet the specifications of reformulated gasoline (minimum 2.7 wt% oxygenate content). In general, MTBE is the preferred oxygenate because of its low production cost and convenient preparation route relative to those of other ethers.

3.5 Polymerization and Dimerization

Catalytic polymerization and dimerization in petroleum refining refer to the conversion of FCC light olefins, such as ethylene, propylene, and butenes, into higher octane hydrocarbons for gasoline blending. Polymerization combines two or more identical olefin molecules to form a single molecule, with the same elements being present in the same proportions as in the original molecules. Light olefin feedstock is pretreated to remove sulfur and other undesirable

compounds. In the catalytic process, the feedstock either is passed over a solid phosphoric acid catalyst on silica or comes into contact with liquid phosphoric acid, where an exothermic polymeric reaction occurs. Another process uses a homogenous catalyst system of aluminum-alkyl and a nickel coordination complex. The hydrocarbon phase is separated, stabilized, and fractionated into LPG and oligomers or dimers.

4. HEAVY DISTILLATE PROCESSING

4.1 Catalytic Hydrotreating

Catalytic hydrotreating is a hydrogenation process used to remove approximately 90% of contaminants, such as nitrogen, sulfur, oxygen, and metals, from liquid petroleum fractions. These contaminants, if not removed from the petroleum fractions as they travel through the refinery processing units, can have detrimental effects on the equipment, the catalysts, and the quality of the finished product. In addition, hydrotreating converts olefins and aromatics to saturated compounds. World capacity of all types of hydrotreating stands at approximately 38.3 million barrels/day. Hydrotreating is used to pretreat catalytic reformer feeds, saturate aromatics in naphtha, desulfurize kerosene/jet and diesel, saturate aromatics, and pretreat catalytic cracker feeds. It also includes heavy gas oil and residue hydrotreating as well as posthydrotreating of FCC naphtha.

Hydrotreating for sulfur or nitrogen removal is called hydrodesulfurization or hydrodenitrogenation. Hydrotreating processes differ depending on the feedstock available and the catalysts used. Mild hydrotreating is used to remove sulfur and saturate olefins. More severe hydrotreating removes nitrogen and additional sulfur and saturates aromatics. In a typical catalytic hydrotreater unit, the feedstock is mixed with hydrogen, preheated in a fired heater (315–425°C), and then charged under pressure (up to 68 atm) through a fixed-bed catalytic reactor. In the reactor, the sulfur and nitrogen compounds in the feedstock are converted into H_2S and NH_3 . Hydrodesulfurized products are blended or used as catalytic reforming feedstock. Hydrotreating catalysts contain cobalt or molybdenum oxides supported on alumina and less often nickel and tungsten.

4.2 Fluid Catalytic Cracking

Catalytic cracking is the largest refining process for gasoline production, with a global capacity of more

than 14.2 million barrels/day. The process converts heavy feedstocks such as vacuum distillates, residues, and deasphalted oil into lighter products that are rich in olefins and aromatics. There are several commercial FCC processes that are employed in world refineries with major differences in the method of catalyst handling. FCC catalysts are typically solid acids of fine particles, especially zeolites (synthetic Y-faujasite), aluminum silicate, treated clay (kaolin), bauxite, and silica-alumina. Zeolite content in commercial FCC catalysts is generally in the range of 5–20 wt%, whereas the balance is a silica-alumina amorphous matrix. Additives to the FCC process make up no more than 5% of the catalyst and they are basically used as octane enhancers, metal passivators, and SO_x reducing agents and are used in CO oxidation and for gasoline sulfur reduction.

The FCC unit comprises a reaction section, product fractionation, and a regeneration section. In principle, the reactor (riser) and the regenerator form the catalyst circulation unit in which the fluidized catalyst is continuously circulated using air, oil vapors, and steam as the conveying media. Figure 3 presents a schematic of a typical FCC process. The operating temperatures of the FCC unit range from 500 to 550°C at low pressures. Hydrocarbon feed temperatures range from 260 to 425°C, whereas regenerator exit temperatures for hot catalyst range from 650 to 815°C. Several operating parameters, mainly temperature, affect overall conversion and it is essential to determine which product slate is needed so that process conditions are appropriately selected.

4.2.1 Reaction Section (Riser)

A typical FCC unit involves mixing a preheated hydrocarbon charge with hot, regenerated catalyst as it enters the riser leading to the reactor. Major process variables are temperature, pressure, catalyst/oil ratio, and space velocity. Hydrocarbon feed is combined with a recycle stream within the riser, vaporized, and raised to reactor temperature by the hot regenerated catalyst. As the mixture moves up the riser, the charge is cracked at approximately 110 kPa and the residence time is on the order of 1 s. In modern FCC units, almost all cracking occurs within the riser. The cracking continues until the oil vapors are separated from the catalyst in the reactor cyclones.

Cracking reactions are endothermic; the energy balance is obtained by the burning of catalyst-deposited coke in the regenerator. Both primary and secondary reactions take place during catalytic

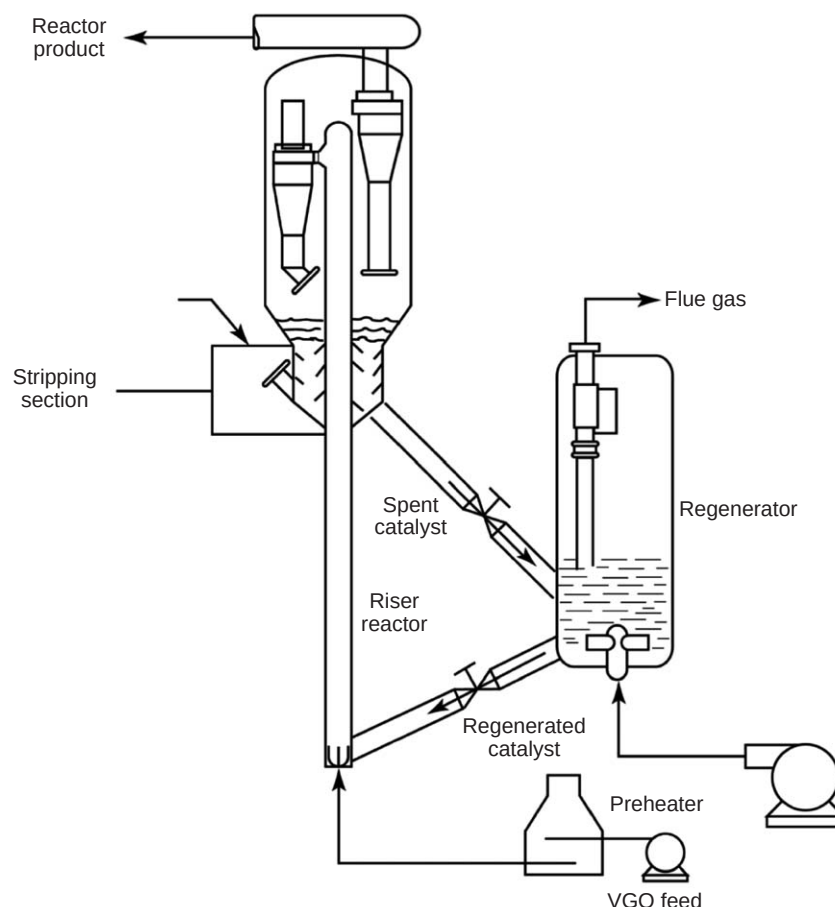


FIGURE 3 Schematic diagram of fluid catalytic cracking process. VGO feed, vacuum gas oil feed.

cracking. Primary reactions are the result of the cracking of paraffins, alkyl naphthenes, and alkyl aromatics. In general, all cracking reactions are characterized by the production of appreciable amounts of corresponding olefins. During the reactions, however, approximately 40% of the sulfur in the FCC feed is converted to H_2S , which is easily removed. Much of the ongoing research is directed to the removal of the remaining sulfur in FCC gasoline.

4.2.2 Product Fractionation

Cracked hydrocarbons are separated into various products. The resultant product stream from the reaction section is charged to a fractionating column, where it is separated into fractions, and some of the heavy oil is recycled to the riser. The main FCC products are LPG, the gasoline fraction, and light cycle oil. By-products include refinery gases, residue (slurry), and coke. Since the FCC unit is the major source of olefins in the refinery (for the downstream alkylation

unit or petrochemical feedstock), an unsaturated gas plant is generally considered a part of it.

4.2.3 Regeneration Section

Spent FCC catalyst is regenerated by burning off deposited coke to carbon dioxide. The catalyst flows through the catalyst stripper to the regenerator, where most of the coke deposits burn off in the presence of preheated air. The carbon content of the regenerated catalyst is generally kept at the lowest level to achieve selectivity benefits. Catalyst circulation and coke yield determine the temperature at which the regenerator is operated. Maximum regenerator temperatures are limited by mechanical specifications or sometimes by catalyst stability. The temperature in the regenerator reaches almost 650°C due to the exothermic nature of coke-burning reactions. Spent catalyst is continuously removed and fresh catalyst is added as makeup to optimize the cracking process. This added catalyst is in effect the main determinant of catalyst activity. The typical

catalyst makeup requirement is approximately 0.1 kg per barrel of total feed.

4.3 Hydrocracking

Catalytic hydrocracking of heavy petroleum cuts is an important process for the production of gasoline, jet fuel, and light gas oils. Some hydrocracking processes also allow the production of a highly purified residue, which can be an excellent base for oils. The process employs high pressure, high temperature, a catalyst, and hydrogen. In contrast to FCC, the advantage of hydrocracking is that middle distillates, jet fuels, and gas oils of very good quality are provided. In general, hydrocracking is more effective in converting gas oils to lighter products, but it is more expensive to carry out.

Hydrocracking is used for feedstocks that are difficult to process by either catalytic cracking or reforming, since these feedstocks are usually characterized by a high polycyclic aromatic content and/or high concentrations of the two principal catalyst poisons, sulfur and nitrogen compounds. These feedstocks include heavy gas oils, FCC cycle oils, deasphalted oil, and visbreaker or coke gas oil. The process depends largely on the nature of the feedstock and the relative rates of the two competing reactions, hydrogenation and cracking. Heavy aromatic feedstock is converted into lighter products under a wide range of very high pressures (70–140 atm) and fairly high temperatures (400–820°C), in the presence of hydrogen and special catalysts.

4.3.1 Hydrocracking Process

Hydrocracking is a two-stage process combining catalytic cracking and hydrogenation, wherein heavier feedstocks are cracked in the presence of hydrogen. The reaction typically involves a reactor section, gas separator, scrubber for sulfur removal, and product fractionator. The reactor section contains a multi-catalyst bed that can be of the fixed-bed or ebullated-bed type and some employ on-stream catalyst addition and withdrawal to maintain catalyst activity.

4.3.2 Hydrocracking Catalysts

The catalysts used in hydrocracking are all of the bifunctional type, combining an acid function and a hydrogenating function. The acid function is carried by supports with a large surface area and having a superficial acidity, such as halogenated aluminas, zeolites, amorphous silica-aluminas, and clays. The hydrogenating function is carried either by one or more transition metals, such as iron, cobalt, nickel, ruthenium, rhodium, palladium, osmium, iridium,

and platinum, or by a combination of molybdenum and tungsten. The conventional catalysts of catalytic hydrocracking are made up of weak acid supports. These systems are more particularly used to produce middle distillates of very good quality and also, if their acidity is very weak, oil bases. Amorphous silica-aluminas serve as supports with low acidity. These systems have a very good selectivity in middle distillates and the products formed are of good quality. The low-acid catalysts among these can also produce lubricant bases.

5. RESIDUAL OIL PROCESSING

5.1 Solvent Deasphalting

Solvent deasphalting (SDA) is a separation process that represents a further step in the minimization of residual fuel. Figure 4 presents a schematic diagram of a typical SDA process. The process takes advantage of the fact that maltenes are more soluble in light paraffinic solvents than asphaltenes. This solubility increases with solvent molecular weight and decreases with temperature. There are constraints with respect to how deep an SDA unit can cut into the residue or how much deasphalted oil (DAO) can be produced. These constraints are related to the DAO quality specifications required by downstream conversion units and the final high-sulfur residual fuel oil stability and quality.

SDA has the advantage of being a relatively low-cost process that has the flexibility to meet a wide range of DAO qualities. The process has very good selectivity for asphaltenes and metals rejection, some selectivity for carbon rejection, and less selectivity for sulfur and nitrogen. It is most suitable for the more paraffinic vacuum residues as opposed to the high-asphaltene-, high-metal-, high-concarbon-containing vacuum residues. The disadvantages of the process are that it performs no conversion, produces a very high viscosity by-product pitch, and where high-quality DAO is required, SDA is limited in the quality of feedstock that can be economically processed.

5.2 Visbreaking

Visbreaking is the most widespread process for noncatalytic mild conversion of residues, with a world capacity of 3.7 million barrels/day. The process is designed to reduce the viscosity of atmospheric or vacuum residues by thermal cracking. It produces 15–20% of atmospheric distillates with proportionate reduction in the production of

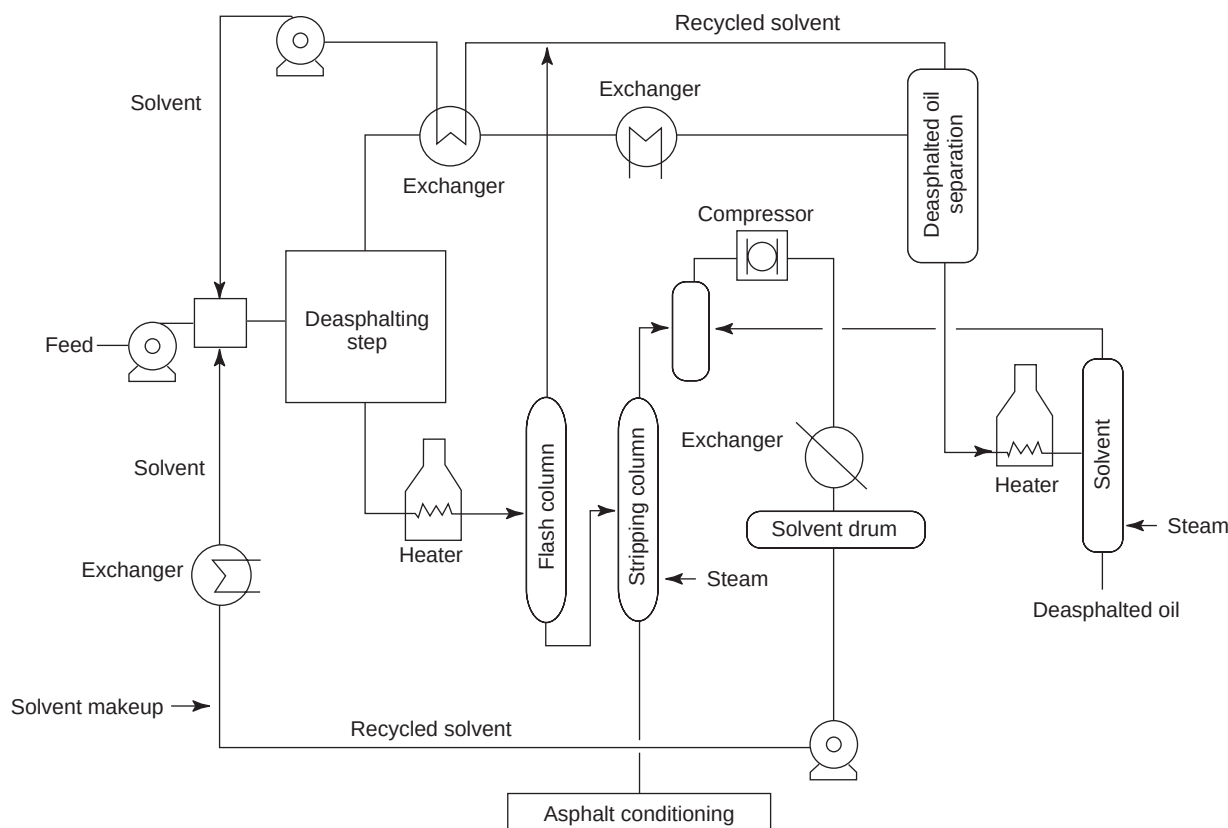


FIGURE 4 Schematic diagram of a typical solvent deasphalting process.

residual fuel oil. Visbreaking reduces the quantity of cutter stock required to meet fuel oil specifications and, depending on fuel oil sulfur specifications, typically reduces the overall quantity of fuel oil produced by 20%. In general, visbreakers are typically used to produce vacuum residues.

The process is available in two schemes: coil cracker and soaker cracker. The coil cracker operates at high temperatures during a short residence time of approximately 1 min. The soaker scheme uses a soaking drum at 30–40°C at approximately 10–20 min residence time. The residue is rapidly heated in a furnace and then cracked for a specific residence time in a soaking zone under proper conditions of pressure and temperature. The soaking zone may be within the heater or in a separate adiabatic soaking drum. The cracked residue leaves the soaking zone after the desired degree of reaction is reached and is quenched with gas oil to stop the reaction and prevent coking.

5.3 Coking

Approximately 90% of coke production comes from delayed coking. The process is one of the preferred

thermal cracking schemes for residue upgrading in many refineries, mainly in the United States. The process provides essentially complete rejection of metals and concarbon while providing partial or complete conversion to naphtha and diesel. World capacity of coking units is 4.2 million barrels/day (approximately 54% of this capacity is in U.S. refineries) and total coke production is approximately 172,000 tons/day. New cokers are designed to minimize coke and produce a heavy coker gas oil that is catalytically upgraded. The yield slate for a delayed coker can be varied to meet a refiner's objectives through the selection of operating parameters. Coke yield and the conversion of heavy coker gas oil are reduced, as the operating pressure and recycle are reduced and, to a lesser extent, as temperature is increased.

5.4 Residue Hydrotreating and Residue FCC

Refineries that have a substantial capacity for visbreaking, solvent deasphalting, or coking are faced with large quantities of visbreaker tar, asphalt or

coke, respectively. These residues have high viscosity and high organic sulfur content (4–6 wt%), with primary consequences reflected in the potential for sulfur emissions and the design requirements for a sulfur removal system. Sulfur content is also important from the standpoint of corrosion, which requires proper selection of design materials and operating conditions. Other properties of residues include their high heating value due to the high level of fixed carbon that results in a higher yield of syngas per ton of residue processed. Moreover, residues have low volatile matter and ash content as well as little to no oxygen content, resulting in low reactivity.

Residue hydrotreating is another method for reducing high-sulfur residual fuel oil yields. Atmospheric and vacuum residue desulfurization units are commonly operated to desulfurize the residue as a preparatory measure for feeding low-sulfur vacuum gas-oil feed to cracking units (FCC and hydrocrackers), low-sulfur residue feed to delayed coker units, and low-sulfur fuel oil to power stations. Two different types of processing units are used for the direct hydrotreating of residues. These units are either a down-flow, trickle phase reactor system (fixed catalyst bed) or a liquid recycle and back-mixing system (ebullating bed). Economics generally tend to limit residue hydrotreating applications to feedstocks containing less than 250 ppm nickel and vanadium.

Residue FCC (RFCC) is a well-established approach for converting a significant portion of the heavier fractions of the crude barrel into a high-octane gasoline-blending component. In addition to high gasoline yields, the RFCC unit produces gaseous, distillate, and fuel oil-range products. The RFCC unit's product quality is directly affected by its feedstock quality. In particular, unlike hydrotreating, RFCC redistributes sulfur, but does not remove it from the products. Consequently, tightening product specifications have forced refiners to hydrotreat some, or all, of the RFCC's products. Similarly, in the future, the SO_x emissions from an RFCC may become more of an obstacle for residue conversion projects. For these reasons, a point can be reached where the RFCC's profitability can economically justify hydrotreating the RFCC's feedstock.

6. TREATING PROCESSES

6.1 Hydrogen Production

Refineries are experiencing a substantial increase in hydrogen requirements to improve product quality and process heavy sour crudes. Hydroprocessing and

saturation of aromatics and olefins will accelerate the demand for hydrogen within the refinery. Catalytic naphtha reforming alone is not able to meet refinery hydrogen requirements. A survey on world refining indicated that the capacity of supplementary refinery hydrogen, produced mainly by steam reforming of methane, reached 337 million m^3/day (11,880 million ft^3/day -MMcfd) in 2002 compared to 110 million m^3/day in 1990. There is a growing recognition that there will be a significant future shortage of refinery hydrogen supply. Specific hydrogen production units, such as steam methane reformers or those carrying out partial oxidation of heavy residues, will have to be built.

6.2 Residue Gasification

The gasification of refinery residues into clean syngas provides an alternative route for the production of hydrogen and the generation of electricity in a combined turbine and steam cycle. Compared to steam-methane reforming, gasification of residues can be a viable process for refinery hydrogen production when natural gas price is in the range of \$3.75–4.00/million British thermal units (MMBtu). The largest application of syngas production is in the generation of electricity power by the integrated gasification combined cycle (IGCC) process. Consumption of electricity in the modern conversion refinery is increasing and the need for additional power capacity is quite common, as is the need to replace old capacity. The design of a residue gasification plant requires a careful matching and integration of the various process steps to ensure optimum performance of the whole system. In general, the IGCC plant consists of several steps: gasification, gas desulfurization, and a combined cycle. The technologies of the gasification and the combined cycle are well known; the innovation, however, is their integration in order to maximize the overall IGCC efficiency.

6.3 Aromatics Extraction

BTX aromatics are high-value petrochemical feedstocks produced by catalytic naphtha reforming and extracted from the reformate stream. Whether or not other aromatics are recovered, it is sometimes necessary to remove benzene from the reformate in order to meet mandated specifications on gasoline composition. Aromatics production in refineries reached 1.2 million barrels/day in 2002. Most new aromatic complexes are configured to maximize the yield of benzene and paraxylene and sometimes orthoxylene. The solvents used in the extraction of

aromatics include dimethylformamide, formylmorpholine, dimethylsulfoxide, sulfolane, and ethylene glycols.

6.4 Sweetening

Sweetening is the removal of contaminants such as organic compounds containing sulfur, nitrogen, and oxygen; dissolved metals and inorganic salts; and soluble salts dissolved in emulsified water from petroleum fractions or streams. A variety of intermediate and finished products, including middle distillates, gasoline, kerosene, jet fuel, and sour gases, are sweetened. Treating can be accomplished at an intermediate stage in the refining process or just before the finished product is sent to storage. Choices of a treating method depend on the nature of the petroleum fractions, the amount and type of impurities in the fractions to be treated, the extent to which the process removes the impurities, and end-product specifications. Treatment materials include acids, solvents, alkalis, oxidizing agents, and adsorption agents.

6.5 Sulfur Recovery

Sulfur recovery converts hydrogen sulfide in sour gases and hydrocarbon streams to elemental sulfur. Total sulfur production in world refineries reached approximately 64,000 tons/day in 2002 compared to approximately 28,000 tons/day in 1996, corresponding to a yearly growing recovery rate of 20%. In other words, in 2002 an average refinery recovered 0.8 kg sulfur from one processed barrel of crude oil compared to less than 0.4 kg sulfur recovered in 1996. This indicates the increasing severity of operations to meet stringent environmental requirements. The most widely used sulfur recovery system is the Claus process, which uses both thermal and catalytic conversion reactions. A typical process produces elemental sulfur by burning hydrogen sulfide under controlled conditions. Knockout pots are used to remove water and hydrocarbons from feed gas streams. The gases are then exposed to a catalyst to recover additional sulfur. Sulfur vapor from burning and conversion is condensed and recovered.

6.6 Acid Gas Removal

Amine plants remove acid contaminants from sour gas and hydrocarbon streams. In amine plants, gas and liquid hydrocarbon streams containing carbon dioxide and/or hydrogen sulfide are charged to a gas absorption tower or liquid contactor, where the acid contaminants are absorbed by counterflow amine solutions [i.e., monoethanol amine (MEA), diethanol

amine (DEA), methyl diethanol amine (MDEA)]. The stripped gas or liquid is removed overhead and the amine is sent to a regenerator. In the regenerator, the acidic components are stripped by heat and reboiling and are disposed of, and the amine is recycled.

7. ENVIRONMENTAL AND FUTURE ISSUES

7.1 Environmental Issues

Refiners are faced with many environmental, economic, and operational issues. Environmental legislation is a growing concern, driving changes in product specifications, product markets, and refinery operating practices. Strict product quality specifications and severe emission and discharge limits have economic impact on the average refiner. In the near future, the following environmental trends will continue to grow, but they will not create significant changes in oil consumption patterns: (1) the production of clean transportation fuels according to new specifications and (2) refinery operation within strict emissions regulations.

The configuration of many refineries has changed substantially, mainly due to the declining quality of crude oil supply and environmental regulations. In retrospect, refinery changes brought about by the variations in crude supply and composition were evolutionary, whereas environmental regulations were revolutionary.

7.1.1 Clean Transportation Fuels

Since 1990, government agencies have imposed strict environmental restrictions on transportation fuels to improve product quality specifications. Fuel reformulation is being discussed all over the world. Automotive manufacturers are demanding lower gasoline sulfur levels and lower driveability indices. Refiners must improve air quality by delivering clean products that minimize emissions of toxic and hazardous hydrocarbons. Gasoline and diesel formulations have been already changed in many countries and will change even more in the coming years.

Refining is faced with huge investments to meet new stringent specifications for sulfur, aromatics, and olefin content. Gasoline sulfur reduction is centered around the FCC unit employing feed pretreatment or gasoline posttreatment. The reduced demand for ethers, such as MTBE in gasoline for oxygenate content, necessitates the utilization of branched paraffin isomer products of alkylation and isomerization. For diesel fuel, this means a sulfur content

less than 30 or even 15 ppm, an increase of the cetane number, a reduction in polyaromatic content, and T_{95} point limitation. To fulfill all these legislative and regional requirements, refiners must either revamp existing units or invest in new hydroprocessing and hydrogen production units.

7.1.2 Refinery Emissions

Refiners must comply with various environmental regulations to reduce all types of pollutants in their waste gas as well as wastewater systems. Most concerns involve the emissions of SO_x , NO_x , CO, hydrocarbons, and particulates. The oxides are present in flue gases from furnaces, boilers, and FCC regenerators. Tail gas treatment and selective catalytic reduction units are being added to limit SO_2 and NO_x emissions. Water pollutants include oil, phenol, sulfur, ammonia, chlorides, and heavy metals. New biological processes can be used to convert H_2S or SO_x from gaseous and aqueous streams. Spent catalysts and sludges are also of concern to the refinery in reducing pollution. Some spent FCC catalysts can be used in cement but other catalysts that contain heavy metals need special treatment before proper disposal.

7.2 Future Refining Issues

World refining has been adapting to ongoing product changes and environmental challenges. Transportation fuels with approximately free sulfur will be needed to satisfy the demand of the automotive industry to reduce emissions from internal combustion engines. There will be an increased demand for alkylate and isomerate gasoline as well as deep-desulfurized diesel. This will increase the hydrogen content in gasoline, enhance combustion, and reduce the levels of carbon dioxide emissions.

The introduction of fuel cells as a feasible way to fuel zero-emission vehicles is a major challenge to oil companies and refiners. Virtually every major automotive manufacturer has a fuel-cell program and most claim production readiness by 2005. Refiners need to adapt to this technology in the future, especially regarding new fuels needed for fuel cells. Fuel-cell vehicles need hydrogen generated on-board or carried in either compressed or liquid form. The latter calls for a global hydrogen infrastructure. The use of hydrocarbons and specifically gasoline to generate hydrogen offers many economic advantages

such as the availability of a ready-made global fueling infrastructure.

The huge technological challenges associated with the transfer to a hydrogen economy necessitate an efficient and better use of hydrocarbon resources to compete with renewable energy sources. Refiners need to enhance and integrate their business with chemical production and power generation. In the long run, the refinery will produce not just fuels, but also chemicals and electricity.

SEE ALSO THE FOLLOWING ARTICLES

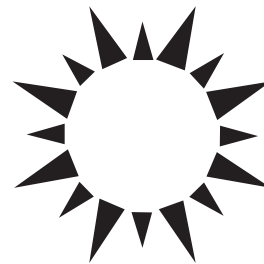
Coal Preparation • Oil and Natural Gas Drilling • Oil and Natural Gas Exploration • Oil and Natural Gas: Offshore Operations • Oil Pipelines • Oil Recovery • Petroleum System: Nature's Distribution System for Oil and Gas

Further Reading

- Aitani, A. (1995). Reforming processes. In "Catalytic Naphtha Reforming" (G. Antos *et al.*, Eds.), pp. 409–436. Dekker, New York.
- Farrauto, R., and Bartholomew, C. (1997). "Fundamentals of Industrial Catalytic Processes," pp. 519–579. Blackie Academic and Professional, London.
- Gary, J., and Handwerk, G. (2001). "Petroleum Refining Technology and Economics." 4th ed. Dekker, New York.
- Heinrich, G. (1995). Introduction to refining. In "Petroleum Refining" (J. P. Wauquier, Ed.), pp. 365–413. Editions Technip, Paris.
- Hoffman, H. (1992). Petroleum and its products. In "Riegel's Handbook of Industrial Chemistry" (J. Kent, Ed.), 9th ed., pp. 490–496. Van Nostrand Reinhold, New York.
- Le Page, J. P., Chatila, S., and Davidson, M. (1992). "Resid and Heavy Oil Processing." Editions Technip, Paris.
- Maples, R. (2000). "Petroleum Refinery Process Economics." 2nd ed. PennWell Books, Tulsa, OK.
- Martino, G., and Wechem, H. (2002). "Current Status and Future Developments in Catalytic Technologies Related to Refining and Petrochemistry," Review and Forecast Paper, 17th World Petroleum Congress, Rio de Janeiro, Brazil, September 2002.
- Meyers, R. (1997). "Handbook of Petroleum Refining Processes." 2nd ed. McGraw-Hill, New York.
- Penning, T. (2001). "Petroleum Refining: A Look at the Future, Hydrocarbon Processing," February, pp. 45–46.
- Silvy, R. (2002). Global refining catalyst industry will achieve strong recovery by 2005, Oil & Gas Journal, pp. 48–56. September 2, 2002.
- Speight, J., and Ozum, B. (2002). "Petroleum Refining Processes." Dekker, New York.
- Stell, J. (2002). Worldwide refining survey. Oil & Gas Journal, December 23, 2002, pp. 68–70.

Oil Sands and Heavy Oil

WORLD ENERGY COUNCIL
London, United Kingdom



-
1. History
 2. The Resource Base
 3. Extraction and Conversion Technology
 4. Economic Issues
 5. Environmental Issues
 6. Country Profiles

Glossary

estimated additional reserves The amount, expressed as tonnage of recoverable synthetic crude oil or raw bitumen (additional to the proved recoverable reserves), that is of foreseeable economic interest. Speculative amounts are not included.

natural bitumen Comprises bitumen or other petroleum with very high viscosity (contained in bituminous sands, oil sands, or tar sands), which is not recoverable by conventional means; the petroleum is obtained either as raw bitumen (through *in situ* recovery) or as a synthetic crude oil (via an integrated surface mining plus upgrading process).

proved amount in place The tonnage of natural bitumen that has been carefully measured and assessed as exploitable under present and expected local economic conditions with existing available technology.

proved recoverable reserves The tonnage of synthetic crude oil or raw bitumen that has been carefully measured and assessed as recoverable under present and expected local economic conditions with existing available technology.

Oil sands are deposits of bitumen, viscous oil that must be treated in order to convert it into an upgraded crude oil before it can be used in refineries to produce gasoline and other fuels. Natural bitumen and heavy oil are closely related types of petroleum, differing from each other, and from the petroleum from which they are derived, only to the degree by which they have been degraded. This alteration, through bacterial attack, water washing, and inspis-

sations, has resulted in severe loss of the light ends of the petroleum, notably the paraffins, and subsequent relative enrichment of the heavy molecules, leading to increased density and viscosity. Of these molecules, the asphaltenes are very large and incorporate nonhydrocarbons such as nitrogen, sulfur, oxygen, and metals, in particular nickel and vanadium. The result of this chemistry is an array of problems beyond those encountered with conventional petroleum with respect to exploitation, transportation, storage, and refining. This, of course, is reflected in the increased cost of extraction and processing and the physical limitations on production capacity.

1. HISTORY

The Bible has numerous references to the use of pitch, which is asphalt or bitumen in its soft state. It also is called “slime” in some translations of the Bible. Noah waterproofs the Ark with pitch: “And God said to Noah, ‘... Make yourself an ark of Gopher Wood; make rooms in the ark, and cover it inside and out with pitch.’” (Genesis 6:13–18 RSV). The reed basket that carries the infant Moses into the Nile River is waterproofed with pitch: “... And when she could hide him no longer she took for him a basket made of bulrushes, and daubed it with bitumen and pitch; and she put the child in it and placed it among the reeds at the river’s brink.” (Exodus 2:1–4 RSV)

Indeed, throughout ancient times, the Elamites, Chaldeans, Akkadians, and Sumerians mined bitumen. Mesopotamian bitumen was exported to Egypt, where it was used for various purposes, including the preservation of mummies. The Dead Sea was known as Lake Asphaltites (from which the term asphalt was derived) because of the lumps of semisolid petroleum that were washed up on its shores from underwater seeps. Bitumen was used as mortar for bricks, as caulking for ships, in road building, for

bonding tools, in inlaid work and jewel setting, and for waterproofing baskets and mats. Artistic and religious objects were carved from bitumen-impregnated sands.

The fur trader Peter Pond is the first European to describe the vast oil sands (tar sands) of Athabasca, Canada. In 1788, he wrote: "At about 24 miles from the fork (of the Athabasca and Clearwater Rivers) are some bituminous fountains into which a pole of 20 feet long may be inserted without the least resistance. The bitumen is in a fluid state and when mixed with gum, the resinous substance collected from the spruce fir, it serves to gum the Indians' canoes. In its heated state it emits a smell like that of sea coal." In 1913, Sidney Ells, an engineer for the Canadian Federal Department of Mines, advocates the hot water flotation method of separating bitumen from the tar sands of Athabasca. He is the first to bring out samples for laboratory testing as a road paving material. Oil sands cannot compete economically with imported asphalt and the project is dropped. His is one of several unsuccessful attempts to develop the Athabasca sands. Decades later, oil sands will re-emerge as a viable energy source. In 1967, the first barrel was shipped from the Great Canadian Oil Sands Project, the world's first commercially successful operation to tap the rich Athabasca oil sands in Alberta, Canada. The project,

owned by the Sun Oil Company (later Suncor Energy), ushered in an era of rapid development of the oil sand resource base.

2. THE RESOURCE BASE

Table I lists estimates of the magnitude of oil sand and heavy oil resource base. It is estimated that there are 1.7 trillion barrels of oil in the oil sands of Canada and that 15% (255 billion barrels) of the total oil in place is recoverable. Canada accounts for approximately 75% of the world's oil sand resources. Other countries and regions that have significant, but more modest, resources include the United States, China, Eastern Europe and the former Soviet Union, the Caribbean Basin, and Pakistan (see Fig. 1). Approximately 700,000 barrels per day of Canadian oil sands are being produced. This supply is divided into two categories: "oil sands *in situ*" (often referred to as bitumen) and "oil sands mining." These two categories reflect the method of recovery. The bitumen is extracted by injecting very hot steam into the rock formation to heat the oil, lower its viscosity, and allow it to flow more like conventional oil. Slightly more than one-half (approximately 400,000 barrels per day) of Canadian oil sands production is derived from the more expensive oil sands mining method.

TABLE I

Natural Bitumen: Resources, Reserves, and Production in 1999

	Recovery method	Millions of tons			
		Proved amount in place	Proved recoverable reserves	Estimated additional reserves	Production in 1999
North America					
Canada		45,300	979		30.0
	Surface	9000	747		15.9
	<i>In situ</i>	36,300	232		14.1
United States of America	<i>In situ</i>			4231	
South America					
Venezuela	<i>In situ</i>	3880	373	118	3.3
Europe					
Romania	<i>In situ</i>	4	1		
Middle East					
Jordan	Surface	40	5		

Note. The data shown above are those reported by World Energy Council Member Committees in 2000/2001. They thus constitute a sample, reflecting the information available in particular countries; they should not be considered as complete, or necessarily representative of the situation in each region. For this reason, regional and global aggregates have not been computed. World Energy Council Member Committees have been unable to provide sufficient numerical data for extra-heavy oil to be included in this table.

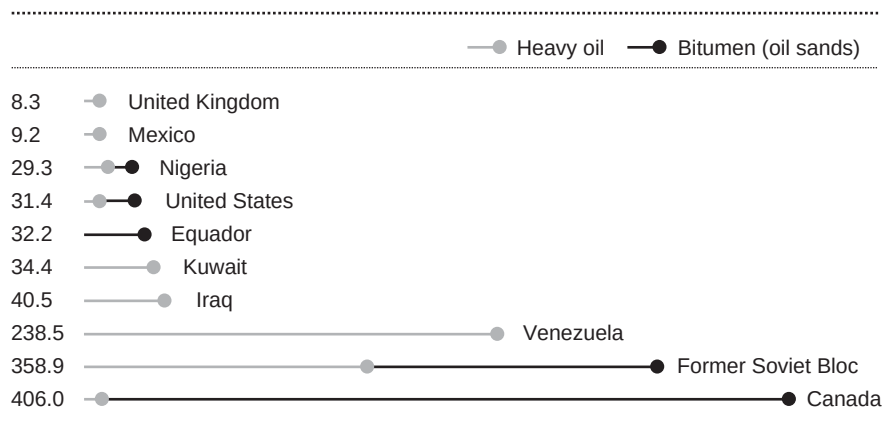


FIGURE 1 World oil sands (bitumen) and heavy oil resources, in billions of cubic meters. Reprinted from Petroleum Communication Foundation, Alberta, Canada, with permission.

Those deposits that are close enough to the surface are actually mined.

The heavy crude oil deposit of the Orinoco Oil Belt, a part of the Eastern Venezuela basin, represents nearly 90% of the known heavy oil in place. Between them, these two deposits, each located up-dip against a continental craton, represent approximately 3600 billion barrels of oil in place. This is only the remaining, degraded remnant of petroleum deposits once totaling as much as 18,000 billion barrels.

Heavy oil is recorded in 219 separate deposits; some of these are different reservoirs in a single field, some are producing, and some are abandoned. The deposits are found in 30 countries and in 54 different geological basins, with 11 of the deposits located offshore and 5 partially offshore. The following data are average values. Most of the reservoirs are sandstone at depths of 5400 ft, thicknesses of 126 ft, porosities of 21%, and permeabilities of 1255–6160 millidarcies. The American Petroleum Institute (API) gravity is 8° and viscosity is 22,700 cP. The viscosity varies from 6000 at 70°F, to 4600 at 100°F, to 1400 at 130°F, and the gas/oil ratio is only 431. The chemical data for the whole crude demonstrate the processing difficulties with heavy crude. The Conradson Carbon is 11.5 wt%, asphaltenes are 16 wt%, sulfur is 4.69 wt%, nickel is 260 ppm, and vanadium is 972 wt%. Especially significant are the residuum yield of 62 vol% and its specific gravity of 1.06 and Conradson Carbon of 17.8 wt%. These data suggest for extra-heavy crude a content of 56 wt% asphalt and 23 wt% coke.

Oil resource data are very incomplete but those that are available for heavy crude, especially data from Canada, the United States, and Venezuela, are

as follows, all in millions of barrels: original oil in place, 2,133,912; cumulative production, 17,214; annual production, 432; reserves, 45,575; and probable reserves, 193,203.

3. EXTRACTION AND CONVERSION TECHNOLOGY

Oil sands recovery processes include extraction and separation systems to remove the bitumen from the sand and water. There are two basic methods of extraction. To depths of approximately 150 ft, the bitumen and rock may be surface-mined, with the bitumen subsequently separated from the rock by a hot water process. The mining operation begins with clearing trees and brush from the site and removing the overburden that sits above the oil sands deposit. This includes the topsoil, sand, clay, gravel, and muskeg (a swamp or bog in which there are many layers of decaying matter, often covered with sphagnum moss). The topsoil and muskeg are stockpiled so that they can be replaced as sections of the mined-out area are reclaimed. The rest of the overburden is used to reconstruct the landscape.

At the processing plant, the mixture of oil, sand, and water goes into a large separation vessel. A thick froth is skimmed off, mixed with a solvent, and spun in a centrifuge to remove water and dissolved salts from the bitumen. The solvent is recycled. The sand and water are waste products, known as tailings, that fall to the bottom of the separation vessel.

The sand is returned to the mine site to fill in mined-out areas. Water from the extraction process, containing sand, fine clay particles, and traces of

bitumen, goes into settling ponds. Some bitumen may be skimmed off the ponds if it floats to the surface. The sand sinks to the bottom and bacteria digest the remaining bitumen, but the fine clay particles stay suspended for some time before slowly settling. Water is recycled back to the extraction plant for use in the separation process.

For deeper deposits, *in situ* extraction is required. In cyclic steam stimulation, high-pressure steam is injected into the oil sands formation for several weeks. The heat softens the bitumen and the water vapor helps to dilute and separate the bitumen from the sand grains. The pressure also creates channels and cracks through which the bitumen can flow to the well. When a portion of the reservoir is thoroughly saturated, the steam is turned off and the reservoir “soaks” for several weeks. This is followed by the production phase, when the bitumen flows, or is pumped up, to the same wells to the surface. When production rates decline, another cycle of steam injection begins. This process is sometimes called “huff and puff” recovery.

A second *in situ* method is steam-assisted gravity drainage (SAGD), in which pairs of horizontal wells are drilled near the base of the bitumen deposit. Steam is injected into the injector well that is placed approximately 5 m above the producer well. The steam rises and heats the bitumen, which flows down under the force of gravity to the lower producer well from which it is pumped to the surface. The bitumen is either upgraded at the site, or at a regional upgrader, or mixed with diluent and shipped to a refinery. Several pilot projects have tested the SAGD process and several commercial-scale projects are in the construction, engineering design, and regulatory approval stages.

4. ECONOMIC ISSUES

How much does recovery from oil sands cost? Supply costs are expressed as “full cycle” costs. They include all costs associated with exploration, development, and production; capital costs; operating costs; taxes and royalties; and a 10% real rate of return to the producer. Capital costs average \$5 to \$9 per barrel and operating costs average \$8 to \$12 per barrel. Such costs are presented as a range, reflecting the variance in reservoir quality, depth, project size, and operating parameters. The remainder of the supply cost is dominated by the cleaning and upgrading methods that are required to turn a very low quality hydrocarbon into more conventional oil that can be

accepted by a refinery. Such methods include the removal of sulfur, heavy metals, and noncombustible materials, as well as conversion to a more hydrogenated and lighter hydrocarbon. These costs are typically in the \$3 to \$5 per barrel range. None of the aforementioned costs include transportation to market. Suncor Energy opened the upgrading units of its “Project Millennium” in Alberta with production costs of approximately \$9 per barrel. The company’s near-term goal is to lower production costs to \$5.50 per barrel, which would make Suncor the lowest cost oil producer in North America.

5. ENVIRONMENTAL ISSUES

The extraction and processing of oil sands require huge volumes of water, creating the potential for various forms of water pollution, including possible contamination of ground and surface water with dissolved solids and toxic metal compounds produced from processed materials. Processing operations also release various air pollutants, such as sulfur oxides, hydrogen sulfide, nitrogen oxides, particulate matter, and carbon dioxide. Mining and processing disturbs large amounts of land. Approximately 2 tons of oil sand must be extracted and processed to make one 159-liter barrel of crude oil.

6. COUNTRY PROFILES

6.1 Canada

The National Energy Board (NEB) distinguishes between two types of nonconventional oil obtained from deposits of oil sands, defining them as follows:

Bitumen (also known as crude bitumen) is “a highly viscous mixture, mainly of hydrocarbons heavier than pentanes. In its natural state, it is not usually recoverable at a commercial rate through a well.”

Upgraded crude oil (also known as synthetic crude) is “a mixture of hydrocarbons similar to light crude oil derived by upgrading oil sands bitumen.”

Canada’s “discovered recoverable resources” of oil sands bitumen are quoted by the NEB as 49 billion m³ (over 300 billion barrels), of which approximately 475 million m³ had been produced by the end of 1999. Of the remainder (shown as “proved amount in place” in Table I), 9650 million m³ (9 billion tons) consists of synthetic crude recoverable through mining projects and 38,850 million m³

(36.3 billion tons) consists of crude bitumen recoverable through *in situ* extraction.

Within these huge resources, the “remaining established reserves” at the end of 1999 (shown as “proved recoverable reserves” in Table I) have been assessed by the Canadian Association of Petroleum Producers (CAPP) as 799.9 million m³ (equivalent to approximately 747 million tons) of mining-integrated synthetic crude oil and 248.1 million m³ (approximately 232 million tons) of *in situ* bitumen.

The major deposits are in four geographic and geologic regions of Alberta: Athabasca, Wabasca, Cold Lake, and Peace River.

Although the existence of oil sands deposits was noted in the 18th century, it was not until 1875 that a complete survey was undertaken and it was only in the 20th century that exploitation was embarked on. The deposits range from being several hundred meters below ground to surface outcroppings. The extraction of bitumen from the oil sands was initially based on surface mining but *in situ* techniques became necessary in order to reach the deeper deposits.

There was much experimentation with oil sands technology in the first half of the 20th century but it was not until the effects of the economic climate of the 1950s and early 1960s began to be felt that commercial development became viable. The Government of Alberta’s oil sands development policy was announced in 1962 and the Great Canadian Oil Sands Project (GCOS) was conceived and approved. The ownership of GCOS passed to Sun Oil Company and in 1967 the world’s first integrated oil sands production and upgrading plant was started up by Suncor (formerly Sun Oil).

Suncor’s area of operation, 40 km north of Fort McMurray, is within the Athabasca deposits. The processing capability of the original Oil Sands Plant has steadily increased and the expansion of the Steepbank Mine (on the opposite side of the Athabasca River) resulted in record production of 105,600 barrels/day in 1999. At the beginning of 1999, the company announced its Project Millennium, a phased series of expansions to the Steepbank mine, adding bitumen extraction plants and increasing upgrader capacity. The first phase was expected to increase production to 130,000 barrels/day by 2001 and the second phase to 225,000 barrels/day in 2003. In 2000, the establishment (subject to the necessary approval) of an *in situ* project at Firebag (40 km northeast of the Oil Sands Plant) was announced. It is planned that Firebag, in conjunction with the open-pit mining operation, will result in

production reaching 260,000 barrels/day in 2004. Through a combination of mining and *in situ* development, Suncor envisages an oil sands production of 400,000–450,000 barrels/day in 2008.

Syncrude, a joint venture with 10 participants (Imperial Oil, a subsidiary of Exxon, is the majority shareholder with 25% of holdings), operates the Lake Mildred plant, also 40 km north of Fort McMurray. Production began in 1978 and, using open-pit mining methods, the shallow deposits are recovered for bitumen extraction and the production of upgraded crude oil. Gross production was 223,000 barrels/day in 1999. A new project, the Aurora mine, a 35 km extension from Lake Mildred, opened in August 2000. The mine’s output is partially processed on-site and then pipelined to the upgrader for further treatment. In 1999, a major expansion to Syncrude’s upgrading capacity was approved by the federal government; construction began in 2001. It is planned that the work under development will result in a capacity of approximately 350,000 barrels/day in 2004.

The Cold Lake oil sands deposits area is operated by Imperial Oil. The company began commercial development in 1983 and has since gradually expanded facilities; the total production of bitumen in 1999 was 132,000 barrels/day. Imperial Oil plans to bring further expansion on-stream so that as of late 2002, bitumen production would have increased by 30,000 barrels/day.

Commercial production of Shell Canada’s Peace River *in situ* deposits (northwestern Alberta) began in 1986. Bitumen production capacity is set at approximately 12,000 barrels/day although during the year 2000 the actual production from existing wells was considerably lower. In an attempt to boost declining bitumen production, Shell announced in late 2000 that it would drill 18 new wells.

Albian Sands Energy, a joint venture, has been created to build and operate the Muskeg River Mine on behalf of its owners: Shell Canada (the majority shareholder, with 60% of holdings), Chevron Canada (20% of holdings), and Western Oil Sands (20% of holdings). The mine, already under construction, is located 75 km north of Fort McMurray (Athabasca). In addition, a pipeline is to be constructed to link the mine to an upgrader to be built next to Shell’s Scotford refinery. The start-up of the project was scheduled for late 2002, with anticipated production of 155,000 barrels/day of bitumen.

Taking into account all operations, total output from Canadian oil sands in 1999 was 323,000 barrels/day of synthetic crude and 244,000 barrels/day of crude bitumen from the *in situ* plants; together

these represented 22% of Canada's total production of crude oil and natural gas liquids.

6.2 Trinidad and Tobago

The famous Pitch Lake at La Brea (named after the Spanish word for tar or pitch) was reputedly discovered at the end of the 16th century. Trinidad Lake Asphalt, a semisolid emulsion of soluble bitumen, mineral matter, and other minor constituents (mainly water), was mined and used as a road surfacing material as long ago as 1815.

The lake contains 10 million tons of reserves, which at the current rate of extraction are expected to last for another 400 years. Lake Asphalt of Trinidad and Tobago (1978) Ltd. (TLA), a state-owned company, produces between 10,000 and 15,000 tons per annum and exports most of this amount, after removal of water, etc.

In combination with bitumen (asphalt) from refined crude oil, the product has featured significantly in the road construction industry over a long period of time and in many countries. In addition to mining it, TLA distributes the natural bitumen and has incorporated it into a range of paints and coatings. The company has also developed a process for making cationic bitumen emulsions. Production of these emulsified bitumen, water, and soap solutions began in late 1996 and they are used widely throughout the industrialized world in place of solvent-based bitumen emulsions.

6.3 United States of America

Distillation of tar sands, occurring as a surface outcrop in California, was carried out in the 1870s. During the following century, efforts were periodically made to establish the industry in both California and various other states, but the availability of low-priced, indigenous conventional oil meant that there was never a persistently strong incentive for the development of tar sands deposits.

The United States classifies tar sands as being measured or demonstrated ("the bitumen resource based on core and log analyses") or as being speculative or inferred ("the bitumen that is presumed to exist from reported tar shows on drillers' lithological logs and/or geological interpretations").

The tar sands resource of 58.1 billion barrels (22.4 "measured" and 35.7 "speculative") is widely distributed throughout the country, with 33% located in Utah, 17% in Alaska, and the remaining 50% in California, Alabama, Kentucky, Texas, and other states. There are eight giant (>1 billion barrels)

deposits of natural asphalt *in situ*, which represent nearly 80% of the total U.S. demonstrated and inferred resources.

The geological conditions of the Utah deposits have meant that recovery is difficult and expensive. Likewise, the Texan deposits, mostly deep and relatively thin, have also proved difficult to recover. The only state where small volumes of tar sand hydrocarbons are being produced from subsurface deposits (associated with heavy oil) is California.

Gilsonite (a naturally occurring solid hydrocarbon) is being produced by three companies from a number of vertical veins in the Green River Formation and overlying Eocene Uinta Formation in Uintah County, eastern Utah. Production figures for the gilsonite district are not available, but probably total several hundred thousand tons per year. Gilsonite is used in a variety of specialty products, such as printing inks, paints, and protective coatings, drilling and foundry sand additives, and briquetting.

6.4 Venezuela

There are vast deposits of bitumen and extra-heavy oil in the Orinoco Oil Belt (OOB) in eastern Venezuela, north of the Orinoco river. The original-oil-in-place of the extra-heavy oil reservoirs of the OOB has been estimated at 1200 billion barrels, with some 270 billion barrels of oil recoverable. Venezuela's total proved reserves of crude oil (76.8 billion barrels at the end of 1999) include 35.7 billion barrels of extra-heavy crudes.

There are four joint ventures for the exploitation of extra-heavy crude. *Petróleos de Venezuela* (PDVSA), the state oil company, has a minority interest in all four and all are at different stages of development:

The Hamaca project (a joint venture between Phillips Petroleum, Texaco, and PDVSA) has been delayed, owing to financing problems, but is planned to produce 190,000 barrels/day.

The Sincor Project (a joint venture between TotalFinaElf, Statoil, and PDVSA) was reported to have started bitumen production in December 2000, with its upgrading plant having been scheduled to come on-stream a year later. The project is planned to produce 180,000 barrels/day.

Production from the Petrozuata project, a joint venture between Conoco and PDVSA, has begun and had reached its target of 120,000 barrels/day by February 2001. Work to enable production to increase to 150,000 barrels/day is under way. An upgrader processes the 120,000 barrels/day of 9° API

oil, turning it into 103,000 barrels/day of lighter, synthetic crude, some of which is used as refinery feedstock to obtain gasoline and diesel for the domestic and export markets. Beginning early in 2001, the remainder was shipped to the United States for processing into higher value products.

The Cerro Negro is a joint venture project between ExxonMobil, Veba, and PDVSA. Output was expected to rise from 60,000 barrels/day in 2000 to 120,000 barrels/day by March 2001, following the completion of a new coking unit.

In the early 1980s, Intevep, the research affiliate of the state oil company PDVSA, developed a method of utilizing some of the hitherto untouched potential of Venezuela's extra-heavy oil/natural bitumen resources. Natural bitumen (7.5°–8.5° API) extracted from the reservoir is emulsified with water (70% natural bitumen, 30% water, <1% surfactants), the resulting product being called Orimulsion. Orimulsion can be pumped, stored, transported, and burnt under boilers using conventional equipment with only minor modifications. Initial tests were conducted in Japan, Canada, and the United Kingdom and exports began in 1988.

Orimulsion is processed, shipped, and marketed by Bitúmenes del Orinoco S.A. (Bitor), a PDVSA subsidiary, but with the fuel's relatively high sulfur content and its emission of particulates, Intevep continues to seek improvements in its characteristics in order to match increasingly strict international environmental regulations. Bitor operates an Orimulsion plant at Morichal in Cerro Negro, with a capacity of 5.2 million tons per year. The company hopes to produce 20 million tons per year by 2006.

Following manufacture at the plant, the Orimulsion is transported by pipeline approximately 320 km to the José export terminal for shipment. During the 1990s, other markets were developed and Barbados, Brazil, Canada, China, Costa Rica, Denmark, Finland, Germany, Guatemala, Italy, Japan, Lithuania,

Northern Ireland, the Philippines, South Korea, Taiwan, Thailand, and Turkey either consume or are considering consuming the product. In 1999, 4.9 million tons of Orimulsion were exported, bringing the cumulative total to greater than 27 million tons.

In addition to being used in conventional power plants using steam turbines, Orimulsion can be used in diesel engines for power generation, in cement plants, as a feedstock for Integrated Gasification Combined Cycle, and as a "reburning" fuel (a method of reducing NO_x by staging combustion in the boiler).

Acknowledgments

This article is adapted from the chapter on Natural Bitumen and Extra-Heavy Oil in the "World Energy Council 2001 Survey of Energy Resources." Reprinted by permission of World Energy Council, London, United Kingdom.

SEE ALSO THE FOLLOWING ARTICLES

Gas Hydrates • Natural Gas Resources, Unconventional • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Recovery • Oil Shale

Further Reading

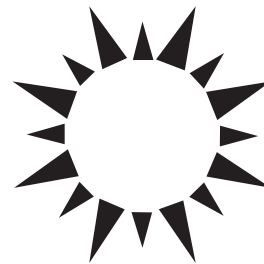
- Alberta Environment. (1999). "Guidelines for Reclamation to Forest Vegetation in the Alberta Oil Sands Region." Conservation and Reclamation Information Letter. C&R/IL/99-1. Calgary.
- Ferguson, B. G. (1986). "Athabasca Oil Sands: Northern Resource Exploration, 1875–1951." Canadian Plains Research Center Regina, Saskatchewan.
- Shih, S.S., and Oballa, M. C. (Eds.). (1991). "Tar Sand and Oil Upgrading Technology." American Institute of Chemical Engineers, New York.

Oil Shale

JOHN R. DYNI

U.S. Geological Survey

Denver, Colorado, United States



-
1. Introduction
 2. What Is Oil Shale?
 3. Age and Origin of Oil Shale
 4. Oil Shale Types
 5. Retorting Methods
 6. Oil Shale Deposits
 7. Summary of Resources
 8. Oil Shale Production
 9. Environmental Concerns
 10. Future Trends

Glossary

by-products Mineral commodities and other products that are produced from an oil shale operation. These may include products from retorting oil shale, such as sulfur and ammonium sulfate, or extraction of metals from the raw or spent shale.

char The noncombusted portion of the organic matter that remains with the mineral residue of oil shale after retorting.

Fischer assay A laboratory method of determining the quantity of synthetic oil that a sample of crushed oil shale will produce upon heating in a laboratory bench-scale retort.

hydroretort A retort in which oil shale is thermally decomposed in the presence of hydrogen under pressure.

kerogen The same as organic matter as defined in this glossary.

in situ retort A method of retorting oil shale underground that offers the advantages of minimum materials handling and reduction in atmospheric pollutants.

lacustrine pertains to lakes; a depositional environment for some oil shales.

oil shale resource The total amount of shale oil in a deposit, including those portions that are not economically or technologically recoverable at present.

organic matter The portion of an oil shale composed mainly of carbon, hydrogen, and oxygen that can be retorted to yield oil, gas, water, and char.

overburden The unconsolidated sediments and bedrock that overlie an oil shale or mineral deposit.

overburden ratio The ratio of the thickness of overburden to the thickness of a mineable bed of oil shale or other mineral.

retort A closed vessel wherein crushed oil shale is heated to a temperature at which the oil shale will thermally decompose the organic matter into oil, gas, and water.

spent shale The mineral waste from retorting oil shale that usually includes some char derived from the organic matter.

Resources of oil shale are abundant in the world. Oil shale deposits range from Cambrian to Tertiary in age and were formed in a variety of geologic environments, including continental bogs and swamps, commonly in association with coal deposits, lacustrine (lake), and marine environments. Some deposits are too small to have economic value, whereas others with great potential value underlie thousands of square kilometers and are hundreds of meters thick. Some deposits of oil shale are interbedded and commingled with minerals that have current or potential economic value, including trona ($\text{Na}_2\text{CO}_3 \cdot \text{NaHCO}_3 \cdot 2\text{H}_2\text{O}$), nahcolite (NaHCO_3), and dawsonite [$\text{NaAl}(\text{OH})_2\text{CO}_3$]. Other deposits contain large, but generally low-grade, resources of metals, notably uranium, molybdenum, vanadium, copper, zinc, chromium, and nickel.

1. INTRODUCTION

Oil shale is a low-unit-cost energy commodity that requires mining and retorting thousands of tons of oil shale per year to achieve profitability. For example, one short ton of a commercial grade oil shale from the Green River Formation in Colorado assays about 35 U.S. gallons of shale oil per short ton of rock (approximately 145 liters per metric ton), or about 0.83 U.S. barrel of oil. About one-fifth of the rock by weight is organic matter that can be retorted to obtain oil and gas. The remaining four-fifths of the

oil shale is inert mineral matter that is mined with the organic matter, passed through the retort, and disposed of in an environmentally acceptable way.

Developing a commercial oil shale operation includes opening a mine, building rock-crushing facilities, constructing a retort (or several retorts) of 10,000 or more bbls per day, and perhaps adding a hydrotreating plant for upgrading the shale oil. The costs for such a facility would be many millions of dollars. Additionally, acquiring water rights and bonus bids for oil-shale tracts on Federal lands increase start-up costs. Various factors preclude shale oil from developing a major share of the today's fossil energy market, including using today's technology, high capital costs, relatively low shale-oil production rates, yields of low-hydrogen shale oil from the retort that needs upgrading for shipment to an oil refinery, disposal of spent shale, and high operating costs. However, new innovative technologies of mining and retorting may reduce the operating costs to a point where oil shale would become competitive in price with petroleum and natural gas.

However, some countries have oil shale industries, including Australia, Brazil, China, Germany, and Estonia. The largest oil shale operation by far is in Estonia, where approximately 12–13 million metric tons of oil shale is mined per year. Approximately 80% is burned as fuel in two large electric power plants located in northeastern Estonia, 15% is retorted for shale oil, and the remainder is used to manufacture cement. Other countries, including Sweden, Great Britain (Scotland), and Canada, have had oil shale industries in the past, but these have closed owing to the availability of cheaper supplies of petroleum. Some countries, including Israel, Jordan, and Morocco, which have sizable deposits of oil shale but lack other fossil fuels, are interested in developing their oil shale resources.

This article defines oil shale and discusses its origin, reviews the resources of world oil shale deposits, and outlines the methods of extraction of combustible products from oil shale. Environmental problems associated with the commercialization of oil shale are briefly reviewed. Future trends in the utilization of oil shale as a resource of energy, petrochemicals, and other by-products are discussed.

2. WHAT IS OIL SHALE?

Oil shale is commonly defined as a fine-grained sedimentary rock containing organic matter that will yield abundant amounts of oil and gas upon

destructive distillation. Most of the organic matter is insoluble in ordinary organic solvents; therefore, it must be decomposed by heating to release oil and combustible gases.

Oil shales range widely in organic content and oil yield. Commercial grades of oil shale as determined by the yield of shale oil range from approximately 100 to 200 liters per metric ton (mt) of rock. The U.S. Geological Survey has used a lower limit of 40 liters/mt, whereas other researchers have suggested a limit as low as 25 liters/mt.

How much oil and gas (or heat energy) an oil shale can produce depends not only on the amount of organic matter present in the rock but also on the composition of the organic matter as well as the geothermal history of the deposit. Some deposits have been so altered by geothermal heating that the organic matter has been degraded to the point that it will yield little oil and gas. Commercial deposits, therefore, are those that have not been subjected to strong geothermal alteration by deep burial depth or thermally altered by magmatic sources.

2.1 Determination of Oil Shale Grade

When oil shale is heated in a closed vessel (i.e., a retort) in the absence of air to approximately 500°C, the organic matter thermally decomposes into oil, gas, water, and char. The grade of an oil shale is commonly determined by the Fischer assay method. This method is a laboratory procedure for measuring the amounts of oil, water, "gas plus loss," and residue composed of mineral matter and char that a 100-g sample of crushed oil shale will produce in a small retort when heated under a set of standard conditions. Although the Fischer assay has been widely used to determine the grade of many oil shale deposits, it has several drawbacks. The amount of char in the spent shale as well as the composition of the gases produced by the Fischer assay are not determined. Typically, the gases consist mainly of methane, with some heavier alkanes, hydrogen, and carbon dioxide. The char may be burned for heat energy. Therefore, the Fischer assay does not measure the total available energy in the oil shale.

An additional complication is that certain types of retorts or retorting processes can produce more oil than that reported by Fischer assay. One such process is hydroretorting, whereby the crushed oil shale is retorted in a hydrogen atmosphere under pressure. This method has been demonstrated in the laboratory to produce more than 200% of Fischer assay yields for some types of oil shale.

2.2 Products from Oil Shale

Oil obtained by retorting oil shale is essentially a synthetic crude oil that requires refining, like any petroleum crude oil, to produce transportation fuels, heating oil, asphalt, and other petrochemicals. Some oil shales are burned directly as a low-grade fuel in electric power plants or are even utilized in the manufacture of cement. By-products obtained during retorting of oil shale include sulfur and ammonium sulfate.

Some oil shales have notably high concentrations of potentially valuable metals and minerals. The Swedish Alum shale (an oil shale), which contains Europe's largest low-grade resource of uranium, was mined for alum salt from the 1600s to the early 1900s. During World War II oil was produced from the Alum shale for military use. Shale oil production continued until 1966, when the operation ceased due to import of cheaper supplies of petroleum. The Alum shale was also mined for uranium from 1965 to 1969.

A resource estimated at 29 billion metric tons of water-soluble nahcolite (NaHCO_3) is commingled with thick, high-grade deposits of oil shale in the Green River Formation in Colorado. The nahcolite (but not the oil shale) is solution mined for sodium bicarbonate and soda ash (Na_2CO_3). Enormous deposits of water-soluble trona interbedded with oil shale, marlstone, and mudstone are present in the same formation in southwestern Wyoming. The Wyoming trona deposits, the world's largest known mineral resource of soda ash, are mined by several companies. Sodium bicarbonate and soda ash are important industrial chemicals that are used domestically and are exported to many countries for the manufacture of glass, detergents, and many other industrial products.

3. AGE AND ORIGIN OF OIL SHALE

Oil shale deposits range from Cambrian to Tertiary in age and are found in many countries. Oil shales are formed in a variety of geologic environments, including swamps and bogs (often in association with coal deposits), lakes of small to very large size, and marine settings commonly covering hundreds to thousands of square kilometers.

Much of the organic matter in oil shale is derived largely from various types of freshwater and marine algae. For example, much of the organic matter in the lacustrine Green River oil shale of Eocene age in the western United States is thought to be derived from

blue-green prokaryotic algae (cyanobacteria). Probably other bacteria including sulfate reducers were also abundant, as evidenced by the presence of iron sulfide minerals and the absence of sulfate salts that should have precipitated during the evaporative stages of the Green River lakes. Algal precursors also make up much of the organic matter in marine oil shales. Specific types of algae comprise the bulk of the organic matter in some oil shales. Fragments of land plant material derived from stems, leaves, pollen, spores, and cuticles can be carried into a lake or sea by streams and be deposited with the remains of the aquatic microorganisms. Such fragments can often be recognized under the microscope and can help determine the paleogeographic setting and identify precursors of the organic matter in an oil shale deposit.

Some conditions favorable for the formation and preservation of the organic matter include bodies of water in which current and wave action are minimal and the oxygen content of the bottom waters is low. The waters are commonly thermally or chemically stratified and may contain relatively high levels of H_2S , which is toxic to many bottom-feeding organisms.

4. OIL SHALE TYPES

In early studies many local, often colorful, names were applied to specific deposits of oil shale. Some examples include kukersite, kerosene shale, bituminite, alum shale, cannel coal, wollongite, albertite, and candle coal. With modern analytical tools including blue fluorescence microscopy, it has been found that lipid-rich organic components of the organic matter fluoresce in bright colors, making it possible to identify some of the organisms from which the organic matter was derived. The method has proved useful in developing the following rational classification of oil shales.

Organic-rich sedimentary rocks can be classified as bitumen-impregnated rocks, humic coals, and oil shales. Based of their origin, three major types of oil shale are recognized: terrestrial, lacustrine (lake), and marine oil shale. These types can be subdivided in part on the basis of organic petrography into cannel coal (terrestrial oil shale), lamosite and torbanite (lacustrine oil shale), and kukersite, tasmanite, and marinite (marine oil shale) (Fig. 1).

Cannel coal (also known as sapropelic coal) is derived from planktonic debris, spores, pollen, and other very fine detritus of aquatic plants deposited in more or less open bodies of water in a swamp or bog under

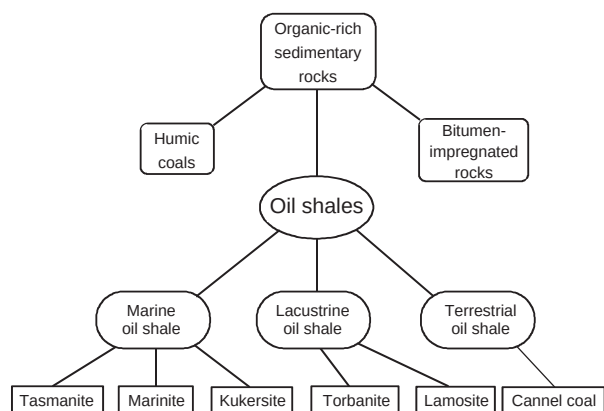


FIGURE 1 Classification of oil shales.

anaerobic conditions. The organic matter is rich in lipid and proteinaceous material, giving rise to a high oil yield upon retorting. Cannel coals can be closely associated with humic coals that originate from peat derived from terrestrial plants growing in swamps.

Lacustrine oil shales are derived largely from algae and associated phytoplankton in lakes whose waters range from fresh to saline under dysaerobic to anaerobic conditions. Two types of lacustrine oil shale are recognized: lamosites and torbanites. Lamosites are oil shales derived from lacustrine algae and other phytoplankton. Some of the organic matter may include leaf and stem fragments, pollen, and spores derived from land plants carried by streams into the lake. Fossil remains of clams, mollusks, fish, and other aquatic fauna can be found in some deposits. However, such organisms are usually sparse in parts of the oil shale further from shore, where the waters were stagnant and deficient in oxygen. Thin, wavy laminae and streaks of lipid-rich algal remains are common. Some thinly laminated black algal mats may also be present. The Green River oil shale deposits of the western United States are lamosites.

Torbanite is a lacustrine oil shale in which a substantial amount of the organic matter is derived from the alga *Botryococcus* or related species of lacustrine algae. Terrestrial plant debris is a common constituent of the organic matter. Some torbanites form very high-grade oil shales yielding as much as 45–55% shale oil by weight of the rock.

Marine oil shales include marinite, tasmanite, and kukersite. These are formed in restricted seas in which current and wave action is minimal, and the waters are commonly low in oxygen content. The most common type of marine oil shale is marinite. Tasmanite and kukersite are specific to certain types of algae.

Marinites are derived mainly from marine phytoplankton. Some aquatic organisms that can be identified under the microscope include tasmanitids, *Pediastrum* (a marine green alga), and spores. A significant portion of the organic matter in marinites consists of very fine-grained, poorly fluorescing material called bituminite. Its origin is not clear, but it may consist of degraded algal and bacterial detritus. Land plant detritus may be more or less abundant depending on the location of the accumulating organic-rich sediments in the sea relative to land areas.

Tasmanite is a variety of marine oil shale containing a large organic component derived from tasmanitids, a thick-walled, unicellular, marine green alga. Kukersite is another variety of marine oil shale whose organic matter is derived mainly from a specific type of green alga. There is an important deposit of kukersite in northeastern Estonia.

5. RETORTING METHODS

Extraction of energy (oil, gas, and heat) from oil shale takes essentially two forms: (i) the recovery of shale oil and combustible gas by retorting or (ii) burning the oil shale directly for heat energy. Two basic methods of recovering oil from oil shale are surface retorts and *in situ* methods. Surface retorts include horizontal or vertical units in which oil shale, crushed to an appropriate size (commonly 1- to 10-cm particles), is heated to approximately 500°C. The organic matter thermally decomposes into vapors of oil and combustible gases, which are recovered. Part of the organic matter forms a char that remains with the discharged spent shale after retorting. The char can be burned for heat energy useable elsewhere in an oil shale facility (Fig. 2).

A variety of surface retorts have been used to recover shale oil. Some are batch retorts, whereby the retort is charged with a given amount of crushed shale, which is retorted, and oil vapors are collected from a section of the retort and cooled and condensed to liquid. The spent shale is removed from the retort before the next batch of raw oil shale is introduced. Other retorts operate on a continuous feed basis whereby a stream of crushed shale moves through the retort and the spent shale exits at the other end of the retort. One type of experimental “hydroretort” has been designed to operate under a hydrogen atmosphere at an elevated pressure. Hydroretorting increases yields for some oil shale deposits by 200% or more compared to yields obtained by Fischer assay.

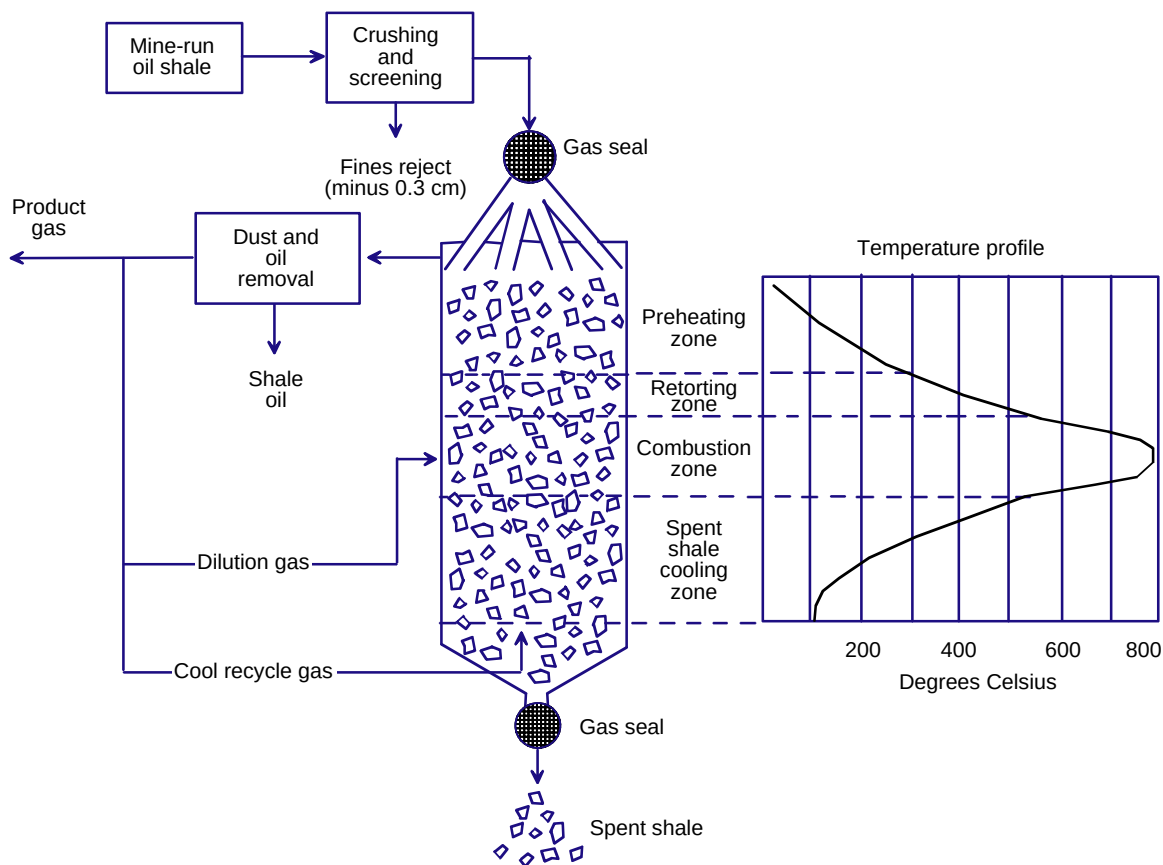


FIGURE 2 Schematic cross section of typical oil shale retort.

In Estonia, two types of retorts are in use: the Kiviter and Galoter retorts. The former retort can process 2.5- to 12.5-cm particles of oil shale but not smaller ones. The Galoter retort was developed to process smaller sized particles ranging up to 2.4 cm. The Tasiuk retort, originally designed to process Canadian tar sands, is used to process oil shale from the Stuart deposit in Queensland, Australia. It is a horizontal unit that uses part of the spent shale for preheating the raw shale. In Brazil, the vertical Petrosix retort, based on an early U.S. Bureau of Mines design, produces oil from the Iratí oil shale. In Colorado, the Tosco II horizontal retort utilized preheated ceramic balls to retort the crushed oil shale. The Unocal II retort, a vertical unit, was unusual in that the crushed shale was pumped into the bottom of the retort and the spent shale was discharged at the top.

Shale oil may also be recovered by *in situ* methods, such as retorting the oil shale through boreholes drilled into an oil shale deposit. The borehole is a retort in which the oil shale is heated by one of a variety of methods to thermally

decompose the organic matter to oil and gas. The oil shale may be retorted by electrical resistance heating, radio-frequency heating, propane-fired burner units, or heat exchangers whereby the source of the primary heat for the exchanger may be electrical heating by fossil fuels or by a small nuclear reactor.

Modified *in situ* retorting involves mining galleries above and below a column of oil shale several tens of meters wide by several hundred meters deep and rubblizing the column of oil shale between the upper and lower galleries by explosives. The column of broken oil shale, which serves as a retort, is ignited at its top. Air is pumped into the top of the retort to sustain a fire front that moves downward through the rubblized oil shale. The shale oil vapors that form in the heated zone ahead of the fire front move downward and collect as oil in the bottom gallery, where it is pumped to the surface.

Several problems have been encountered with field tests using modified *in situ* retorting. The distribution of particle sizes of oil shale after blasting is not optimal because the largest particles are not completely retorted. The fire front does not move

uniformly downward through the rubblized shale but follows channels through the retort and bypasses unretorted shale. Further field experiments are needed to resolve these problems.

6. OIL SHALE DEPOSITS

Oil shales are widely distributed throughout many countries. Although the total world resource is unknown, estimates are as high as 500 or more billion metric tons of shale oil. How much of these resources can be recovered with today's mining and processing technologies is highly speculative, but if as much as 20% of these resources is recoverable, 100 billion metric tons of shale oil would still represent an enormous fossil energy resource.

Some deposits have been relatively well investigated by geologic mapping and by core drilling. Others are only poorly known, but some of these may prove to be major world-class deposits when they are more thoroughly explored. Some better known deposits are described in detail here to better acquaint the reader with the wide variety of oil shale deposits that exist.

6.1 United States

The United States is endowed with two exceptionally large resources of oil shale: the oil shale deposits of the lacustrine Green River Formation of Eocene age in Colorado, Wyoming, and Utah and the marine Devonian black shales of the eastern United States. Smaller deposits of oil shale of submarginal economic value are found elsewhere in the United States.

6.1.1 Green River Formation

Lacustrine sediments of the Green River Formation were deposited in two large inland lakes that occupied as much as 65,000 km² in several structural-sedimentary basins in northwestern Colorado, northeastern Utah, and southwestern Wyoming during Early through Middle Eocene time. Lake Gosiute, which occupied the Green River and Washakie Basins in southwestern Wyoming, was separated from Lake Uinta to the south in Colorado and Utah by the Uinta Mountain Uplift and its eastern extension, the Axial Basin Anticline (Fig. 3). The two lakes were in existence for more than 10 million years during a period of warm temperate to subtropical climatic conditions. Episodic volcanism introduced much volcanic ash to the lakes and their drainage basins. The warm alkaline lake waters

together with water-soluble nutrients derived from the ash, provided excellent conditions for the abundant growth of blue-green algae (cyanobacteria). At times, the lake basins were closed and drier conditions prevailed, which led to precipitation of billions of metric tons of soluble sodium salts, including halite, nahcolite, trona, and shortite (Na₂CO₃ · 2CaCO₃). In the Piceance Creek Basin in northwestern Colorado, thick high-grade oil shale commingled with 29 billion tons of nahcolite was deposited.

At times, the Green River lakes hosted a variety of fishes, rays, bivalves, gastropods, ostracods, and other aquatic fauna. Land areas peripheral to the lakes supported a large and varied assemblage of land plants, insects, amphibians, turtles, lizards, snakes, crocodiles, birds, and numerous mammalian animals.

Numerous attempts to establish an oil shale industry in the western United States, especially in northwestern Colorado, have been unsuccessful. The Unocal oil shale operation was the last major project to suspend operations in 1991. Construction costs for the oil shale facility were \$650 million. The company produced 657,000 tons of shale oil, which was shipped to Chicago for refining under a price-support program subsidized by the U.S. government. The average rate of production during the last months of operation was approximately 857 tons of shale oil per day. Several major oil companies continue to hold large tracts of oil shale lands and water rights in northwestern Colorado for future development.

Shale oil resources in the Green River Formation in Colorado are estimated at 171 billion tons of shale oil. The Green River shale oil resources in northeastern Utah are estimated at 48 billion tons. In southwestern Wyoming, the Green River shale oil resources are estimated to be ≥35 billion tons. Additional oil shale resources are also present in the Washakie Basin east of Rock Springs, Wyoming.

6.1.2 Eastern Devonian Black Shales

Black organic-rich marine shale and associated sediments of late Devonian and early Carboniferous age underlie approximately 725,000 km² in the eastern United States. These shales have been exploited for natural gas for many years, but they have also been considered as a potential resource of shale oil and uranium.

The Devonian black shales were deposited in a large epeiric sea that covered much of the middle and eastern United States east of the Mississippi River (Fig. 4). This area included the broad, shallow, interior

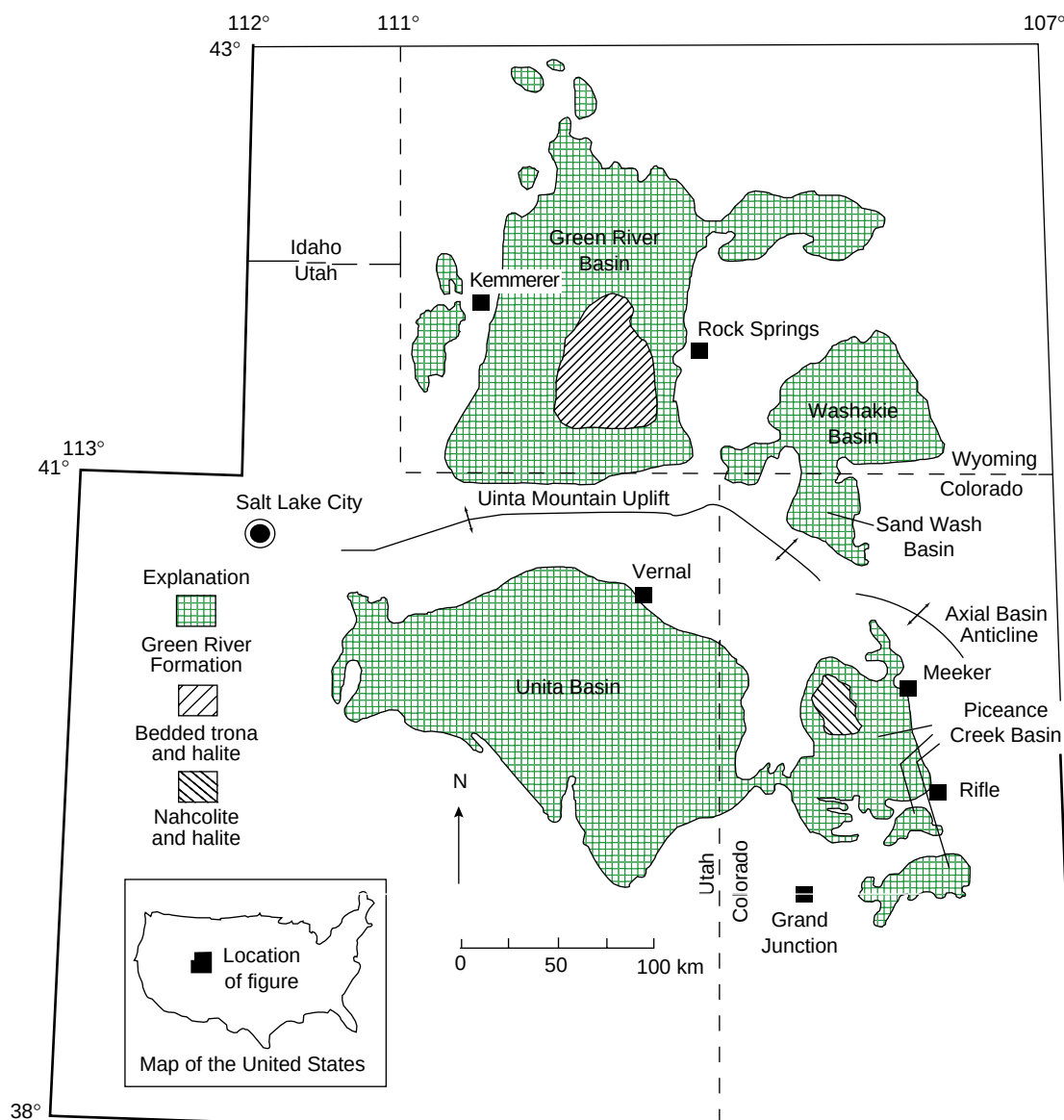


FIGURE 3 Areal distribution of the Green River Formation containing oil shale and sodium carbonate resources in Colorado, Wyoming, and Utah.

platform to the west, which grades eastward into the Appalachian Basin. The depth to the base of the Devonian black shales ranges from zero for surface exposures to 2700 m along the depositional axis of the Appalachian Basin. The oil shale resource is in that part of the basin where the organic content is highest and where the black shales are closest to the surface.

The late Devonian sea was relatively shallow with minimal wave and current activity. A large part of the Devonian black shales is amorphous nonfluorescing material, although a few structured marine fossil organisms, including *Tasmanites*, *Botryococcus*, and

Foerstia, have been recognized. Conodonts and linguloid brachiopods are sparingly distributed through some beds. Although much of the amorphous organic matter (bituminite) is of uncertain origin, it is generally believed that it was derived from planktonic marine algae.

In distal parts of the Devonian sea, the organic matter accumulated very slowly with very fine-grained clayey sediments in poorly oxygenated waters free of burrowing organisms. Thirty centimeters of such oil shale, on a compacted basis, are estimated to represent as much as 150,000 years of sedimentation.

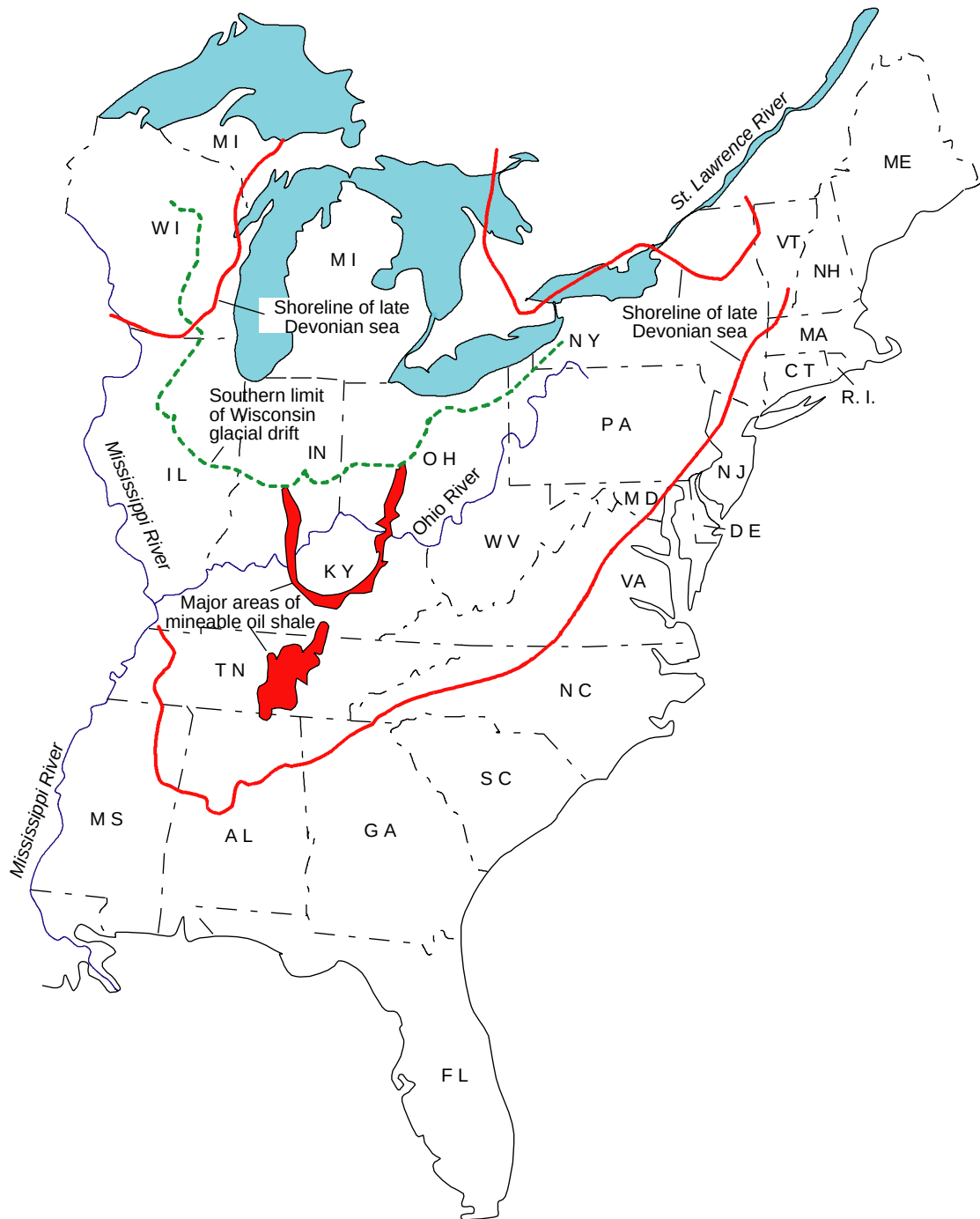


FIGURE 4 Map of the eastern United States showing the approximate limit of the Devonian sea, the southern limit of Pleistocene glacial drift, and the area of near-surface mineable Devonian oil shale.

The black shales thicken eastward into the Appalachian Basin due to increasing amounts of clastic sediments shed into the sea from the Appalachian highland east of the basin. Pyrite and marcasite are abundant accessory minerals, but

carbonate minerals comprise only a minor fraction of the black shales.

The organic matter in the Devonian black shales yields only half as much oil as an equal weight of organic matter in the Green River oil shale. This

difference is attributable to the type of organic carbon found in each oil shale. The organic matter in the Devonian oil shale has a higher ratio of aromatic to aliphatic carbon than that of the Green River oil shale. As a consequence, laboratory analyses show that the Devonian shale produces much less oil and more methane and char than the Green River oil shale. However, the oil yield from the Devonian shale can be increased by more than 200% by hydrotretorting.

The mineable shale oil resource of the Devonian black shales (Fig. 4) is estimated to be 423 billion barrels (~ 61 billion tons). This figure is based on the following criteria: organic content ≥ 10 weight percent, overburden ≤ 60 m, stripping ratio $\leq 2.5:1$, thickness ≥ 3 m, mining by open pit, and processing by hydrotretorting.

6.2 Other Deposits

Some examples of oil shale deposits in other countries, including several countries that have oil shale industries, follow.

6.2.1 Australia

The oil shale deposits of Australia range from small noneconomic occurrences to deposits large enough to be exploitable. The demonstrated oil shale resources for 12 deposits are estimated to be 58 billion tons, from which 24 billion barrels (~ 3.1 billion tons) of shale oil are estimated to be recoverable. The Australian deposits, which range in age from Cambrian to Tertiary, are of diverse origin. These deposits are located in eastern Queensland, New South Wales, South Australia, Victoria, and Tasmania. Those having the best potential for development are the Tertiary lacustrine oil shale deposits in Queensland. One of these, the Stuart deposit, contains an estimated 36.4 billion barrels (~ 5.2 billion tons) of shale oil. An open-pit mine and demonstration retorting facility have been constructed. The first shipment of Stuart shale oil to market was made in 2001, and as of February 2003, a total of 700,000 barrels ($\sim 100,000$ tons) of shale oil have been produced and marketed.

6.2.2 Brazil

At least nine occurrences of oil shale ranging from Devonian to Tertiary in age have been reported in different parts of Brazil. Of these, oil shale of the Permian Iratí Formation has the greatest potential for development because of its accessibility, grade, and widespread distribution. The formation crops out in the form of a large "S" extending from the

northeastern part of the state of São Paulo southward for 1700 km to the southern border of Rio Grande do Sul and into northern Uruguay. In the state of Paraná, in the vicinity of the town of São Mateus do Sul, two economic beds of oil shale have been identified. The upper one, 6.5 m thick, and the lower one, 3.2 m thick, are separated by approximately 8 m of barren shale and limestone. The two beds are mined by open pit. An *in situ* resource of approximately 11.5 billion tons of shale oil has been identified.

The Iratí oil shale is dark gray, brown, and black, very fine grained, and laminated. Clay minerals comprise approximately 70% of the shale, with minor contributions of detrital quartz, feldspar, and pyrite; carbonate minerals are sparse. Unlike the Devonian black shales of the eastern United States, the Iratí oil shale is not notably enriched in metals. The oil shale is high in sulfur (approximately 4%) and the moisture content is approximately 5%. Gross heating value (dry basis) is approximately 1500 kcal/kg of shale.

The origin of the Iratí oil shale is unclear. Some researchers have argued that the oil shale was deposited in a fresh to brackish water lacustrine environment. Others have suggested that the oil shale originated in a partially enclosed intracontinental marine basin of reduced salinity connected to the open sea that formed after the close of late Carboniferous glaciation.

Development of the Iratí oil shale began at São Mateus do Sul, where a prototype retort and processing plant were constructed. The facility began operations in 1972 with a design capacity of 1600 tons of oil shale per day. In 1991, an industrial-size vertical retort, 11 m in diameter by 60 m high, with a design capacity of 6200 tons of oil shale per day, was put into operation. Between 1980 and 1998, approximately 1,550,000 tons of shale oil was produced, with by-products of sulfur, liquefied petroleum gas, and fuel gas.

6.2.3 Estonia

In Estonia, two kukersite deposits are recognized: the Middle Ordovician Estonia deposit, which is continuous with the Leningrad deposit in Russia, and the somewhat younger Middle Ordovician Tapa deposit. Together, they occupy more than 50,000 km².

In the Estonia deposit, there are as many as 50 beds of kukersite, commonly 10–40 cm in thickness, and kerogen-rich limestone that alternates with fossiliferous fine-grained limestone. The organic

content of individual beds of kukersite can reach 40–45 weight percent of the oil shale and yield as much as 320–500 liters of oil per ton of rock. In the operating mines, the ore zone is approximately 2.5 m thick and has a heating value of 2440–3020 kcal/kg. Most of organic matter in kukersite is derived from the fossil marine green alga *Gloeocapsomorpha prisca*, which has affinities to the modern cyanobacterium *Entophysalis major*, an extant species that forms algal mats in intertidal to very shallow subtidal marine waters. The mineral fraction of the kukersite consists of calcite (>50%), dolomite (<10–15%), and quartz, feldspar, illite, chlorite, and pyrite (<10–15%).

The interbedded limestone and kukersite sequence was deposited in a shallow subtidal marine basin adjacent to a shallow coastal area on the north side of the Baltic Sea near Finland. The abundance of marine macrofossils and low pyrite content of the sequence suggest oxygenated waters, unlike most depositional environments for oil shale. The widespread continuity of the even thin beds of kukersite suggests negligible bottom currents.

The kukersite deposits of Estonia have been known since the 1700s, but their exploitation did not start until World War I as a result of fuel shortages brought about by the war. Full-scale mining began in 1918 when 17,000 tons of oil shale was mined by open pit. By 1940, annual production of oil shale had reached 1.7 million tons. However, it was not until after World War II, during the Soviet era, that production increased dramatically, peaking in 1980 when 31 million tons of oil shale was mined by 11 open-pit and underground mines. The annual production of oil shale decreased to 14 million tons in 1995 and has ranged from 10 to 15 million tons until 2000. Approximately 80% of the mined kukersite is burned as fuel in two large electric power plants, 15% is retorted for shale oil to make a variety of petrochemicals, and a few percent is used to manufacture cement and other products. In 1997, the Estonian oil shale industry was sustained in part by state subsidies amounting to approximately \$10 million. Mineable reserves of kukersite are estimated to be 6 billion tons. Since 1917, 850 million tons of oil shale has been mined.

The Tapa deposit is still in the prospecting stage and the resources are yet to be evaluated. Available data suggest that the deposit is of lower grade than the Estonia deposit.

The marine *Dictyonema* shale of Early Ordovician age underlies most of northern Estonia and is

probably equivalent to the upper part of the Alum shale of Sweden. Until recently, little was known about the *Dictyonema* shale because it was covertly mined and processed for uranium at a secret plant located at Sillamäe in northeastern Estonia during the Soviet era. The shale ranges from less than 0.5 to more than 5 m in thickness. A total of 22.5 tons of uranium was produced from approximately 272,000 tons of *Dictyonema* shale.

6.2.4 Israel

Twenty marinite deposits of Late Cretaceous age have been found in Israel and approximately 12 billion tons of oil shale reserves has been identified. The organic content of the marinites is relatively low, ranging from 6 to 17 weight percent, and the oil yield is only 60–71 liters/ton. The moisture content is high (~20%), as is the carbonate content (45–70% calcite) and sulfur content (5–7%). Some of the deposits are close enough to the surface to be mined by open pit. A commercially exploitable bed of phosphate rock, 8–15 m thick, underlies the Misor Rotem oil shale mine in south central Israel. The oil shale from this mine has a heating value of 650–1200 kcal/kg. It was burned in a fluidized bed boiler to provide steam to operate a turbogenerator in a 25-MW electric power plant. The plant began operating in 1989 but is now closed.

6.2.5 Jordan

Approximately 26 occurrences of oil shale are found in Jordan, but relatively few have been studied in detail or explored by core drilling. The deposits commonly occur in fault-bounded synclines (grabens) covered by unconsolidated gravel and silt with lenticular beds of marl and limestone and, locally, by volcanic rocks. The oil shales are marinites of Late Cretaceous to Early Tertiary age. Eight deposits that have been explored to some extent yield approximately 7–11% oil by weight, and their average thickness ranges from 30 to 136 m. Resources in terms of shale oil for 6 of these 8 deposits range from 74 million to 2.15 billion tons.

Overburden ratios range from 0.4:1 to 2.2:1. The oil shales are low in moisture, averaging approximately 2–5.5 weight percent. The sulfur content for some of the oil shales is approximately 0.3–4.3 weight percent. Metals of possible economic value, including uranium, molybdenum, vanadium, chromium, and nickel, occur in concentrations ranging from approximately 20 to 650 parts per million. Surface water supplies are scarce in Jordan; therefore, groundwater aquifers will need to be

tapped for an oil shale industry. Because Jordan has few resources of oil and gas and no coal, it has been actively exploring its oil shale deposits for their possible use as fuel in electric power plants.

6.2.6 China

Two noteworthy deposits of oil shale are found at Fushun in Liaoning Province in northeastern China and at Maoming in the western part of Guangdong Province approximately 350 km west of Hong Kong. At Fushun, the oil shale overlies a thick deposit of subbituminous to bituminous coal of Eocene age. The oil shale is lacustrine in origin. The oil shale ranges from 48 to 190 m in thickness. At the West Open Pit coal mine, it is 115 m thick. The lower 15 m consists of low-grade, light-brown oil shale and the upper 100 m consists of richer grade, brown to dark brown finely laminated oil shale. The contact between the underlying coal and oil shale is gradational, suggesting a peat swamp environment that gradually subsided and was replaced by a lake in which the oil shale was deposited. The oil yield ranges from 5 to 16 weight percent. The oil shale resources in the vicinity of the mine are estimated to be 260 million tons, of which 90% is thought to be mineable. The total resource of the Fushun deposit is estimated to be 3.6 billion tons of oil shale. The first production of shale oil at Fushun began in 1930. In the mid-1990s, 60 retorts, each with a capacity of 100 tons of oil shale per day, produced 60,000 tons of shale oil per year.

The deposit of oil shale at Maoming is also Tertiary in age and lacustrine in origin. Reserves estimated at 5 billion tons are reported in an area 50 km long, 10 km wide, and 20–25 m thick. The oil yield of Maoming oil shale is 4–12 weight percent and averages 6 or 7%. The ore is yellowish brown with a bulk density of approximately 1.85 g/cc. The oil shale contains 72% mineral matter, 11% moisture, and approximately 1% sulfur, and it has a heating value of approximately 1750 kcal/kg (dry basis).

6.2.7 Russia

More than 80 deposits of oil shale have been identified in Russia. The kukersite deposit near St. Petersburg, which is continuous with the Estonian kukersite deposit, is burned as fuel in the Slansky electric power plant near St. Petersburg. Oil shale in the Volga–Pechersk oil shale province was used as fuel in two electric power plants but is no longer burned due to high SO₂ emissions.

7. SUMMARY OF RESOURCES

Estimating the total world resources of oil shale presents a number of problems. Many deposits are poorly explored and analytical data are sparse, making it difficult to determine the grade and size of the resource. Core drilling and laboratory analyses of oil shale samples are expensive, which lessens the likelihood of conducting exploration programs, especially in Third World countries with limited budgets. However, there has been considerable exploration in some countries with good deposits of oil shale but that are not well endowed with other fossil fuels, such as Australia, Jordan, Israel, and Morocco.

Resource data are reported in a variety of ways: as volumetric data (i.e., gallons and liters per metric ton), gravimetric data (U.S. short tons and metric tons), and heating values (British thermal units per pound of oil shale, kilocalories per kilogram, or megajoules per kilogram of oil shale). To bring some uniformity to the data, Table I lists the resources only in U.S. barrels and metric tons of shale oil. To do this, it was necessary to convert volumetric to gravimetric units, which requires knowledge of the specific gravity of the shale oil. With some exceptions, when the specific gravity is unknown, a value of approximately 0.910 was assumed.

The quality of the resource estimates given in Table I ranges from poor to satisfactory. Therefore, the estimates should be regarded with caution as to their accuracy. The best estimates are those given for deposits that have been explored by moderate to extensive core drilling, including the Estonia kukersite deposit; some of the Tertiary deposits in eastern Queensland, Australia; several deposits in Israel, Jordan, and Morocco; and the Green River deposits in the western United States.

The total shale oil resources in 39 countries is estimated at 487 billion tons (3 trillion barrels). Because many deposits are not included in this estimate, this total is a conservative number that should increase significantly when lesser known deposits have been explored more thoroughly.

How much of the world oil shale resources can be recovered is based on numerous factors. The size, depth, thickness, and grade of a deposit are, of course, important. However, conflicts with surface uses of lands overlying a deposit, such as towns and parks, could restrict areas that could be mined. Improved mining and processing technologies, such as *in situ* energy recovery and hydrotreating methods, could make some deposits attractive for development that would otherwise be subeconomic.

TABLE I

Shale Oil Resources of Some World Oil Shale Deposits

Country, region, or deposit	In-place shale oil resources		Country, region, or deposit	In-place shale oil resources	
	× 10 ⁶ bbls	× 10 ⁶ tons		× 10 ⁶ bbls	× 10 ⁶ tons
Argentina	400	57	Yarmouk ^a	Large	
Armenia	305	44	Kazakhstan	2837	400
Australia			Madagascar	32	5
New South Wales	40	6	Mongolia	294	42
Queensland	31,041	4432	Morocco		
South Australia	600	86	Tarfaya, zone R	42,145	6448
Tasmania	48	7	Timahdit	11,236	1719
Austria	8	1	New Zealand	19	3
Belarus			Poland	48	7
Pripyat Basin	6988		Russia		
Brazil	82,000	11,734	Central Basin	70	10
Bulgaria	125	18	Olenyok Basin	167,715	24,000
Burma	2000	286	St. Petersburg kukersite	25,157	3600
Canada			Timano–Petshorsk Basin	3494	500
Manitoba–Saskatchewan	1250	191	Turgai and Nizneiljsk deposit	210	30
New Brunswick	286	40	Volga Basin	31,447	4500
Nova Scotia	1705	244	Vyshegodsk Basin	195,967	2800
Ontario	12,000	1717	Other deposits	210	30
Chile	21	3	South Africa	130	19
China	16,000	2290	Spain	280	40
Egypt	5700	816	Sweden	6114	875
Estonia			Syria	Large?	
Dictyonema shale	12,386	1900	Thailand	6400	916
Estonia deposit	3900	594	Turkey	1985	284
France	7000	1002	Turkmanistan and Uzbekistan		
Germany	2000	286	Amudarja Basin	7687	1100
Hungary	56	8	United Kingdom	3500	501
Iraq			United States		
Yarmouk ^a	Large		Eastern Devonian shale	189,000	27,000
Israel	4000	550	Elko Formation	228	33
Italy	73,000	10,446	Green River Formation	1,734,000	258,000
Jordan			Heath Formation	180,000	25,758
Attarat Umm Ghudran	8103	1243	Phosphoria Formation	250,000	35,775
El Lajjun	821	126	Uzbekistan		
Juref ed Darawish	3325	510	Kyzylkum Basin	8386	1200
Sultani	482	74	Zaire	100,000	14,310
Wadi Maghar	14,009	2149			
Wadi Thamad	7432	1140	Total	3,323,910	487,002

^aThe Yarmouk deposit lies across the Jordanian–Iraqi border.

8. OIL SHALE PRODUCTION

A number of countries have had oil shale industries in the past. Some countries have restarted oil shale

operations following periods of no activity. Australia, for example, produced shale oil from several deposits in the 1800s, but these closed when the high-grade ore was worked out. During the past

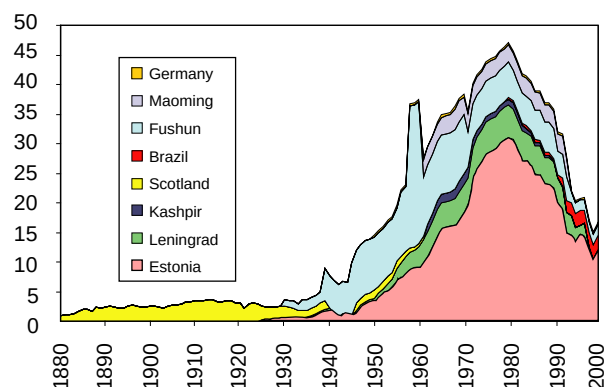


FIGURE 5 Production of oil shale in millions of metric tons from Estonia (Estonia deposit), Russia (Leningrad and Kashpir deposits), Scotland (Lothians), Brazil (Irati Formation), China (Maoming and Fushun deposits), and Germany (Dotternhausen).

several years, a new attempt has been made to establish an oil shale industry in Queensland. Brazil has been operating an oil shale industry since the 1980s. Estonia, however, has one of the longest continuous oil shale operations, beginning during the World War I. Despite a major effort to establish an oil shale industry in Green River deposits in the western United States by the federal government, the effort failed in part to low petroleum prices and the lack of a technology to profitably recover shale oil from the deposits.

Figure 5 shows the production of oil shale from eight deposits in six countries from 1880 to 2000. World production, much of it from the Estonian kukersite deposit, peaked in 1980 when approximately 47 million tons of oil shale was mined. Production declined to a low of 15 million tons in 1999.

9. ENVIRONMENTAL CONCERNS

Open-pit mining and surface retorts present the most problems of environmental concern. Dust from mining operations can be a significant air pollutant and hazardous to workers, although wetting mine roads and piles of waste rock reduces the amount of dust. Crushing operations may also give rise to dusty conditions, but again, wetting the materials during handling reduces the amount of dust from such operations. Spent shale from some retorts contains significant quantities of phenols and other types of soluble organic compounds that are potential health hazards. Surface and groundwaters infiltrating waste piles of spent shale and overburden materials can leach soluble organic compounds, sulfates,

chlorides, and carbonates that could pollute fresh-water streams and lakes. Some trace elements found in some oil shales may be toxic to humans and animals if they are released to the atmosphere or into streams and lakes.

In situ operations offer the possibility of reducing air and water pollutants by leaving the overburden intact as well as leaving most or all of the spent shale underground. A detailed understanding of the subsurface geology is required to ensure that the potentially hazardous waste products of an *in situ* operation remain underground and will not surface later through springs in contact with the spent shale. Other favorable aspects of *in situ* retorting include minimum disturbance of the surface, a smaller labor force, reduced capital costs for plant construction and operation because surface retorts are not required, and possibly higher quality shale oil from subsurface retorting.

10. FUTURE TRENDS

Estonia, which is endowed with important resources of oil shale, will probably continue to operate an oil shale industry for years. Strides have been made to reduce air pollution from burning and retorting oil shale and from leaching deleterious substances from huge waste piles of spent shale in Estonia. Oil shale operations in China, Germany, Brazil, and Australia may also continue for some years and possibly expand if economic conditions are favorable.

Although current oil shale operations constitute a small industry on a worldwide basis, the enormous resources of oil shale suggest that it will find a place in the production of energy, petrochemicals, and by-products in the future. New mining and processing technologies, especially *in situ* techniques that minimize the handling of large amounts of raw and spent shale, could provide the impetus for oil shale development.

SEE ALSO THE FOLLOWING ARTICLES

Gas Hydrates • Natural Gas Resources, Unconventional • Oil and Natural Gas Drilling • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitudes and Distribution • Oil Crises, Historical Perspective • Oil Recovery • Oil Refining and Products • Oil Sands and Heavy Oil • Strategic Petroleum Reserves

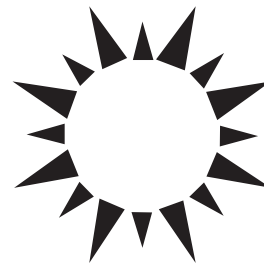
Further Reading

- Allred, V. D. (1982). "Oil Shale Processing Technology." Center for Professional Advancement, East Brunswick, NJ.
- Crisp, P. T., *et al.* (1987). "Australian Oil Shales; A Compendium of Geological and Chemical Data." University of Wollongong, Departments of Geology and Chemistry, New South Wales, Australia.
- Dyni, J. R., Anders, D. E., Rex, and R. C., Jr. (1990). Comparison of hydrotretorting, Fischer assay, and Rock-Eval analyses of some world oil shales. *In* "Proceedings of the 1989 Eastern Oil Shale Symposium." pp. 270–286. Institute for Mining and Minerals Research, University of Kentucky, Lexington.
- "Eastern Oil Shale Symposium Proceedings" (1981–1993). Institute for Mining and Minerals Research, University of Kentucky, Louisville. [Includes papers on geology, mining, and processing of oil shale with emphasis on the Devonian black shales of the eastern United States.]
- Hutton, A. C. (1987). Petrographic classification of oil shales. *Int. J. Coal Geol.* 8, 203–231.
- Kattai, V., Saadre, T., and Savitski, L. (2000). "Estonian Oil Shale." Eesti Geoloogiakeskus, Tallinn. [in Estonian with an extended English summary].
- "Oil Shale" (serial publication). Estonian Academy, Publishers, Tallinn. [Includes papers, mostly in English, on geology, mining, and processing of oil shale.]
- "Oil Shale Symposium Proceedings" (1964–1992). Colorado School of Mines, Golden. [Includes papers on geology, mining, and processing of oil shale with emphasis on the Green River oil shale deposits of Colorado, Wyoming, and Utah.]
- Roen, J. B., and Kepferle, R. C. (eds.). (1993). "Petroleum geology of the Devonian and Mississippian black shale of eastern United States, Bulletin 1909, Chapters A–N." U.S. Geological Survey, Denver, CO.
- Russell, P. L. (1990). "Oil Shales of the World, Their Origin, Occurrence, and Exploitation." Pergamon, New York.
- Smith, J. W. (1980). Oil shale resources of the United States. *Mineral Energy Resour.* 23(6), 1–20.

OPEC, History of

FADHIL J. CHALABI

Centre for Global Energy Studies
London, United Kingdom



1. The Oil Industry Structure under the “Seven Sisters”
2. OPEC’s Formative Years
3. A Short Period of Cooperation between Producers and Oil Companies
4. The Oil Price Shocks
5. OPEC as the Last Resort Oil Supplier: The Quota System
6. The Backlash
7. OPEC’s Limitations in Oil Supply Management
8. OPEC and a Future of Uncertainty

Glossary

downstream In the oil industry, refers to refineries and the distribution of oil products.

horizontal integration A system that secured cooperation among the major oil companies in the upstream phase of the industry and allowed for the planning of crude oil production according to the needs of each company.

oil concessions A system that prevailed in the Middle East and Venezuela, whereby the major oil companies were given the exclusive right to explore, find, develop, and export crude oil and where the companies jointly owned an “operating company” in each country and held a quasi-monopoly among the producing/exporting companies.

OPEC basket price A “basket” of seven crudes as a reference price, namely Sahara Blend of Algeria, Dubai Crude, Saudi Arabia’s Arab Light, Minas of Indonesia, Bonny Light of Nigeria, Tia Inana of Venezuela, and Isthmus of Mexico.

per barrel income Oil revenue per barrel of oil after accounting for the cost of production.

posted price A fictitious fixed price, set by the major oil companies until 1960, that served as a reference point for calculating taxes and royalties paid by them to the host countries.

“Seven Sisters” Refers to the seven major international oil companies that, until the early 1970s, rigorously controlled the world oil industry outside the United States and former Soviet Union (British Petroleum, Mobil, Socal, Jersey [Exxon], Shell, Gulf, and CFP [Total]).

upstream In the oil industry, refers to crude oil production, that is, investments in exploring, finding, and developing crude oil.

vertical integration A system by which each operating company obtained crude oil, destined for its own refineries, so as to plan crude oil production in line with the needs of those refineries.

Delegates from five major oil-producing countries—Iran, Kuwait, Saudi Arabia, Venezuela, and Iraq—met in Baghdad and announced the foundation of the Organization of Petroleum Exporting Countries (OPEC) on September 16, 1960. At that time, no one could have foreseen that this event would play such a crucial role, some 10 years or so later, in reshaping the world energy scene and would have such an impact on the world economy. The event at the time passed nearly unnoticed, except perhaps by the specialized petroleum media. It took place after two successive reductions in what was then called the posted price of oil, which previously was set and posted unilaterally by the major international oil companies, usually referred to as the “Seven Sisters.” The first cut, in February 1959, was by as much as U.S. \$0.18/barrel (or ~10%) for Gulf oil, from \$2.04 to \$1.86/barrel. The second price reduction of 7% was decided by Exxon in August 1960, and the other companies followed suit. The OPEC delegates’ immediate and prime objective was to safeguard their member countries’ oil revenue against any further erosion as a result of the companies’ deciding to cut prices further. According to the profit-sharing agreements put into force during the early 1950s between the holders of oil concessions in these Middle East areas, on the one hand, and their host countries, on the other, the latter’s per barrel revenue of exported oil was determined at 50% of the official (i.e., “posted”) price minus half of the cost of producing that barrel. Therefore, any fluctuation in that price, whether upward or downward, changed

the per barrel revenue accordingly. When the oil companies undertook price cuts, the per barrel income of these countries was reduced by more than 15%, compared with the price before the two cuts. Until the late 1950s, the official posted price was more of a tax reference price, on which host countries received their taxes, than a real market price resulting from the exchange of oil between buyers and sellers in a free market. For the companies, on the other hand, the tax paid to host countries, together with the production costs, determined the operators' tax-paid cost, meaning the cost incurred by the oil companies as a result of producing one barrel of oil plus tax and royalties paid to the host governments. For this reason, the first resolution adopted by the newly formed organization emphasized that the companies should maintain price stability and that prices should not be subjected to fluctuation. It also emphasized that the companies should not undertake any change of the posted price without consultation with the host countries. Furthermore, the resolution accentuated the necessity of restoring prices to the pre-1959 level. Other wider objectives were also articulated and incorporated into the organization's five resolutions, namely that it would endeavor to have a price system that would secure stability in the market by using various means, including the regulation of production, with a view to protecting the interests of both oil consumers and oil producers and guaranteeing stable oil revenues for the latter. Moreover, OPEC announced that the real purpose of the organization was to unify the oil policies of member countries so as to safeguard their interests individually and collectively.

1. THE OIL INDUSTRY STRUCTURE UNDER THE "SEVEN SISTERS"

Prior to the formation of the Organization of Petroleum Exporting Countries (OPEC), and until the late 1950s, the international oil industry had been characterized mainly by the dominant position of the major multinational oil companies through a system of oil concessions granted to the major oil-producing countries, according to which these companies were interlinked in the "upstream" phase of the industry, that is, exploration, development, and production of crude oil. Through joint ownership of the holding companies that operated in various countries, they were able to plan their

production of crude oil according to their requirements. The oldest example of this kind of "horizontal integration" (i.e., with the companies being interlinked in the upstream phase) was the Iraqi Petroleum Company (IPC), formerly the Turkish Petroleum Company, which was a non-profit-making holding and operating company owned by the Anglo-Iranian Oil Company (British Petroleum [BP]), Compagnie Française Pétrole (CFP or Total), Royal Dutch Shell, Standard Oil of New Jersey (Exxon), Mobil, and the Gulbenkian Foundation (known as "Mr. 5 percent," Gulbenkian had been instrumental in negotiations for an agreement, prior to World War I, between the Ottoman Empire and German and British interests involved in the exploitation of Iraqi oil).

Each of these companies was a shareholder in other countries in the Middle East. For example, BP owned half of the Kuwait Oil Company and all the oil of pre-Musadeq Iran (although its holding was reduced to 40% in the consortium that was founded after Musadeq) as well as shares in Qatar Petroleum Company and Abu Dhabi Petroleum Company, whereas the American company Esso (Exxon) had 30% of Aramco in Saudi Arabia as well as shares in Qatar Petroleum, Abu Dhabi Petroleum, and the like. This type of interlinking enabled them to control and manage crude oil supplies worldwide, along with the bulk of oil exports from the major oil-producing countries, so that oil trading became a question of intercompany exchange with no free market operating outside the companies' control whereby crude oil was exchanged between sellers and other buyers. At the same time, each of these "sisters" had its own "downstream" operations—transportation, refining, oil products, and distribution networks—that made them "vertically" integrated. This compact system of horizontal and vertical integration allowed the companies to plan for their future crude oil requirements in line with their downstream requirements, that is, the amount of oil products needed by each according to its market outlets in the countries to which crude oil was shipped.

However, by the second half of the 1950s, this compact system of complete control of production and intercompany exchanges began to weaken with the appearance of independent oil companies searching for access to cheaper crude oil. This new development led to the creation, albeit on a limited scale, of a free market for buying and selling crude oil outside the control of the major companies. The state-owned Italian company, ENI, was able to find

its way in investing beyond the control of the Seven Sisters by offering different and better financial terms than the oil concessions. Similarly, many independent oil companies began to invest outside the sisters' control by offering apparently better financial terms than the oil concession system.

At the same time, oil was discovered in new producing areas, as in the case of Libya, where in addition to Exxon Mobil, Occidental was an important American independent company. These developments led to a greater amount of crude oil being produced outside the system, controlled by the "majors" and offered for sale to independent refiners with no access to crude oil. Furthermore, crude oil from the then Soviet Union began to make its way into a free market for sale at competitive prices, and this country needed to offer attractive discounts to move its oil in the market. Other factors also played roles. One was a measure taken by the United States to limit the entry of crude oil into the market so as to protect its oil industry. This policy evolved toward the end of the 1950s, when the U.S. government's oil import policy predicated itself on the premise that foreign oil should be a supplement to, and should not supplant, American oil. This deprived many independent American oil companies that had access to foreign crude oil from entering the U.S. market, a situation that led to their having to sell surplus crude outside the major oil companies' system. A market price began to shape up in the form of discounts off the posted price set by the major oil companies. On the other hand, a new pattern of tax relationship emerged following the entry of newcomers investing in oil in new areas such as Libya, where the government's per barrel share was calculated not on the official posted price (which was a fixed price) but rather on the basis of a price realized in the free market (which was always below the posted price). In this way, when the realized price fell, the host government's per barrel share fell accordingly and the tax cost for the new producer would correspondingly be lower, thereby providing a good margin of profit with which to reinvest in new production capacity for crude oil.

In this way, the free market, where crude oil is bought and sold physically and freely, started to grow. Obviously, the larger this market, the less it was dominated by the major oil companies, whose share in world production outside the United States and Soviet Union now came under threat from the newcomers. As explained by Edith Penrose in her book, *The Large International Firm in Developing Countries: The International Petroleum Industry*, in

1950 the Seven Sisters owned 85% of the crude oil production in the world outside Canada, the United States, the Soviet Union, and China. By 1960, this share had fallen to 72%, to the benefit of the newcomers, whose share had correspondingly grown from 15% to 28%. The fall in the Majors' share of refinery capacity was even more pronounced: from 72% to 53% during the same period – again to the benefit of the independent oil companies. Against this atmosphere of increasing competition and the threat to the dominant position of the major companies by newcomers and the Soviet Union, (the latter of which was selling crude at massive discounts), the major oil companies had to cut the price posted by them to reduce the tax paid cost. This was actually pushed by Esso (Exxon) and followed by the other companies; hence, the cuts in the posted price by 1959 and 1960 were a countermeasure to the newcomers' increased market share. In other words, if this mounting competition with price cuts in a free market had not been curbed by restrictive measures, this situation would have led to a harmful level of competition and continued price cuts.

2. OPEC'S FORMATIVE YEARS

The timely formation of OPEC put an end to this increasingly damaging level of competition with its adverse effect on the position of the major oil companies and on the oil industry in general. The first step that OPEC took was to enlarge its membership, and it was only shortly after its foundation that Qatar joined the organization, followed by Indonesia, Libya, the Emirates of Abu Dhabi, Algeria, Nigeria, Ecuador, and Gabon (although Ecuador and Gabon left OPEC during the 1990s). The enlargement of the OPEC base at this stage served to strengthen its negotiating power vis-à-vis the oil companies. In addition to this enlargement, OPEC took certain measures that helped greatly in putting an end to price erosion. These involved, among other things, the unification of the tax system, that is, levying taxes from the oil companies that operated in member countries' territory so as to make them conform to those applied in the Gulf (i.e., by using the fixed official posted price as a reference for the calculation of tax and royalties and not a realized market price). This measure put an end to the tax system in new areas where tax and royalties were paid by investors on the basis of realized market prices and not fixed posted prices, a system that tends to put new companies in a better competitive position than the

major companies. In so doing, OPEC succeeded in curbing the newcomers' ability to invest in expanding capacity and acquire an enlarged market share at the expense of the majors.

According to OPEC resolutions, Libya amended its tax system and put it in line with that of the Gulf. Obviously, this measure, adopted by Libya in response to OPEC, had the effect of limiting the expansion of new companies' investments as they began paying higher taxes than they did previously. This OPEC measure, while serving the producing exporting countries by consolidating price stability, also benefited the major oil companies by preventing their dominant competitive position in the oil industry from eroding further.

Another measure taken by OPEC that was of no less significance in consolidating the price structure was the amendment to the concessions' tax regime by which the payment of "royalties" (paid by investors to the owners of oilfields) was treated separately from the taxation or what was called the "expensing of the royalty" (i.e., the royalty to be paid by the concessionaires but considered as part of the production costs for the purpose of calculating taxes). Among the established rules in the oil industry, as implemented in the United States since the early part of the 20th century, royalties had to be paid to the landowners regardless of whether profits were realized or not. Taxes are levied on profits that have to be paid in addition to royalties. When the 50/50 profit sharing agreements between host governments and the major oil companies were concluded during the 1950s, the 50% belonging to the host countries covered both the royalty at 12.5% and the tax at 37.5% of the posted price.

OPEC asked the companies to implement this principle in the oil concessionary system operating in the areas, with 50% tax being paid after the host countries had received the royalty of 12.5% of the posted price. The companies agreed with this, provided that the royalty would be considered part of the cost of production. This meant that only half of the royalty would be payable to the host countries from royalties; hence, the total government share increased from 50 to 56.25% of the posted price. Although this increase had to be implemented gradually by allowing temporary decreasing discounts to the companies, it nevertheless had the effect of consolidating the price structure by increasing the tax paid cost. This measure had the effect of strengthening the level of realized market prices because the higher the tax paid cost, the higher the price floor, making price cuts more difficult.

It is clear that the mere existence of OPEC, whose *raison d'être* was to put an end to price cuts by the companies, helped not only to safeguard the producers' interests but also to strengthen the price structure and put an end to the expansion of the competitive market. No less important were the consequential measures taken by OPEC that had the effect of serving the mutual interests of the oil producers and major oil companies. It was because of these measures that the majors were able to maintain their leading role in the industry and to weaken the role of the newcomers. During the first 10 years of its life, OPEC placed a limitation on the expansion of the free market, which was the source of price instability. Without OPEC and its measures during those early years, other changes taking place in the structure of the oil industry at that time would have further encouraged the newcomers, thereby increasing instability in the market.

In fact, the position of the major companies improved after the creation of OPEC, and to quote Penrose again, "The share of the major oil companies in the upstream part of the industry increased from 72% in 1960 to 76% in 1966. Similarly, in the downstream operations, their share increased from 53% to 61%, respectively." This meant that the shares of the newcomers had to decline from 28 to 24% in upstream activity and from 47 to 39% in downstream activity. In effect, OPEC had reversed the trend of increasing the share of the newcomers and limiting their power.

Although OPEC had failed to restore the posted prices to their levels prior to the 1959 and 1960 price cuts, as was the initial objective of the organization when it was founded in Baghdad, Iraq, in 1960, the formation of OPEC and the measures it took simply constituted a form of amendment of oil concessions. Prior to OPEC, host countries were not partners in their own industry and their role did not exceed the levying of taxes from the oil companies, the latter of which, according to the terms of the concessions, had the absolute freedom to decide whatever was suitable to their own interests in all matters relating to the oil industry in territory covered by concessions.

Another episode that would later play a pivotal role in changing the structure of the oil industry and the nature of the organization was OPEC's "declaratory statement of petroleum policy in member countries" in June 1966, which became a historical turning point for OPEC in that it laid down basic principles that would later influence all of the major events of the 1970s. The statement emphasized the right of producing countries to fix unilaterally the

price of their own oil in an endeavor to give the states a greater role in the development of hydrocarbon resources and to accord the states the right to participate in concession-holding agreements. The statement focused on amending the concession system in the light of changing circumstances.

The first 10 formative years of OPEC constituted a change in the relationship between the oil producers and the oil companies in a way that served both interests in the changing conditions of the industry. Although it is true that OPEC was still a dormant partner during those years, within 10 years this partnership had evolved into an active partnership, shifting from the function of simple tax collectors into that of active negotiating partners in determining prices.

3. A SHORT PERIOD OF COOPERATION BETWEEN PRODUCERS AND OIL COMPANIES

After these first 10 years, things took another turn. At the OPEC Conference in Caracas, Venezuela, in December 1970, it was decided to enter collectively into negotiations with the oil companies with a view to raising the price of oil and the tax ratio. Negotiations with the oil companies were first conducted in Tehran, Iran, by delegates from the six OPEC member countries bordering the Gulf: Iran, Iraq, Kuwait, Saudi Arabia, Qatar, and the United Arab Emirates. After several rounds of negotiations, they succeeded in taking a firm collective position to raise the fixed price by approximately \$0.35 and increasing the tax ratio from 50 to 55%.

The Tehran Agreement was concluded in mid-February 1971 but was applied retrospectively from January 1 of that year. It was meant to be for 5 years, after which it was to be renegotiated. In addition to price and tax amendments, the Tehran Agreement provided for the adjustment of the price upward (on an annual basis) to reflect the impact of world inflation, together with a fixed annual increase. Later, and following the two successive devaluations of the U.S. dollar in 1972 and 1973, two agreements were negotiated with the companies and signed by the same signatories of the Tehran Agreement to adjust oil prices and, thus, to preserve the purchasing power of the other currencies relative to the U.S. dollar. The fact that all prices are denominated in U.S. dollars means that any depreciation of the value of the U.S.

dollar could lead to a loss in real value of oil revenues when used to pay OPEC countries' imports from non-dollar countries. The dollar negotiations took place in Geneva, Switzerland, and the Geneva Agreements (known as Geneva 1 and Geneva 2) laid down the formulas with which to correct the oil price (agreed to in the 1971 Tehran Agreement), whether upward or downward, in relation to the movement of the value of the U.S. dollar vis-à-vis other major currencies (U.K. sterling; French, Swiss, and Belgian francs; German Deutschmark; Italian lira; and Japanese yen).

4. THE OIL PRICE SHOCKS

However, after $2\frac{1}{2}$ years of price stability and cooperation between OPEC and the oil companies, the history of OPEC and the oil industry took a dramatic turn, unleashing huge price volatility. In June 1973, OPEC decided to reopen negotiations with the companies to revise upward the Tehran Agreement price in the light of prevailing market conditions. By that time, market prices had reached such high levels that the Tehran Agreement price was left lagging behind, and OPEC argued that the additional windfalls gained by the companies ought to have been shared with the producers. Accordingly, negotiations took place in Vienna, Austria, in early October 1973 but failed as the oil companies refused OPEC's demand for an increase in the price set by the Tehran Agreement. This provoked OPEC into fixing the price of its oil unilaterally and independently of the oil companies' participation. At a meeting in Kuwait on October 16, 1973, that was attended by the same OPEC countries bordering the Gulf, a decision was announced to increase the oil price by 70% so that the OPEC price was fixed at \$5.40/barrel.

At the time of this decision, war was raging between Egypt and Syria, on the one hand, and in the state of Israel, on the other. When Egyptian forces succeeded in crossing the Suez Canal into occupied Sinai, the reaction of the United States and The Netherlands was to rally to the side of Israel. This led to mounting pressure from Arab public opinion to use Arab oil as a "political weapon" to coerce the United States into changing its hard-line policy of continued support for Israel and culminated with the Arab member countries of OPEC meeting to decide to place an oil embargo on both the United States and The Netherlands and to undertake successive cuts in production. It is little wonder that, following that fateful decision, the oil market prices

soared—unchecked and uncontrolled. Meanwhile, the Shah of Iran began a campaign within OPEC to raise the price substantially. At this time, the shah, having enjoyed the support of the West, appointed himself as the “policeman of the Gulf,” a role that he saw as justifying his need to raise huge sums of cash with which to arm his country to the teeth. The reality was that the shah wished to control the area and present Iran as the most dominant power in the whole of the Gulf region.

This drove the Shah of Iran at first to put pressure on the oil companies to raise production until he realized that the most effective means of achieving this was by simply pushing for ever higher prices. He couched this in terms of “a new discourse of oil”: “a noble commodity that should not be burnt” but instead should be reserved for “higher, nobler purposes such as petrochemicals.”

By the time that OPEC’s ordinary (winter) meeting was convened in Tehran in December 1973, market prices had already exceeded \$15/barrel as a result of the Arabs’ so-called oil embargo. The Shah of Iran took every advantage of this and succeeded in increasing the price from \$5.40 to \$10.84/barrel (or an increase of 140%). Relative to the Tehran Agreement and the oil prices during the summer of 1973, this constituted a 400% price increase, causing an enormous shock across the market. Various measures were taken by the consumer nations to slow down oil consumption such as limiting vehicular speed and implementing other energy-saving measures.

The vast majority of OPEC members tended to emulate the Shah of Iran in his quest for price escalation, but without fully comprehending the long-term reaction of such a sudden dramatic increase in oil prices and the negative impact on world demand for OPEC oil. For their part, the Saudis, aware of the economic implications and concerned about the consequential negative impact of price escalation on their own oil, were worried about such sharp oil price increases. One who was particularly worried was Sheikh Ahmed Zaki Yamani, the then Saudi oil minister, who consistently expressed his opposition to any further price increase. However, in OPEC price discussions during the late 1970s, especially during the oil market developments in the wake of the Iranian Revolution, the Saudi kingdom became more amenable to political pressures exerted by the others and, hence, finally acquiesced to most of the price decisions.

The price shock proved to be an inducement for greater efficiency in fuel use so that the amount of gasoline used per mile driven was in continuous

decline. Coupled with this and consumer countries’ fiscal policies, which involved dramatically increasing taxes on gasoline and other petroleum products (especially in Europe), oil became an extremely expensive commodity for end consumers. The resultant dramatic slowdown in the growth of oil consumption indicated how the impact of the price shock took its toll on Western Europe to a far greater extent than on the rest of the world. During the 1960s and early 1970s, oil consumption in Western Europe had grown at an average rate of more than 8% per year, but after the price shock, oil consumption fell from a peak of 15.2 million barrels/day in 1973 to 13.5 million barrels/day 2 years later, settling at approximately 14 million barrels/day thereafter. Conversely, the impact of higher prices on oil consumption in the United States and Japan was far less than in Western Europe, mainly due to those two countries’ far lower level of taxation on oil products.

With a lead time, the price shock heralded a process of structural change in the world oil industry. Significantly, the emerging energy-saving campaigns led to diminishing oil consumption while achieving the same level of economic growth. Prior to the price shock, an increase in gross domestic product (GDP) had entailed an equal increase in oil consumption, but after the shock, the relationship between economic growth and oil consumption changed. A process of gradual “decoupling” of oil consumption from economic growth would not be attenuated later with the advent of even higher oil prices; that is, less oil consumption per unit of GDP became a permanent feature.

No sooner had the world started to live with the high oil prices than another, even greater price shock occurred in June 1979, the effect of Ayatollah Khomeini’s revolution in Iran and, prior to that, the all-out strike of Iran’s petroleum industry, causing a sudden interruption of more than 4 million barrels/day of Iranian oil and consequent market chaos with unparalleled high prices. Although the disappearance of Iranian oil from the market was soon compensated for by the increase in production from Saudi Arabia, Kuwait, and Iraq, the oil market continued to be turbulent.

OPEC’s initial reaction to the rocketing prices on the world market showed some degree of rationality. Because market prices were soaring well beyond the OPEC price, and so were generating great windfalls for the oil companies, the organization (assuming that the situation was only temporary and that production could be increased from other Gulf countries to compensate for the disappearing Iranian oil)

considered adding temporary market premia, intended to be abolished once the market had returned to normal conditions. However, the continued raging markets caused many OPEC members to clamor for changing the initially rational decision and to instead add the premia increase to OPEC's fixed price on a permanent (rather than temporary) basis.

By the time OPEC met in Caracas in December 1979, OPEC's official prices had soared to more than \$24/barrel, and with the markets still in turmoil, OPEC's members continued to add price increments, so that when OPEC met in Bali, Indonesia, a year later, the majority decided to fix the price of OPEC oil at \$36/barrel as of January 1981. However, Saudi Arabia, aware of the negative effects of higher prices on oil consumption and the world economy, refused to raise the price beyond \$32/barrel. Thus, OPEC arrived at a situation whereby a two-tiered pricing system prevailed.

The second oil price shock had an even greater negative impact on world oil consumption, and the soaring oil price served only to promote further world oil supplies from outside OPEC as well as alternative sources of energy. An ineluctable backlash lay in store for OPEC when the oil price across the world market began to fall, led by non-OPEC producers, in particular the North Sea.

5. OPEC AS THE LAST RESORT OIL SUPPLIER: THE QUOTA SYSTEM

OPEC had to either stick with its official price by reducing its output or else follow the market and reduce its price to retain its share of the world oil market. OPEC chose the first option in resorting to the "quota" system, based on the concept of its member countries, as a whole, producing oil in volumes designated only to fill the gap between, on the one hand, total world demand and, on the other, world supplies from outside OPEC; that is, OPEC chose to be the last resort producer or what was termed the "swing producer." This meant that at a certain demand level, the higher the non-OPEC supplies, the lower the OPEC total quotas. Given increasing oil supplies from outside OPEC in the face of weakening world demand, the organization's market share would correspondingly fall naturally and continuously.

Although first discussed in 1982, the OPEC quota system was formally adopted in March 1983 at an extraordinary meeting held in London that set a

production ceiling of 17.5 million barrels/day, reducing the OPEC price to a unified \$28/barrel. This system proved to be even more anomalous given that Saudi Arabia did not have a quota share and was left to be the "swing" producer within a range of 5 million barrels/day. OPEC decided to fix its price and take a reference price based on the crude of Arab Light API 34 f.o.b. Ras Tanura, which Saudi Arabia became committed to charging the oil companies. Meanwhile, the prices of other crudes were to be fixed upward or downward of that reference price, taking into account the quality of the crude and its geographical location in Saudi Arabia. In other words, OPEC now became the swing producer, bearing the burden of price defense at the expense of its world market share; whereas Saudi Arabia bore the brunt by becoming the swing producer within OPEC, absorbing this fall in world demand for OPEC oil far more than did the other OPEC members. This created a situation in which the 12 member countries of OPEC were producing practically the amount of oil commensurate with their allocated quotas, whereas Saudi Arabia was forced to swing downward relative to the fall in the call on total OPEC production, seeing its production in 1985 fall to approximately 2.5 million barrels/day, or roughly one-quarter of its production capacity (leaving three-quarters unproduced). The fall in Saudi Arabia's production was so great that it caused a problem of insufficiency of associated gas for the generation of electricity for water desalination plants.

With this continued decline in the call on its oil, OPEC saw its own market share fall from 62% during the mid-1970s to 37% in 1985, all as a result of increasing supplies from outside OPEC coupled with the fall in demand as a result of high prices and improving energy efficiency in industrial nations. For example, North Sea production (both British and Norwegian) leaped to 3.5 million barrels/day in 1985, up from 2.0 million barrels/day in 1975, and continued to rise thereafter. The reason behind this was that OPEC's high prices provided such a wide profit margin for oil investors that high-cost areas such as the North Sea became not only feasible but highly lucrative areas for continued reinvestment. Because OPEC adhered to the system of fixed price and swing production, any additional oil coming from outside OPEC would first capture its share in the market before buyers resorted to OPEC oil. Also, the greater the supplies of non-OPEC oil, the less OPEC oil that was on the market to meet world demand.

Furthermore, the two oil price shocks triggered enormous investments in scientific research to

improve the efficiency of technology in finding, developing, and producing oil and, thus, in reducing the high cost of upstream operations in these new areas, thereby adding for the oil companies an even higher margin of profit with which to reinvest in high-cost areas.

By 1985, OPEC production was indicative of the backlash suffered by OPEC at the hands of its own price policies. In less than 6 years, OPEC's total production had fallen from 31 million barrels/day to roughly half that amount, and this led to a situation that was increasingly difficult to sustain. Consequently, things got out of hand to the point where Saudi Arabia decided to drop its system of selling oil at fixed prices in accordance with the 1973 quota system and instead adopted market-oriented pricing, or "net-back value," meaning that it obtained the price of its crude from prevailing prices of oil products in the main consuming areas minus the cost of refining, transporting, and handling oil. Saudi Arabia's production started to rise quickly, and by the summer of 1986, a "free-for-all" situation prevailed.

6. THE BACKLASH

This led to a severe collapse in the oil price so that during the summer of 1986 Arab Light was selling at less than \$8/barrel. However, this price collapse generated mounting pressure on Saudi Arabia to forgo its refusal to adopt the quota system. The pressure came first from the OPEC members that had seen their revenues fall so drastically. A large pressure group formed within OPEC (led by Iran, Algeria, Libya, and Venezuela) that was politically powerful enough to campaign against Saudi Arabia, which was supported (within OPEC) only by Kuwait. Meanwhile, pressure came from external quarters when the high-cost producers in the United States were forced to shut down their expensive oil wells, and Texas, where oil is the spur behind all economic activity, sank into a deep recession. Investments in high-cost North Sea oil were similarly jeopardized by the drastic fall in oil prices. This additional external pressure finally pushed Saudi Arabia into changing its policy and agreeing to readopt the quota system with a fixed price for its oil at \$18/barrel.

In January 1987, OPEC changed its pricing system of taking Saudi Arabia's Arab Light as a reference price for other OPEC crudes and replaced it with a "basket" of seven crudes as a reference price: Sahara

Blend of Algeria, Dubai Crude, Saudi Arabia's Arab Light, Minas of Indonesia, Bonny Light of Nigeria, Tia Inana of Venezuela, and Isthmus of Mexico. However, by 1988, OPEC had effectively abandoned the fixed price system, substituting it with an agreement on target prices so that supplies could be regulated to maintain that target. This new system saved OPEC from the headache of having to agree to a fixed price for all other crudes in the light of their quality and geographical locations. However, OPEC has kept to its system of quotas and a total ceiling corresponding to the difference between total global demand and world oil supplies outside OPEC. There is no doubt that this new pricing system made decisions on the OPEC price more flexible than the earlier system, which had caused several litigious problems and endless discussions among member countries about the appropriate price for each crude relative to the reference price.

In deciding to set the price of oil, OPEC had taken on the heavy burden of oil price management with little success in achieving price stability. Yet this had been the original objective on which the organization, when founded, predicated itself. Instead, when OPEC took over oil pricing control in 1973, enormous price volatility ensued, as is clearly shown in Fig. 1. For example, the price of Arab Light crude skyrocketed from \$2 to \$5/barrel during the summer of 1973 to approximately \$40/barrel only 7 years later and then plummeted to roughly \$8/barrel 5 years after that.

7. OPEC'S LIMITATIONS IN OIL SUPPLY MANAGEMENT

In OPEC's management of oil supplies, its limitations were to prove to be serious, with the first and foremost drawback being the intense politicization of the organization that became especially evident once OPEC took over the pricing of oil, at total variance with the original objective when the organization was formed. OPEC oil became a political commodity subject to political pressures and maneuvers. In general, politicians seek immediate or short-term benefits with little regard for the long-term consequences of their decisions. The politicization of OPEC first became evident when the Shah of Iran sought to escalate prices without due regard for the impact of higher prices on world oil demand and the supplies from outside the organization. The heads of state of OPEC's members became involved in the

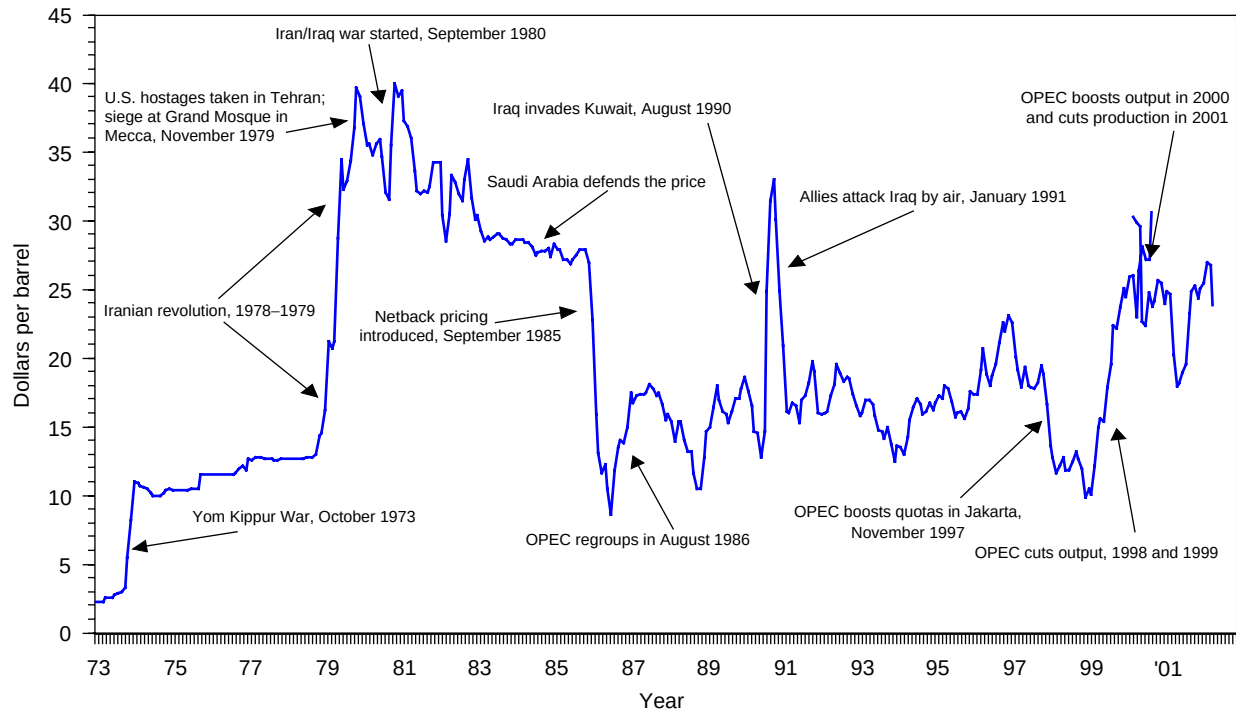


FIGURE 1 Arab Light Spot Prices, 1973–2002 (dollars/barrel). Data from British Petroleum, Center for Global Energy Studies, and Middle East Economic Survey.

organization's oil policies without due regard for each member country's relative importance in terms of oil reserves, in a way as to allow low-reserve countries (e.g., Algeria, Libya) to put political pressure on their high-reserve fellow members to comply with their demands. Thus, OPEC's decision-making process had little to do with sound economics.

OPEC's second limitation is that its decisions on the oil price are pushed by political events that invariably lead to an interruption of oil supplies. Figure 1 shows how the wide fluctuations in the oil price have been caused primarily by political events rather than by market forces. World oil markets can be seen as simply reacting to the interruption of supplies caused by political events. As shown in the figure, the first price shock was the result of the political decision of the Arab oil embargo in 1973, and the second price shock was caused by the Iranian Revolution in 1979. Prices then rose steeply again during the Iraqi invasion of Kuwait in 1990. In 2003, the war with Iraq, followed by lack of security of Iraqi oil pipelines and supplies, once again posed a renewed threat of oil shortfall, forcing market prices upward.

A third and major limitation placed on OPEC's role as an instrument for price stabilization lies in the disintegration of the organization's crude oil production within the international oil industry. This began

when OPEC first wrested control of the oil industry from the international oil companies, creating a predicament quite contrary to the situation that prevailed when the oil companies controlled the upstream in response to downstream requirements, all of which were controlled by the major oil companies in a way as to create a stable market by avoiding a shortage or surplus of crude oil. In OPEC's case, decisions on crude oil production are not organically related to downstream requirements. OPEC production is based simply on the difference between world demand and the production from outside the organization, without any knowledge of the downstream exigencies of consumer countries. This kind of disintegration has inevitably led to price volatility in the world oil market.

OPEC's fourth limitation is that OPEC members are entirely dependent on their oil revenues to meet their foreign currency requirements. This dependence, which amounts to more than 90% of their external trade, means that OPEC's decisions are, in reality, dictated by the short-term financial requirements of its member countries. Moreover, during the 30 years or so since the first price shock, OPEC countries have done practically nothing to diversify their economies and reduce their dependence on oil. These characteristics continue to subject OPEC's

decisions to the pressing short-term needs of its member countries' budgets without looking into any form of economic rationale.

Fifth among OPEC's limitations as an efficient oil supply manager is the frequent failure of some member countries to abide by the quota system. It is noticeable that there is always a difference between the actual OPEC production quotas and the official ones. The degree of quota compliance within OPEC is reinforced when oil prices are very low, affecting its member countries' budgets; conversely, the degree of compliance falls off when oil prices are high.

Perhaps an even more significant limitation to OPEC's decision making is the organization's heterogeneous nature. Not having a homogenous entity has proved to be a major weakness because there are enormous differences among the member countries. For example, those with low oil reserves always seek higher prices so as to maximize their oil revenues by increasing the per barrel income; also, because their production capacity is so limited, they do not even care about the market share or the long-term effects of high prices on demand and supply. Conversely, member countries with large reserves (e.g., Saudi Arabia) in principle have regard for their market share to maximize income from their larger volume and, thus, higher market share. OPEC has in the past tried to formulate a long-term strategy but has never succeeded due to these conflicts of interest within the organization.

8. OPEC AND A FUTURE OF UNCERTAINTY

What future awaits OPEC? So far, OPEC has managed to survive many crises, even those of the 1980s (Table I). Regardless of the problems facing the organization's quota system, OPEC continues to function. Despite differences in the long-term interests of its members, what unites and motivates OPEC, enabling it to reach unanimous decisions, is short-term oil revenue and financial gain. However, its power has been eroded gradually over time as a result of its own policies. OPEC price practices, which amount to deriving short-term gain at the expense of long-term benefit to the oil industry as a whole, have altered the entire world energy structure and made oil far less dominant in its share of energy consumption, to the benefit of natural gas and nuclear power. This is especially evident in the cases of both Western Europe and Japan. Table II shows the extent to which

oil has been losing its share in total energy consumption. Prior to OPEC's price shocks, oil had accounted for 62% of Western Europe's total consumption, yet by the year 2000, oil's share had fallen to 42%. Meanwhile, natural gas's share of total energy increased from 10 to 22%, and nuclear power's share increased from 1 to 13%, during the same period. The shift is more dramatic in the case of Japan, where oil's share of total energy has diminished from 77 to 50%. Meanwhile, its share of natural gas increased from 2 to 13%, and its share of nuclear power increased from 1 to 14%, during the same period.

In other words, OPEC's price shocks triggered a process in the industrialized world of gradually shifting from oil to alternative energies. This shift will accelerate in the future for reasons related to both environment and technological development. Although the United States still refuses to ratify the Kyoto Protocol (the international agreement that aims at a drastic reduction in carbon dioxide emissions through policy measures to reduce the consumption of fossil fuels), there is mounting pressure in both Europe and Japan to comply and, thus, reduce the consumption of oil in favor of more environmentally friendly resources, in particular natural gas, nuclear power, and (in the future) renewable energy resources. Meanwhile, technology favors these shifts because it reduces the costs of otherwise expensive alternatives. For example, the hybrid vehicle consumes much less gasoline than does the standard combustion engine, and eventually fuel cell-driven vehicles will further shrink the demand for oil.

Moreover, OPEC's high price policies have led to heavy investment in oil resources in high-cost areas outside OPEC while technological progress in the upstream has reduced the costs of finding and developing oil, thereby making investments in these alternative high-cost areas very lucrative. The amount of oil that has been produced outside OPEC has been increasing enormously. Excluding the former Soviet Union, non-OPEC oil supplies increased from 18,904 million barrels/day in 1973 to 35,569 million barrels/day in 2000. After 1982, when OPEC had opted to be the last resort supplier, the increase in non-OPEC oil supplies has been at the expense of OPEC's market share, the decline in which has occurred specifically during the same period. The less its market share, the more difficult it is to manage oil supplies at a certain stabilized price level. In addition to the price effects, geopolitics has played a significant role in changing the map of world oil supplies.

TABLE I

OPEC: A Chronology of Events

1960 (September 16)	In Baghdad, Iraq, the foundation of OPEC was announced by the five major oil-producing countries (Iran, Iraq, Kuwait, Saudi Arabia, and Venezuela)
1961	Qatar joined OPEC, followed by Indonesia and Libya (1963), the Emirates of Abu Dhabi (1968), Algeria (1969), Nigeria (1971), Ecuador (1973), and Gabon (1975). The latter two members left OPEC during the 1990s
1966 (June)	OPEC issued a “declaratory statement of petroleum policy in member countries,” a historical turning point, as it laid down basic principles that would later influence all major events of the 1970s. The statement emphasized the right of producing countries to fix unilaterally the price of their oil, in an endeavor to give the states a greater role in the development of hydrocarbon resources and to give the states the right to participate in concession-holding agreements. The emphasis was on amending the concession system in the light of changing circumstances
1970 (December)	In Caracas, Venezuela, the OPEC Conference was held. A decision was made to negotiate collectively with the oil companies to establish a general increase in posted prices and to raise the tax ratio to a minimum of \$55
1971 (February)	In Tehran, Iran, the Tehran Agreement (to be of 5 years’ duration and then renegotiated) was the outcome of this first round of negotiations with the oil companies, conducted by delegates from the six OPEC member countries bordering the Gulf (Iran, Iraq, Kuwait, Saudi Arabia, Qatar, and the United Arab Emirates). In addition to price and tax amendments, the Tehran Agreement provided for the adjustment of the price upward (on an annual basis) to reflect the impact of world inflation, together with a fixed annual increase
1972 and 1973	Two successive devaluations of the U.S. dollar led to two further agreements (in Geneva, Switzerland) to adjust prices and preserve the purchasing power of the other currencies relative to the U.S. dollar. Known as the Geneva Agreements (“Geneva 1” and “Geneva 2”), these two agreements laid down the formulas with which to correct the oil price (agreed in the 1971 Tehran Agreement), whether upward or downward, in relation to the movement of the value of the U.S. dollar relative to other major currencies (U.K. sterling; French, Swiss, and Belgian francs; German Deutschemmark; Italian lira; and Japanese yen)
1973 (June)	OPEC decided to reopen negotiations with the companies to revise upward the Tehran Agreement oil price in the light of prevailing market conditions, where oil prices had reached such high levels that the Tehran price was left lagging behind, and OPEC also argued that additional windfalls gained by the companies ought to have been shared with the producers
1973 (July)	In Vienna, Austria, OPEC negotiations with the oil companies failed, with the latter refusing OPEC’s demand for an increase in the price set by the Tehran Agreement. This provoked OPEC into fixing the price of its oil unilaterally and independently of the oil companies’ participation
1973 (October 16)	In Kuwait, the same OPEC Gulf members decided to increase the oil price by 70% so that the OPEC price was fixed at \$5.40/barrel
1973 (October)	The Arab oil embargo was implemented
1973 (December)	In Tehran, OPEC’s ordinary (winter) meeting was held. Spurred on by the Arab oil embargo, the Shah of Iran pushed for increasing the price yet further by 140%, from \$5.40 to \$10.84/barrel, a 400% increase relative to the Tehran Agreement
1979	During early part of year, an all-out petroleum industry strike took place in Iran, forcing oil prices upward to an unparalleled level
1979 (June)	The Ayatollah Khomeini’s revolution and political coup occurred in Tehran
1979 (December)	In Caracas, with OPEC’s official prices having soared to more than \$24/barrel, OPEC members continued with price increments
1981 (January)	In Bali, Indonesia, the majority of OPEC members decided to fix the price of OPEC oil at \$36/barrel. Saudi Arabia refused to raise the price beyond \$32/barrel, giving rise to a two-tiered pricing system
1983 (March)	In London, an extraordinary OPEC meeting was held. Instead of substantially reducing its price to retain its share of the world oil market, OPEC stuck with its official price by reducing its output through a “quota” system. OPEC opted to be the last resort supplier with a production ceiling of 17.5 million barrels/day. The official OPEC price was now set at a unified \$28/barrel
1985	The extent of the petroleum backlash was evident when OPEC total production had already fallen by half from 31 million barrels/day
1986	A “free-for-all” situation prevailed with a rapid rise in Saudi Arabia’s production, culminating in an oil price collapse in July, with Arab Light selling at less than \$8/barrel
1986	Pressure groups form within OPEC and among other countries to push Saudi Arabia into readopting the quota system with a fixed price for its oil at \$18/barrel

continues

Table 1 continued

	OPEC changed its pricing system (of taking Arab Light as a reference price for other OPEC crudes) and replaced it with a "basket" of seven crudes as a reference price (Sahara Blend of Algeria, Dubai Crude, Saudi Arabia's Arab Light, Minas of Indonesia, Bonny Light of Nigeria, Tia Inana of Venezuela, and Isthmus of Mexico). However, by 1988, OPEC had effectively abandoned the fixed price system, substituting it with an agreement on "target prices" so that supplies could be regulated to maintain that target
1987 (January)	
1990 (August)	Iraq invaded Kuwait. Oil prices rose steeply. UN sanctions were imposed on Iraq
1991 (January)	The Gulf War with Iraq took place
1997 (September)	The Kyoto Protocol compelled signatory nations to set a deadline for reducing carbon dioxide emissions to diminish global warming. This favored the adoption of alternative cleaner energies
2003 (March 20–April 17)	The second Gulf War involving Iraq took place

TABLE II

Percentage Shares of Total Energy Consumption 1973–2001

	Western Europe			Japan			North America		
	1973	1985	2001	1973	1985	2001	1973	1985	2001
Oil	62	45	42	77	57	48	45	41	39
Natural gas	10	16	22	2	10	14	30	26	25
Coal	20	25	14	15	20	20	18	24	23
Nuclear	1	10	13	1	10	14	1	6	8
Hydro	7	3	8	5	2	4	6	3	5

Source. Data from British Petroleum.

The most critical issue currently is the question of the security of Gulf oil supplies, with this area being considered prone to political instability, a factor affecting oil supplies and prices. This geopolitical factor had added momentum to greater investment in oil supplies from outside the Gulf. The understanding reached in 2002 between the United States and the Russian Federation indicates how U.S. policy is oriented toward reducing dependence on the Gulf in favor of Russian and Caspian Sea oil as well as toward obtaining increased oil supplies from West Africa (mainly Nigeria, Angola, and Equatorial Guinea) and from offshore oil production in the Gulf of Mexico.

An even greater contributory factor to the weakening of OPEC's control over oil prices is the huge investment that some member countries (e.g., Nigeria, Algeria) have designated for themselves to increase their production capacities. Once these expansion programs come on-stream, it is very difficult to perceive that OPEC can continue to subject these countries to the restrictions of its production quota regime for the sake of shoring up the oil price.

Perhaps even more significant will be the question of Iraq, a country extremely rich in oil deposits that have been discovered but have remained undeveloped and that could provide a huge increase in that country's production capacity. It has been estimated that, given favorable political and financial conditions, Iraqi production could exceed 6 million barrels/day and could even reach 8 million barrels/day. It is obvious that the organization would be unable to restrict Iraqi production after the country's dire economy has been so severely held back since the Gulf War and UN sanctions.

From the preceding picture, the prospects for the future of OPEC may appear grim. However, in the event of the organization's collapse, oil prices would immediately collapse across the world markets, affecting all investments in high-cost areas outside OPEC. Such a scenario would renew the demand for cheaper OPEC oil, and for geopolitical reasons, this would not be favored by the developed nations. For this reason, it is very difficult to make any accurate prediction concerning the future of OPEC and the political and economic forces that may reshape the future status of the organization.

The fall of Saddam's Ba'athist regime in Iraq will have far-reaching effects on OPEC and on the oil industry in general. Iraq's oil potential is huge and could ultimately reach production capacity levels near those of Saudi Arabia. During the 1970s, many important giant oilfields were discovered but remained undeveloped due to Saddam's wars and the UN sanctions. In a matter of a few years, it would be technically feasible to reach a capacity of more than 6 million barrels/day. But meanwhile, the current administration is primarily concerned with restoring production and exports from the existing oilfields together with plans to rehabilitate the oil sector.

The first phase of developing Iraqi oil is to restore Iraq's capacity to its 1990 pre-UN sanctions level of 3.5 million barrels/day. Even with huge investments, this may take at least 2 years. It is not yet clear what policy would be adopted for the next phase, which is the expansion of Iraq's oil capacity to 6 million barrels/day or even to 8 million barrels/day. It goes without saying that foreign investment in Iraq would not be possible without first establishing a credible and stable Iraqi government with a sound legal system that would safeguard the interests of foreign investors.

SEE ALSO THE FOLLOWING ARTICLES

Economic Geography of Energy • Geopolitics of Energy • Global Energy Use: Status and Trends • International Energy Law and Policy • National Energy Policy: United States • Nationalism and Oil • National Security and Energy • Oil Crises, Historical Perspective • Oil-Led Development: Social, Political, and Economic Consequences • Oil Industry, History of

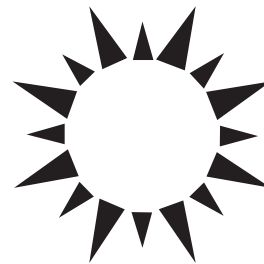
Further Reading

- Amoco, B. P. (1970–2001). “Annual statistical reviews.” – BP Amoco, London.
- Center for Global Energy Studies. (1990–2002). “Global Oil Reports.” CGES, London.
- Chalabi, F. J. (1980). “OPEC and the International Oil Industry: A Changing Structure.” Oxford University Press, Oxford, UK.
- Chalabi, F. J. (1989). “OPEC at the Crossroads.” Pergamon, Oxford, UK.
- Penrose, E. (1968). “The Large International Firm in Developing Countries: The International Petroleum Industry.” Allen & Unwin, London.
- Sampson, A. (1975). “The Seven Sisters.” Hodder & Stroughton, London.

OPEC Market Behavior, 1973–2003

A. F. ALHAJJI

Ohio Northern University
Ada, Ohio, United States



-
1. Introduction
 2. Overview
 3. Oligopoly Models
 4. Competitive Models
 5. OPEC and the Cartel Status
 6. Conclusion

Glossary

cartel A group of producers that collude to reduce production and increase prices in order to maximize the wealth of the group.

monopoly A market with a single producer that sells a product having no close substitute.

nationalization The host government takeover of foreign companies' operations.

netback pricing A method of determining the wellhead price of oil and gas by deducting from the downstream price, the destination price, the transportation price, and other charges that arise between wellhead and point of sale.

oligopoly A market with few competing producers.

OPEC basket price The average price of seven marker crudes.

OPEC quota A specified maximum amount of oil production assigned by OPEC to each of its members.

Organization of Petroleum Exporting Countries (OPEC) Historically, it consisted of 13 countries: Saudi Arabia, Kuwait, Iraq, Iran, Qatar, Abu Dhabi, Libya, Algeria, Nigeria, Venezuela, Indonesia, Gabon, and Ecuador. Ecuador and Gabon withdrew from the organization in 1992 and 1995, respectively.

posted price A price of a barrel of oil that the host nations used as a base to collect taxes from the international oil companies, regardless of the market price. In most cases, posted prices were higher than market prices by a wide margin. After nationalization, countries decided to make it the market price.

price band The minimum and maximum for the average OPEC basket price, set by OPEC in 2000.

Seven Sisters Refers to the seven international oil companies that dominated the world oil industry in the 1930s and 1940s: Exxon, British Petroleum, Shell, Gulf, Texaco, Socal, and Mobil.

user's cost The cost of replacing oil that has already been produced. It is a concept used only with nonrenewable resources.

Two supply factors are known to influence oil prices and dominate research activities: scarcity and market power. Scarcity increases prices above marginal cost by the discount rate, even in a competitive market. In noncompetitive markets, prices are usually higher than marginal cost. However, events of the past 80 years indicate that the market power of the Texas Railroad Commission, international oil companies, and Organization of Petroleum Exporting Countries (OPEC) has had a more significant impact on oil prices than scarcity. For this reason, most researchers focus on OPEC when they discuss world oil markets of the past 30 years. Nevertheless, this focus on market power without regard to depletion and technical issues has polarized academic and political circles to the extent that it is almost impossible to reach a consensus on OPEC and its role in world oil markets. It is also difficult to discuss OPEC behavior without putting analyses in their historical context. It is necessary to understand the political, economical, social, and philosophical factors that led to the establishment of OPEC and its role today. Oil is not only a nonrenewable resource but also a political and strategic commodity.

1. INTRODUCTION

Some researchers argue that oil is a special commodity that requires intervention and depletion management.

The oil industry is characterized by booms or busts, which calls for additional controls not provided by market forces. This extra control seeks to avoid or minimize the boom-or-bust cycle. Standard Oil, the Texas Railroad Commission, the Seven Sisters, and the Organization of Petroleum Exporting Countries (OPEC) have all exercised it with varying degree of success. OPEC has been the least successful and failed to manage depletion. Some researchers even argue that U.S. embargos on some of the oil-producing countries are a form of oil market management to support oil prices. OPEC does not have grand strategies and plans. It pragmatically reacts to various political, economical, and natural factors that influence its members, but the impact of these factors varies among members at any given time. Thus, OPEC must reconcile the divergent interests of its membership through meetings and negotiations.

The focus of this article is OPEC market behavior. Therefore, only market models introduced in the past 30 years are discussed. An overview of the many oil market models developed from 1973 to present shows that these models divide easily into oligopoly models and competitive models. Most researchers use oligopoly models to explain OPEC behavior and assume that OPEC, or a subset of it, is a cartel that seeks monopoly profits by changing production and influencing prices. Others argue that the world oil market is competitive and that factors other than cartelization better explain oil price increases. For each side, there are various models that explain OPEC behavior.

A few researchers argue that OPEC is a wealth maximizer. Most researchers argue that OPEC functions not to maximize wealth but to maximize revenues and achieve other social and political purposes. These contrasting views resulted from the emphasis of most researchers on market power rather than on market failure and depletion management.

Section 2 presents an overview of OPEC and non-OPEC production, world oil demand, oil prices, and OPEC's actions. Section 3 reviews and critiques the oligopoly models of OPEC. The competitive models are presented in Section 4. Section 5 provides various points of view regarding the application of the term cartel to OPEC.

2. OVERVIEW

Figures 1 and 2 show oil prices, OPEC production, non-OPEC production, and world demand for oil between 1973 and the first quarter of 2003. Major

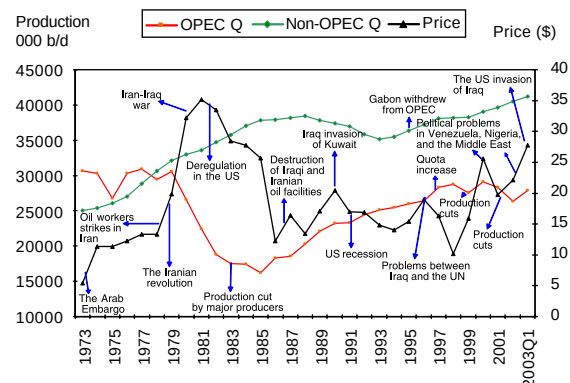


FIGURE 1 Oil prices, OPEC and non-OPEC production (thousands b/d), and major events, 1973–2003. Data from *Monthly Energy Review*, Energy Information Administration, July 2003. Data for 2003 are for the first quarter.

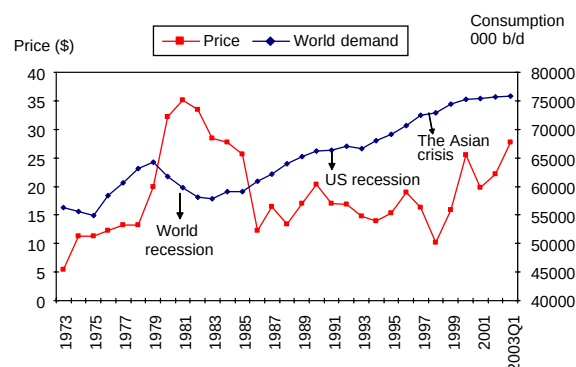


FIGURE 2 Oil prices and world demand (thousands b/d), 1973–2003. Data From “BP Statistical Review of World Energy” (2003) and Energy Information Administration (2003). Data for 2003 are estimates for the first quarter.

political events and some OPEC actions are plotted in Fig. 1 to illustrate their impact on prices and oil production. OPEC as an organization did not cause energy crises, oil price spikes, and price volatility. However, various political incidents that involved individual OPEC members did contribute to energy crises and market instability. Researchers provide various models and explanations for such changes in production and prices. From the information in Figs. 1 and 2, the period between 1973 and 2003 can be divided into six phases: 1973–1978, 1979–1981, 1982–1986, 1987–1991, 1992–1997, and 1998 to the present.

2.1 The First Oil Crisis: 1973–1978

Oil prices increased in 1973 when, on October 19, several Arab countries—all of them allies of the

United States—imposed an embargo on the United States and The Netherlands for their support of Israel in the 1973 war. Although the embargo led to a mere 5% decrease in world oil supply and lasted only 2 months, oil prices increased fourfold. Theory predicts that oil prices will return to their original level when an embargo is lifted, but prices in 1974 remained high. World oil demand decreased at first because of substitution, conservation, and lower economic growth. However, it continued its upward trend once the economies of the consuming countries absorbed the impact of the energy shock.

This production cut by some OPEC members and the subsequent spike in prices led some researchers and observers to conclude that OPEC had become a powerful cartel. They point out that non-OPEC members behaved differently and increased production to take advantage of the higher oil prices. Others believe that different factors led to the production cut and that price increases were not related to cartelization, including changes in property rights, the limited capacity of the economies of OPEC members to absorb productive investment, environmental concerns about natural gas flaring by some OPEC members, panic and stockpiling in the consuming countries, and U.S. wellhead price control. This last factor kept U.S. oil prices below the world price, decreased U.S. oil production, increased U.S. demand for oil, and increased oil imports.

2.2 The Second Oil Crisis: 1979–1981

Oil prices increased again in 1979 as a result of the Iranian revolution. Only Saudi Arabia and some non-OPEC members increased their production to compensate for the lost Iranian oil. Prices continued to rise when the Iraq–Iran war started in September 1980. Iraqi and Iranian oil production came to a complete halt by the end of 1980 and the world lost 9 million barrels per day (mb/d). Coupled with panic and stockpiling, spot oil prices reached historic levels. However, OPEC contract prices were much lower, especially those of Saudi Arabia. These prices were set by OPEC members, not the market.

Conservation, substitution, and a worldwide recession brought a substantial decline in world oil demand in 1980 and 1981. Meanwhile, non-OPEC production continued to increase. As a result, OPEC output and market share decreased. Two major factors affected oil prices and OPEC behavior in 1981: price deregulation in the United States and price agreement among OPEC members. The agreement set price differentials for each OPEC crude

relative to the Saudi marker. This agreement indicated that Saudi Arabia was not willing to defend prices alone. It wanted other members to share this burden. The agreement ensured that if the price of the Saudi marker declined, the price of other member's crudes would decline too. Therefore, countries had to cooperate with Saudi Arabia and cut output to preserve prices. This was the first serious coordinated attempt by OPEC to influence world oil markets. Market management requires price control, output control, or both. OPEC chose the first method in 1981. By the time OPEC agreed on prices, a decrease in demand led Nigeria unilaterally to cut its prices by \$5.50/b. First Iran and then other OPEC members followed. As a result, the price agreement collapsed.

2.3 The Oil Glut: 1982–1986

A decline in world oil demand and an increase in non-OPEC production led to a sharp decrease in OPEC market share between 1982 and 1986. The fundamentals of world oil markets changed drastically to the extent that massive production cuts by some OPEC members were not enough to prevent an oil price collapse, as shown in Fig. 1.

The breakdown of the first price agreement in 1981 led OPEC to consider another method of market control: production management. For the first time, OPEC established a loose quota system in March 1982. In this agreement, Saudi Arabia was not assigned a quota with the understanding that it would behave as a swing producer to balance the world oil market. In addition, Iraq was allowed to produce more than its quota. Many countries were dissatisfied with the system, which broke down almost immediately. OPEC members violated their quota by more than 2 mb/d by the end of 1982.

The 1983–1986 period witnessed a sharp decrease in OPEC market share, especially that of Saudi Arabia; an increase in cheating by OPEC members; a decline in world demand; and an increase in non-OPEC production. Consequently, OPEC was forced to reduce its posted marker prices and lower its production ceiling. Saudi Arabia continued to defend oil prices by cutting production until its output declined to only 2 mb/d, an amount insufficient to sustain the Saudi economy. For the first time since the establishment of the quota system, OPEC assigned Saudi Arabia a quota of 4.35 mb/d in order to protect Saudi Arabia's market share. This figure was less than half of what Saudi Arabia was producing in 1980. In addition, OPEC abandoned the posted price

mechanism. It abandoned Saudi Arab Light as the marker crude in favor of market-oriented netback prices. Prices were no longer administered by OPEC through the posted prices mechanism. OPEC decided to take the market price and produce accordingly. As a result of these developments, the price plummeted to single digits in August 1986.

This collapse reversed historical trends of increased substitution and conservation. It stimulated the world demand for oil, as shown in Fig. 2. It also decreased non-OPEC production by shutting down stripper wells in the United States and other countries.

2.4 Calm before the Storm: 1987–1991

Four factors brought about increased oil prices in 1987: major damage to Iraqi and Iranian oil facilities, less cheating by OPEC members, reduction in non-OPEC output, and an increase in oil demand. Netback pricing was abandoned by OPEC members by 1988 in favor of a sophisticated pricing formula that linked OPEC crude prices to market prices.

Oil prices improved by 1989, and OPEC regained market share because of a decline in non-OPEC production and an increase in demand. Prices increased to more than \$40/b when Iraq invaded Kuwait in August 1990. The world lost 4.5 mb/d, but Saudi Arabia, along with other oil-producing countries, was able to compensate for the loss of Iraqi and Kuwaiti oil. Although oil prices remained relatively high due to the Gulf War, the decline in world demand and the U.S. postwar recession resulted in reduced oil prices in 1991. A United Nations (UN) embargo on Iraqi oil went into effect in 1991. Iraqi oil did not find its way into world markets legally until 1996. The UN lifted the embargo in April 2003 after the U.S. invasion of Iraq.

2.5 Post Gulf War: 1992–1997

A combination of factors caused oil prices to plunge in 1993, including the expected return of Iraqi exports to world oil markets, surging North Sea output, and weak demand. At the end of 1997, OPEC decided to increase its quota by 10% for the first half of 1998 on the erroneous assumption that Asian oil demand would continue to grow. Unfortunately, the Asian financial crisis of 1997 proved OPEC wrong. Oil prices began to decrease. OPEC ignored many signals at that time. For example, the increase in oil prices in 1996 was caused not only by an increase in Asian demand but also by the delay in signing the oil-for-food program between Iraq and the UN.

2.6 Recent Developments: 1998–Present

Despite very strong economic growth in the United States and many European countries in 1998, oil prices continued to decrease until they reached single digits in the first quarter of 1999. Analysts cited four reasons for this decline: the Asian financial crisis, the increase in OPEC and non-OPEC production, the increase in Iraqi exports, and two consecutive mild winters in the consuming countries.

OPEC efforts to halt the 1998 decline in oil prices failed despite two announced cuts and the promise of output cuts by some non-OPEC members. The collapse of oil prices in the first quarter of 1999 forced OPEC and non-OPEC members to announce production cuts by 2.104 mb/d in March 1999. OPEC decided to continue the cuts when it met in September of that year, and oil prices continued to rise, as shown in Fig. 1.

The sharp increase in oil prices in 2000 forced OPEC to increase its quotas in March, June, September, and October. However, these increased quotas did not translate into actual production. Many OPEC members lacked the spare capacity. Three factors characterized 2000: frequent meetings, frequent quota increases, and the unofficial creation of a price band. In October, OPEC activated the unofficial price band for the first time and increased production by 500,000 b/d.

OPEC set up a price band mechanism during its March 2000 meeting. According to the mechanism, OPEC would increase production automatically by 500,000 b/d if the average OPEC basket price remained above \$28/b for 20 consecutive trading days. Production would be cut by a similar amount if the basket price remained below \$22/b for 10 consecutive trading days. The automatic adjustment was later abandoned so that OPEC members could adjust production at their discretion. The creation of the price band is OPEC's fourth serious attempt to manage world oil markets. At first, OPEC tried to control prices through creating price differentials for each OPEC crude in 1981. The failure of this pricing mechanism led OPEC to adopt a quota system to control output in 1982. The failure of this move led OPEC to adopt netback pricing in 1985. Theoretically, netback pricing with the quota system in place should have enabled OPEC to control both prices and outputs. However, netback pricing, a market-oriented pricing mechanism, annulled the quota system, which is a form of market management. OPEC failed to control the market for the third time. Creating the price band may have enabled OPEC for

the first time to use both outputs and prices to manage world oil markets.

The sharp increase in oil prices in 1999 and 2000 led to lower demand and higher non-OPEC production in 2001. Consequently, oil prices declined and OPEC was forced to cut production twice—in January by 1.5 mb/d and in March by 1 mb/d. The demand for oil declined again at the end of 2001 after the September 11 terrorist attack on the United States. The decline in air travel and lower economic growth were the primary reasons for the decline in world oil demand. Consequently, OPEC agreed in November 2001 to cut production by 1.5 mb/d effective January 1, 2002, but only if non-OPEC producers cut their output by 500,000 b/d as well. With non-OPEC cooperation, the cut was implemented, although data indicate that non-OPEC cooperation may have been merely symbolic.

Despite lower world economic growth in 2002 and higher world oil production, political problems pushed oil prices higher. Iraq halted its official oil exports in April for 1 month to protest the Israeli incursion into the West Bank. Meanwhile, an attempted coup forced Venezuelan President Hugo Chavez to resign on April 12, only to resume his presidency on April 14.

Oil prices had been increasing since May 2002. A general fear that the war with Iraq would disrupt the Middle Eastern oil supplies initiated the increase. Labor strikes in Venezuela that halted oil exports forced prices to increase to more than \$30/b in December 2002. Both Iraq and Venezuela are major oil producers and OPEC members.

OPEC met in December 2002 and decided to take unusual action—to cut oil production by increasing quotas and cutting production at the same time. OPEC members were violating their quota by 3 mb/d. OPEC decided to increase quotas by 1.3 mb/d and cut production by 1.7 mb/d. When prices continued to increase, OPEC met a month later and increased production by 1.5 mb/d.

Oil prices continued to increase in 2003. The possibility of a U.S. invasion of Iraq and political tensions in several oil-producing countries kept prices high despite the substantial increase in OPEC production and low economic growth in the consuming countries. Analysts attributed the increase in prices to the “war premium,” which is the difference between market prices and what prices would have been without the threat of war in Iraq.

Contrary to expectations, oil prices started to decline when President Bush issued an ultimatum to the president of Iraq, Saddam Hussein. He gave him

48 h to leave Iraq or face war. Prices continued to decline during the war. Several factors contributed to this decline. First, the war premium disappeared as traders became certain and optimistic about the outcome of the war. Second, several countries started selling their oil that was stored in the Caribbean and on large tankers in the hope that prices would be higher during the war. Third, several countries increased their oil production, especially Saudi Arabia, as shown in Fig. 3. Fourth, traders were convinced that the Bush administration would use the Strategic Petroleum Reserves if prices increased. Fifth, traders expected Iraq to start producing in large quantities very soon after the end of the war.

Oil prices did not decline substantially as many experts expected. The delay in Iraq exports and lower inventories kept prices at approximately \$30/b in the United States. Despite talk of production cuts, only Saudi Arabia cut production by 200,000 b/d by May, whereas most members increased production.

3. OLIGOPOLY MODELS

Production by non-OPEC members, which comprises approximately 60% of world production, eliminates the monopoly models for OPEC. Monopoly in the literature refers to a single producer controlling the whole market. Hence, the focus is on oligopoly models, in which the world oil market is shared between a dominant producer and the competitive fringe. Most models that assign market power to OPEC or to some subset of its members are categorized as one of three models: OPEC as a dominant producer, the OPEC core as a dominant producer, or Saudi Arabia as a dominant producer. Although OPEC literature frequently contains expressions such as “monopoly,” “monopoly power,” and “monopoly profit,” researchers are actually endorsing oligopoly models.

3.1 OPEC as a Dominant Producer

This model assumes that OPEC is a monolithic cartel in which countries have unified goals and collectively set the price for oil. Non-OPEC oil producers are the competitive fringe. The demand for OPEC’s oil is the residual demand, the difference between world demand and non-OPEC supply. OPEC sets the price where its marginal revenue equals its marginal user’s cost. The competitive fringe equates that price with their marginal user’s cost and supplies accordingly; OPEC supplies the rest.

The model assumes that OPEC became a powerful cartel in 1973. OPEC was able to cut production, increase prices, and transfer wealth from the oil-consuming countries to the oil-producing countries.

This model may explain changes in OPEC and non-OPEC production, world demand, and changes in oil prices from 1973 to the present. As shown in Fig. 1, OPEC initially cut production and prices continued to increase between 1973 and 1981. The higher price was a self-defeating strategy for OPEC. It decreased demand through conservation, substitution, and fuel utilization. In addition, it increased non-OPEC supplies. The remaining residual demand decreased. OPEC lost its market share in the early 1980s. Consequently, oil prices and OPEC oil export revenues declined.

The lower price in the mid-1980s stimulated world demand and slowed the growth of non-OPEC production. Such development increased the residual demand for OPEC's oil and increased oil prices. The same situation explains the behavior of world oil markets from the 1990s to the present.

3.1.1 *Criticism of the Model*

This model does not apply to OPEC for the following reasons:

1. OPEC does not meet the characteristics of a cartel as stated in the economic literature.
2. The model assumes that production decisions are made by OPEC authority and not by the countries. However, several examples show that countries have acted unilaterally.
3. Historical evidence indicates that production cuts in the 1970s were related to the deterioration of technology after nationalization in several countries, such as Libya and Venezuela.
4. Statistical tests do not support this model.

3.2 The OPEC Core as a Dominant Producer

The failure of the previous model to fit OPEC behavior led researchers to introduce more sophisticated models. To be more realistic, researchers acknowledge that OPEC has a market power and its members have different interests. Three models have been introduced: the two-part cartel, the three-part cartel, and the core as monolithic cartel.

3.2.1 *Two-Part Cartel Model*

This model assumes that OPEC is a nonuniform cartel. OPEC is divided into two groups—saver and

spender countries. The saver countries are the United Arab Emirates (UAE), Iraq, Kuwait, Libya, Qatar, and Saudi Arabia. The spender countries are Algeria, Ecuador, Indonesia, Iran, Nigeria, and Venezuela. The two groups can compete or collude with each other. In the case of competition, the saver countries will act as a dominant producer and set the price that maximizes profit. The spender countries will act as a competitive fringe. In the case of collusion, both groups will form a monolithic cartel. The outcome will be similar to the outcome in the monolithic cartel model discussed previously.

Applying the model to Figs. 1 and 2, the saver countries that set the world price made production cuts, whereas spender countries continued to increase their production. Once the quota was implemented in 1982, the saver countries cut production, whereas the spender countries violated their quotas. As world demand decreased and non-OPEC production increased, the core's market share shrank. Later, a trend reversal led to an increase in prices and an increase in the core's share. A similar explanation can be used to illustrate changes in output and prices up to the present.

3.2.2 *Three-Part Cartel*

OPEC is divided into three groups:

1. The cartel core countries: Kuwait, Libya, Qatar, Saudi Arabia, and the UAE. These countries have vast oil reserves, small populations, and flexible economic development. The core acts as the dominant firm and sets the price of oil. The cartel core nations carry excess capacity in order to maintain prices during emergency demand.
2. The price-maximizing countries: Algeria, Iran, and Venezuela. These countries have relatively low oil reserves and large populations with potential for economic development. These countries aim for higher prices by cutting production.
3. The output-maximizing countries: Ecuador, Gabon, Indonesia, Iraq, and Nigeria. They have limited reserves, large populations, and a pressing need for economic development. These countries sell at any price, even in a weak market.

Applying this model to Figs. 1 and 2 explains OPEC behavior in a way similar to that of the previous model. However, this model shows that some countries pushed for higher prices when prices declined. This situation may explain the reversal of oil prices in 1987, despite the fact that Saudi Arabia had more than doubled its production by that time.

3.2.3 Core as a Monolithic Cartel

This model differs from the previous two models in that it does not assume that OPEC is a cartel, but that the core countries behave as a cartel that sets the price while all other producers, including the rest of OPEC, behave as a competitive fringe. The choice of core members (Kuwait, Qatar, Saudi Arabia, and the UAE) is not based on reserves and population but on economic and political factors. This model may explain OPEC behavior in the same manner as before, especially in the OPEC dominant producer model.

3.2.4 Criticism of the Model

This model does not match actual OPEC behavior for the following reasons:

1. The division of OPEC into two and three groups was subjective. Researchers did not base the division on theoretical, behavioral, or statistical analyses. They used factors that are not directly related to production decisions, such as reserves and population. Theoretically, production capacity, not reserves, has an impact on production. Statistically, various tests indicate that neither reserves nor population has a direct impact on production levels and prices.

2. Historical data do not support this model. Countries within each group did not act in unison, and some countries switched positions occasionally.

3. According to the three-part cartel model, the price-maximizing countries cut production with higher oil prices. At the same time, the core increases its production to offset the reduction, and vice versa in the case of low prices. Historical data do not support such conclusions.

4. For the core to be a cartel, it must operate as a dominant producer. Statistical tests do not support such behavior.

5. Given the sheer size of Saudi production relative to the production of other core members, it is very difficult to distinguish between the impact of the core and the impact of Saudi Arabia.

3.3 Saudi Arabia as the Dominant Producer

This model assumes that OPEC is not a cartel. Saudi Arabia acts as the dominant producer. All other producers behave as a competitive fringe. Saudi Arabia sets the price. All other producers expand their output to the point at which that price equals

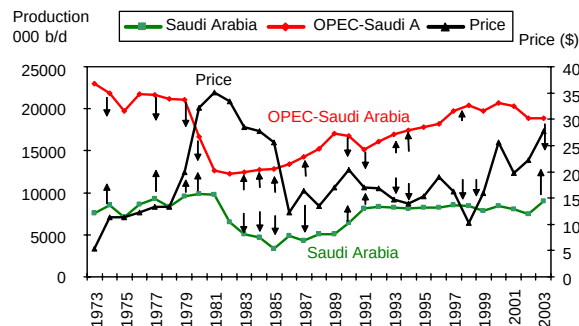


FIGURE 3 Saudi Arabia and other OPEC countries production (thousands b/d) and oil prices. Arrow direction indicates that Saudi Arabia's actions differ from those of other OPEC members. Source: Energy Information Administration (2003, January).

their marginal user's cost. Saudi Arabia then supplies the rest to meet the world demand.

This model could be the most plausible for explaining OPEC's behavior in the past 30 years. It gained support over the years even from researchers who first concluded that OPEC was a cartel. Looking at Figs. 1 and 2, this model can explain the changes in the world oil market in a similar way as the OPEC as a dominant producer model. The only difference is that production cuts and increases are made by Saudi Arabia, not by OPEC.

There are several factors that lend support to this model. First, historical data indicate that only Saudi Arabia has large excess production capacity. It is the only country to voluntarily reduce capacity and production. The 1 mb/d increase in Saudi production before and during the invasion of Iraq led to a \$9/b price decline from preinvasion levels. The price during the invasion of Iraq was at least \$15/b less than the price that analysts expected.

Second, Saudi Arabia's output is negatively correlated with that of other OPEC members, as seen in Fig. 3. The opposite arrows point to the direction of change in production. For example, in 1998, Saudi production decreased (arrow is pointing down) and the production of the rest of OPEC members increased (arrow is pointing up).

Third, Saudi Arabia was the only OPEC member not assigned a quota when OPEC implemented the 1982 quota system, so Saudi Arabia could act as "swing producer" to stabilize the world oil market.

Finally, statistical tests indicate that Saudi Arabia has operated on the elastic part of its demand curve since 1973. Operation on the elastic part of the demand curve is a requisite for the dominant producer model to apply.

3.3.1 Criticism of the Model

This model does not apply to OPEC for the following reasons:

1. The model assumes that the behavior of other OPEC members is similar to that of non-OPEC members. Historical data do not support this assumption.
2. The model does not explain why the Saudis pressed especially hard to convince non-OPEC members, such as Mexico, Norway, and Russia, to cooperate in production cuts. However, political factors may explain such behavior.
3. Figure 3 indicates that the model may not fit after 1999. Data show concerted efforts between Saudi Arabia and other OPEC members to cut production. Although Fig. 3 shows that Saudi Arabia increased production while the rest of OPEC decreased production in the first quarter of 2003, the decrease occurred because of the U.S. invasion of Iraq. Data indicate that all remaining OPEC members increased production during this period.
4. Saudi political objectives contradict this model. Political scientists who try to explain OPEC behavior claim that Saudi Arabia usually deviates from its dominant producer position in order to achieve political goals. These models assume that political interaction among OPEC members is important for cartel stability. These models are based on the idea that wealth maximization, security, and political influence are to some degree substitutes for each other. Thus, the decision to choose, for example, more security necessarily implies less wealth. In this case, political models and wealth maximization models offer divergent interpretations and predictions.

4. COMPETITIVE MODELS

The lack of support for various oligopoly models led researchers to introduce a variety of models that assume a competitive world oil market. Based on this view, five different models are presented to explain OPEC behavior.

4.1 Competitive Model

Oil prices increased in 1973 not only because of the embargo but also because of panic, stockpiling, and a spot market in which refiners sold futures crude oil. Rather than refining it, they used the crude as a commodity to gain quick profit.

Panic, stockpiling, spot markets, and speculation decreased supply and increased demand, which resulted in very high oil prices. In the same manner, oil prices increased in 1979, 1990, 1996, 2000, 2002, and 2003: A political problem led to a series of actions that lowered supply and increased demand. According to this model, there was no role for OPEC as an organization in the output disruption.

Applying this concept to Figs. 1 and 2, the boycott of 1973 shifted the supply curve to the left and increased oil prices. A domino effect followed: Panic and speculation shifted the demand to the right, refiners sold the crude and shifted supply farther to the left, and prices continued to increase. The Iranian revolution and the Iran–Iraq war caused the same chain reaction, which led to the second oil crisis.

In the early 1980s, the demand decreased and world supply increased, which led to lower oil prices. Increased demand and lower world supply, mostly from non-OPEC countries, resulted in increased oil prices in the second half of the 1980s. In the same manner, this model explains changes in prices and production in 1990, 1996, 1998, and 2000–2003.

4.1.1 Criticism of the Model

This model does not apply to OPEC for the following reasons:

1. In a competitive market, output will increase with higher prices. This is not the case for OPEC. It either cut or maintained output when prices were increasing, as shown in Fig. 1.
2. The model does not explain why some OPEC members, mainly Saudi Arabia, cut production in the mid-1980s while non-OPEC countries increased production.
3. It does not explain why the behavior of Saudi Arabia is different from that of other countries.
4. Statistical tests reject this model.

4.2 Property Rights Model

Production cuts by some OPEC members in the early 1970s were associated with changes in property rights when the host countries claimed ownership of the oil reserves either through nationalization or gradual ownership of the operating international oil companies (IOCs). This change in ownership led to uncoordinated production cuts. The host countries' discount rates were lower than those of the IOCs. In other words, foreign companies wanted to get as much money as possible from their concessions

before they left, which resulted in increased production. When the host countries took over, they wanted to produce less in order to conserve the resource for future generations and to maximize efficient rate of production per oil field.

Applying this model to Figs. 1 and 2, one sees that 1973–1974 was a transition period. It reflected a sharp increase in government ownership during which a sharp decrease in production caused a switch to a higher price path. Even during the 1979 price shock, some researchers argued that production cutbacks were the result of ownership transfers from companies to host countries during that period.

Results of statistical tests show “mixed performance” for OPEC members. The property rights model cannot be rejected for Iran, Kuwait, Libya, Nigeria, Qatar, and Venezuela. The model was rejected for the rest of OPEC members.

Researchers who introduced this model acknowledge that production cuts led to a panic in the market and an increase in demand, which increased oil prices even further. They also acknowledge that the limited absorptive capacity of OPEC members sustained prices after the transfer of ownership. As shown for the next model, this limitation led to additional production cuts.

4.2.1 *Criticism of the Model*

This model does not apply to OPEC for the following reasons:

1. The model cannot explain changes in production and prices after property transfer ended in the late 1970s. However, one might argue that oil-producing countries changed their discount rates as a result of “property reevaluation” or a change in “demand expectations.”

2. Some researchers claim that historical facts contradict this model. For example, production increases in the 1950s and 1960s may have resulted from the host countries’ desire to increase production. However, this criticism ignores the fact that IOC payments to the host governments during that period were a function of production levels. Countries collected more revenues as production increased. Therefore, they pushed the IOCs to increase production.

3. Some researchers argue that the production cut was not related to a lower discount rate by the countries. Rather, lack of technology caused it. The IOCs were deploying the latest technology in the early 1970s to enhance production and take advantage of rising oil prices. After nationalization, state-owned companies

were not able to deploy the advanced technology that had been used by the IOCs, which resulted in lower production, especially in old oil fields.

4.3 The Target Revenue Model

This model was introduced to explain why oil prices were sustained at their postembargo levels despite the official lifting of the embargo in March 1974. According to this model, political events and changes in property rights in 1973 caused prices to increase. Oil prices did not decline after the end of the embargo because countries, without coordination or collusion, cut their production further as prices increased.

Most OPEC members had primitive economies that could not absorb the additional revenues from higher oil prices. Countries had two options: transfer the extra revenues to foreign financial markets or keep the extra dollars in cash for future use. For three reasons, OPEC members chose to do neither. First, in the 1970s, real returns on investments in the West were very low, if not negative. Second, some members feared asset confiscation by Western governments, especially after the freezing of Iranian assets by the United States. Third, dollar devaluation and inflation would have reduced the value of any cash revenues saved. These reasons led some oil-producing countries to cut production. They believed that oil in the ground was “the world’s best investment.”

The result of output cuts is a backward-bending supply curve for prices higher than those needed to support the target revenue. If there are a sufficient number of countries with backward-bending supply curves, then the market supply curve will take the same shape.

Most researchers agree that this model may explain OPEC behavior in the short term but only in some periods during the 1970s. Recent research indicates that, even today, the model applies only to centrally planned and isolated oil-producing countries, regardless of their OPEC membership. Statistical tests show that this model applies only to two OPEC members, Libya and Iran. Both countries are centrally planned and isolated by UN and U.S. embargoes.

4.3.1 *Criticism of the Model*

This model does not apply to OPEC for the following reasons:

1. The backward supply curve can hold only for a very short period of time. Prices will continue to move along the supply curve.

2. Some researchers argue that the model is not stable. If we have a backward supply curve, an increase in the negatively sloped demand will lead to two points of equilibrium with different production levels and different prices. However, some researchers argue that the equilibrium is stable. Producers in this case will choose the lower production level to conserve the resource.

3. Historically, some researchers rejected this model because of its assumption that foreign investment is not a viable alternative to domestic investment. They cited Kuwait's success in building a financial empire in the West. They also cited data illustrating that every OPEC member has an overseas investment. This objective ignored the possibility that political reasons may prevent some countries from acquiring access to foreign financial markets.

4. All statistical tests in the literature reject this model for OPEC as a whole. However, it cannot be rejected for some countries, such as Iran and Libya.

4.4 The Purchasing Power Explanation

Although not a formal model, it explains why some OPEC members behave competitively, violate their quota, and decline to coordinate production with Saudi Arabia. According to this explanation, some OPEC members increase production and violate their quota if the purchasing power of their oil exports is decreasing as a result of variations in major world currency exchange rates and higher export prices of industrial countries.

If oil production cuts lead to higher oil prices and higher export prices for the products of industrial countries, it is not in the interest of the member to participate in production cuts. The purchasing power of its oil exports would decline. Although this concept may apply to all imports of goods and services, the situation is more severe if the oil-producing country is importing large amounts of petroleum products. If an OPEC member must import petroleum products, production cuts will increase the cost of imports and force the real value of oil exports to decrease.

Although Indonesia is an OPEC member, its large population and relatively small reserves force it to import crude oil and other refined products. Imports of crude and refined products represent approximately 8% of its total imports. During the fourfold increase in oil prices in the 1970s and the threefold increase in 1980, the value of these imports increased substantially. As a consequence, the real value of Indonesian oil exports decreased. This decline may

have forced Indonesia to increase its oil production and ignore its OPEC quotas in order to increase its purchasing power. Data indicate that Indonesia did not coordinate production with other OPEC members. Studies indicate that the importation of crude oil and other oil-intensive products was the primary reason for the decrease in the purchasing power of Indonesia's oil exports.

Historically, seven OPEC members imported crude oil and petroleum products on a scale large enough to affect their purchasing power: Algeria, Ecuador, Gabon, Indonesia, Nigeria, the UAE, and Venezuela. These imports may explain the competitive behavior of these countries.

4.5 The Foreign Investment Explanation

This theory was introduced to explain the behavior of Kuwait. It assumes that OPEC is not a cartel and attempts to explain why Kuwait does not coordinate production with Saudi Arabia, as indicated by the production data of both countries.

Kuwait's behavior can be explained by its foreign investments. The amount of foreign investment does not play a role in oil production decisions. The type of investment does. At first, Kuwait invested in upstream oil operations in the North Sea. At that time, it was in Kuwait's interest to cut production or sustain it to obtain higher prices for its oil in Kuwait and the North Sea. Not until the 1980s did Kuwait extend its investment to downstream operations to include thousands of gas stations throughout Europe and refineries in Singapore.

Since governments control retail prices of petroleum products in Europe, it was in Kuwait's interest to violate its OPEC quota and increase production. Lower crude oil prices from increased production enabled Kuwait to generate a large profit margin from retail outlets, where product prices did not change. The profit from lower crude oil prices outweighed the loss from selling the Kuwaiti crude at a lower price. In addition, Kuwait's return on investment exceeded its revenues from crude exports throughout the 1980s.

5. OPEC AND THE CARTEL STATUS

Despite the widespread use of the word *cartel* by media outlets and politicians, most researchers today, even those who assigned cartel status to OPEC in the 1970s, believe that OPEC is not a "monolithic" cartel. However, events since 1999 show that OPEC's

behavior may have changed now that a price band is in place. It holds almost monthly meetings instead of the usual quarterly meeting and makes frequent output adjustments. Figure 3 indicates that since 1999, production adjustment by Saudi Arabia matched that of other OPEC members. Some researchers believe that the recent changes in OPEC behavior constitute a *de facto* cartel. Others argue that Saudi Arabia is still the dominant player that initiates most output changes. They point out that production cuts by OPEC and non-OPEC members were not voluntary but rather the result of technical, political, and natural factors. The arguments of each group are summarized next.

5.1 OPEC as a Cartel

Those who argue that OPEC is a cartel cite the following reasons:

1. Historical records indicate that OPEC was modeled after the Texas Railroad Commission. Even today, its objective is to coordinate production among members.
2. Since its inception, OPEC members have been meeting periodically to coordinate policies, production, and prices.
3. Historically, OPEC cut production to increase prices, as indicated in Fig. 1.
4. OPEC currently uses a quota system and a price band, two important cartel characteristics.
5. The price of oil is much higher than marginal cost.
6. OPEC members have been able to transfer massive amount of wealth from the consuming countries since 1973.

5.2 OPEC as a “Commodity Producer Club” But Not a Cartel

The cartel theory states that there are seven characteristics that must exist in a group of producers in order to be labeled a cartel: A cartel must assign quotas to its members, monitor members to avoid violations, punish violators, target a minimum price, take action to defend the price, have a large market share, and divide votes among members based on their market share. OPEC did not meet any of these characteristics during its peak of power between 1973 and 1981. OPEC assigned quotas in 1982 but failed to enforce them. OPEC did not set a price band until 2000. It might have been

able to defend prices since 2001 with the help of technical, political, and natural factors that prevented many oil-producing countries from increasing production. OPEC never had a punishment mechanism or an effective monitoring system such as that of the Texas Railroad Commission, which used the National Guard and Texas Rangers to enforce its rules and cut production in Texas and Oklahoma. In addition, OPEC's market share is relatively small (approximately 40%). Even today, each OPEC member has one vote regardless of its production and reserves. Unlike the Seven Sisters, OPEC did not divide the world market and did not control investment. The market share of the Seven Sisters was more than double that of OPEC.

By U.S. standards, OPEC does not fit the description of a cartel. Monopolization has two elements: the acquisition of monopoly position and the intent to monopolize and exclude rivals. Official U.S. judicial records indicate that a monopoly position has been associated with companies whose market share approached or exceeded 80%. This is double the current OPEC market share and much larger than the 54% market share achieved in 1974. OPEC has thus failed, by a wide margin, to acquire a monopoly position. As for intent to acquire a monopoly and to exclude rivals, OPEC has never taken the steps mentioned previously to be a successful cartel. OPEC was labeled a cartel because it supposedly increased prices, but how could OPEC exclude rivals by increasing oil prices? Figure 1 shows that higher oil prices bring higher production, more producers, and more competition. OPEC did not act as a monopolist. OPEC has increased production many times in recent years to stabilize prices. It has become an active participant in the world community through negotiations with consuming countries and active aid programs to developing countries. Many researchers today, including lawyers and policy makers, view OPEC as an agent of stabilization in a volatile world. In fact, the Bush administration praised OPEC for its efforts to prevent prices from increasing to record levels during the invasion of Iraq.

None of the statistical tests in the literature support the cartel model for OPEC. Researchers who conducted these tests concluded that OPEC is a “partial,” “weak,” “clumsy,” or “loose” cartel. Such conclusions are mere journalistic expressions. No known theory in the economic literature supports them. The results of so-called statistical tests are not acceptable. For example, applying the same tests to non-OPEC countries shows that non-OPEC producers fit the

cartel model better than OPEC members. In addition, these tests focus on parallel behavior, not on cartelization. Parallel behavior may exist in any market, including competitive ones.

The cartel model requires OPEC to operate on the elastic portion of its demand curve. Studies show that OPEC operated on the inelastic portion of its demand curve during its power peak in the 1970s.

Coordination among governments and commodity producer clubs is well-known. This coordination is not labeled as cartelization, despite periodic meetings and clear goals to curb production and maintain prices.

Researchers who attempt to prove that OPEC is a cartel by focusing on the difference between marginal cost and the price of oil ignore the basic principles of natural resource economics. The price of a non-renewable resource in a competitive market reflects the marginal cost and the discount rate chosen by the producer. Therefore, in a perfectly competitive market the price of any natural resource is higher than its marginal cost, unlike renewable resources, the price of which equals the marginal cost. Hence, focusing on the difference between marginal cost and price to prove market power is conceptually wrong unless the researcher has perfect estimates of the user's cost.

Researchers who believe that OPEC is not a cartel argue that there are other explanations for OPEC wealth transfer that are not related to cartelization. Some researchers argue that it was both the market power of Saudi Arabia and U.S. price controls that led to wealth transfer to OPEC. Various studies show that U.S. oil price controls suppressed U.S. production, increased the world demand for oil, and raised Saudi Arabia's output. This unintended effect enabled OPEC members, mainly Saudi Arabia, which increased its production substantially in the late 1970s, to transfer wealth from the consuming countries to the producing countries.

Historical records indicate that OPEC was reacting to economic and political events rather than being a proactive cartel that consistently followed policies that would maximize its wealth. OPEC reacts pragmatically to compromise on the conflicting interests of its members. In addition, OPEC pricing has always followed the spot market.

6. CONCLUSION

There are two prominent supply factors that explain the behavior of OPEC: scarcity and market power.

The focus of most researchers on market power without regard to scarcity and technical issues and their political and social implications has polarized academic and political circles to the extent that it is almost impossible to reach a consensus on OPEC and its role in world oil markets.

Review of various models introduced to explain OPEC behavior indicates the following unequivocal conclusions:

1. Significant production by non-OPEC members rules out the use of monopoly models.
2. Historical data and research studies indicate that the competitive model does not apply to OPEC.
3. Only oligopoly models may explain OPEC behavior. However, theory, historical data, and statistical tests do not support the monolithic cartel model for OPEC.
4. No single model can explain OPEC behavior even at particular periods of time.

One way to explain the behavior of OPEC and non-OPEC producers, world oil demand, and oil prices is to synthesize these models based on the results of various studies. Such a synthesis indicated that OPEC is driven by political and technical influences as much as by the inexorable forces of economics and the market.

This synthesis must tell the story of the past 30 years without violating logic, theory, data, or history. The behavior of OPEC can be explained using the dominant producer model for Saudi Arabia or a dynamic core that changes members periodically and by using various models to explain the behavior of other OPEC members. In this case, other OPEC members follow various models at various points of time, such as the property rights model, the target revenue model, the purchasing power explanation, the foreign investment explanation, and even pure competitive behavior. At different times, OPEC, or a subset of it, may look like a monolithic cartel when, coincidentally, various models necessitate the same action. In this case, countries may cut or increase production jointly. For this reason, single-equation models, correlation tests, and cointegration tests indicate parallel action but not cartelization.

However, one should not forget the role of political and natural events, panic, stockpiling, and speculation and their impact on oil markets. OPEC, like the Texas Railroad Commission, was established to protect property rights and to bring order to a chaotic situation that was leading to tremendous wastage and losses. OPEC aid to Third World

countries makes OPEC look more like a world-class intergovernmental organization than merely a cartel that cuts production and increases prices. Production increases by OPEC members to combat the increasing prices that resulted from the labor strikes in Venezuela and Nigeria and the war in Iraq indicate the significant role that OPEC plays in stabilizing oil prices, especially given that political turmoil, not OPEC production cuts, caused world energy crises.

The shape and role of OPEC may change in the future. Some members may leave the organization when they become net oil importers. Others may leave if they privatize their oil industry and become unable to force private companies to follow OPEC quotas. On the other hand, new members may join the organization. Some current OPEC members may increase their capacity substantially. Others may even build excess capacity and compete with Saudi Arabia. Regardless of these changes, OPEC will keep playing an important role in world oil markets. Lower oil prices and the withdrawal of some members will not lead to the demise of this powerful organization.

The impact of the growing importance of natural gas on OPEC remains to be seen. Four OPEC members are among the world's largest natural gas players: Iran, Algeria, Qatar, and Libya. The emphasis on natural gas may either weaken or strengthen OPEC. Only time will tell.

Acknowledgments

I thank Michael Bunter, Collin J. Campbell, Juan Pablo Perez Castillo, Melaku Gebaye Desta, Dermot Gately, Luis Lugo, Adam

Sieminski, Thomas Walde, and an anonymous referee for their helpful comments on the manuscript.

SEE ALSO THE FOLLOWING ARTICLES

Economic Geography of Energy • Geopolitics of Energy • Global Energy Use: Status and Trends • National Energy Policy: United States • Nationalism and Oil • National Security and Energy • Oil Crises, Historical Perspective • Oil Industry, History of • Oil-Led Development: Social, Political, and Economic Consequences • War and Energy

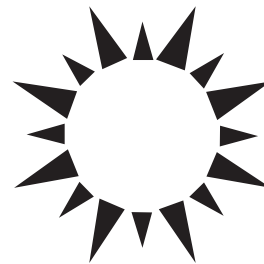
Further Reading

- Adelman, M. A. (1995). "The Genie out of the Bottle: World oil since 1970." MIT Press, Cambridge, MA.
- Alhajji, A. F., and Huettner, D. (2000a). OPEC & other commodity cartels: A comparison. *Energy Policy* 28(15), 1151–1164.
- Alhajji, A. F., and Huettner, D. (2000b). Crude oil markets from 1973 to 1994: Cartel, oligopoly or competitive? *Energy J.* 21(3), 31–60.
- Alhajji, A. F., and Huettner, D. (2000c). The Target revenue model and the world oil market: Empirical evidence from 1971 to 1994. *Energy J.* 21(2), 121–144.
- Crèmer, J., and Salhi-Isfahani, D. (1991). "Models of the Oil Market." Harwood, New York.
- Griffin, J., and Teece, D. (1982). "OPEC Behavior and World Oil Prices." Allen & Unwin, Boston.
- Mabro, R. (1998). OPEC behavior 1960–1998: A review of the literature. *J. Energy Literature* 4(1), 3–26.
- Noring, Ø. (2001). "Crude Power, Politics and the Oil Market." Tauris, London.
- Stevens, P. (2000). "The Economics of Energy," Vols. 1 and 2. Elgar, Cheltenham, UK.

Origin of Life and Energy

RONALD F. FOX

Georgia Institute of Technology
Atlanta, Georgia, United States



1. Stellar Nucleosynthesis of the Elements
2. Relative Abundance of the Elements
3. Gibbs Free Energies of Formation
4. The Monomer-to-Polymer Transition: No Life to Life
5. Energy: Redox Reactions, Thioesters, and Phosphate
6. Banded Iron Formations in the Geological Record
7. The Importance of D-Orbitals
8. Thermal Energy and Synthesis
9. The RNA World

Glossary

dehydration condensates Polymers in which the monomer linkages are equivalent to the removal of a molecule of water between the joined monomers.

Gibbs free energy A type of thermodynamic energy containing internal energy, entropic energy, and pressure volume energy that is applicable for systems at constant temperature and pressure.

macromolecular Pertaining to molecules of large size having masses in the thousands to millions of Daltons (1 Dalton is 1 atomic mass unit). All biologically important macromolecules are polymers, molecules made from linking together monomers by means of dehydration linkages. There are three major types: proteins made from amino acid monomers, polysaccharides made from simple sugars, and polynucleotides made from mononucleotides.

micells Self-assembled monolayered closed structures made from fatty acids dissolved in water.

phosphate A molecule made from phosphorus, oxygen, and hydrogen that is an ion when dissolved in water with the formula HPO_4^{2-} .

protein biosynthesis The complex mechanism by which proteins are synthesized, amino acid by amino acid, in accord with a sequence of nucleic acid bases in a messenger RNA molecule that has been transcribed from a DNA gene. Many proteins and RNA molecules are involved in the apparatus that performs this function.

redox energy The energy of electrons transferred with them during electron transfer reactions.

stromatolites Layered fossil rocks believed to be the remains of mats of bacterial colonies topped and fed by phototrophic organisms, perhaps the ancestors of contemporary cyanobacteria.

thioester An energy-rich compound involving sulfur covalently linked to the carbon of a carbonyl group.

vesicle A bilayer, self-assembled, closed structure made from glycerophospholipids possessing two fatty acid side chains.

The phrase origin of life refers to those natural geophysical processes that may have occurred on primitive Earth approximately 3.5–4.0 billion years ago that gave rise to life. Presumably there are no extant representatives of the earliest forms of life since they were surely driven to extinction by more advanced forms very early in the evolution of life. Fossils do not preserve macromolecular structures inside cells so that no direct fossil evidence exists. Even putative cellular fossils from 3.5 billion years ago, the stromatolites, are not absolutely proven to be what they appear to be. On the other hand, their similarity to contemporary formations deposited by bacterial mats topped with cyanobacteria, which may not be too different from the primitive organisms of 3.5 billion years ago, suggests that the metabolic and protein biosynthesis macromolecular machinery of contemporary cells was already developed at that time. How these macromolecular structures came into being is the problem of the origin of life.

In this context, energy refers to a variety of forms, both chemical and physical, that are relevant in the context of the origin of life. These include (i) fusion energy released during the formation of the nuclei of elements during stellar nucleosynthesis; (ii) ultraviolet (UV) light energy of the sun that excites electrons into excited states, thereby providing them with redox energy, for example, during the oxidation of iron from ferrous iron (Fe^{2+}) to ferric iron (Fe^{3+});

(iii) thioester compounds that are rich in Gibbs free energy so that they spontaneously engage in a variety of reactions and syntheses; (iv) phosphate compounds that serve to activate intermediates in the synthesis of polymers; and (v) thermal energy or heat. The problem for the origin of life and energy is to understand how these different forms of energy are connected and transduced from one kind to another and promote the origin and evolution of the macromolecular processes of living matter. In contemporary organisms there is also the fundamental importance of chemiosmotic proton energy, but this form of energy likely evolved relatively late in the origin of the first cells.

1. STELLAR NUCLEOSYNTHESIS OF THE ELEMENTS

Understanding why life is made out of the elements that are observed to be most abundantly present in organisms is aided by considering the origin of the elements during stellar nucleosynthesis. These elements are hydrogen (H), carbon (C), nitrogen (N), oxygen (O), phosphorous (P), sulfur (S), chlorine (Cl), sodium from natrium (Na), magnesium (Mg), potassium from kalium (K), calcium (Ca), and iron from ferrum (Fe). Although several other elements are essential for many contemporary organisms, such as iodine (I), selenium (Se), and cobalt (Co), primitive life almost certainly could have arisen based on the primordial dozen elements listed previously. These elements are the available building materials for life because they are the most abundant naturally occurring elements.

During the lifetime of a typical star, tremendous changes in pressure, mass density, and temperature take place inside the star. Temperatures between 10^7 and 10^9 K occur as well as densities from 10^2 to 10^5 g/cm³ (ordinary liquid water has a density of 1.0 g/cm³). These extreme conditions promote a series of nuclear fusion processes that are organized around the nucleus of helium, called an alpha particle, that contains two protons and two neutrons. This is a particularly stable configuration of nucleons (protons or neutrons) and its formation from free but very hot protons (the H nucleus) releases vast amounts of energy, thereby driving the formation process. Subsequent fusion involves the incorporation of more protons into the alpha particles and tends to give rise to nuclei that are simple integer multiples of alphas, the nuclei of C, O, Mg, silicon (Si), S, Ca, and Fe.

Notice the presence of Si and the absence of N, P, Na, K, and Cl. The processes of synthesis of the alpha multiple nuclei generate these additional elements, albeit in lower amounts than the true alpha multiples. As discussed later, the d-orbital properties of Si make it unsuitable for the basic metabolic constituents of cells, and it instead has a very strong tendency to combine with O to form enormous three-dimensional inert silicate compounds such as quartz. Fusion energy release leads to the elements of the first four periods of the periodic table, the most abundant elements in the cosmos, and these are the building blocks of living matter.

2. RELATIVE ABUNDANCE OF THE ELEMENTS

By weight, three-fourths of the earth's crust is made of O (50%) and Si (25%), primarily as silicates. The oceans are mostly O (86%) and H (11%), primarily as water. Although abundance is important, fitness is just as important. Si is 135 times more abundant than C anywhere on Earth even remotely suitable for life, and it shares with C the tendency to form four covalent bonds, but it is highly unfit as a constituent of living matter compared to C. C can combine with other types of atoms to form myriads of small molecules, some of which involve double bonds. It combines to make strong stable bonds. Si, however, combines most strongly with O to make highly polymeric inert structures called silicates. Si-Si bonds are highly susceptible to attack by H₂O, NH₃⁺, and O₂, unlike C-C bonds, which are stable. Only when structural material in the form of spines and spicules is used by living organisms such as cacti and sponges does Si actually occur in living matter. The metabolic pathways, however, are devoid of Si, using C compounds instead. These differences between C and Si have to do with their location in the second and third periods of the periodic table, respectively.

H, C, N, and O make up approximately 99% of the constituents of all living organisms. As members of the first and second periods, they are the most abundant elements and they combine to make short, strong, stable bonds. P and S, although less abundant, are nevertheless more abundant than heavier elements, and because they reside in the third period they possess unfilled d-orbitals. These orbitals confer on them the ability to engage in transfer reactions that mediate energy changes. S in the form of the thiol group, -SH, contributes to thioester-based energy

transactions, and P combines with O to form phosphate, which contributes to energy activation for the promotion of polymerizations. The most stable nucleus of all, and the end product of the fusion processes in stellar nucleosynthesis, belongs to Fe. This element is ideally suited as a conveyor of redox energy. Generation of redox energy, the energy of excited electrons, is caused by sunlight.

3. GIBBS FREE ENERGIES OF FORMATION

Once the elements have formed and planetary systems have accreted around Sun-like stars, temperatures become low enough for the spontaneous formation of many types of small molecules. This spontaneity is measured by the Gibbs free energy of formation from the elements. In a spontaneous process at constant temperature and pressure, the Gibbs free energy must decrease. This says nothing about rates. Without catalysts, a long period of time may be required before the minimum in free energy is achieved. The previous statement is the second law of thermodynamics expressed in terms of the Gibbs free energy rather than the entropy. This difference is a result of applying the second law to a system in contact with both temperature and pressure reservoirs and applying it to an isolated system, respectively.

In Table I, the Gibbs free energy of formation from the elements for a variety of compounds is listed for the standard conditions of a temperature of 298 K and atmospheric pressure. The table lists the name of the substance; its empirical formula; its molecular weight; its free energy of formation as $-\Delta G_{298}^0$, where the minus sign is explicitly included (a positive listing in the table means a negative free energy of formation); and its free energy of formation per gram as $-\Delta G_{298}^0/g$. By convention, the elements are given the value 0.0.

Notice how many important biological compounds are listed in Table I. The strongest tendencies toward formation from the elements (i.e., the most positive entries in the table) occur for aqueous ions, quartz, calcium phosphate, and water; however, most entries have values between 0.5 and 2.0. The table lists simple sugars, amino acids, a purine (xanthine), and numerous compounds that occur in metabolism. Clearly, life evolved from materials that were naturally occurring spontaneous products of the most abundant elements. Their formation from the elements is driven by a decrease in Gibbs free energy.

4. THE MONOMER-TO-POLYMER TRANSITION: NO LIFE TO LIFE

Life as we know it is distinguished by the presence of macromolecules: proteins, polysaccharides, and polynucleotides. These provide structural elements, catalysts, energy storage, and information storage and processing components for the cell. The dynamical cellular processes inside the cell that we recognize as living properties are almost exclusively associated with these polymers.

Polymers are chains of monomeric constituents that have been joined together by covalent bonds. In the cases of proteins and polynucleotides, these chains are linear, whereas in the case of polysaccharides they are branched. In all cases, the chemical structure of the linkages that join monomers into polymers is that of dehydration condensates. This means that the linkage between any two adjacent monomers anywhere in the chain can be broken by hydrolysis (i.e., cleavage by a molecule of water; literally “hydro-lysis”). Most cells are 80–95% water, which means that the tendency toward hydrolysis is strong, although perhaps not rapid unless catalysts are available. An acid environment and digestive enzymes are needed by our stomachs to process food because the spontaneous rates of hydrolysis are too slow, but these processes are spontaneous because they are attended by a decrease in Gibbs free energy. Thus, the problem for organisms is the reverse—that is, the synthesis of polymers from monomers, an uphill process thermodynamically that requires the input of Gibbs free energy. It is for this reason that the metabolic pathways are organized around the generation of energy-rich compounds, particularly adenosine triphosphate (ATP), the almost universal currency of free energy for the cell.

The transition of matter from no life to life is in essence the transition from monomers to polymers. The monomers are formed spontaneously from the elements and have negative free energies of formation relative to the elements. The polymers, however, have positive free energies of formation relative to the monomers from which they are synthesized. Most would appear in Table I with a value of 0.9 ± 0.15 . Making the “peptide” bond between amino acids in proteins costs approximately 2 or 3 kcal/mol, the “glycosidic” bond between sugar monomers in polysaccharides costs approximately 3 or 4 kcal/mol, and the “phosphodiester” bond between mononucleotides in polynucleotides costs approximately 5 kcal/mol. In each case, activation is essentially a phosphate activation, slightly disguised

TABLE I

Gibbs Free Energy of Formation for a Variety of Compounds^a

Compound	Empirical formula	Molecular weight (g/mol)	$-\Delta G_{298}^0$ (kcal/mol)	$-\Delta G_{298}^0/g$ (kcal/g)
Acetaldehyde	C ₂ H ₄ O	44	33.4	0.759
Acetic acid	C ₂ H ₄ O ₂	60	94.7	1.578
Acetate (aq)	C ₂ H ₃ O ₂ ⁻	59	89.0	1.508
Acetyl CoA	C ₂₃ H ₃₉ O ₁₈ N ₇ P ₃ S	825	89.4	0.108
Alanine	C ₃ H ₇ O ₂ N	89	88.7	0.997
Ammonium ion (aq)	NH ₄ ⁺	18	19.0	1.056
Arginine	C ₆ H ₁₅ O ₂ N ₄	175	126.7	0.724
Asparagine	C ₄ H ₈ O ₃ N ₂	132	125.8	0.953
Aspartate	C ₄ H ₆ O ₄ N ⁻	132	167.4	1.268
Calcium (c)	Ca _n	40 _n	0.0	0.000
Calcium (aq)	Ca ²⁺	40	132.2	3.305
Calcium hydrogen phosphate	CaHPO ₄	138	401.5	2.909
Carbon (c, graphite)	C _n	12 _n	0.0	0.000
Hydrogen carbonate (aq)	CHO ₃ ⁺	61	140.3	2.300
Carbon dioxide (g)	CO ₂	44	94.2	2.141
Carbon dioxide (aq)	CO ₂	44	92.3	2.098
Chlorine	Cl ₂	71	0.0	0.000
Citrate	C ₆ H ₅ O ₇ ³⁻	189	278.7	1.475
Creatine	C ₄ H ₉ O ₂ N ₃	131	63.1	0.482
Cysteine	C ₃ H ₇ O ₂ NS	121	81.2	0.671
Dihydroxyacetone phosphate	C ₃ H ₇ O ₆ P	170	308.9	1.817
Erythrose 4-phosphate	C ₄ H ₉ O ₇ P	200	343.8	1.719
Ethanol	C ₂ H ₆ O	46	43.4	0.943
Formaldehyde	CH ₂ O	30	31.2	1.040
Formic acid	CH ₂ O ₂	46	85.1	1.850
Formate (aq)	CHO ₂ ⁻	45	83.8	1.862
Fructose	C ₆ H ₁₂ O ₆	180	218.7	1.215
Fructose 6-phosphate	C ₆ H ₁₃ O ₉ P	260	420.0	1.615
Fructose biphosphate	C ₆ H ₁₄ O ₁₂ P ₂	340	621.3	1.827
Fumarate	C ₄ H ₃ O ₄ ⁻	115	144.3	1.255
Galactose	C ₆ H ₁₂ O ₆	180	220.6	1.226
Glucose	C ₆ H ₁₂ O ₆	180	219.1	1.217
Glucose 6-phosphate	C ₆ H ₁₃ O ₉ P	260	420.5	1.617
Glutamate	C ₅ H ₈ O ₄ N	146	166.5	1.140
Glutamine	C ₅ H ₁₀ O ₃ N ₂	146	125.4	0.859
Glycerol	C ₃ H ₈ O ₃	92	116.7	1.268
Glycerol phosphate	C ₃ H ₉ O ₆ P	172	319.2	1.856
Glycine	C ₂ H ₅ O ₂ N	75	90.0	1.200
Glyceraldehyde 3-phosphate	C ₃ H ₇ O ₆ P	170	307.1	1.806
Hydrogen (g)	H ₂	2	0.0	0.000
Hydrogen cyanide	HCN	27	-28.7	-1.063
Hydrogen sulfide	H ₂ S	34	6.5	0.191
Hydronium ion (aq)	H ₃ O ⁺	19	56.7	2.984
Hydroxyl	HO ⁻	17	37.6	2.212
Iron (c)	Fe _n	55.8 _n	0.0	0.000
Iron (II) (aq)	Fe ²⁺	55.8	20.3	0.364
Iron (III) (aq)	Fe ³⁺	55.8	2.5	0.045

continues

Table I continued

Compound	Empirical formula	Molecular weight (g/mol)	$-\Delta G_{298}^0$ (kcal/mol)	$-\Delta G_{298}^0/g$ (kcal/g)
Isocitrate	$C_6H_5O_7^{3-}$	189	277.1	1.466
Isoleucine	$C_6H_{12}O_2N$	131	82.2	0.627
α -Ketoglutarate	$C_5H_4O_5^{2-}$	144	190.7	1.324
Lactate	$C_3H_5O_3^{-}$	89	123.4	1.387
Lactose	$C_{12}H_{22}O_{11}$	332	362.0	1.090
Leucine	$C_6H_{13}O_2N$	131	85.1	0.650
Magnesium (c)	Mg_n	24.3 n	0.0	0.000
Magnesium ion (aq)	Mg^{2+}	24.3	109.0	4.486
Methane	CH_4	16	12.1	0.756
Methanol	CH_4O	32	41.9	1.309
Nitrate ion (aq)	NO_3^{-}	62	26.4	0.426
Nitrite ion (aq)	NO_2^{-}	46	8.2	0.178
Nitrogen	N_2	28	0.0	0.000
Oxalate	$C_2O_4^{2-}$	88	161.2	1.832
Oxaloacetate	$C_4H_2O_5^{2-}$	130	190.4	1.465
Oxygen	O_2	32	0.0	0.000
Phenylalanine	$C_9H_{11}O_2N$	165	49.5	0.300
Dihydrogen phosphate ion (aq)	$H_2PO_4^{-}$	97	271.2	2.825
Hydrogen phosphate ion (aq)	HPO_4^{2-}	96	261.4	2.723
Phosphate ion (aq)	PO_4^{3-}	95	245.1	2.580
Phosphoric acid	H_3PO_4	98	-274.1	-2.797
Phosphorus (c, white)	P_n	31 n	0.0	0.000
Phosphoenolpyruvate	$C_3H_5O_6P$	136	303.3	2.230
Potassium (c)	K_n	39.1 n	0.0	0.000
Potassium ion (aq)	K^{+}	39.1	67.5	1.726
Pyruvate	$C_3H_3O_3$	87	113.4	1.303
Quartz (c)	$(SiO_2)_n$	60 n	192.4 n	3.207
Ribose 5-phosphate	$C_5H_{11}O_8P$	230	382.2	1.662
Ribulose 5-phosphate	$C_5H_{11}O_8P$	230	381.7	1.660
Sedoheptulose 7-phosphate	$C_7H_{15}O_{10}P$	290	457.1	1.576
Serine	$C_3H_7O_3N$	105	122.1	1.163
Silicon (c)	Si_n	28 n	0.0	0.000
Sodium (c)	Na_n	23 n	0.0	0.000
Sodium ion (aq)	Na^{+}	23	62.6	2.722
Sodium chloride (c)	$NaCl$	58.5	91.8	1.569
Succinate	$C_4H_4O_4^{2-}$	116	164.9	1.422
Succinyl CoA	$C_{25}H_{40}O_{20}N_7P_3S$	882	164.0	0.186
Sucrose	$C_{12}H_{22}O_{11}$	342	370.7	1.084
Sulfate ion (aq)	SO_4^{2-}	96	177.3	1.847
Sulfite ion (aq)	SO_3^{2-}	80	118.8	1.485
Sulfur (c, rhombic)	S_n	32 n	0.0	0.000
Threonine	$C_4H_9O_3N$	119	122.9	1.033
Tryptophan	$C_{11}H_{12}O_2N_2$	204	29.9	0.147
Tyrosine	$C_9H_{11}O_3N$	181	92.5	0.511
Urea	CH_4ON_2	60	48.7	0.812
Valine	$C_5H_{11}O_2N$	117	86.0	0.735
Xanthine	$C_5H_5O_2N_4$	153	33.3	0.218
Xylulose	$C_5H_{10}O_5$	150	178.7	1.191
Water	H_2O	18	56.7	3.150

^a Abbreviations used: aq, aqueous; c, crystalline; g, gaseous.

one way or another. Indeed, for proteins the intermediate-activated species is the aminoacyl-adenylate, a phosphate that is part of adenine monophosphate (AMP). The adenine of ATP can be substituted for by uracil, thymine, cytosine, and guanine, generating the activated monomers UTP, TTP, CTP, and GTP, from which, along with ATP, all DNA and RNA chains can be made. In these substances, the sugar ribose occurs for the RNA case and the sugar deoxyribose occurs for the DNA case. When deoxyribose is involved, the notation dATP, dGTP, dCTP, and dTTP is used. The activation of sugar monomers involves uridine diphosphate (UDP) derivatives. Thus, in all three cases—proteins, polysaccharides, and polynucleotides—the activated monomers are phosphorylated compounds. A single phosphate is used for aminoacyl-adenylates that are AMP derivatives, a diphosphate is used for UDP monosaccharides that are UDP derivatives, and triphosphates are used for UTP, ATP, dATP, TTP, CTP, dCTP, GTP, and dGTP that ultimately are all generated from ATP by phosphate transfers.

The required extra free energy is ultimately provided by ATP through formation of activated monomers that, once activated, spontaneously condense into polymers but only with the aid of complex catalytic molecular machinery. Thus, dehydration and activation are universal structural and synthetic themes for all life as we know it. Moreover, the use of energy-rich phosphate compounds to achieve activation is also universal. The problem with regard to the origin of life is how these phosphate compounds came into being before there were enzymes to catalyze energy metabolism. From this perspective, this is the fundamental “chicken-and-egg” problem.

Elaborate macromolecular structures have evolved that make activation possible in all extant cells. Many enzymes support a complex chain of oxidation reactions that harvests redox energy from high-energy electrons—not high energy in the physicist’s sense but high energy in the sense of biochemistry. The electrons may be excited initially by the intense UV irradiating the nascent Earth. Also, reduced compounds with high electrical potentials (in Volts) can be generated in gentle thermal processes that dry out the reaction mixture. Drying means that water is leaving the system as steam and is taking with it the thermodynamic barrier to dehydration condensations. In the redox case, the electrons engage in a series of electron transfers, called redox reactions, that can couple to other energy-requiring processes such as monomer activation.

The purest form of redox process is of the type exemplified by the ferrous–ferric transition:



There are many instances of this type of redox step in the electron transport chains of bacteria and in mitochondria and chloroplasts, the energy-transducing organelles of eukaryotic cells. Two major classes of Fe-based transfers are found in all forms of life: those involving iron–sulfur proteins and those involving cytochrome, a heme-Fe-containing protein. These different species of Fe-containing proteins contain individuals with redox potentials spanning the range from free Fe (0.771 V) to ferredoxin-Fe (−0.420 V). This permits evolution of a chain of many relatively small steps in the total decrease of Gibbs free energy, affording many opportunities for coupling to other processes.

The most rudimentary alternative to the simple electron transfers of Fe is cotransfer of a proton with an electron (i.e., a hydrogen atom in pieces). The ubiquitous ubiquinone species, found in all organisms possessing electron transport chains, are of this type and are responsible for the mechanism of chemiosmosis in membrane-bound electron transport chain complexes. When the fundamental electron–proton, current–current coupling occurs in mitochondria, a decrease of Gibbs free energy during the electron transfer is partially used to increase Gibbs free energy for a steepening concentration gradient (that for a proton means a pH gradient) and an increasing transmembrane electrical potential because the proton is positively charged. This generates the chemiosmotic “proton-motive force,” which is actually an electrical potential for protons (in Volts), not a force. Many processes, most notably transport of all sorts of molecules across the membrane, are driven by this form of cellular energy rather than by ATP. Indeed, ATP is synthesized chemiosmotically in bacteria, mitochondria, and chloroplasts by rotary enzyme complexes. ATP is ideally suited for activating monomers for the synthesis of polymers. However, there are some ATP-driven ion transporters as well. Since chemiosmosis probably evolved after life began, phosphate energy was the universal currency of energy for the origin of life, but only with the aid of another sort of energy discussed later.

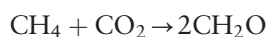
The evolution of the cell membrane is unknown. Contemporary cellular membranes are primarily made from the class of lipids called glycerophospholipids, which contain two fatty acid side

chains and a phosphorylated alcohol such as choline, ethanolamine, inositol, or serine. The fatty acid side chains contain 14–24 carbon atoms. An elaborate enzyme complex catalyzes a long sequence of reactions in order to make these highly specialized lipids. Published models for an early evolution of a primitive precursor to these glycerophospholipids, in an abiotic geochemical thermal process, make an earlier evolution of chemiosmosis more plausible.

Glycerophospholipids are needed to form lipid bilayer membranes. These structures, the essence of the cellular envelope, spontaneously self-assemble from aqueous solutions of lipid molecules. The more easily generated fatty acids, or single-tailed lipids, produced by early Earth's geochemophysical processes only form micells instead of the larger and more versatile bilayer vesicles. Did the cellular membrane structure evolve before life evolved, during the prebiotic chemical evolution phase? Did it require a sophisticated metabolic machinery before its molecular constituents could be generated in sufficient quantities to comprise cells? Wouldn't such an apparatus necessarily have to be contained within a membranous boundary in order to keep enzyme and substrate together? This is another chicken-and-egg question.

Evidence suggests that before there was life as we know it, there were morphological structures resembling cells, the bilayer vesicles, and there was a source of chemical energy through coupling to the electron transfer processes. These electron transfers run downhill spontaneously but through couplings to other processes, they are generative of high-energy chemicals. The oxidation of carbohydrates by molecular O_2 releases enormous amounts of Gibbs free energy (approximately 700 kcal/mol for 1 mol of glucose and 6 mole of O_2), approximately 40% of which is harvested by the pathways of glycolysis, generation of acetyl CoA, the citric acid cycle, and the electron transport chain.

Carbohydrates are made by plants using the energy of sunlight, now in the visible part of the spectrum rather than in the primordial UV. From Table I, it is seen that carbohydrates, sugars, have less negative Gibbs free energies of formation than an equivalent amount of methane and carbon dioxide. The composition of many sugars can be written as a simple integer multiple, n , of formaldehyde [i.e., as $(CH_2O)_n$]. Thus, the point is made by the fact that the process



is attended by an increase in Gibbs Free energy of approximately 43.9 kcal/mol. It can be argued that sugar formed by whatever means can spontaneously convert back to methane and carbon dioxide by the inverse of the process shown previously, a process referred to as carbon disproportionation. Since this releases energy, it has been argued that it is a possible source of prebiotic energy. O_2 is also made by plants. This means that glycolysis (that does not directly involve O_2) is more primitive than the citric acid cycle or the electron transport chain (that are directly linked to O_2). Interestingly, glycolysis couples the carbon disproportionation of glucose to the generation of high-energy phosphate in the form of ATP. The dimer, pyrophosphate, is the energy-rich portion of ATP and is sufficiently energy rich that it alone can drive synthetic processes such as polymerizations. Pyrophosphate, as well as other polyphosphates, can also be made abiotically by simply using gentle heating to dryness or by coupling to thioester generation, as will be discussed later.

There might well have been myriads of micron-sized bilayer microspheres along with abundant reducing potential from high-energy electrons and abundant pyrophosphate before there was true cellular polymer synthesis. This micron-scale, energy-rich environment appears ideally suited for the evolution of polymerization machinery. Only within such a simple energy-rich environment can the required variety of polymers form at all. Before true macromolecular polymers emerged, a smaller class of mixed oligomers, the coenzymes, appear to have emerged. Coenzymes are the active catalytic moieties of many enzymes. They contain small molecular components called vitamins as well as ribose, phosphate, and various other constituents, usually approximately six units of various kinds per coenzyme. The presence of phosphate, indeed pyrophosphate, and sulfur in many of them suggests that these species are relics of an early stage of chemical evolution when energy transactions were already dominated by thioesters and phosphates. Since the coenzymes are thermodynamically uphill from their constituents, because they contain exclusively dehydration linkages, they could come into being only if the chemical milieu was one rich in energy-transducing species. With their emergence, a rudimentary metabolism, still devoid of modern protein catalysts, could have also emerged. This chemical environment, perhaps housed in a membranous vesicle, would provide the setting for the ultimate development of polymer synthesis.

know it is consistent with the proposition that it is these naturally occurring compounds that form the basis for life.

It is popular to refer to various stages of the evolution of life as “worlds” (e.g., the RNA world, discussed later). The earlier stage of chemical evolution described previously can be called the iron–thioester–pyrophosphate world. It represents the fundamental significance of energy transductions underlying the ultimate development of polymers and the overwhelming importance in energy transductions of three elements: P, S, and Fe. However, because P is so scarce, this world must have been preceded by the iron–thioester world. This world lasted for some time and provided a mechanism for the slow but steady accumulation of phosphate. Once the pyrophosphate world emerged, which was really the iron–thioester–pyrophosphate world, the origin of coenzymes could leave its signatures.

6. BANDED IRON FORMATIONS IN THE GEOLOGICAL RECORD

Banded iron formations are worldwide layered sedimentary deposits rich in iron that range in age from 1.5 to 3.8 billion years old, the age of the earliest known rocks. They contain 30–60% Fe^{3+} . No molecular oxygen was present in the earth’s atmosphere until 1.5–2.0 billion years ago. Thus, the oxidant earlier could not have been O_2 . O_2 is now the ultimate oxidant for nearly all organisms, but its function as such depends on a complicated enzyme complex. In contrast, the ability of UV to oxidize Fe^{2+} is geophysically natural and even autocatalytic. It has been demonstrated in the laboratory at acid, neutral, and alkaline pH. Since Fe^{3+} can serve as an oxidant for other molecules, Fe^{2+} can be regenerated. This creates an iron cycle. Depending on the stage of chemical evolution for all other molecular species, the amount of Fe^{3+} in any geological era could have gone through cycles as well, leading to the layers of magnetite minerals as the banded iron formations.

In an aqueous environment, the UV oxidation of Fe^{2+} forms Fe^{3+} and is followed by coprecipitation of ferric and ferrous oxides (Fe_2O_3 and FeO , respectively). These oxides form from the iron ions and water molecules releasing protons as a by-product as well as the mineral precipitate, magnetite:



The released protons are available for the reduction of a large variety of substances and could have been responsible for the production of such species as H_2 , CH_4 , HCN , and NH_3 as long ago as 3.8 billion years. These reduced species are otherwise not readily produced in other abiotic processes and are essential for the subsequent chemical evolution of amino acids and nucleoside bases, the purines and pyrimidines. It should be noted that UV energy was one of the most abundant forms of energy on the primitive earth.

Fe^{3+} may well have served as an early electron acceptor and thereby supported the oxidative synthesis of thioesters, as was argued previously. Iron–sulfur proteins are among the most ancient known proteins and may be relics of an earlier iron–thioester–pyrophosphate world but clearly not of the iron–thioester world.

7. THE IMPORTANCE OF D-ORBITALS

It is thought that the singular importance of P and S in energy transactions and the noteworthy unfitness of Si for metabolic chemistry are due to the properties of elements in the third period of the periodic table and also to d-orbitals. These electron states appear in the third period of the periodic table but are unoccupied in Si, P, and S. Moving from left to right within a period, atoms actually get smaller even though they get larger as the period increases. Thus, all atoms in the third period are larger than those in the second period, but within the third period the size order is $\text{Si} > \text{P} > \text{S}$. Like C, N, and O, P and S can form double bonds, but Si is too large to do so. Thus, Si cannot form the vast variety of compounds that C does and is restricted to single bonds, forcing it into immense three-dimensional polymers of $(\text{SiO}_2)_n$. Moreover, its unoccupied d-orbitals afford an opportunity for attack of the relatively open Si–Si bonds by nucleophiles such as O_2 , H_2O , and NH_3 . In contrast, the bonds of P and S are tighter and their relatively smaller size allows them to form double bonds. The same d-orbitals that are such a liability for Si are instead an asset for P and S. For phosphates in particular, the result is a chemical group that is easily transferred while relatively impervious to chemical attack. In contrast, the smaller atoms of H, C, N, and O produce very stable bonds that would not be suitable for easily transferred groups. Thus, phosphate appears to be a natural choice for the activation group in synthetic biochemistry. The d-orbitals make it pentavalent and this contributes

to the resonance enhancement of its energy transactions that confer on it its high-energy status.

8. THERMAL ENERGY AND SYNTHESIS

Thermal energy is of central importance for the origin of life in two distinct ways. At the molecular level, it gives rise to a very robust energetic activity called Brownian motion. The collisions between water molecules and other molecules are very energetic and are responsible for the Brownian motion of other molecules. This Brownian motion makes proteins fluctuate vigorously in conformation and promotes the catalytic activities of these molecules. Indeed, without the Brownian motion, the protein enzymes would lose their functionality. Brownian motion is also responsible for very rapid mixing of species within a micron-sized environment. The other importance for thermal energy is that it can overcome the thermodynamic barrier to dehydration polymerizations. Since this barrier is a result of the overwhelming presence of water molecules and the fact that all polymer linkages are dehydration bonds, thermal energy (i.e., heat) can remove the water molecules if applied in a gentle enough manner. Too much heating will simply destroy the molecules and create a gooey, tarry mess. For origin of life considerations, it is this second aspect of thermal energy that is of synthetic importance.

A prime example of the synthetic power of thermal energy is the thermal synthesis of polyamino acids, called thermal proteinoids. This type of polymer synthesis works especially well if the amino acid mixture is rich in aspartic acid and glutamic acid, not unlike the composition of many contemporary biological proteins. It is also assisted by the presence of phosphate, which by itself yields polyphosphates when heated. Probably, the polyphosphates formed in the amino acid mixture help promote condensation of the amino acids. Such phosphate-enhanced reaction mixtures yield thermal proteinoids at temperatures as low as 60°C if heating is allowed for 150 h. It is also of interest that when the thermal proteinoids are returned to water after synthesis, they spontaneously form uniform populations of micron-diameter microspheres. Since many amino acid residues are lipophilic, these thermal polymers have some properties in common with true glycerophospholipids. In particular, the proteinoid microspheres are bilayered vesicles rather than

micells. The proteinoids also possess a number of relatively weak catalytic abilities covering the basic reaction steps that make up metabolism. All of this together strongly suggests a natural mechanism for the formation of catalytic microsphere environments in which the iron–thioester–pyrophosphate world could gain a foothold. Although the proteinoid microsphere membranes are not nearly as impermeable as the membranes of modern organisms, they would nevertheless trap any polymers formed within them from monomers that could freely enter them from the environment. With the evolution of a more sophisticated metabolism, incorporation of primitive lipid analogs would make the membranes much less permeable and permit the establishment of chemiosmotic mechanisms. The proteinoid model has received a great deal of criticism in the literature. However, a close study of the primary literature presenting the experimental results that support the model clearly refutes much of the criticism.

9. THE RNA WORLD

Since the discovery of ribozymes (RNA molecules with enzymatic activity) in the early 1980s, it has been very popular to speak of the origin of life in terms of the RNA world. The basic notion is that RNA simultaneously has the ability to serve as the genetic material and as the first enzymes. Thus, the answer to the question, Which came first, proteins or polynucleotides?, receives the emphatic answer, polynucleotides, specifically RNA. This question is an old chicken-and-egg question, and opinion has alternated between proteins first and polynucleotides first for some years. A more evenly balanced view seems warranted given the difficulties surrounding the chemical evolution of mononucleotides, especially their important ingredient, the ribose molecule, that so far is not easily produced in abiotic experiments. Previously, it was clearly established that the spontaneous production of polymers is thermodynamically inhibited. RNA polymers require phosphodiester linkages, the dehydration linkages for polynucleotides, and these are uphill in Gibbs free energy. This situation is doubly a problem for polynucleotides because they are made from monomers, the mononucleotides, which are themselves oligomers (i.e., very small polymers). ATP, for example, contains adenine, ribose, and three phosphates. Each component is linked to its neighbors by dehydration linkages for a total of four in the case of ATP or any other mononucleotide triphosphate.

Thus, even the oligomeric monomers for polynucleotide synthesis are uphill in Gibbs free energy relative to their constituents. Given these thermodynamic barriers, coupled with the difficulty explaining an easy natural origin for ribose, it is highly unlikely that an RNA world was the beginning stage for life as we know it. For this reason, the RNA world section of this article is the last section, not the first.

In the preceding sections, it was argued that an energy-transducing chemical evolution preceded the true emergence of life. Before the RNA world could have emerged, the iron–thioester–pyrophosphate world had to be in place, perhaps already encapsulated inside microspheres with proteinoid and/or lipid precursor membranes. Recall that phosphate, an essential component of polynucleotide phosphodiester backbones, is scarce on the earth and only through the aid of thioesters does it seem likely that it was recruited from the geophysical environment into the biological realm, as described previously. If the question is asked, Which came first, the RNA world or the iron–thioester–pyrophosphate world?, then it should be clear that it was the latter.

Once the essential energy requirements for the transition from monomers to polymers are established, the RNA world becomes possible. In this sense, the origin of life, the transition from monomers to polymers, coincides with the emergence of the RNA world. The problem then shifts from one of origins to one of evolution. The key issue is that of protein biosynthesis by a gene-directed mechanism. The evolution of this mechanism must account for the emergence of ribosomes, transfer RNA (tRNA), messenger RNA (mRNA), ribosomal RNA, and the aminoacyl-tRNA synthetases (aaRSs). The aaRSs of contemporary organisms are a diverse and complex group of enzymes responsible for attaching the cognate amino acid to a tRNA with a specific anticodon. The mechanism of this recognition process in contemporary organisms is far from completely understood, and all of the structural and evolutionary evidence strongly points to a long evolution of the mechanism and its components. This unsolved problem constitutes the big gap in our comprehension of the contemporary genetic mechanism.

Models for how these complex components could have evolved from much more primitive precursors have been proposed. They purport to explain (i) why the genetic code is a three base code with much degeneracy in the third codon and why the codons for the amino acids are what they are; (ii) why the N-terminus to C-terminus direction of the gene-directed protein is colinear with the 5' → 3' direction

of the cognate mRNA; (iii) why L-amino acids and D-ribose are used (i.e., whether this is a relative relationship that could as easily have been D-amino acids with L-ribose instead or an absolute necessity); and (iv) how gene-directed synthesis of proteins got started before there was the complex array of protein and RNA catalysts that currently make the process work. There are many other related issues. One published model proposes that the first genes and mRNAs were one and the same molecule, an RNA, and that protein synthesis took place directly on this RNA when activated amino acids interacted directly with the 2'-OH groups of the ribose moieties to form amino acyl ribose esters. Through a subsequent conformation change, the esterified amino acids link up to form small proteins and release the RNA for another round of synthesis. The RNA has served as both gene and messenger in this model. The activated amino acids might be carboxyl phosphates generated from pyrophosphate or they might be thioesters. Subsequent evolution of DNA, lacking the 2'-OH groups, would separate the roles of information storage in DNA and information translation in the RNA. This very simple genetic system would be poised for the subsequent evolution of ribosomes, tRNAs, and aaRSs. Such models provide conceptual insight into origins and evolution and also provide a number of challenges for experimentalists. However, advances in biotechnology make possible a number of experiments to test the ideas just presented.

SEE ALSO THE FOLLOWING ARTICLES

Earth's Energy Balance • Ecosystems and Energy: History and Overview • Heterotrophic Energy Flows • Lithosphere, Energy Flows in • Ocean, Energy Flows in • Photosynthesis and Autotrophic Energy Flows • Sun, Energy from • Thermal Energy Storage • Work, Power, and Energy

Further Reading

- Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). "Biochemistry." Freeman, New York.
- Calvin, M. (1969). "Chemical Evolution." Oxford Univ. Press, New York.
- Deamer, D., Kuzina, S. I., Mikhailov, A. I., Maslikov, E. I., and Seleznev, S. A. (1991). Origin of amphiphilic molecules and their role in primary structure formation. *J. Evol. Biochem. Physiol.* 27, 212–217.
- de Duve, C. (1991). "Blueprint for a Cell: The Nature and Origin of Life." Patterson, Burlington, NC.
- Fox, R. F. (1988). "Energy and the Evolution of Life." Freeman, New York.

- Fox, S. W., and Dose, K. (1972). "Molecular Evolution and the Origin of Life." Freeman, San Francisco.
- Fowler, W. A. (1967). "Nuclear Astrophysics." American Philosophical Society, Philadelphia.
- Harold, F. M. (1986). "The Vital Force: A Study of Bioenergetics." Freeman, New York.
- Joyce, G. F., and Orgel, L. E. (1999). Prospects for understanding the origin of the RNA world. *In* "The RNA World" (R. F. Gesteland, Ed.), 2nd ed., pp. 49–77. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Schopf, J. W. (1999). "The Cradle of Life: The Discovery of the Earth's Earliest Fossils." Princeton Univ. Press, Princeton, NJ.
- Wald, G. (1962). Life in the second and third periods; Or why phosphorus and sulfur for high-energy bonds? *In* "Horizons in Biochemistry" (M. Kasha and B. Pullman, Eds.), pp. 156–167. Academic Press, New York.
- Weber, A. L. (1998). Prebiotic amino acid thioester synthesis: Thiol-dependent amino acid synthesis from formose substrates (formaldehyde and glycolaldehyde) and ammonia. *Origins Life Evol. Biosphere* 28, 259–270.

Passenger Demand for Travel and Energy Use



ANDREAS SCHAFER

Massachusetts Institute of Technology
Cambridge, Massachusetts, United States

-
1. Determinants of Passenger Travel Energy Use
 2. Travel Demand and Mode Choice
 3. Vehicle Technology
 4. Human Factors
 5. Passenger Travel Energy Use and Emissions
 6. Reducing Passenger Travel Energy Use and Emissions
 7. Outlook

Glossary

energy intensity Energy use per unit service (here passenger-kilometer).

light-duty vehicle Individual passenger transport modes consisting of mainly automobiles and those light trucks (e.g., minivans, vans, sport utility vehicles, pickup trucks) used for passenger travel.

occupancy rate The number of occupants in a vehicle weighted by the driving distance, that is, the ratio of passenger-kilometers and vehicle-kilometers.

passenger-kilometer The product of the number of passenger trips and the average distance traveled, where 1 passenger-km corresponds to one passenger traveling a distance of 1 km and so on.

vehicle-kilometer The product of the number of vehicle trips and the average distance traveled, where 1 vehicle-km corresponds to one vehicle traveling a distance of 1 km and so on.

Among the many factors determining travel demand, growth in income and growth in population are the two single most important drivers of human mobility. In addition, on a highly aggregate, world regional level that is necessary to estimate world passenger travel energy use, mainly four constraints determine travel demand and mode choice. Two of the constraints are so-called travel budgets; although a country's population consists of many individuals

with very different preferences, the average expenditure shares of money and time are roughly constant. Travel demand and mode choice also depend on transportation prices, that is, the unit cost of travel to the consumer. As this article shows, for some modes, especially aircraft, these unit costs fell sharply during the second half of the 20th century, contributing to a rapidly rising demand for that mode. Finally, travel demand is also determined by land use settings; humans in densely populated areas rely more heavily on mass transit modes, whereas those living in countries with very low-density urban areas enjoy greater use of automobiles.

1. DETERMINANTS OF PASSENGER TRAVEL ENERGY USE

Getting from travel demand to energy use requires multiplying the product of per capita passenger-kilometer (pkm/cap) and population (pop) with energy use per vehicle-kilometer traveled (E/vkm) and dividing by the occupancy rate (pkm/vkm). Equation (1) describes the overall identity of passenger travel energy use (E) as it may apply to, for example, automobile transport:

$$E = \frac{E}{vkm} \cdot \frac{vkm}{pkm} \cdot \frac{pkm}{cap} \cdot pop. \quad (1)$$

Note that most of the factors in Eq. (1) are a function of additional variables. For example, energy use per vehicle-kilometer traveled (E/vkm) is determined by vehicle size, technology, payload, and driving characteristics. Also note that many of the factors in the equation are interrelated. For example, increasing the number of persons sharing a vehicle (i.e., decreasing vkm/pkm) reduces E/pkm but increases E/vkm because the vehicle is now transporting a

heavier load. Simultaneously, a greater number of people sharing a car reduces the number of vehicles on the road, and this—as a first-order effect—increases traffic flows and, thus, reduces E/vkm . In the subsequent sections, this article examines each of these components more closely. It begins with the growth in per capita travel demand (pkm/cap), moves on to vehicle technology (E/vkm), and finally examines trends in human factors that are reflected in the inverse vehicle occupancy rate (vkm/pkm) but also in the choice of technology and driving characteristics (E/vkm).

2. TRAVEL DEMAND AND MODE CHOICE

2.1 Factors Affecting Travel

Growth in income and growth in population are the most important drivers of passenger mobility. Higher income allows households to afford motorized transportation and, ultimately, to own automobiles that enable more trips and longer distances within a given amount of time compared with most other modes of transport. In contrast, population represents a scaling factor of total mobility. In addition, travel demand is determined by socioeconomic variables, including age, gender, and profession. For example, males travel more than females, and people in their 20s and 30s are more mobile than other age groups. Further determinants include land use settings, supply variables (e.g., accessibility to transport systems), prices for (transportation-related) goods and services, and political and institutional forces. Although this list of factors affecting travel demand could be extended, quantifying total energy use for, and emissions from, passenger travel requires a more aggregate perspective of travel demand.

Hence, in addition to the two driving forces, growth in gross domestic product (GDP) per capita and growth in population, aggregate passenger travel demand depends mainly on four variables: the share of money dedicated to transportation, the share of time dedicated to transportation, land use characteristics, and prices for travel by mode. Yacov Zahavi was first to hypothesize constant levels of time and money expenditure shares per traveler (i.e., a person who conducts at least one mechanized trip per day) and per household, respectively. Zahavi has used the two budgets in urban transportation models. Subsequently, other researchers have generalized them to the average person and found similar levels of

stability at national levels. More recently, both budgets were applied to project future levels of travel demand worldwide.

Figures 1A and 1B illustrate the stability of the two travel budgets for a wide range of income levels. As shown in Fig. 1A, people dedicate roughly 5% of their daily time (or 1.2 h) to transportation on average, independent of whether they live in African villages or in the highly industrialized world. Although travel time and money expenditures are highly variable on an individual level (reflecting widely differing preferences and constraints across travelers), mean values of these expenditures are much more stable for groups of people. Thus, the variability of travel time expenditures is in part dependent on the aggregation level. Variability is highest across cities (empty circles in the figure), already smaller across countries (full circles), and should be negligible on a world regional level. In addition to the constant travel time budget, humans spend a predictable share of income on travel, ranging from 3 to 5% in the case where they predominantly depend on public transportation modes (i.e., the developing world). This share increases to 10 to 15% if they own at least one automobile (Fig. 1B). Mainly because of these budgets' skewed distribution over the population, attempts to explain the underlying forces for constant averages have not been convincing until recently.

Because of their fundamental nature, the two budgets govern the daily travel behavior of virtually each person irrespective of cultural, geographical, and economic settings. While these budgets are very similar across national boundaries, economic and other settings can vary substantially and thus cause differences in averages of travel demand indicators. For example, in very low-income rural areas of developing countries, nearly all trips are done within a distance of 5 km—the distance that can be covered on foot within the travel time budget. With rising income, humans can afford faster modes of transport and, thus, increase the geographical area of daily interaction. In the high-income, automobile-dominated industrialized countries, more than 95% of all trips occur within a distance of 50 km—the distance that can be covered by an automobile within the travel time budget.

In addition to the two travel budgets, travel behavior is determined by land use characteristics. Extreme differences in population densities can have a sensible impact on the choice of transportation mode. Because of severe space constraints and the

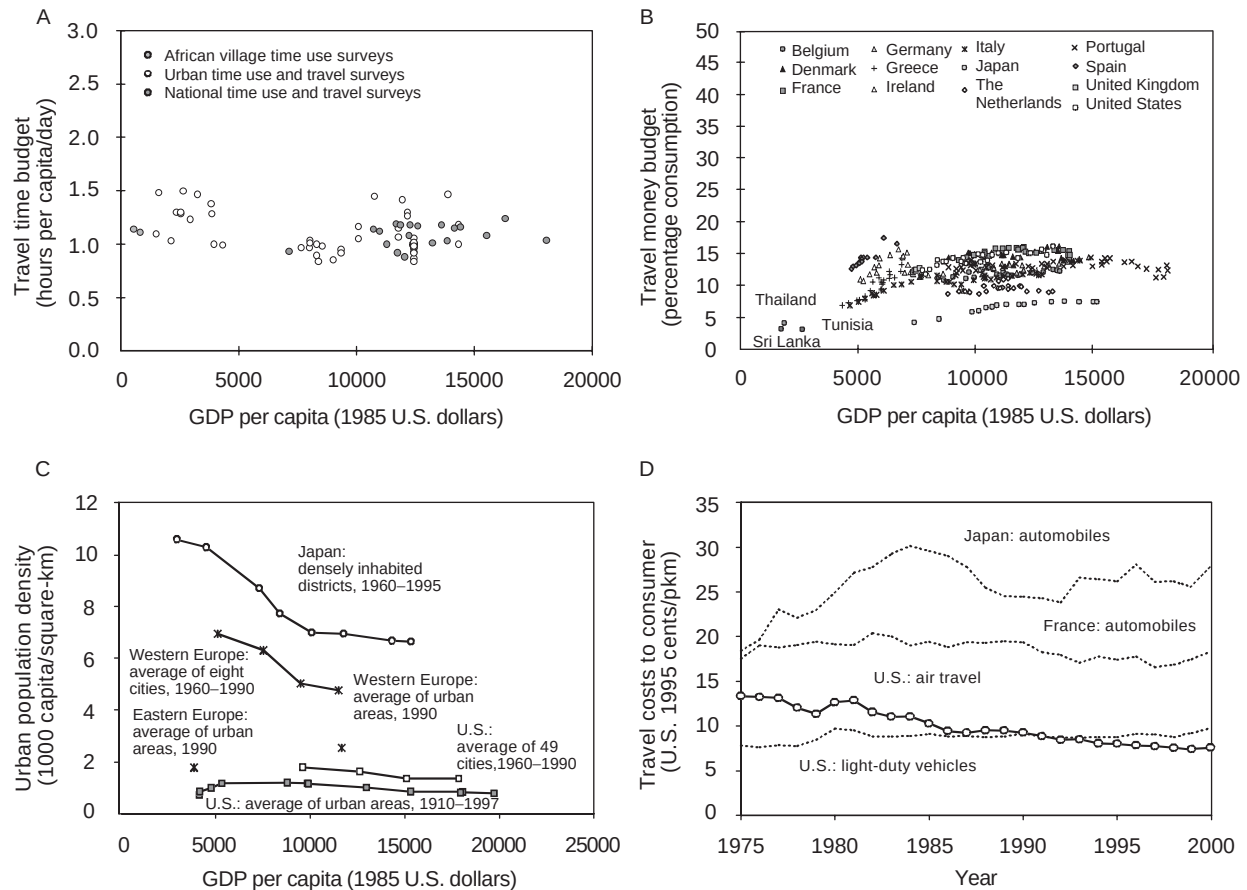


FIGURE 1 Fundamental determinants of total travel demand and mode choice. (A) Average travel time expenditures per capita and day (travel time budget). (B) Share of consumption dedicated to transportation on average (travel money budget). (C) Declining urban population densities in various parts of the industrialized world. (D) Average price of travel per passenger-kilometer for major transportation modes in the United States, France, and Japan.

associated high costs of operating a private motor vehicle, very high-population density areas are served mainly by public transportation modes. In contrast, lower density cities are dominated by private vehicles, which provide the highest levels of flexibility in that type of environment. Thus, the ongoing process of suburbanization in higher income countries, enabled mainly by the automobile, necessarily reduces the relative importance of public transport modes. Figure 1C illustrates the long-term process of urbanization and the subsequent sprawl in the United States beginning after World War II, the period of mass motorization. Suburbanization has occurred in all other parts of the industrialized world, mostly at significantly higher population densities.

The final determinant of aggregate travel behavior is travel costs to the consumer. Figure 1D reports time trends for air travel (United States) and automobile transport (France, Japan, and United

States). While absolute automobile travel costs have remained stable in France and the United States, real per capita income has increased by more than 50% during the same time. In Japan, where automobile travel costs have increased by nearly 30% since 1975, real per capita income has increased by 80%. Together the trends in travel costs and income imply that automobile travel has become increasingly affordable everywhere. From a cross-country perspective, differences in driving costs remain related to the availability of land and associated prices. Costs per passenger-kilometer for automobile travel are lowest in the United States, the country with the lowest urban population densities, and are highest in highly dense Japan. (The low gasoline taxes in the United States have contributed to the comparatively low cost of driving and the associated low population densities.) Whereas automobile travel costs have remained roughly constant in the United States, costs

of air travel have actually declined, from U.S. 12.1 cents/pkm in 1978 (the beginning of airline deregulation) to 7.5 cents/pkm in 2000.

Together, the two driving forces (income and population) and the four variables describe a sufficiently complete picture of worldwide travel demand and mode choice.

2.2 Total Mobility

Translating income into a roughly proportional amount of travel expenditures, a stable travel money budget implies that total travel demand rises approximately in proportion to income. (The initially increasing share of the travel money budget, which results in travel expenditures rising faster than income, is used to finance the transition from lower cost public modes to higher cost automobiles [Fig. 1]; thus, travel demand also roughly grows in proportion to income during the period of rising money budgets.) Figure 2 illustrates the tight relationship between income (approximated by GDP per capita) and per capita traffic volume (e.g., automobiles, buses, railways, aircraft) for 11 world regions and the world from 1960 to 1990. Although the illustrated growth relationship holds for all regions, differences in economic structures, prices, land use settings, culture, and values can cause differences in per capita travel demand at a given GDP per capita level. For example, lower population densities along with lower prices for owning and operating an automobile, but also

differences in cultural values such as higher labor flexibility, have caused per capita traffic volume to be nearly 40% higher in North America than in Western Europe at a GDP per capita level of U.S. \$10,000.

Table I reports traffic volume per capita and totals for all 11 world regions for 1960 and 1990. Historically, most of the world's traffic volume has occurred in the industrialized world, accounting for slightly more than half of the world passenger traffic volume in 1990. However, because the highest growth in income and population has occurred in the developing world, culminating with a nearly 12-fold rise in the Other Pacific Asia region (the rapidly developing Asian economies), those countries are increasing their share in world mobility. Currently, most of the traffic in the developing world occurs in rapidly growing urban areas (where income is higher), including the many "megacities," where the vehicle fleet grows much faster than transportation infrastructure, leading to extreme levels of congestion, urban air pollution, accidents, and other transportation-related problems.

2.3 Mode Choice

Whereas a constant travel money budget translates rising income into growing travel demand, a constant travel time budget requires the growing traffic volume to be satisfied by faster modes of transport. The associated continuous shift toward increasingly fast transportation modes with rising per capita

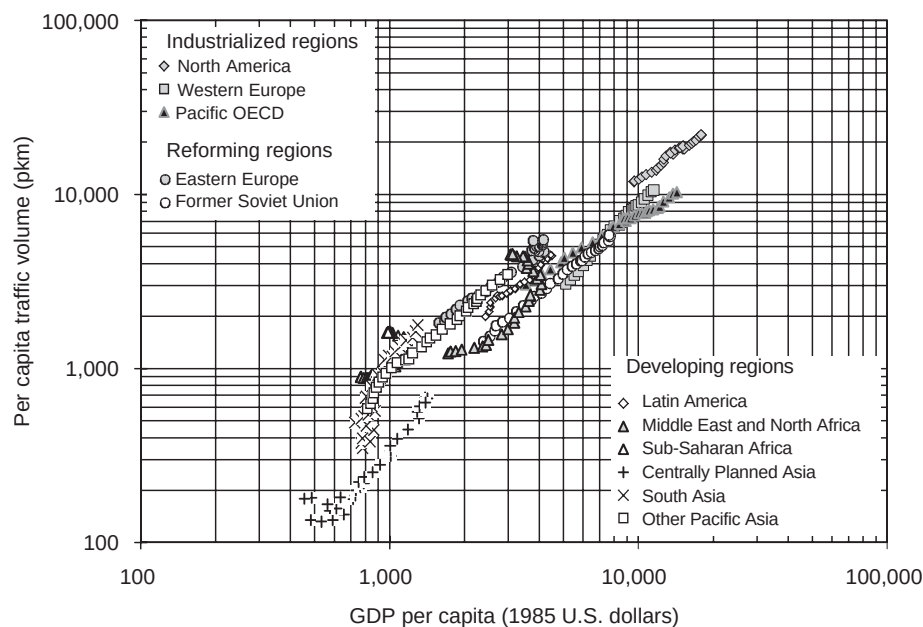


FIGURE 2 Evolution of per capita traffic volume over GDP per capita for 11 world regions, 1960–1990.

TABLE I

Traffic Volume (per Capita and Total) in 11 World Regions, 3 Meta-regions, and the World: 1960 and 1990

	1960			1990		
	pkm/cap	Absolute		pkm/cap	Absolute	
		Billion pkm	Percentage world		Billion pkm	Percentage world
North America	11854	2384	43.5	22078	6193	26.7
Pacific OECD	3074	1106	20.2	10622	4696	20.2
Western Europe	3025	323	5.9	10294	1482	6.4
Total industrialized	4400	3813	69.6	14276	12372	53.3
Eastern Europe	1824	181	3.3	5389	666	2.9
Former Soviet Union	1419	295	5.4	5796	1631	7.0
Total reforming	1550	477	8.7	5672	2297	9.9
Centrally Planned Asia	1222	140	2.6	4546	1244	5.4
South Asia	898	193	3.5	1614	811	3.5
Other Pacific Asia	152	109	2.0	637	805	3.5
Sub-Saharan Africa	349	200	3.7	1778	2015	8.7
Middle East and North Africa	587	125	2.3	3470	1459	6.3
Latin America	1980	424	7.7	5094	2228	9.6
Less developed countries	582	1191	21.7	2125	8562	36.9
World	1814	5481	100.0	4382	23231	100.0

traffic volume is shown in Fig. 3 for five world regions. The modal trajectories in this figure report three stages of the evolution of motorized mobility.

During the first stage, at low levels of economic development and travel demand, low-speed collective modes (railways and buses) account for virtually all motorized transport services, as seen in the South Asia region. At that initial stage, a substantial share (> 50%) of all passenger-kilometers are covered by nonmotorized modes. Individual and high-speed means of transport still account for less than 10% of total passenger-kilometers and serve only a small share of the population. (Due to the lack of data, nonmotorized modes and motorized two- and three-wheelers were excluded from that analysis.)

During the second stage, at higher levels of income and travel demand (~2000 passenger-km per capita), the relative importance of low-speed collective modes saturates and declines due to the strongly rising demand for increasingly affordable automobile travel. Without a significant share of automobile transport, these elevated demand levels could not be reached within a fixed travel time. In Eastern Europe, the rise of the automobile occurred despite constrained access to individual modes of transport in a socialist system. Apparently, these fundamental dynamics of the transportation system could not be

easily eliminated by transportation policies. High-speed transport systems still have a small share.

During the third stage of mobility development, above demand levels of 6000 to 10,000 passenger-km per capita, the share of automobile traffic volume saturates and subsequently declines. The high demand levels now require a substantial share of air traffic and/or high-speed ground transportation systems. The rising significance of high-speed travel was enabled by continuously declining air travel costs (Fig. 1D). In North America, where all passenger traffic is split between automobiles and aircraft, the latter, offering still greater mobility through significantly higher speeds, is strongly increasing its market share at the expense of passenger cars. At these high levels of per capita traffic volume, ordinary rail and bus services maintain a low share in special market niches (mainly high-density population areas).

The specific modal shares are also determined by urban land use characteristics. In North America, where urban population density is low (~800 people/square-km in U.S. urban areas, as reported in Fig. 1C), the share of automobile traffic volume was saturated at nearly 90% in 1960. In more densely populated Western Europe, with about four times the density levels of U.S. cities, automobiles achieved only a 72% share of total traffic volume. Correspondingly, in the Pacific Organization for Economic Cooperation

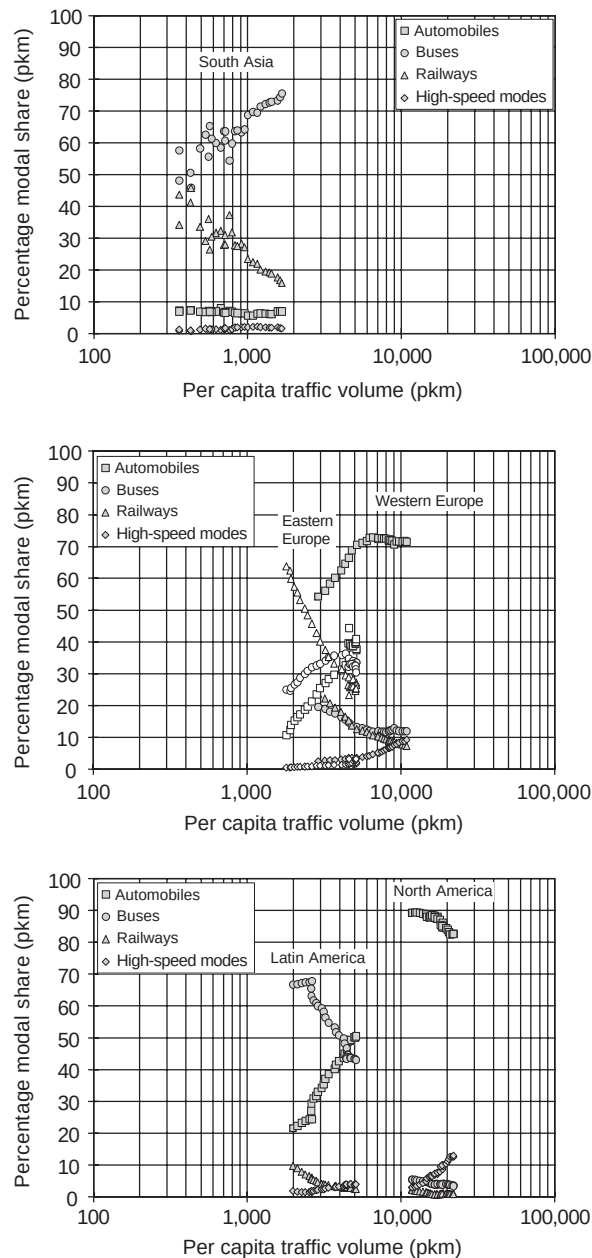


FIGURE 3 Three stages characterizing the evolution of motorized mobility: relative importance of transport modes (in percentage passenger-kilometers) over passenger-kilometers per capita, 1960–1990.

and Development (OECD) region, which is dominated by Japan and its high population density (five to six times the U.S. level), the relative importance of automobile traffic volume is likely to peak below the Western European level. Primarily as a consequence of different types of land use, investments in transportation infrastructures also have been different across regions (e.g., a high share of railways in high-

density Pacific OECD [essentially Japan], a low share of railways in low-density North America).

On a lower level of aggregation, the previously described mode shifts occur within two separate transportation markets. Whereas the substitution of automobiles for public transportation occurs mainly in expanding urban areas, the shift from automobiles to high-speed modes of transport occurs exclusively in longer distance intercity travel.

2.4 Travel Purposes

Nearly all travel is done for satisfying a purpose, whether it is getting to work, pursuing personal business-related activities such as shopping, or enjoying leisure. Figure 4 shows that the observed growth in travel demand follows regular patterns when disaggregated into its purposes. At low levels of income and low levels of daily distance traveled, people undertake less than two trips per day (e.g., Delhi suburbs during the late 1970s). The associated per capita traffic volume is less than 15 passenger-km per day (5500 passenger-km per year). At such low mobility levels, travel serves satisfying basic needs nearly exclusively. One trip in a day is dedicated to a combination of work (short-term survival) and education (longer term well-being), and roughly half a trip is dedicated largely to personal business (essentially shopping at local markets), on average. With rising per capita traffic volume and automobile ownership, additional trips for personal business (e.g., shopping, health care, religious services) and leisure (including holidays) are undertaken. At the high income levels comparable to those of OECD countries, people make more than three trips per day, devoting approximately one trip to work or education, one to two trips to personal business, and one trip to leisure. The highest trip rate can be observed for the United States, with the highest level of passenger-kilometers per capita. Here, personal business trips account for nearly half of all trips made, and work trips only account for 20% of all trips. At such high mobility levels, some travel is also done for its own sake. According to the 1995 U.S. Nationwide Personal Transportation Survey, pleasure driving accounted for 28 billion passenger-km or 0.7% of total passenger traffic volume.

3. VEHICLE TECHNOLOGY

Any movement requires energy. Basic laws of physics imply that vehicle energy requirements rise with the

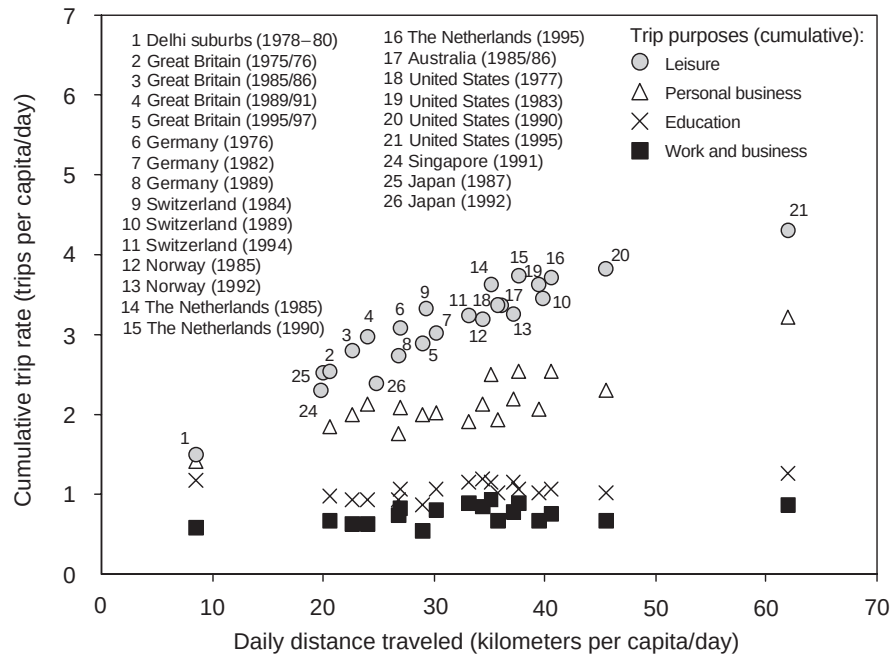


FIGURE 4 Growth in trips per capita per day, by purpose, over daily distance traveled.

TABLE II

Major Characteristics of Transportation Modes in the United States

	Mean travel speed (km/h)	Mean trip distance (km)	Mean occupancy rate (pkm/vkm)	Mean energy use/intensity	
				(MJ/vkm)	(MJ/pkm)
Walking	4.7	0.8	N/A	N/A	0.21
Bicycle	11	2.1	1.0	0.07	0.07
Motorcycle	52	16	1.2	1.6	1.4
Buses	24	14	16.8	18	1.1
Railways	10–46	6–168	23.4	45	1.9
Automobile	54	15	1.6	3.7	2.3
Personal truck	54	15	1.6	4.6	2.9
Aircraft	550	1216	91.1	220	2.4

Note. Mean travel speeds and mean trip distances are derived from the 1995 U.S. Nationwide Personal Transportation Survey. Mean occupancy rates, energy use, and energy intensity for motorized modes are derived from aggregate statistical references and related to 2000. The conversion between energy use and energy intensity does not account for extra passenger weight. Ranges in speed and mean trip distance for railways reflect tramways (lower number) and intercity rail (high number). Other figures are averages.

square of speed, whereas vehicle power requirements (i.e., energy use per unit time) increase with the cube of speed. Because travel speeds can differ substantially across transport modes, understanding differences in vehicle energy use requires taking into account vehicle operation characteristics. Table II reports major operation characteristics and the approximate level of energy use for various transport modes. As would be expected, energy intensity rises with speed provided that comparisons are

made on a consistent basis. The mean vehicle speed and distance, here derived from the 1995 U.S. Nationwide Personal Transportation Survey, cover a range of two to three orders of magnitude. Although the coefficient of variation of the mean trip distance is much greater than 1.0 for all motorized surface transport modes, these modes cannot be easily substituted due to very different service characteristics (including speed) and human factors.

The energy use and intensity figures of the mechanized modes in Table II relate to the United States in 2000 and cannot reflect the enormous progress that was achieved in vehicle technology. Although the fundamental principles of surface vehicle and aircraft propulsion have remained unchanged over many decades, the performance of these systems and their integration into the vehicle have improved continuously. These improvements were compounded by ever further gradual reductions in all types of vehicle resistance, resulting in significant reductions in fuel consumption overall. An illustrative case is the automobile. If adjusting for changes in vehicle mass, fuel consumption per kilometer traveled and vehicle empty mass of the average new light-duty vehicle (LDV) sold have declined significantly—in the United States, from 8.2 to 5.5 L/100 ton-km between 1975 and 2000 alone. However, human factors did not allow a full translation of these gains into real reductions of transportation energy intensity. These and other barriers are described in the following section.

4. HUMAN FACTORS

Human factors, here defined as how technology is being used, generally include a wide range of behavior-related activities, from declining vehicle occupancy rates to the process of decision making when buying a new vehicle in a market offering models with a wide range of fuel consumption. This section briefly describes the most important components.

4.1 Declining Vehicle Occupancy Rates

In the United States, an average of 2.20 persons were sharing an automobile in 1969. That number declined to 1.90 in 1977, 1.75 in 1983, 1.64 in 1990, and 1.59 in 1995. Not taking into account secondary effects, that 28% decline in occupancy rate caused a 38% increase in passenger travel energy use, with all other factors being equal. Such continuous decline in occupancy rates was caused by the rising share of women in the labor force, the increase in vehicle fleets, and the decrease in average household size. These somewhat interrelated forces were strongest during the second half of the 20th century. Between 1969 and 1995 alone, the share of women in the labor force increased from 38 to 46%, the growth in the number of vehicles per household increased from 1.16 to 1.78 (reducing the need for sharing a vehicle), and average household size

declined from 3.16 to 2.65 (reducing the opportunity for sharing a vehicle). Given that these trends' rate of change will diminish in future, vehicle occupancy rates will necessarily level off at a rate well above 1.0. A further factor that contributes to stabilizing occupancy rates is the inherently higher rate of leisure-based trips, the most rapidly growing trip purpose. Whereas automobiles carry only 1.2 occupants in work-related travel on average, the occupancy rate is 2.0 for leisure travel. These U.S. figures and trends are very consistent with those in other parts of the industrialized world.

Occupancy rates have also declined for buses and trains, due mainly to the shift toward faster and more flexible modes. For example, the average U.S. transit bus carried 12.9 passengers per bus in 1970, and that number declined to 9.2 in 2000. Occupancy rates have increased only in air traffic and currently account for approximately 70% in U.S. domestic travel. That comparatively high rate can be attributed to preferable service characteristics (high speed) in combination with declining fares (Fig. 1D).

4.2 Shift toward Larger and More Powerful Vehicles

The average 1975 model year U.S. LDV had an empty mass of 1.84 tons and an engine power of 102 kW, translating into a vehicle power/mass ratio of 55 kW/ton, and fuel consumption was 15.1 L/100 km. As a result of higher oil prices and tight fuel economy standards, the curb weight of the average new LDV declined to 1.43 tons and the power/mass ratio increased only slightly to 60 kW/ton by 1987, while fuel consumption declined by more than 40% to 8.8 L/100 km. However, with the onset of declining oil prices and the absence of a further tightening of fuel economy standards after 1987, curb weight rose to 1.66 tons, the power/mass ratio rose to 79 kW/ton, and fuel consumption rose to 9.2 L/100 km in 2000. The renewed growth in these performance indicators after 1987 is due mainly to the rising share of personal trucks, which already account for half of all LDVs sold in the United States today. In addition, shifts toward larger vehicles within the automobile segment further contributed to the growth in vehicle mass and engine power after 1987. These shifts can be attributed to consumer desires for increased safety, acceleration capability, and convenience (e.g., space, climate control, entertainment and navigation systems). Obviously, such an increase in consumer attributes, which has also been observed in many

other industrialized countries (albeit to a lesser extent), partly compensates the reduction potential in vehicle fuel consumption.

4.3 Other Factors

In addition to the decline in occupancy rates and the shift toward larger and more powerful vehicles, further human factors exist. Among those, perhaps the most significant one is driving behavior. Although little empirical evidence exists in that area, the shift toward more powerful vehicles generally allows more rapidly changing driving maneuvers that increase vehicle energy use per unit distance traveled. It is likely that a significant fraction of drivers adapt their driving style to the changing technology.

4.4 The Offsetting Effect of Human Factors

As a simplified illustration of the offsetting effect of human factors, energy intensity of the U.S. LDV fleet (the largest single energy consumer in the U.S. transportation sector) is decomposed into reductions in vehicle fuel consumption, relative growth in vehicle size classes, and declining occupancy rates.

LDV energy intensity can be formulated as follows:

$$\left(\frac{E}{pkm}\right)_{LDV} = \sum_i \frac{E_i}{vkm_i} \cdot \frac{vkm_i}{vkm} \cdot \frac{vkm}{pkm}, \quad (2)$$

where, E_i/vkm_i is vehicle energy use of automobiles and light trucks, vkm_i/vkm is the share of each of the two segments in kilometers driven, and vkm/pkm is the inverse average occupancy rate, assumed to be identical for all LDVs. Differentiating Eq. (2) and reorganizing the terms results in the percentage change of LDV energy intensity

$$\left[\frac{d\left(\frac{E}{pkm}\right)}{\frac{E}{pkm}}\right]_{LDV} = \sum_i \frac{d\left(\frac{E}{vkm}\right)_i}{\left(\frac{E}{vkm}\right)_i} \cdot \frac{E_i}{E} + \sum_i \frac{d\left(\frac{vkm_i}{vkm}\right)}{\frac{vkm_i}{vkm}} \cdot \frac{E_i}{E} + \frac{d\left(\frac{vkm}{pkm}\right)}{\frac{vkm}{pkm}}. \quad (3)$$

The application of Eq. (3) to the U.S. LDV fleet from 1970/1971 to 1999/2000 is depicted in Fig. 5. During the entire 31-year period, the change in human factors (here explicitly the relative growth in travel demand and decline in vehicle occupancy rates) was greater than zero; thus, human factors at least partly offset LDV fuel efficiency improvements. Until 1978, human factors more than offset fuel efficiency

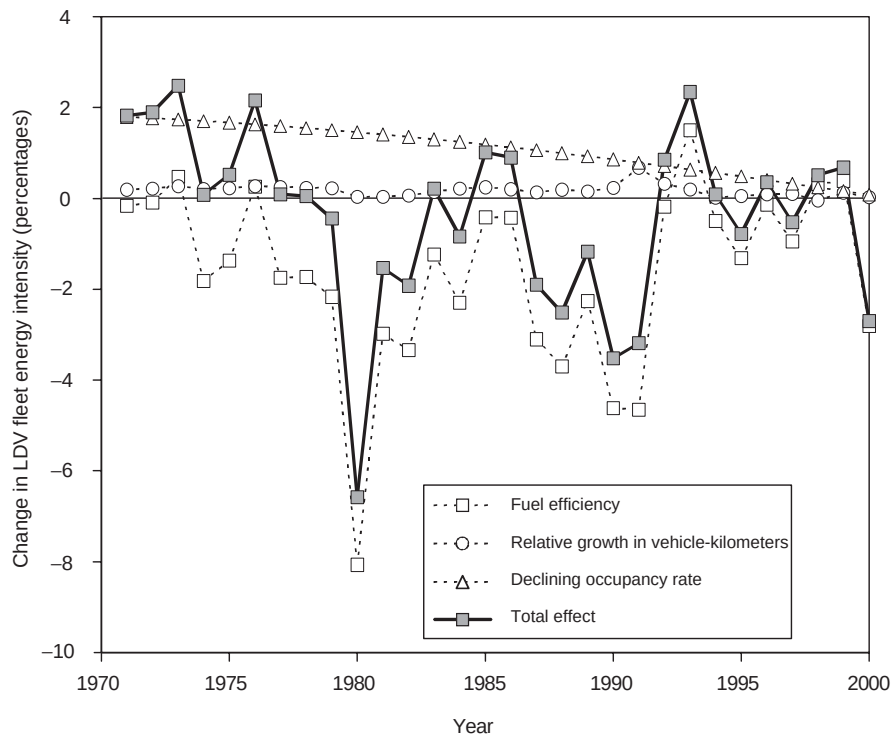


FIGURE 5 Annual percentage change in light-duty vehicle (LDV) energy intensity as the total of changes in technology improvements, shift in vehicle segments, and occupancy rate.

improvements occurring in the vehicle fleet, raising its average energy intensity. Because of drastic vehicle fuel efficiency improvements caused by the combination of the second oil shock and tight fuel economy standards, average energy intensity declined between 1979 and 1984. However, during that period, human factors offset that decline in energy intensity by more than 1%. Starting in 1985, LDV energy intensity declined only gradually, increasingly due to only modest improvements in energy efficiency.

5. PASSENGER TRAVEL ENERGY USE AND EMISSIONS

Figure 6 reports total passenger travel energy intensity (all modes) for France, Germany, and the United States over time. In France and Germany, where passenger travel energy intensity has remained roughly constant, improvements in fuel efficiency have been completely offset through a combination of human factors and shifts toward faster and more energy-intensive transportation modes. In contrast, in the United States, passenger travel energy intensity declined from 3.2 MJ/pkm in 1977 to 2.5 MJ/pkm in 1991, due mainly to drastic LDV and passenger aircraft fuel efficiency improvements. Multiplying average passenger travel energy intensity by growth in passenger-kilometers leads to total passenger transport energy use. Thus, in light of a nearly 90% increase in passenger traffic volume, U.S. passenger travel energy use grew from 10.4 EJ in 1970 to 17.8 EJ in 1990, an increase of more than 70%.

On a global scale, total transportation final energy use grew from 35 EJ in 1971 to 75 EJ in 2000, with

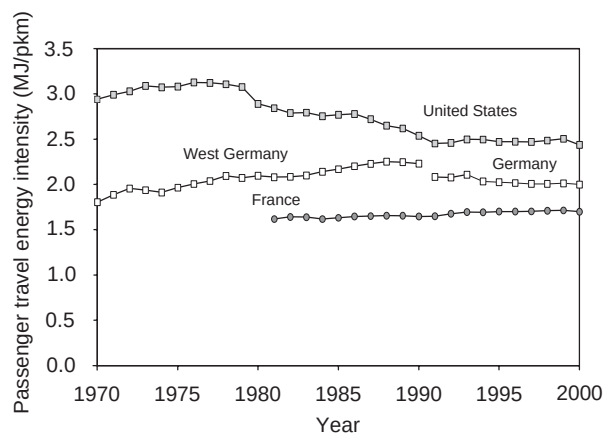


FIGURE 6 Evolution of total passenger travel energy intensity in France, Germany, and the United States.

approximately 60% of that amount corresponding to passenger travel. In contrast to criteria pollutants (e.g., carbon monoxide, lead, nitrogen oxide, ozone, particulate matter, sulfur dioxide), which can be decoupled from energy use through adjustments in fuel composition, engine combustion settings, and exhaust gas catalysts, carbon dioxide emissions are strongly linked to energy use via the fuel's carbon content. Thus, because oil products supplied approximately 97% of transportation energy, passenger travel carbon emissions increased in the same proportion as did energy use, that is, from approximately 0.4 billion tons to 0.9 billion tons.

6. REDUCING PASSENGER TRAVEL ENERGY USE AND EMISSIONS

According to Eqs. (1) and (2), passenger travel energy use can be controlled through improving energy efficiency and shifting toward less energy-consuming modes (reducing E/vkm) and reducing vehicle-kilometers traveled (reducing vkm). Various kinds of emissions, most notably greenhouse gases (GHGs), can be further reduced through burning less carbon-containing fuels or through burning zero-carbon fuels.

6.1 Reducing Vehicle-Kilometers Traveled

Transportation demand management (TDM) aims at reducing passenger travel through increasing the use of mass transit systems, increasing vehicle occupancy rates, substituting travel by electronic means, and reducing travel needs by land use planning. In addition, TDM includes a range of administrative measures, including more flexible work schedules, auto-restricted zones, and parking management. Common to most of these measures individually is their small potential, resulting in part from fundamental travel behavior and societal countertrends. The small potential of TDM is also partly inherent; electronic substitutes offer only a small reduction potential in traffic and associated energy use, even if assuming a relatively high percentage of the labor force working a significant amount of the time at home. (If 20% of the entire labor force worked an average of 2.5 days per week at home, work travel would be reduced by only 10%, and in light of the small 20% share of work trips in the United States, total vehicle trips would be reduced by a mere 2%.) Although the potential of each individual measure is limited, a sensible bundle of measures promises a significantly larger impact on travel and energy use.

Such packages of measures will be most effective if they contain pricing instruments that directly charge travelers for using transportation infrastructure, for example, through time- and day-dependent measures (e.g., road pricing) or rigid measures (e.g., variabilization of fixed costs, i.e., reducing vehicle taxes and/or insurance payments but increasing fuel costs). Unfortunately, such direct measures often meet public resistance and, thus, require long time periods for implementation.

6.2 The Promise of Technology

In the past, technology change has proven to be the most promising tool for reducing energy use and emissions. However, because human factors in combination with the growth in total travel demand can easily offset gains achieved by more fuel-efficient technology, noticeable reductions in passenger travel energy use require radical technology change. Such drastic change was offered by the three-way catalyst that has reduced vehicle tailpipe emissions of carbon monoxide, unburned hydrocarbons, and nitrous oxide by two orders of magnitude. Because reductions of that magnitude are not possible for energy use (given that a large share of a country's population would have to move on foot or by bike, as can be seen from Table II), noticeable reductions in energy use also require appropriate policy measures not only to push more fuel-efficient technology into the market but also to mitigate other human factors and demand growth at least in part.

Despite significant technological progress that has already occurred in improving fuel efficiency of all transport systems, a substantial potential in reducing transport sector energy use still exists. Perhaps the largest potential is offered by LDVs, which currently consume approximately 80% of U.S. passenger transport energy use. LDV fuel efficiency can be increased by increasing the efficiency of the drivetrain (engine plus transmission) and by reducing the amount of energy necessary to move the vehicle (driving resistances). Although the single largest reduction in energy use is expected to come from the drivetrain, the largest total potential will result from a combination of drivetrain measures and a reduction in driving resistances. Following such a holistic approach, individual energy savings potentials are compounded. If maintaining consumer attributes of current vehicles, fuel consumption can be reduced by 50 to 70%. Slightly lower potentials for reducing fuel consumption exist for urban buses and, over the longer term, for passenger aircraft.

6.3 Barriers

In addition to the human factors discussed previously, there are several barriers to reducing passenger travel energy use. One barrier reflects the higher retail price of more fuel-efficient transport technology. In the case of LDVs, this retail price increase can be drastic. If dividing into thirds the difference in fuel consumption between the average new vehicle on the road today and a highly fuel-efficient vehicle consuming 3 to 4 L/100 km, the first third of the potential for reducing energy use comes at a comparatively low cost. Whereas the second third already requires an increase in the vehicle retail price by up to 20%, the final third comes at a very high price (up to 30%). Because empirical evidence suggests that consumers may be willing to amortize extra costs for a more fuel-efficient vehicle over a few years only, penetrating highly fuel-efficient technologies into the market requires government action such as higher gasoline prices, incentives for purchasing more fuel-efficient technology, and some kind of regulation.

However, even if tight policy measures force more fuel-efficient technology into the transportation sector, increasingly slow fleet turnover significantly delays the translation of new vehicle fleet energy use to that of the entire vehicle fleet. Along with the above-discussed decline in vehicle energy use per ton-kilometer, continuous technological progress increased the average vehicle lifetime from 11 years for a 1966 model to 15 years for a 1990 model. (As a result, the median age of the U.S. automobile fleet nearly doubled from 4.9 years in 1970 to 8.8 years in 2000.) Thus, the associated time for a (nearly) complete phase-out of the entire vehicle fleet increased from 18 years (1966 fleet) to 21 years (1990 fleet). In light of these increasingly long time scales, reductions in fuel consumption need to be carried out rather soon to experience midterm effects.

7. OUTLOOK

Taking into account the previously discussed determinants of passenger demand for travel and energy use as well as the barriers for introducing more fuel-efficient vehicles, passenger travel energy use will continue to increase worldwide. Growth will be largest in those parts of the developing world where income and population increase most rapidly. Building on a continuation of the illustrated trends in travel demand, mode choice, and energy intensity, one recent study projected world passenger travel demand

to rise by a factor of 4.5 through 2050. During the same period, energy use and carbon dioxide emissions would increase by a factor of 3.5. To prevent such drastic growth in energy use and GHG emissions, technology for drastically reducing passenger travel energy use must be made available on a large scale during the coming years. Because of the extra investments for fuel-saving technologies and the high consumer discount rate, their large-scale introduction requires government action. The long turnover rates of transportation technology require such reductions to be implemented rather soon to experience noticeable reductions in energy use and (cumulative) GHG emissions during the coming decades. However, whether technology change alone can be the ultimate answer to reducing passenger travel energy use and GHG emissions remains to be seen.

SEE ALSO THE FOLLOWING ARTICLES

Alternative Transportation Fuels: Contemporary Case Studies • Bicycling • Intelligent Transportation

Systems • Internal Combustion Engine Vehicles • International Comparisons of Energy End Use: Benefits and Risks • Leisure, Energy Costs of • Lifestyles and Energy • Motor Vehicle Use, Social Costs of • Transportation and Energy, Overview • Vehicles and Their Powerplants: Energy Use and Efficiency

Further Reading

- Davis, S. C. (various years). "Transportation Energy Data Book." Oak Ridge National Laboratory, Oak Ridge, TN. www-cta.ornl.gov/cta/data/index.html.
- Greene, D. L. (1996). "Transportation and Energy." Eno Transportation Foundation, Washington, DC.
- Rennie, J., *et al.* (eds.). (1997). The future of transportation [special issue]. *Sci. Am.* 277 (4).
- Schafer, A., and Victor, D. (2000). The future mobility of the world population. *Trans. Res. A* 34(3), 171–205.
- Hellman, K. H., and Heavenrich, R. M. (2003). "Light-Duty Automotive Technology and Fuel Economy Trends: 1975 through 2003." EPA 420-R-03-006. U.S. Environmental Protection Agency, Washington, DC.

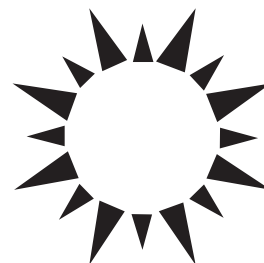
Peat Resources

ANNE JELLE SCHILSTRA

University of Groningen
Groningen, The Netherlands

MICHEL A. W. GERDING

Historian, Province of Drenthe
Assen, The Netherlands



-
1. Peat: A Short Introduction
 2. Peat as an Energy Source
 3. Environmental Concerns
 4. The History of the Use of Peat for Fuel

Glossary

- bog** (raised) A mire raised above the surrounding landscape and fed only by precipitation.
- boreal** Biogeographical zone between the temperate and the subarctic zones.
- fen** Peatland, which in addition to precipitation, receives water that has been in contact with mineral soil or bedrock.
- mire** A peatland where peat is currently being formed and accumulating.
- peat** Vegetable matter accumulated and preserved in anaerobic waterlogged areas.
- turf** The material noun for dried peat.

Peat is the dead vegetable matter that has accumulated in waterlogged anaerobic layers over periods of often thousands of years. The slow diffusion of oxygen in water limits the decay of organic matter so that the cycle of carbon fixation by photosynthesis and usual decomposition of the material after dying is incomplete. The mires or peatlands where peat is formed are quite extensive in Canada, Alaska, Siberia, northern Europe, and some tropical countries. In boreal areas especially, *Sphagnum* species contribute to peat formation, increasing the thickness of the peat layer annually by 0.2–4 mm. In the tropical peatlands, rainforest trees produce the litter and roots that may accumulate in permanently waterlogged forest floors.

1. PEAT: A SHORT INTRODUCTION

A traditional distinction is made between bogs and fens: bogs rise above their surroundings and are fed by rainwater only. Due to the capillary properties of peat, the water level can be drawn up so that soggy lens-shaped domes hundreds of meters wide and up to 12 m deep can be formed. By weight, the peat from bogs may consist of 98% water and only 2% solid matter. Peat that is formed in shallow lakes and that may eventually fill them completely is usually called a fen. In suitable circumstances, the peat blanket can grow sideways and can eventually even cover nearby hills. The solid matter may constitute approximately 10% of the peat. See Fig. 1. Undisturbed peatlands are ecosystems where the wet environment and the lack of nutrients have led to the evolution of a characteristic specialized flora and fauna.

Boreal peatlands may date from the end of the last ice age, approximately 10,000 years ago. But obviously the upper parts and the fringes can be of quite recent origin. As a high groundwater level is crucial for the continuing existence of peatlands, interference with the water level has dramatic consequences. Drainage will end the accumulation of new peat, but in contrast may enhance the growth of other plants; for example, trees that remain rather stunted on intact peatlands due to the shallowness of the aerobic top layer (the acrotelm) can grow normally when, after drainage, the top of the formerly anaerobic layer (the catotelm) is oxygenated. Another consequence of drainage is that the layers now freshly exposed to oxygen will start to decompose. Whereas the methane (CH₄) emissions that characterize many healthy mires are halted, the emission of carbon dioxide (CO₂) greatly increases: the mire has changed from a net carbon sink to a

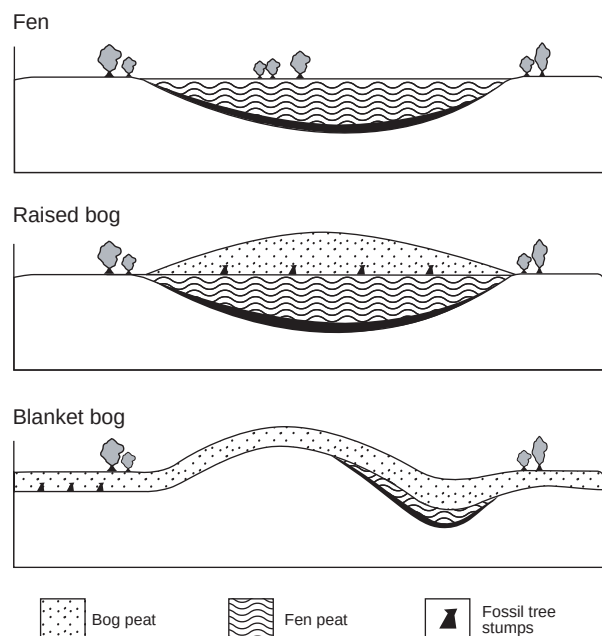


FIGURE 1 A fen can develop into a raised bog. Only precipitation feeds the living surface. A bog may also grow sideways and cover nearby hills. The bottom of the bog may be thousands of years old.

carbon source. When forest grows on a drained peatland, it will temporarily fixate CO_2 faster than the peat underneath oxidizes. But eventually, when the forest approaches maturity and the net uptake of CO_2 diminishes, CO_2 emissions from the peat will out-compete the forest uptake.

The question of whether or not peat is a fossil fuel has been around for some time. Much younger than brown coal or lignite, peat is rather “old” when compared to renewable energy sources, such as wood from production forests. Designating peat as a renewable or fossil fuel may have implications; the power it generates may be labeled “green” or not. It can convincingly be argued, however, that peat is neither a fossil fuel nor a renewable energy source. Some agreement has been found by labeling it a “slowly renewable biofuel.”

2. PEAT AS AN ENERGY SOURCE

Estimations on the total area of peatlands are rather varied, in part because countries use different definitions of peatlands. A rough estimation would be almost 4 million km^2 (3% of the earth’s land area), not including the more than 2.4 million km^2 of terrestrial wetlands. Canada, Alaska, Scandinavia,

and Russia have huge boreal peatlands and in Indonesia and South America extensive tropical peatlands can be found. Estimations of the amount of carbon stored in peatlands also varies, but a safe guess would be 270–370 Gt, (1 Gt = 1 gigaton = 10^9 tons), approximately one-third of the global soil (nonfossil), carbon store or two-thirds of the atmospheric carbon pool. Annual addition to this carbon store is approximately 40–70 million tons, small compared to the 100 Gt that is exchanged annually worldwide between the atmosphere and the terrestrial ecosystems. In Europe and Africa, more than one-half of the original peatlands have been used or destroyed, whereas in the Netherlands, for example, only 1% is left. Agriculture (50%), forestry (30%), and peat extraction (10%) are the main causes for global anthropogenic peatland losses.

In Fig. 2, some properties of peat and other fuels are depicted. As can be seen, the heating value of dry peat is between those of wood and brown coal. The water content of peat is an important factor for its practical heating value. Peat taken from undrained peatlands may have a water content like that of a very wet sponge. Preparing peat for industrial use then amounts largely to the efficient drying of the substrate. Peatlands to be harvested are usually drained for several years by parallel ditches 20 m apart, cleared of growth, and leveled. The soil is then roughened for a couple of centimeters and left to dry in the wind and the sun. After a suitable time, the loose peat is collected by giant pneumatic harvesters or brushed into ridges, after which the next layer is prepared to dry. After additional drying when necessary, the peat is moved by truck or narrow-gauge railway to a collecting area or directly to the power plant. Clearly, the annual harvest strongly depends on the weather; a given area may yield in a good year several times more peat than in a bad year. For example, in 1997 Finland harvested 10.1 million tons of peat for energy production but in 1998 it harvested only 1.5 million tons.

In Finland, Ireland, and Sweden, peat is also used for strategic reasons. With little or no local fossil fuel deposits, fuel diversification that includes local sources seems important. Also, in remote areas where peat is easily available, it is sometimes the obvious choice for an energy source. An additional argument for the use of peat for power generation is found as a means to stimulate local economy and labor opportunities. The Edenderry peat electricity plant (127 MW) in central Ireland is an example of this policy. Modern large-scale power plants are based on fluidized bed combustion technology that

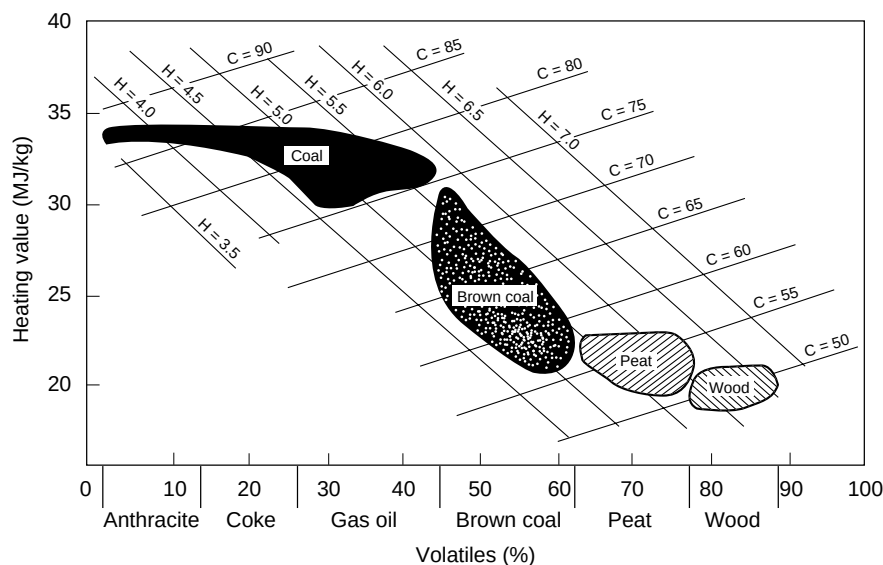


FIGURE 2 Fuel properties of dry peat compared to other fuels. Carbon and hydrogen content by weight is indicated by the sloping lines. Natural gas (CH_4 , by weight 75% carbon and 25% hydrogen) lies outside the frame.

results in higher efficiencies, lower emissions, and the capability to include other fuels such as wood chips.

Finland, Ireland, and Sweden are the only industrialized countries that use peat with a state of the art technology. In Ireland, 11% of the electrical power is generated by peat-fueled stations; in Finland, approximately 6% of the electrical power is generated by peat-fueled stations, and wood is often co-fired as well. In Sweden, less than 1% of the electrical power is generated by peat-fueled stations.

3. ENVIRONMENTAL CONCERNS

The use of peat for energy has a number of environmental problems in common with the use of fossil fuels. Of global concern is the emission of CO_2 . Plants that are the source of renewable biofuels recycle atmospheric CO_2 so that solar energy can in principle be harvested and used in a carbon-neutral way. Growing and harvesting peat similarly would require a cycle length of thousands of years, quite impractical. Nor is the fact that other living peatlands continue to accumulate atmospheric CO_2 faster than peat is converted to CO_2 in power stations a valid argument for "sustainability." The sulfur content of peat is low but varies; 0.2% by weight of dry matter is a representative value. Local effects are extensive. Harvesting of peat is akin to open-cast mining. After harvesting, which can take 20 years, the remaining area can be put to good use,

for rewetting and regrowth of peat, for forestry, or for agriculture. In the Netherlands, for example, in major parts of the west of the country, agriculture is actually "after use" of peatlands where most peat was removed centuries ago. Only small areas where accumulating peat is left (a total of $\sim 10 \text{ km}^2$) have to be managed and protected in order to keep its natural values from further deteriorating. Largely due to the oxidizing of the former peat cover, approximately 26% of this country lies below sea level and 66% would be inundated regularly without the protection from dikes and dunes. The harvesting of peat as a natural resource generations ago has as a consequence that the Dutch today must face the risk of flood disasters and the costs of preventing them.

4. THE HISTORY OF THE USE OF PEAT FOR FUEL

Fens and bogs, peatlands, are in the words of Julius Caesar "land that can not be walked upon and water that can not be sailed on." The fact that it can be used for fuel was first recorded by Plinius the Elder, who, in his *Naturalis Historia* (77 AD), mentions people who dug lumps of earth that they dried in wind and sunshine and then set fire to in order to cook their food and warm their houses. He writes about a people living in the delta of the river Rhine, now the Netherlands. Apart from this singular quote, almost nothing is recorded about the use of peat for

fuel until well into the Middle Ages, and then only from the Low Countries (Flanders and the Netherlands). Yet millions of hectares of peatlands existed north of the Alps.

In the history of the use of peat for fuel, it is necessary to make a distinction between peat digging for domestic use only and peat production on a larger scale in a businesslike operation intended for larger markets. This section concentrates on the latter. Outside the Netherlands, the literature on peat and its history appears to be very scarce or virtually nonexistent. A comprehensive overview can be given only of the Low Countries.

The reason for the widespread use of peat for fuel in the Netherlands and Flanders lies in the lack of firewood on the one hand and the abundance of waterways (rivers and canals) on the other hand. The relatively low cost of transportation of the bulky material combined with the engineering skills of the Dutchmen in controlling the water and managing waterways were crucial in the proliferation of the use of peat for fuel. From the early Middle Ages until well into the 20th century, peat (turf) was the source of "fossil" energy in the Netherlands.

The Flanders cities of Antwerp, Ghent, and Brugge were the first to invest in large-scale peat production from ca. 1250 AD onward. From there it spread northward, so that in the 14th century the fens in a large part of Holland, the most important province of the Netherlands (with cities such as Amsterdam, Leiden, Delft, and Haarlem), were brought into production. A crucial step in the development was the introduction in 1513 of the "dredger" (bagger-beugel), a bucket on a long wooden pole of approximately 4 m. Now it was no longer necessary to drain the fens first before the peat could be cut and dried. The wet peatmud could be dredged directly from the water and stored in small boats and then spread out on nearby fields to be cut and dried. Production soared and coincided with the strong economic and demographic growth of the Netherlands in the 16th and 17th centuries ("Golden Age"). It has been argued that the Dutch Golden Age was born from peat. If the total energy supply had relied on wood, one-quarter of the present area of the country would have to have been covered completely and permanently with production forest. Furthermore, the absence of waterways would have required approximately 100,000 horses for the transport of fuel and a further 600,000 ha of cultivated area to feed them. Thus, the energy attainability gave the Dutch Republic a tremendous advantage.

The dredging of low-peat turf had one major disadvantage: the loss of land was enormous. Over time, lakes replaced pastures and the unstable banks threatened the nearby countryside. Strict regulations were issued and taxes were levied to form funds from which new polders could be formed. This, and the demand for energy, caused the Dutch merchants to look elsewhere in the country. In the northeastern part of the country, 140,000 ha of good-quality black peat was still available. From the middle of the 17th century onward, this was the most important production area. In contrast to the dredging of low-peat turf, the high-peat areas had to be drained first. Thus, an infrastructure of waterways had to be built in order to drain the bogs and thereafter to transport the turf. This required large investments, which the rich merchants of Holland had in abundance. Also, local nobility and city and regional governments took part in the peat companies ("Veencompagnieën"), which were in effect mainly canal companies. Most of the work in the peateries itself was performed by subcontractors and private entrepreneurs (Fig. 3).

The contribution of peat from the north of the country to the national energy supply is a good example of the importance of peat in the Dutch energy supply. The darker areas are former peat bogs, the lighter areas are fens that have been excavated since the 16th century.

In the period between 1550 and 1950, peat digging played an important role in the northern parts of the Netherlands. At the outset, this was limited, but from the 17th century onward, its role increasingly grew in importance. During these 400 years, approximately 100,000 ha of raised bog peat and 42,000 ha of fenland peat was converted to dry peat fuel. Approximately 16.6×10^{11} turfs were dug, representing a combustion value of 3.5×10^9 GJ, which is comparable with 8300 million tons. This peat was excavated from 62 raised bog areas and 11 fenland areas. An average depth of 2 m (12 layers) of winnable peat was obtained.

The most important reason for the long-term development of production is the growth in population. Until 1850, peat was by far the most important source of energy for just about all sectors of society, not just for domestic needs such as cooking and heating, but also for crafts and industry, especially those industries directly linked to the growth in population such as brick and tile production, brewing, baking, and distilling. Until 1650—it can be assumed—when peat played an important role in the export industry, production was not directly related



FIGURE 3 Peat-containing soils in the Netherlands. Most of the country was covered by peat in Roman times. Ages of use have removed most of the peat, leaving a country where 26% of its area lies below sea level. The darker areas are former peat bogs, the lighter areas are fens that have been excavated since the 16th century.

to the internal growth of the population. From the 19th century onward, coal began to assume a more important role in the satisfaction of energy requirements, but until at least World War I the growth in population was so strong that peat continued to prosper. Thereafter no longer. The continuous growth from 1750 onward was over and peat began to play an increasingly marginal role, leading to a complete dependence on coal. Nevertheless, peat production during the interwar years, in addition to production in many crisis years, underwent many years of reasonable growth, especially in the second half of the 1930s. After a temporary boom during World War II, the role of peat was completely finished by 1950.

4.1 Other Countries: Germany

Dutch peateries were taken as an example by the German governments of the northwest part of the country, especially Niedersachsen, where enormous amounts of peat bogs were to be found also. Yet the

German initiatives never reached the scale of the Dutch. But the cities of Bremen, Oldenburg, and Emden in the 18th and 19th centuries managed to be self-sufficient in their energy supply because they could benefit from nearby peat deposits. The city of Hamburg lacked these advantages and constituted an important market for turf from Ostfriesland and Groningen. The oldest peat colony in Germany was Papenburg, founded in the 17th century. Other colonies followed, but never on the grand scale of the Dutch examples. Instead, Germany had in the 18th and 19th centuries a policy of erecting "Moorkolonien," settlements of paupers from the cities on the edge of bog areas who were forced to become small subsistence farmers.

4.2 British Isles

Although many peat bogs and fens were to be found in England and especially in Scotland, peat digging never developed on a businesslike scale in these

countries. Late in the 19th century, reports were published to promote such activities, for instance, the erection of electricity plants in Scotland fueled by peat, but they never came into being.

4.3 Ireland

In Ireland, the case was quite different. The country has a long tradition of using peat for fuel. There is evidence that it was already used as such in prehistoric times. Peat cutting increased in the 17th century due to the disappearance of forests and turf was the primary source of energy for country dwellers as well as city folk. Production reached its peak in the first half of the 19th century before the great potato famines of the 1840s. Until the early decades of the 20th century, all of the production was manual and carried out by individual turf cutters, who sold their product to boat-owners who transported the turf to Dublin by way of the Grand Canal, which was extended through the great Bog of Allen in 1786. In the west also, much of the turf was carried to the towns by boat. Many people in the Irish countryside as well as in the cities had individual rights to cut turf for their own domestic use, although they were not allowed to sell it. In 1912, 15,000 tons of turf was brought to Dublin on boats that could carry turf loads of 30–60 tons. In the first four decades of the 20th century, cooperative turf societies were established in order to promote the production and harvesting of hand-cut turf and to facilitate its direct sale by the producers. During World War II, turf became of strategic importance when coal supplies dried up and the entire fuel needs of the country depended on turf. In these years, thousands of Dubliners made their way into the Dublin mountains to cut their own peat.

After the war, the Bord na Móna was formed and took over all turf-cutting schemes at the government's request. The company developed into the world's second largest peat company. Most of the fuel peat goes to power plants in the country, which take up approximately 80% of the total peat production. Between 1946 and 1995, Bord na Móna produced ca. 170 million tons of peat for fuel.

4.4 North America

Although a great deal of Canada consists of peatlands, peat for fuel has never been an issue there or in the United States either. Peat is harvested only for horticultural purposes. The same can be said about Central and South America.

4.5 Baltic and Scandinavia

In most of the Baltic states, peat was used for fuel on a very small scale and then only for domestic use. Businesslike activities such as those in Germany, Ireland, and the Netherlands were absent. Peat production for horticultural purposes was, and still is, much more widespread. Only in Finland has there been production of peat for fuel since the 1970s. Since the oil crisis of 1973, the Finnish government labeled peat as a strategic good. In the middle of the country, a power station is fueled by peat.

4.6 Russia

In Russia, formerly the largest user of peat for energy, peat consumption represents only 0.3% of the primary energy source.

SEE ALSO THE FOLLOWING ARTICLES

Biomass Combustion • Biomass for Renewable Energy and Fuels • Coal Resources, Formation of

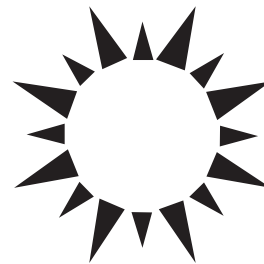
Further Reading

- de Zeeuw, J. J. (1978). Peat and the Dutch Golden Age: The historical meaning of energy-attainability. A.A.G. Bijdragen, Vol. 21, pp. 3–33. University of Wageningen, Wageningen, the Netherlands.
- Gerding, M. A. W. (1995). Vier eeuwen turfwinning (Four ages of peat digging). A.A.G. Bijdragen, Vol. 35, University of Wageningen, Wageningen, the Netherlands.
- Joosten, H., and Clarke, D. (2002). "Wise Use of Mires and Peatlands: Background and Principles Including a Framework for Decision-making." International Mire Conservation Group and International Peat Society. Jyväskylä, Finland.
- Lappalainen, E. (1996). "Global Peat Resources." International Peat Society, Jyskä, Finland.
- Schilstra, A. J. (2001). How sustainable is the use of peat for energy production? *Ecol. Econ.* 39, 285–293.

Petroleum Property Valuation

JAMES L. SMITH

Southern Methodist University
Dallas, Texas, United States



-
1. Impact of Property Valuations on the Petroleum Industry
 2. Discounted Cash Flow (DCF) Analysis: The Problem Simplified
 3. Special Characteristics of Petroleum Properties
 4. Incorporating the Value of Real Options
 5. Portfolio Analysis: The Combined Value of Multiple Properties
 6. Conclusion

Glossary

capital asset Any equipment, facility, or plant capable of generating a long-lived stream of future income.

discounted cash flow (DCF) A method for estimating the present value of future cash flows that adjusts for the time value of money and the degree of uncertainty surrounding future receipts.

discount rate The factor by which expected receipts in a future period are reduced to reflect the time value of money and unpredictable variation in the amount ultimately received.

diversifiable risk A source of unpredictable financial performance that varies randomly from one investment to another and therefore averages out, rather than accumulates, over all of an investor's holdings.

fair market value The price expected to be paid for any asset in a voluntary exchange between an independent buyer and independent seller.

Monte Carlo analysis A method for assessing the magnitude and implications of risk by simulating possible outcomes via random sampling from the probability distribution that is assumed to control underlying risk factors.

net present value (NPV) A measure of the value of a project obtained by discounting the stream of net cash flows (revenues minus expenditures) to be received over the entire life of the project, based on the time profile and riskiness of net receipts.

nondiversifiable risk A cause of unpredictable financial performance that tends to impact all of an investor's holdings in the same direction or manner.

proved reserves The volume of petroleum resources in a developed field that are reasonably expected to be recoverable given current technology and prices.

real options The general phenomenon by which the value of physical assets depends on, and therefore may be enhanced by, management's ability to modify or postpone investment and operating decisions based on the receipt of new information.

risk premium That portion of the discount rate that compensates the investor for the inherent unpredictability of future financial returns, as opposed to the pure time value of money.

Petroleum property valuation refers to the analytical procedure by which the commercial value of oil and gas fields is assessed. This assessment provides to prospective buyers and sellers, and other interested parties such as lenders and tax assessors, an estimate of the fair market value of underground deposits of oil and gas—the amount for which they might be bought or sold. The value of an underground deposit is directly related to the ultimate value of whatever petroleum may be extracted in the future, but because the future is uncertain, the value of the property is subject to various sources of risk that stem from geological as well as economic factors. To be useful, the valuation procedure must take proper account of the unpredictable fluctuations that would cause field development, operations, and performance to deviate from the expected outcome. This task represents a complex analytical problem that has challenged traditional valuation methods. To meet the challenge, new and highly innovative techniques have been developed and are finding widespread use.

1. IMPACT OF PROPERTY VALUATIONS ON THE PETROLEUM INDUSTRY

The petroleum industry's reliance on accurate and reliable valuation methods is apparent. The risk of paying too much for an acquired property, or selling for too little, is always present. The frequency and size of property transactions underscores the importance of getting the valuations right. Since 1979, for example, more than 5000 parcels of existing oil and gas fields have been sold in the United States, bringing more than \$600 billion of revenue to the sellers. The negotiations that support these exchanges hinge on finding valuation methods that both sides deem acceptable. In addition, investments are made by the petroleum industry to acquire leases in raw acreage where no oil or gas field is known to exist. The U.S. government is a major source of such properties and has raised in excess of \$60 billion since 1954 (when the federal leasing program was initiated) by auctioning petroleum exploration and development rights on federally owned lands and the outer continental shelf. The petroleum industry makes comparable investments on a regular basis to acquire oil and gas leases from private landowners and the states, as well. The ability to value petroleum properties accurately therefore plays a critical role in determining the financial success or failure of oil and gas producers.

2. DISCOUNTED CASH FLOW (DCF) ANALYSIS: THE PROBLEM SIMPLIFIED

In some respects, an oil field is no different than any other capital asset and valuation techniques for petroleum properties are therefore similar to procedures used in other sectors of the economy. A capital asset represents any long-lived investment in productive facilities that have the potential to generate a future stream of earnings. If those earnings are not large and predictable enough to justify the initial expenditure, the investment should not be made. Intuitively, the value of the capital asset may be thought of as the extent to which anticipated cash receipts outweigh the initial expenditure. Measuring and weighing the projected cash flows therefore forms the heart of the valuation problem.

2.1 Projecting Cash Flows

The projected cash flow stream is a composite forecast that results from many separate assumptions concerning physical attributes of the oil field and the economic environment in which it will be produced. The number of wells and size of facilities required to delineate and develop the field, in conjunction with the presumed cost level for drilling services and oil field equipment, will roughly determine the scope and timing of initial expenditures. The magnitude and duration of cash inflows (sales revenue minus operating cost) are determined by a further set of assumptions regarding the flow rate from individual wells (and the rate at which production will decline as the field is depleted), the quality and price of produced oil and gas, necessary operating and maintenance costs required to keep wells and field plant facilities in order, and the level of royalties and taxes that are due to lessors and governmental authorities. Thus, the projection of net cash flow for the field as a whole is the combined result of many interrelated but distinct cash flow streams. Some components are fixed by contract and can typically be projected with relative certainty (e.g., royalty obligations and rental payments), but others require trained guesswork that leaves a wide margin of error (e.g., future production rates and oil price trends). It seems reasonable that those components of the cash flow stream that are known with relative certainty be given greater weight in figuring the overall value of the field, but as discussed further later in this article, properly executing this aspect of the valuation procedure was hardly practical until so-called options-based valuation methods were developed.

2.2 Discounting Cash Flows

A dollar received (or paid) in the future is worth less than a dollar received (or paid) today because of the time value of money. Cash in hand can be invested to earn interest, and therefore will have grown in value to outweigh an equivalent amount of cash to be received at any point in the future. If the relevant periodic rate of interest is represented by the symbol i (e.g., $i = 10\%$), then the present value of a dollar to be received t periods hence is given by $PV(i, t) = 1/(1+i)^t$. This expression is referred to as the discount factor, and i is said to be the discount rate. The discount factor determines the relative weight to be given to cash flows received at different times during the life of the oil field. Cash flows to be received immediately are given full weight, since

$PV(i, 0) = 1$, but the weight assigned to a future receipt declines according to the amount of delay. Thus, the net present value (NPV) of an arbitrary cash flow stream represented by the (discrete) series of periodic receipts $\{CF_0, CF_1, CF_2, \dots, CF_T\}$ is computed as the sum of individual present values:

$$NPV = \sum_{t=0}^T \frac{CF_t}{(1+i)^t}. \quad (1)$$

It is quite common to perform this computation on the basis of continuous discounting, where the periodic intervals are taken to be arbitrarily short (a day, a minute, ..., an instant), in which case the discount factor for cash to be received at future time t declines exponentially with the length of delay: $PV(i, t) = e^{-it}$. Therefore, when the cash flow stream is expressed as a continuous function of time, NPV is reckoned as the area under the discounted cash flow curve:

$$NPV = \int_0^T CF_t \cdot e^{-it} dt. \quad (2)$$

It is apparent, whether the problem is formulated in discrete or continuous time, that correct selection of the discount rate is critical to the valuation process. This parameter alone determines the relative weight that will be given to early versus late cash flows. Since exploration and development of oil and gas fields is typically characterized by large negative cash flows early on, to be followed after substantial delay by a stream of positive cash flows, the choice of a discount rate is decisive in determining whether the value of a given property is indeed positive. The extent to which discounting diminishes the contribution of future receipts to the value of the property is illustrated in Fig. 1, which shows the time profile of cash flows from a hypothetical oil field development project. With no other changes to revenues or expense, the property's net present value is reduced by a factor of ten, from nearly \$1 billion to roughly \$100 million, as the discount rate is raised from 8% (panel a) to 20% (panel b). These panels also illustrate how discounting affects the payback period for the property in question (i.e., the time required before the value of discounted receipts finally offsets initial expenditures): 7 versus 11 years at the respective rates of discount.

With so much at stake, the selection of a discount rate cannot be made arbitrarily. If the discount rate is not properly matched to the riskiness of the particular cash flow stream being evaluated, the estimate of fair market value will be in error. Because no two oil fields are identical, the appropriate discount rate may vary

from property to property. A completely riskless cash flow stream (which is rare) should be discounted using the risk-free rate, which is approximated by the interest rate paid to the holders of long-term government bonds. Cash flow streams that are more risky, like the future earnings of a typical oil field, must be discounted at a higher rate sufficient to compensate for the owner's aversion to bearing that risk. The degree of compensation required to adequately adjust for risk is referred to as the risk premium and can be estimated from market data using a framework called the capital asset pricing model.

One important implication of the capital asset pricing model is that diversifiable risks do not contribute to the risk premium. A diversifiable risk is any factor, like the success or failure of a given well, that can be diluted or averaged out by investing in a sufficiently large number of separate properties. In contrast, risks stemming from future fluctuations in oil prices or drilling costs are nondiversifiable because all oil fields would be affected similarly by these common factors. The distinction between diversifiable and nondiversifiable risk is critical to accurate valuation, especially with respect to the exploratory segment of the petroleum industry: although petroleum exploration may be one of the riskiest businesses in the world, a substantial portion of those risks are diversifiable, thus the risk premium and discount rate for unexplored petroleum properties is relatively low in comparison to other industries.

The appropriate discount rate for the type of petroleum properties typically developed by the major U.S. oil and gas producers would be in the vicinity of 8 to 14%. This is the nominal rate, to be used for discounting cash flows that are stated in current dollars (dollars of the day). If future cash flow streams are projected in terms of constant dollars (where the effect of inflation has already been removed), then the expected rate of inflation must be deducted from the nominal discount rate, as well. Cash flow streams derived from properties owned by smaller or less experienced producers who are unable to diversify their holdings, or in certain foreign lands, may be deemed riskier, in which case the discount rate must be increased in proportion to the added risk.

Not only does the appropriate discount rate vary according to property and owner, but the individual components of overall cash flow within any given project are likely to vary in riskiness and should, in principle, be discounted at separate rates. It is fair to say, however, that methods for disentangling the separate risk factors are complex and prone to error, and it is common practice to discount the overall net

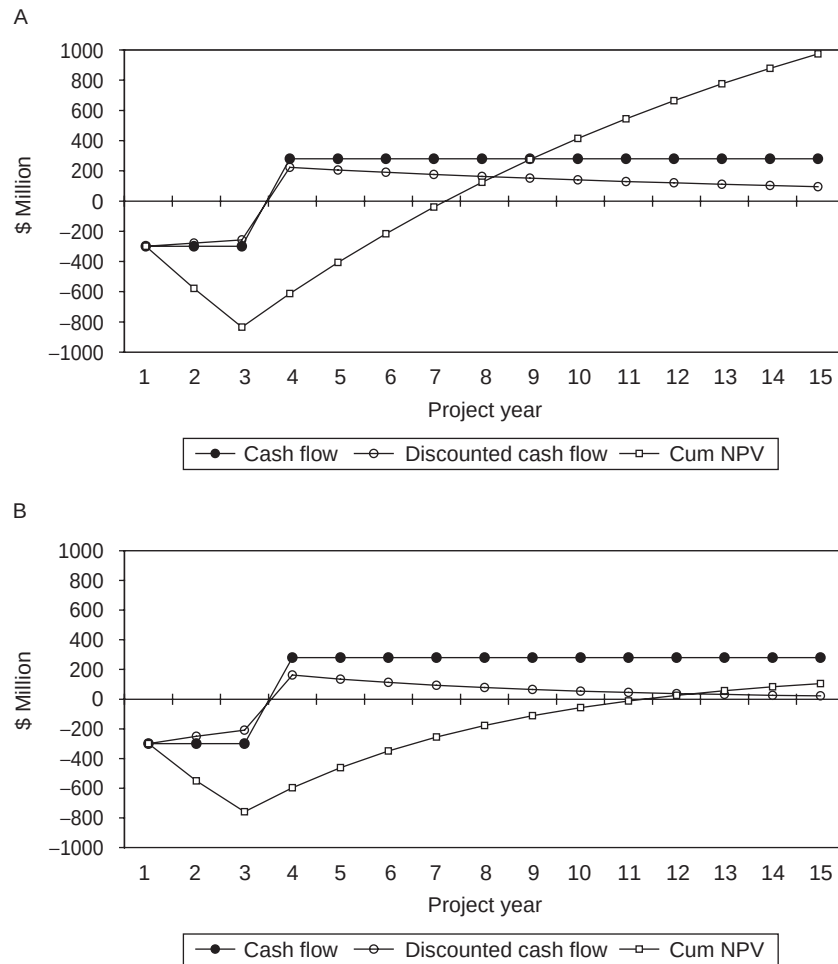


FIGURE 1 Discounting diminishes the contribution of future receipts. The illustrations show a hypothetical net cash flow stream, heavily negative at the outset, followed by consecutive years of positive operating revenues. When discounted at the higher rate, it takes longer for the same revenue stream to offset initial expenditures, resulting in a lower cumulative net present value. (A) Net cash flow and NPV at 8% discount rate. (B) Net cash flow and NPV at 20% discount rate.

cash flow stream at a single rate that reflects the composite risk of the entire project. In many applications, the error involved in this approximation is probably not large. Moreover, new insights regarding the valuation of real options (to be discussed later) provide a procedure in which it is appropriate to discount all cash flow streams at the same rate, which circumvents entirely the problem of estimating separate risk-adjusted discount factors.

3. SPECIAL CHARACTERISTICS OF PETROLEUM PROPERTIES

Most oil and gas fields share certain physical and economic characteristics that strongly influence the

pattern and behavior of cash flows, and therefore value. Although the following factors are not necessarily unique to the valuation of petroleum properties, their influence is of sufficient importance to justify more detailed discussion.

3.1 Exploration and Development Risk

Whether at the stage of exploration or development, investments made for the purpose of exploiting an underground petroleum deposit often go awry. Technical failures or economic circumstances may block the recovery of any resources from the property in question. During the past 30 years, 72% of all exploration wells and 19% of all development wells drilled in the United States have

resulted in so-called dry holes, which is industry parlance for a well that is unsuccessful. The causes of failure are numerous, ranging from barren geological formations, to deficiencies in the quality of the deposit that preclude recovery at reasonable cost, to the technical failure or breakdown of drilling equipment. In all such cases, the initial investment is forfeited—written off as the cost of a gamble.

The risk of dry holes is incorporated in the valuation process directly, by assigning appropriate weight to a zero-payoff outcome. After this modification, the NPV expression given previously, compare Eq. (2), would appear thus:

$$NPV = CF_0 + (1 - p_{DH}) \times \sum_{t=1}^T \frac{CF_t}{(1+i)^t}, \quad (3)$$

where CF_0 represents the cost of drilling, p_{DH} represents the probability of a dry hole, and the $\{CF_1, \dots, CF_T\}$ represent expected future cash flows contingent on success of the well. If the risk of failure is diversifiable, which is certainly true of drilling undertaken by large, publicly owned companies, it does not contribute to the risk premium associated with the property, which means that the discount rate (i) would not be affected by the presence or size of p_{DH} .

Drilling risk can take more complex and subtle forms than the simple dichotomy between success and failure. Outcomes (size of deposit, daily flow rates, gas/oil ratio, etc.) depend on many underlying factors that are not known with certainty but which are amenable to probabilistic analysis. The influence of uncertainty regarding these factors can be quantified via Monte Carlo analysis, wherein projected cash flows and the implied value from Eq. (3) are recomputed under a broad range of possible scenarios. This exercise yields a probability-weighted average outcome, which is the best single indicator of property value. Monte Carlo analysis also reveals the potential range of error in the valuation due to uncertainty in the underlying geological and economic parameters.

A particular nomenclature developed by the petroleum industry permits the degree of exploratory and development risk associated with a given property to be quickly assessed. The “proved reserves” category is the most certain because it includes only those resources that have already been delineated and developed and shown to be economically recoverable using existing technology under prevailing market conditions. Although the outcome of drilling may already have been resolved, proved

reserves are not entirely free of risk due to the continuing possibility that price and cost fluctuations will impact future cash flows. These factors represent nondiversifiable sources of risk that influence selection of the discount rate. In practice, the value of proved reserves reported in a producer’s financial statements must be reestimated and updated each year to reflect the impact of price and cost fluctuations and any new conclusions about reservoir behavior that emerge from ongoing field operations.

Resources categorized as probable and possible are successively further removed from having been tested, let alone proven, by the drill bit. Their valuations are accordingly subject to increasing (and potentially very large) margins of error. In addition, the reader must be warned that, although this resource classification scheme has been endorsed by the World Petroleum Congress, adherence to the definitions may vary in different parts of the world.

3.2 Oil and Gas Equivalents

Many petroleum properties contain natural gas (and various natural gas liquids) as well as oil, a circumstance that complicates the valuation process. Although the thermal energy content of a barrel of oil is roughly six times that of an mcf (thousand cubic feet) of gas, the market values of the two rarely, if ever, stand in that ratio, as illustrated in Fig. 2. Lack of pipeline facilities is one factor that tends to depress the value of gas deposits, which are more difficult and costly to bring to market than oil. The unconcentrated form in which natural gas occurs (low energy density per unit volume) also renders it

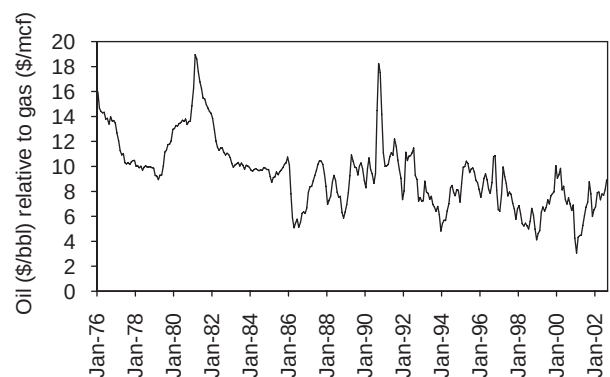


FIGURE 2 Historical fluctuation in the price of oil relative to natural gas. If the two fields traded at parity in terms of raw energy content, one barrel of oil would sell for roughly six times as much as one mcf of natural gas. Historically, oil has tended to trade at a premium, but the ratio is highly variable.

unsuitable for many uses (e.g., as transport fuel), and this tends to further depress its value. Consequently, the relationship between the value of oil and gas deposits is not stable but fluctuates markedly through time and space, depending on the relative demand for, and supply of, the two fuels in regional markets.

The value of any petroleum property will depend on the specific quantities of oil versus gas that are present. While it is common to see properties described in terms of the combined amount of “oil equivalents” (which usually means that each mcf of gas has been counted as one-sixth of a barrel of oil—based on the heat equivalencies of the two fuels), there is no reliable basis for such aggregation, to the extent that the chairman of a leading international oil company has proclaimed that oil- and gas-equivalents simply do not exist. What was meant, of course, is that any calculation of oil equivalents is a garbling of information that only serves to obscure the value of the underlying property. Oil equivalents cannot be compared for purposes of valuation to an equal volume of oil reserves—their values would not be the same. Nor can the oil equivalents of one property be compared to the oil equivalents of any other property where oil and gas reserves are present in different proportions. To do so risks a gross miscalculation of value. As long as consumers continue to distinguish between the two types of hydrocarbons, so must the owners and operators of petroleum properties.

3.3 The Volatility of Commodity Prices and Property Values

Compared to most commodities, oil and gas exhibit highly volatile price movements. Daily, annual, and monthly swings are unusually large relative to the base price levels for both fuels, which puts a large portion of the value of any petroleum property at risk.

The price of a barrel of oil at the wellhead differs, however, from the value of a barrel of oil in the ground, primarily because the reserve cannot be produced and delivered to a buyer instantaneously. This difference is evident in Fig. 3, which contrasts annual changes in the value of oil and gas reserves with corresponding changes in the wellhead price levels of these two commodities. Two things are apparent. First, *in situ* values (i.e., the value of petroleum reserves in the ground) are much smaller than wellhead prices. Over the past 10 years, oil and gas reserves have sold on average for only about

22% and 36% of their respective wellhead prices. Second, the *in situ* values are much more stable than wellhead prices. The year-to-year price change for oil in the ground averages (in absolute value) roughly 13%, versus 20% for changes in price at the wellhead. For gas, the contrast is even stronger: year-to-year price changes averaged 11% in the ground versus 24% at the wellhead.

The relationship between *in situ* values and wellhead prices, although complex and ever-changing, can be understood via a simple model of production from a developed oil field. Let q_0 represent the initial level of production, which is presumed to decline continuously at the rate a due to natural pressure loss in the reservoir as depletion proceeds; thus production at time t is given by $q_t = q_0 e^{-at}$. Assume further that the expected wellhead price of oil remains fixed over the relevant horizon at P per barrel, and that unit production costs amount to C per barrel. Using Eq. (2), the net present value of the property can then be calculated:

$$\begin{aligned} \text{NPV} &= \int_0^T q_0(P - C) \cdot e^{-(a+i)t} dt \\ &\approx \int_0^\infty q_0(P - C) \cdot e^{-(a+i)t} dt \\ &= \frac{q_0(P - C)}{a + i}, \end{aligned} \quad (4)$$

where the approximation is justified by the fact that oil fields are long lived. The volume of reserves (R) in the deposit is given by total production over the life of the property:

$$R = \int_0^T q_0 \cdot e^{-at} dt \approx \int_0^\infty q_0 \cdot e^{-at} dt = \frac{q_0}{a}, \quad (5)$$

which means the rate of extraction from reserves is given by the decline rate: $q_0 = aR$. After substituting this expression for q_0 into (4) and dividing by R , we obtain the *in situ* value (V) of a barrel of reserves:

$$V = \frac{\text{NPV}}{R} = \frac{a}{a + i}(P - C). \quad (6)$$

Equation (6) says quite a lot about the value of a producing property. To be concrete, let us set the production decline rate equal to the discount rate (10% is a realistic number for both), and set production costs equal to one-third of the wellhead price. After simplification, the relationship between *in situ* values and wellhead prices then reduces to $V = P/3$, which illustrates the petroleum industry's traditional one-third rule: The value of a barrel in the ground is worth roughly one-third of the price at the wellhead.

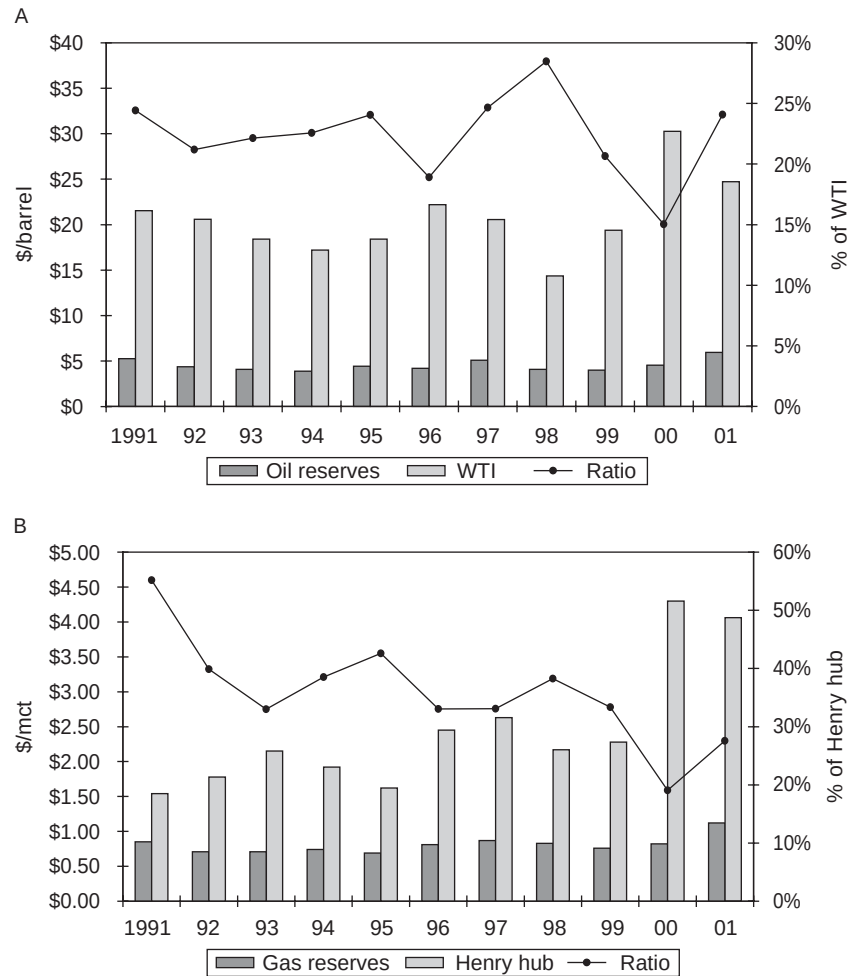


FIGURE 3 Value of oil and natural gas, *in situ* versus wellhead. *In situ* valuations are determined by the average price of reserve transactions tabulated by Cornerstone Ventures, L. P. Wellhead values are determined by the prices of WTI (oil) and Henry Hub (gas) contracts on the New York Mercantile Exchange. From Exhibits 4 and 5, Cornerstone Ventures, “Annual Reserve Report, 2001” (dated February 28, 2002), with permission.

Like any rule of thumb, the one-third rule is often wrong, as the numbers in Fig. 3 demonstrate, but it does point to a general tendency. Moreover, the derivation provided in Eqs. (4) to (6) allows one to anticipate when and why deviations would arise. Reserves that are extracted more rapidly (like natural gas, for example) would tend to sell for more than one-third of the wellhead price. To see why, simply substitute $a > i$ into (6). This confirms a pattern that was evident in Fig. 3: for gas, the value of reserves is consistently a larger fraction of wellhead price than for oil. It is also evident that the value of reserves should move inversely with the level of operating costs, which is why mature fields are eventually abandoned as it becomes more expensive to extract the oil.

What remains to be seen is why *in situ* values are less volatile than wellhead prices. According to the one-third rule, every 10% rise in wellhead price would be matched by a 10% rise in the value of reserves: the volatilities should be the same, but they are not. The explanation stems from the nature of commodity price movements, and the difference between random walk and mean-reverting processes, as illustrated in Fig. 4. A random walk process tends to wander off, rather than to return to its starting point. Any chance departure from the existing price level tends to become permanent. A mean-reverting process tends to be self-correcting; any succession of upward price movements increases the chance of future downward movements, which are required to restore the price to its former level.

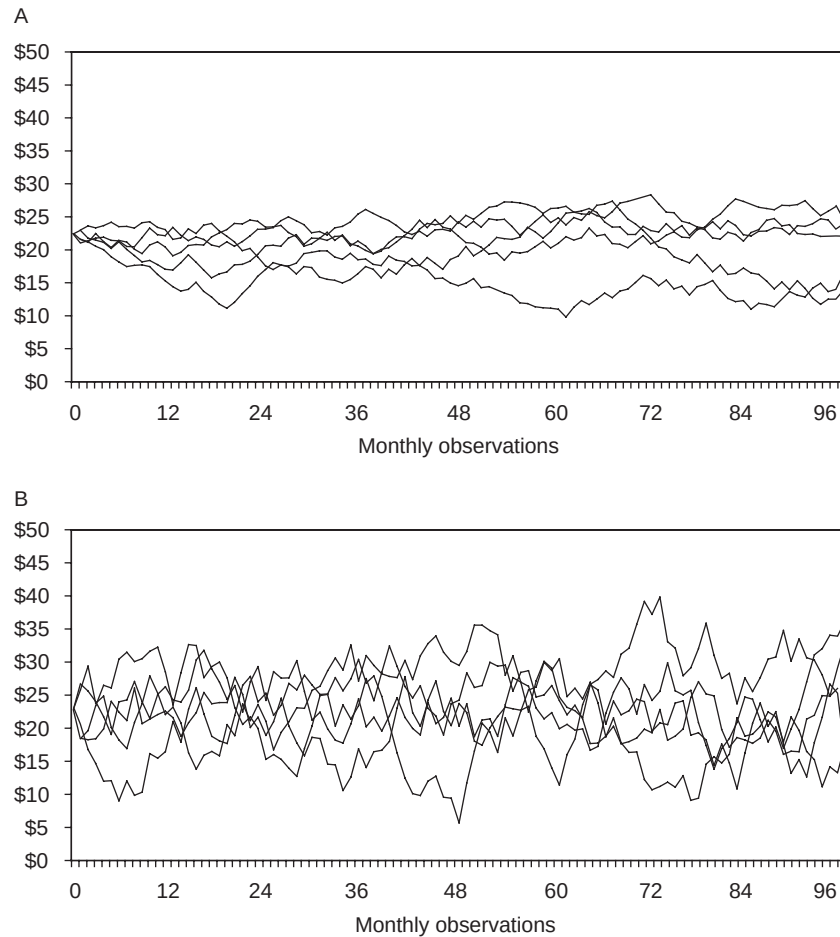


FIGURE 4 The illustrations show five examples each of a random walk sequence and a mean-reverting sequence. Simulations performed by the author. (A) Random walk price trends. (B) Mean-reverting price trends.

Whereas returns on investments in the stock market tend to follow a random walk, the prices of major commodities appear to be mean reverting, which is consistent with the view that the forces of supply and demand tend to keep commodity prices from drifting permanently away from their equilibrium levels. Mean reversion also implies that short-term fluctuations in the price of oil and gas at the wellhead are likely to be reversed in due course. Since the value of a reserve is determined by a combination of current and future prices, the long-term stability provided by mean reversion tends to dampen the impact of short-term commodity price movements. Equation (6), which uses a single value (P) to represent both the current and future commodity price level, is unrealistic in this regard; if prices do not follow a random walk, it gives accurate valuations only when the current wellhead price happens to correspond to the long-term level.

3.4 The Relationship between Reservoir Engineering and Property Valuation

To this point, we have taken the projection of future production, and therefore costs and revenues, as being determined exogenously. Subject to certain physical constraints, however, the rate of production is actually determined by petroleum engineers who design and install facilities with a view to maximizing the value of the field. Property valuation therefore rests implicitly on the assumption that production operations are optimized, and that process of optimization must itself be conducted within the valuation framework.

To illustrate, let us return to the previous example of an oil field subject to exponential decline. Based on our assumption that the discount rate and decline rate both equal 10%, and that operating costs amount to one-third of the wellhead price, we

determined the value of the property to be $P/3$ per barrel of reserves. Now, imagine that our team of reservoir engineers has identified an alternative drilling pattern that would double the extraction rate ($a=20\%$), with no sacrifice in total reserve volume—the same amount of oil would be produced, but faster. However, this alternative development strategy would also require the expenditure of an additional \$2 per barrel of reserves. Should it be adopted by management? If so, what would be the impact on the value of the property?

The valuation framework, represented in this case by Eq. (6), supplies the answer to these questions. After incrementing the extraction rate to 20%, but leaving all else unchanged, we find the value of the reserve to be $4P/9$ under the alternative drilling pattern, an increase of $P/9$ per barrel. This is a sensible strategy only if the gain outweighs the incremental cost of \$2 per barrel. Thus, we are justified in pursuing the faster, but more costly, production program if, but only if, the wellhead price is expected to exceed \$18/barrel.

Although the example may seem overly simplified, it illustrates an essential point: the value of a petroleum property is neither fixed nor guaranteed, and certainly it is not determined by geology and commodity prices alone. Value depends on management's willingness to identify alternative development concepts and production strategies, and the ability to adapt flexibly to changes in the economic environment. Management that falls short in this regard is bound to leave some portion of a property's potential value on the table.

4. INCORPORATING THE VALUE OF REAL OPTIONS

The discounted cash flow (DCF) technique is versatile, but not without limitations. To project and properly discount future cash flows requires a forecast of petroleum prices, some disentangling of myriad risk factors that impinge on individual components of the cash flow stream, and a correct view as to when each step in the enterprise will transpire. If prices were stable, these requirements would be less of a burden. For the petroleum industry, however, and particularly since the rise of OPEC in the 1970s, the degree of guesswork and resulting scope for error can be painfully high.

To alleviate these problems, advances have exploited the options approach, a technique developed in the 1970s as an alternative to the DCF

method. The options approach is a simple but brilliant innovation that was devised initially to value certain types of purely financial assets (e.g., stock market puts and calls). When extended to the problem of valuing tangible investments (bricks and mortar, steel and concrete), the technique is referred to as the real options approach. The profound importance and broad impact of these advances in valuation methodology were quickly recognized, as reflected by the award of the Nobel Memorial Prize in Economics in 1997.

In some situations, the real options approach circumvents all three of the difficulties noted previously: it simplifies the problem of adjusting for risk, provides a suitable forecast of future prices, and dispenses with any rigid or preconceived timeline for project activities. The last aspect is especially critical and gives the method its name. As noted in the preceding section, a portion of the value of any property is dependent on managerial flexibility in the design and timing of project components. The options approach assumes not that management will precommit to a fixed and rigid schedule of drilling and production, but that instead management will react rationally to future events as the project unfolds. Pertinent decisions can be taken at numerous points in the execution of any project, and the essence of the options approach is to recognize that management will make those decisions when the time comes using the information then on hand—and not before.

A simple example gives the flavor of this approach. Consider an owner who holds a 2-year lease on an *undeveloped* oil field, one that has been appraised as holding 100 million barrels of recoverable oil, but the required wells and production facilities have not yet been installed. Suppose installation of productive capacity is estimated to cost \$5.50 per barrel of reserves. The owner of the lease then holds a development option: by incurring an expenditure of \$550 million, the owner will acquire 100 million barrels of developed reserves. If the value of a developed reserve (*in situ*) is expected to exceed \$5.50 per barrel, the expenditure seems justified; otherwise it does not.

As we have seen (see Fig. 3), there is an active market in developed reserves. Suppose those transactions reveal the value of developed reserves to be, say, only \$5 per barrel. Moreover, suppose the historical volatility seen in that market indicates a 50% chance that the value of developed reserves will rise or fall by \$1 each year (a random walk). Thus, the value of developed reserves is expected to remain at \$5 in the

future, albeit with unpredictable variations around that level.

It might appear that development of the reserves in question would be inadvisable, and that a 2-year lease on the property would therefore have no value. It is certain, at least, that immediate development of the reserves would effect a \$50 million loss (\$500 million minus \$550 million), and that prospects for development are not expected to improve on average for the duration of the lease. To conclude from these facts, however, that the lease has no value is to ignore the value of active management. If provisions of the lease do not compel management to develop the reserves, then the lease has considerable option value and should command a relatively high price despite the sometimes unfavorable environment. Indeed, the fair market value of the property would amount to \$31 million.

This result is obtained by a straightforward application of the options approach, as diagrammed in Fig. 5. The figure shows a binomial tree that charts the possible future values of developed reserves. Starting from \$5 per barrel, the value would either rise or fall (with equal probability) to \$6 or \$4 after the first year. Continued volatility could carry it to either \$3 or \$7 by the end of the second year, but the most likely value would be \$5 (because there are two price paths leading to that level). Each box in the tree represents a decision node: a point at which management must decide on the basis of available

information whether or not to exercise the option to develop the reserves.

At the end of the second year, as the lease is about to expire, that decision is straightforward. If the value has reached \$7, then development generates an immediate profit of \$1.50 per barrel. The only alternative is to allow the lease to expire, which would generate no profit. If the value were lower (\$5 or \$3), however, then it would be better to hold off—which guarantees zero profit but avoids a loss. Thus, if the reserves had not already been developed, it is clear how management should proceed when the lease is about to expire. In the diagram, the boldface entry at each decision node reflects the optimal choice of action, either **X** for exercise or **H** for holding off. The number shown beside each symbol represents the profit that would be earned via that course of action.

Knowing the end point is critical because it allows us, by working backward, to evaluate the property at each of the earlier decision nodes as well. Suppose, for example, that we find at the end of the first year the value of developed reserves has risen to \$6. Immediate development would generate a profit of \$0.50 per barrel. The alternative is to hold off, to wait and see what the second year may bring. But we have already evaluated the two possibilities: reserve value will either rise to \$7 (which allows a profit of \$1.50) or fall to \$5 (where profits are \$0). The two possibilities are equally likely, so the expected value is simply the average, \$0.75. Even after discounting at 10% to compensate for the extra year's delay, the present value of holding off at the end of year 1 is \$0.68 ($= \$0.75/1.10$), which exceeds the profit from immediate development. Thus, at the decision node in question, although the option to develop immediately is said to be "in the money," it should not be exercised. This illustrates a more general principle that is not so easy to accommodate within the DCF framework: delay may be advisable even when immediate action appears profitable.

By the same routine, and always working right to left, the other nodes can be completed. Of particular interest is the first node, which represents the property's current valuation based on all available information. While it was evident before we began that immediate development would bring a loss of \$0.50 per barrel, we now see that price volatility (and management's capacity to react appropriately to future price changes) adds value to the property. The value of \$0.31 per barrel that we now assign is the present value of the average of the two outcomes that will be realized by the end of the

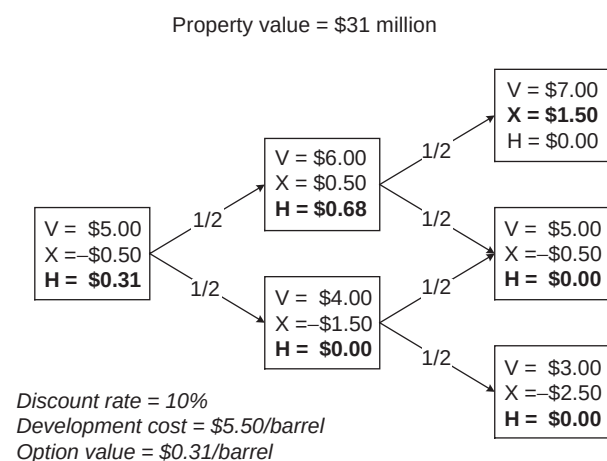


FIGURE 5 Oil field development option. The value of a 2-year field-development lease is calculated by working backward, from right to left, through the binomial tree. The owner must decide, at each node, whether it is more profitable to develop the reserves immediately or to hold off. In this example, the value of developed reserves is assumed to follow a random walk, starting from the level of \$55 per barrel.

first year (either \$0.68 or \$0.00), each discounted at the rate of 10% to allow for one year's delay: $\$0.31 = \frac{1}{2}(\$0.68 + \$0.00)/1.10$.

The options approach provides answers to some additional questions that would be difficult to address via the DCF method. Specifically, the relationship between the length of lease (term to expiration of the development option), the degree of price volatility, and property value is developed explicitly. Extending the lease term (i.e., adding nodes) and/or increasing the degree of future price volatility (i.e., spreading the tree) has the effect of increasing the upside potential of a given property and can only increase (never decrease) its value. It is a simple matter to reconstruct and recompute the binomial tree under varied assumptions and thereby chart the impact of these parameters.

The method could have been illustrated just as well using a trinomial tree, for which the price movement at each node is either up, constant, or down. In practice, the time step between nodes is taken to be relatively small, in which case the final results are invariant to the particular structure of branching that is adopted. Regarding the discount rate and volatility parameters that are required to value the development option, it has been shown that if the analyst follows a certain formulation to measure the volatility (range and probability of future price movements) from historical market prices, then it is appropriate to use the risk-free rate of discount. This aspect of the options approach frees the analyst from the need to separately figure risk adjustment factors for each component of the cash flow stream using the capital asset pricing model, and from the necessity of preparing a subjective forecast of future prices. Many variations on the basic option framework are possible, including, for example, applications to unexplored properties, already-producing properties, and special formulations that are designed to capture random walk, mean-reverting, and other types of price fluctuations in the underlying resource.

5. PORTFOLIO ANALYSIS: THE COMBINED VALUE OF MULTIPLE PROPERTIES

We have so far examined the problem of valuing a single petroleum property—one oil field considered in isolation of other similar properties that might be included in a transaction or already held by a potential buyer. Valuing a collection, or portfolio,

of such properties raises additional issues, some of which push to the very limits of modern techniques. It is useful to distinguish three cases, based on the extent to which the properties are related and whether or not they can be exploited sequentially.

5.1 Properties with Independent Values

If the values of the separate properties are believed to be statistically independent, then the single-property methods described previously can be applied directly and not much else need be said. The value of the whole portfolio would equal the sum of the parts. To satisfy the independence criterion, however, the outcome (i.e., net cash flow) of each property must be uncorrelated with the others. If there are common economic risk factors (e.g., price and cost levels) on which all the properties depend, their values are unlikely to be independent. Common geological factors could also create dependence, as when an exploratory failure on one property is deemed to decrease the probability of success on others.

5.2 Properties with Dependent Values

This case seems more complex, and therefore potentially more interesting. However, dependence among properties does not by itself necessarily require any revision to the valuation method. The whole will still be equal to the sum of the parts, at least if the properties will be exploited simultaneously. By this, we mean that knowledge of the outcome from any one property is not available in time to alter management's plan for exploiting the others. Thus, the properties are operated as if their outcomes are independent, even if an underlying correlation does exist.

5.3 Dependent Properties Exploited Sequentially

If outcomes are dependent and it is possible to exploit the properties sequentially, then the valuation problem changes fundamentally. Management will seek to use whatever information is gleaned from earlier outcomes to enhance subsequent decisions. Statistical dependence implies the existence of relevant information spillovers that facilitate this practice. Thus, it is possible for the value of the portfolio to exceed the sum of the individual parts—the value of acquired information making up the difference. Application of the techniques discussed

previously therefore provides only a lower bound for the combined value.

Valuation models that incorporate information spillovers can become enormously complex as the number of properties increases. The central issue can be illustrated, however, by a simple example. Consider an oil producer who owns two properties, neither of which has been drilled. Let the cost of each exploratory well be \$2 million and the value of each underground deposit (if confirmed by exploratory success) be \$10 million. Finally, we assume that dry hole risk is 40% for each property. What is the value of this portfolio and of its two components?

Based on individual analysis, the value of each component would appear to be \$4 million ($= 0.6 \times \$10 - \$2$). By taking the sum of its parts, the value of the portfolio would then be appraised at \$8 million. If the drilling outcomes of the two properties are independent, or if they are dependent but drilled simultaneously, this is a correct analysis and we are finished. But, suppose the drilling results are highly dependent (perhaps both geological prospects can be traced to a common sedimentary source), such that the outcome of the first foretells the outcome of the second. If the first confirms a deposit, of which the probability is 60%, the other would also be drilled and the second deposit confirmed too, giving a combined value of \$16 million ($= \$10 - \$2 + \$10 - \$2$). On the other hand, if the first well fails, of which the probability is 40%, the second would not be drilled, limiting the combined loss to \$2 million.

By recognizing and exploiting the information spillover, management will on average make \$8.8 million ($= 0.6 \times \$16 - 0.4 \times \$2$), which exceeds by 10% the combined value of the individual properties. To achieve this result, however, the properties must be exploited sequentially, rather than simultaneously. Sequential investment creates an option, while dependence creates the information spillover that gives value to the option. Although our example is very simple, the phenomenon it describes is quite general. Similar results are obtained whether the dependence among properties is complete or partial, symmetric or asymmetric, positive or negative. In all such cases, the sum of property values reckoned individually provides only a minimum valuation for the combined portfolio.

6. CONCLUSION

Uncertainties that range from measurable price volatility to seemingly imponderable geological

gambles constantly beset the petroleum business, and such factors will always challenge the accuracy of even the most advanced valuation methods. Although margins of error are inherently large, it is not too much to ask that petroleum property valuations be correct at least on average. Analytical methods have made some marked progress toward that goal, but the remaining obstacles are not inconsequential. We can reasonably expect the quest for improved valuations of petroleum properties to sustain basic research into some of the most fundamental methods of financial economics well into the future.

Acknowledgments

I sincerely thank G. Campbell Watkins and an anonymous referee for reading an earlier draft of this article and for their many helpful suggestions.

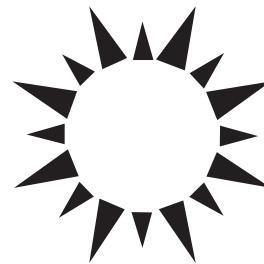
SEE ALSO THE FOLLOWING ARTICLES

Depletion and Valuation of Energy Resources • Discount Rates and Energy Efficiency Gap • Energy Futures and Options • Investment in Fossil Fuels Industries • Markets for Petroleum • Oil and Natural Gas: Economics of Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Value Theory and Energy

Further Reading

- Adelman, M. A. (1990). Mineral depletion, with special reference to petroleum. *Rev. Econ. Stat.* 72(1), 1–10.
- Adelman, M. A., and Watkins, G. C. (1997). The value of United States oil and gas reserves: Estimation and application. *Adv. Econ. Energy Res.* 10, 131–184.
- Dixit, A. K., and Pindyck, R. S. (1995). The options approach to capital investment. *Harvard Business Rev.* 73(3), 105–118.
- Jacoby, H. D., and Laughton, D. G. (1992). Project evaluation: A practical asset pricing method. *Energy J.* 13(2), 19–47.
- Lerche, I., and MacKay, J. A. (1999). “Economic Risk in Hydrocarbon Exploration.” Academic Press, San Diego, CA.
- McCray, A. W. (1975). “Petroleum Evaluations and Economic Decisions.” Prentice-Hall, Englewood Cliffs, NJ.
- Paddock, J. L., Siegel, D. R., and Smith, J. L. (1988). Option valuation of claims on real assets: The case of offshore petroleum leases. *Quarterly J. Economics* 103(3), 479–508.
- Pickles, E., and Smith, J. L. (1993). Petroleum property valuation: A binomial lattice implementation of option pricing theory. *Energy J.* 14(2), 1–26.
- Pindyck, R. S. (1999). The long-run evolution of energy prices. *Energy J.* 20(2), 1–27.

Petroleum System: Nature's Distribution System for Oil and Gas



LESLIE B. MAGOON

U.S. Geological Survey

Menlo Park, California, United States

-
1. Importance of the Petroleum System
 2. Essential Elements and Processes of the Petroleum System
 3. Petroleum System Folio Sheet
 4. The Petroleum System as a Working Hypothesis

Glossary

active source rock A source rock that generates petroleum, either biogenically or thermally. If a source rock was active in the past, it is either inactive or spent in the present.

burial history chart A burial history curve or geohistory diagram constructed to show the time over which hydrocarbon generation occurs. Depicts the essential elements and the critical moment for the system.

critical moment The time that best depicts the generation-migration-accumulation of hydrocarbons in a petroleum system. A map and cross section drawn at the critical moment best shows the geographic and stratigraphic extent of the system.

essential elements The source rock, reservoir rock, seal rock, and the overburden rock of a petroleum system. Together with the processes, essential elements control the distribution of petroleum in the lithosphere.

events chart A chart for a petroleum system showing when the essential elements and processes took place as well as the preservation time and critical moment of the system.

generation-migration-accumulation One petroleum system process that includes the generation and movement of petroleum from the pod of active source rock to the petroleum show, seep, or accumulation. The time over which this process occurs is the age of the petroleum system.

geographic extent The area over which the petroleum system occurs, defined by a line that circumscribes the

pod of active source rock as well as all the discovered petroleum shows, seeps, and accumulations that originated from that pod. The geographic extent is mapped at the critical moment. Also the **known extent**.

level of certainty The measure of confidence that petroleum from a series of genetically related accumulations originated from a specific pod of active source rock. Three levels used are known (!), hypothetical (.), and speculative (?), depending on the level of geochemical, geophysical and geological evidence.

overburden rock The sedimentary rock above which compresses and consolidates the material below. In a petroleum system, the overburden rock overlies the source rock and contributes to its thermal maturation because of higher temperatures at greater depths. An essential element of the petroleum system.

petroleum A mineral oil occurring in subsurface rocks and at the surface which is a naturally occurring mixture of hydrocarbon and nonhydrocarbon compounds. It may occur in the gaseous, liquid, or solid state depending on the nature of these compounds and the existent conditions of temperature and pressure. Common synonyms are **hydrocarbons** and **oil and gas**.

petroleum system The essential elements and processes as well as all genetically related hydrocarbons that occur in petroleum shows, seeps, and accumulations whose provenance is a single pod of active source rock. Also called **hydrocarbon system** and **oil and gas system**.

petroleum system age The time over which the process of generation-migration-accumulation of hydrocarbons in the system takes place on the events chart.

petroleum system name A compound name that includes the source rock in the pod of active source rock, the reservoir rock containing the largest volume of petroleum, and the level of certainty of a petroleum system; for example, the Mandal-Ekofisk(!) petroleum system.

petroleum system processes The two processes of trap formation, and the generation-migration-accumulation. The preservation, degradation, and destruction of

petroleum is omitted as a process as it generally occurs after a petroleum system is formed (preservation time). Together with the essential elements, the processes control the distribution of petroleum in the lithosphere.

pod of active source rock A contiguous volume of source rock that is generating and expelling petroleum at the critical moment and is the provenance for a series of genetically related petroleum shows, seeps, and accumulations in a petroleum system. A pod of mature source rock may be active, inactive, or spent.

preservation time The time after generation-migration-accumulation of petroleum takes place and encompasses any changes to the petroleum accumulations up to the present day.

reservoir rock A subsurface volume of rock that has sufficient porosity and permeability to permit the migration and accumulation of petroleum under adequate trap conditions. An essential element of the petroleum system.

seal rock A shale or other impervious rock that acts as a barrier to the passage of petroleum migrating in the subsurface; it overlies the reservoir rock to form a trap or conduit. Also known as **roof rock** and **cap rock** an essential element of the petroleum system.

source rock A rock unit containing sufficient organic matter of suitable chemical composition to biogenically or thermally generate and expel petroleum. An essential element of the petroleum system.

stratigraphic extent The span of lithologic units that encompasses the essential elements within the geographic extent of a petroleum system. It can be displayed on the burial history chart and cross section drawn at the critical moment.

trap Any geometric arrangement of rock regardless of origin that permits significant accumulation of oil, or gas, or both in the subsurface and includes a reservoir rock and an overlying or updip seal rock. Types are stratigraphic, structural, and combination traps. Trap formation is one of the petroleum system processes.

A petroleum system encompasses a pod of active source rock and all related oil and gas, and it includes all the geologic elements and processes that are essential if a hydrocarbon accumulation is to exist. Petroleum here includes (1) high concentrations of any of the following substances: thermal and biogenic gas found in conventional reservoirs as well as in gas hydrate, tight reservoirs, fractured shale, and coal; and (2) condensate, crude oil, and asphalt found in nature. System describes the interdependent elements and processes that form the functional unit that creates hydrocarbon accumulations. The essential elements include a petroleum source rock, reservoir rock, seal rock, and overburden rock; whereas the two processes are trap formation, and the generation, migration

(primary and secondary), and accumulation of petroleum. These essential elements and processes must occur in time and space so that organic matter included in a source rock can be converted into a petroleum accumulation. As used in this article, petroleum, hydrocarbons, and oil and gas are synonyms and are included under the term petroleum. Petroleum originally referred to crude oil, but was later broadened to include all naturally occurring hydrocarbons, whether gaseous, liquid, or solid. Geochemically, hydrocarbon compounds are those that contain only hydrogen and carbon, such as aromatic hydrocarbons and saturated hydrocarbons. Hydrocarbon compounds are in contrast to nonhydrocarbon compounds, or those that contain nitrogen, sulfur, and oxygen (N, S, O). Hydrocarbon and nonhydrocarbon compounds are both found in crude oil and natural gas, but hydrocarbon compounds usually predominate. Over the past 20 years, whenever the term hydrocarbons has been used without modifiers it is meant to be synonymous with petroleum. When oil and gas are used together, it collectively refers to crude oil and natural gas in any proportion. Condensate is a gas phase in the accumulation and in a liquid phase at the surface; either way it is considered petroleum as is solid petroleum (i.e. natural bitumen, natural asphalt, bituminous sands, etc.).

1. IMPORTANCE OF THE PETROLEUM SYSTEM

Petroleum or oil and gas accumulations mostly occur in sedimentary rocks and less frequently in igneous and metamorphic rocks. These seemingly contradictory occurrences are easily explained when petroleum is viewed in the context of a subsurface fluid distribution network whose origin is an active source rock and whose destination is a trap or seep. An active source rock is sufficiently organic-rich that when heated high enough over sufficient time injects petroleum into the adjacent country rock. This largely thermal process of primary migration or expulsion gives way to the physical process of secondary migration outside the active source rock because of buoyancy and diffusion. Depending on the extent of lateral permeability of the country rock, the expelled petroleum will either stay near the site of expulsion or migrate up dip to the nearest trap. If the nearest trap is filled to the spill point, the spilled petroleum will continue to migrate up dip to the next trap and so on until migration ceases or it seeps to the earth's surface.

The extent, connectivity, and type of the conduits and traps radiating up dip from the active source rock determine the nature and area over which these hydrocarbon fluids migrate. This hydrocarbon distribution network is an integral part of the petroleum system; it can be mapped in space and time using geochemical, geological, and geophysical information.

Even though the petroleum system was introduced in 1972, the concept was largely overlooked until the early 1990s. In the past decade, many petroleum systems have been studied and results published in journals, magazines, and presented at scientific meetings. Petroleum geologists now recognize that the sedimentary basin provides the rock framework in which a hydrocarbon fluid distribution network resides (Fig. 1). It is this fluid system that petroleum geologists need to understand if their exploration of the play and prospect are to be successful.

2. ESSENTIAL ELEMENTS AND PROCESSES OF THE PETROLEUM SYSTEM

2.1 Source Rock

A source rock contains sufficient organic matter of suitable chemical composition to biogenically or thermally generate and expel petroleum. Only one active source rock interval occurs in each petroleum system. The two important elements in organic matter are carbon and hydrogen; the carbon provides the skeleton on which the hydrogen is attached. The oxidation of hydrogen provides the energy so the more hydrogen per unit of carbon, the better quality the source rock.

The higher the concentration of carbon in organic matter, the richer is the source rock. Carbon is

measured in weight-percent of rock matrix. A source rock needs at least 2 wt % of organic carbon before it can expel petroleum. Organic carbon content (TOC) for a petroleum source rock commonly exceeds 3 wt % and can get as high as 10 wt % or more.

Hydrogen content is measured in hydrocarbons per unit of organic carbon (mg HC/g TOC) and is a measure of source rock quality, the higher the hydrogen index (HI), the better the quality. In addition, the HI indicates the type of hydrocarbon fluid expelled. For example, HI greater than 300 (mg HC/g TOC) and less than 900 expel mostly oil with some gas, HI between 200 and 300 expel some light oil but mostly gas, and HI between 50 and 200 is just natural gas. As a source rock thermally matures and expels petroleum, the HI is reduced to below 50 so the difference between the immature HI and HI of the depleted source rock is a measure of the amount of hydrogen or petroleum expelled. For example, if an immature source rock has an HI of 500 and the HI of the spent source rock is 100, the 400 mg HC/g TOC of oil and gas was expelled as compared to an immature HI of 200 and the spent HI is 100, the amount of expelled gas is 100 mg HC/g TOC.

The presence of porphyrins from plant chlorophyll as well as other complex molecules that are the building blocks of single and multicellular plants and animals are proof that the organic matter in source rocks originates from the biosphere. These so-called biomarkers indicate the origin and sometimes the age of the source rock and the origin of the petroleum.

2.2 Overburden Rock

Overburden rock is the sedimentary rock above which compresses and consolidates the material below. In a petroleum system, the overburden rock overlies the source rock and contributes to its thermal maturation because of higher temperatures at greater depths. The age and thickness of the overburden rock determines the burial rate and thereby influences the heating rate. Frequently, the reservoir and seal rocks immediately overlie the source rock and are thus included within the overburden rock. The close proximity of the source and reservoir rocks allow the expelled petroleum immediate access to a migration conduit. The deposition and erosion of the overburden rock through time determines the slope and form of the burial history curve of the source rock through petroleum expulsion. Rapid burial with little erosion is best for the preservation of the petroleum system.

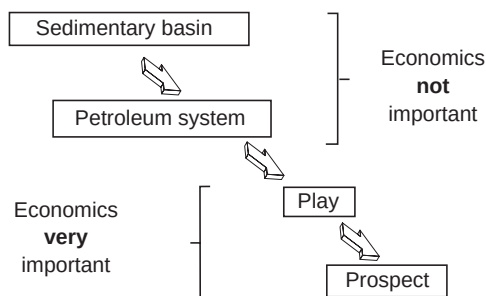


FIGURE 1 Four levels of petroleum investigation. From Magoon, L. B., and Dow, W. G. (1994). The petroleum system—From source to trap. Memoir 60, p. 4 (as Figure 1.1). AAPG © 1994. Reprinted by permission of the AAPG whose permission is required for further use.

2.3 Reservoir Rock

A reservoir rock is a subsurface volume of rock that has sufficient porosity and permeability to permit the migration and accumulation of petroleum under adequate trap conditions. Porosity is a measure in percentage of pore volume or size of holes or vugs per unit volume of rock. For example, a well-sorted sand in a 300 ml container will hold about 100 ml of water in its pore space, or a porosity of 33%. If petroleum is present, it can also occupy this pore space. During burial of this sand, compaction reduces this porosity substantially to where only a small percentage of porosity is left.

Permeability is a measure of the connectivity of pores in the subsurface. The sand in the container has narrow pore throats between the large pores that allows fluid to pass from one pore to another. Permeability is measured in millidarcies (md) or Darcy (1000 md) of these narrow throats. Commonly, permeability in the range of 100 to 500 md are reasonable values for a petroleum reservoir rock. Values over a Darcy are exceptional. Fractures have infinite permeability.

Any lithology can have porosity and permeability. Siliciclastic sandstone is most likely to have vertical and lateral porosity and permeability over a relatively large distance. Frequently, carbonate rock has the most complex patterns of porosity, frequently vuggy porosity with little or no permeability, but dolomitized zones can have lateral porosity and permeability. Volcanic rock can be vuggy where hot gases bubble from molten rock. Metamorphic and other sedimentary rocks can acquire secondary porosity because of weathering or groundwater movement. All these rock types can be shattered so that unconnected pores can be connected with fractures.

Major and minor reservoir rocks are determined from the percentage of in-place petroleum that originated from a particular pod of active source rock. If the volume of in-place petroleum is unavailable, the volume of recoverable hydrocarbons are used. All oil and gas fields included in a petroleum system are listed and the original in-place (recoverable) hydrocarbons are determined by stratigraphic interval. The volumes of in-place hydrocarbons for each stratigraphic interval are added up, and the percentage of in-place hydrocarbons by stratigraphic interval is determined. Usually, one stratigraphic interval contains most of the in-place hydrocarbons, so this interval is the major reservoir rock. Reservoir rocks that contain minor or lesser amounts of in-place hydrocarbons are the minor reservoir rocks.

The major reservoir rock is the second name used in the name of the petroleum system. The major reservoir rock indicates the optimum migration path between the pod of active source rock and the traps that include the major reservoir rock. The minor reservoir rock indicates the least effective migration path, or it indicates a migration path that should be studied for overlooked prospects. Major and minor reservoir rocks should be included on the petroleum system events chart.

2.4 Seal Rock

Seal rock is a shale, evaporite, or other impervious rock that acts as a barrier to the passage of petroleum migrating in the sub-surface; it overlies the reservoir rock to form a trap or conduit. The seal rock is also known as roof rock or cap rock. A dipping reservoir rock overlain by a seal rock provides a migration path for petroleum.

2.5 Trap Formation

Any geometric arrangement of rock regardless of origin that permits significant accumulation of oil or gas, or both, in the subsurface and includes a reservoir rock and an overlying or updip seal rock is a trap. Types are stratigraphic, structural, and combination traps. The formation of traps by deposition, tension or compression is one of the processes needed to create a petroleum system. Trap formation must occur before or during petroleum migration in order to have an accumulation.

2.6 Generation-Migration-Accumulation

One petroleum system process that includes the generation and movement of petroleum from the pod of active source rock to the petroleum show, seep, or accumulation. The time over which this process occurs is the age of the petroleum system. Even though each increment of this process can be studied separately, it is the petroleum in the accumulation, seep, or show in a well that is proof that a petroleum system exists.

3. PETROLEUM SYSTEM FOLIO SHEET

The petroleum-system folio sheet is a graphical way to depict the geographic, stratigraphic, and temporal evolution of the petroleum system (Fig. 2; Table I).

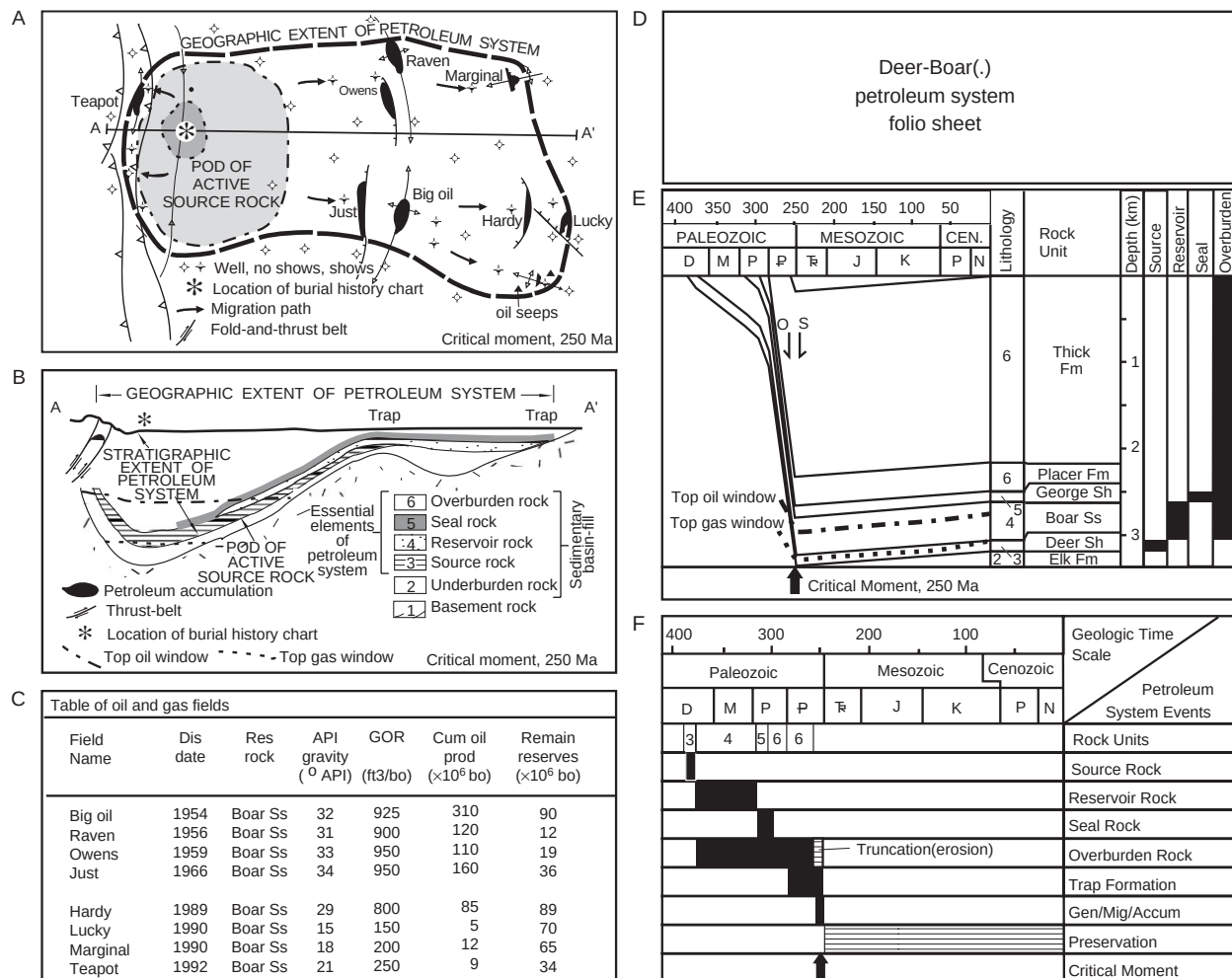


FIGURE 2 (A) Map showing the geographic extent of the so-called Deer-Boar(.) petroleum system at the critical moment (CM, 250 Ma). Thermally immature source rock is outside the oil window. The pod of active source rock lies within the oil and gas windows. (B) Cross section showing the stratigraphic extent of the so-called Deer-Boar(.) petroleum system at the critical moment (250 Ma). Thermally immature source rock lies up dip of the oil window. The pod of active source rock is down dip of the oil window. (C) Table of oil and gas fields in the Deer-Boar(.) petroleum system, or the accumulations related to one pod of active source rock. (D) The name of the petroleum-fluid system. (E) Burial history chart shows the critical moment (250 Ma) and the time of oil generation (260-240 Ma) for the so-called Deer-Boar(.) petroleum system. This information is used on the events chart. All rock unit names used here are fictitious. (F) Events chart shows the relationship between the essential elements and processes as well as the preservation time and critical moment for the so-called Deer-Boar(.) petroleum system. (G) The text and other figures needed to describe the petroleum system. From Magoon, L. B., and Dow, W. G. (2000). Mapping the petroleum system—An investigative technique to explore the hydrocarbon fluid system. Memoir 73, p. 55 (as Figure 1). AAPG © 2000. Reprinted by permission of the AAPG whose permission is required for further use.

The folio sheet includes (1) the petroleum system map (Fig. 2A), (2) the petroleum system cross section (Fig. 2B), (3) a table of genetically related accumulations (Fig. 2C), (4) the petroleum system name (Fig. 2D), (5) a burial history chart located over the pod of active source rock (Fig. 2E), and (6) an events chart to summarize the history of the petroleum system (Fig. 2F). In the ideal case, this folio sheet summarizes the detailed work of many specialists

and provides the supportive information for the petroleum system map needed to evaluate the play and prospect.

3.1 Petroleum System Map

The petroleum system map (Fig. 2A; Table I) shows (1) the geographic extent of the petroleum system; (2) the pod of active source rock; (3) the genetically

TABLE I

Components of a Complete Petroleum System Study

Figure or table	Information required	Purpose
Petroleum system map	Locate petroleum fields included in system; Indicate whether oil or gas; Indicate surface and subsurface oil or gas shows included in system; Indicate direction of petroleum migration; and Indicate distribution of pod of active source rock.	<i>Geographic extent of a petroleum system</i> at critical moment shown by circumscribing the pod of active source rock and the outer limit of migrated hydrocarbons; <i>Source rock name</i> from pod of active source rock; and <i>Petroleum system burial history chart</i> location.
Petroleum system table	List all oil or gas fields by petroleum system; Indicate discovered reserves and in-place petroleum by stratigraphic unit; trap type; reservoir rock name, age, and lithology; and seal rock name, age, and lithology; For oil field indicate GOR, API gravity, sulfur content, and Pr/Ph ratio; and For gas field indicate GOR, $\delta^{13}\text{C}$, and gas wetness $[(\text{C}_1/\text{C}_1 - \text{C}_4) \times 100]$.	<i>In-place petroleum</i> for mass balance calculation; <i>Reservoir rock name</i> from that reservoir rock with the highest percentage of in-place petroleum; <i>Seal rock</i> most commonly occurring in trap; and <i>Oil or gas province</i> from average GOR.
Petroleum system cross section	Structural and stratigraphic information such as deformation style and rock units; Indicate oil window and gas window; Indicate petroleum shows and accumulations; Draw at critical moment; and Indicate direction and conduits for petroleum migration.	<i>Stratigraphic extent of petroleum system</i> at the critical moment by identifying the base of the pod of active source rock or base of hydrocarbon column, whichever is deeper; <i>Geographic extent of petroleum system</i> ; <i>Pod of active source rock</i> shown; <i>Overburden rock</i> shown; and <i>Petroleum system burial history chart</i> location.
Petroleum system burial history chart	Stratigraphic units penetrated in well; Time rock-units were deposited; Thickness of rock units; Names of rock units; Lithology of rock units; Present day thermal maturity profile; Present day geothermal gradient; Computer program to determine time and depth for oil window and gas window; and Indicate essential elements of petroleum system.	<i>Petroleum system events chart</i> information determined from chart, such as onset and end (at uplift) of petroleum generation, and critical moment; <i>Essential elements of petroleum system</i> shown; <i>Oil window</i> depth for cross section at critical moment; and <i>Gas window</i> depth for cross section at critical moment.
Petroleum system events chart	Age of essential elements of petroleum system; Onset and end of trap formation; Onset and end of petroleum generation, migration, accumulation; Preservation time of petroleum system; and Critical moment.	<i>Petroleum system events chart</i> summarizes in one diagram the essential elements and processes of the system, as well as the preservation time and critical moment.
Petroleum-petroleum correlation	Geochemical evidence, such as bulk properties, biological markers, and isotopic data to show that more than one petroleum accumulation came from the same source rock (but not necessarily the same pod of active source rock).	<i>Geographic and stratigraphic extent of a petroleum system</i> is established with this geochemical correlation in concert with the structure and stratigraphy of the pod of active source rock and the adjacent traps.
Petroleum-source rock correlation	Geochemical evidence, such as biological marker and isotopic data, to indicate a certain petroleum originated from a specific source rock.	<i>Level of certainty</i> is established using geological and geochemical evidence and indicates the confidence that a specific source rock expelled a given petroleum.
Mass balance calculation	TOC and Rock-Eval pyrolysis; Source rock density; and Volume of pod of active source rock.	<i>Mass of petroleum generated</i> to determine petroleum system generation-accumulation efficiency (GAE).

Source. From Magoon, L. B., and Valin, Z. C. (1994). Overview of petroleum system case studies. Memoir 60, p. 332 (as Table 20.2). AAPG © 1994. Reprinted by permission of the AAPG whose permission is required for further use.

related oil and gas accumulations, shows, and seeps; (4) the location of the burial history chart; and (5) the location of the petroleum-system cross section. Usually this map is drawn at the present day, but may be refined later to depict the critical moment or that time when most of the hydrocarbons were generated and accumulating, especially if the petroleum system is old. This map is in contrast to a geologic map which depicts rock units and geometry at the surface or an oil and gas field map.

3.1.1 Accumulations, Shows, and Seeps

The genetically related oil and gas fields or petroleum accumulations, by inference, originated from the pod of active source rock. So start with an oil and gas field map of the petroleum province. The accumulations shown on an oil and gas field map need to be grouped into one or more possible petroleum systems based on their geographic and stratigraphic locations, and the bulk and geochemical properties of the fluids in each accumulation (Fig. 2A). For example, the accumulations in the fictitious Deer-Boar(.) petroleum system include Teapot in the thrust belt; Raven, Owens, Just, and Big Oil in the anticlines adjacent to the foreland basin; and Marginal, Hardy, and Lucky, the most distal accumulations. These widely spaced accumulations are in the same stratigraphic interval, the Boar Sandstone, and all have similar geochemical properties, as known from wells that sample these accumulations.

The exploratory wells indicate that some of the wells have oil shows in the Boar Sandstone. One surface seep occurs in the southeast portion of the map where the Boar Sandstone crops out. These oil occurrences have geochemical similarities with the oil accumulations. Based on this information, these oil occurrences are judged to belong to the same petroleum system.

3.1.2 Pod of Active Source Rock

The pod of active source rock is a contiguous volume of source rock that generated and expelled petroleum at the critical moment and is the provenance for a series of genetically related petroleum shows, seeps, and accumulations in a petroleum system. A pod of mature source rock may be active, inactive, or spent. There is only one pod of active source rock for each petroleum system.

The kerogen type for the thermally mature source rock has been shown by numerous investigators to control the type and volume of petroleum expelled. Other investigators have provided explanations of the tools and techniques to characterize and map the

pod of active source rock. The pod of active source rock (also referred to as the kitchen or oil and gas window) is a required feature of the petroleum system map because of its genetic control on petroleum accumulations.

For the fictitious Deer-Boar(.), the pod of active source rock is in the western part of the map area and is mapped using a dashed line (long and short dashes, Figs. 2A and 2B). The Deer Shale is considered the most likely source rock because it is geographically near and stratigraphically below the Boar Sandstone, the reservoir rock for the accumulations, shows, and seep. Thermal maturity data for the Deer Shale indicate it is thermally mature in the foreland basin, but immature in the thrust belt toward the west and on the craton toward the east. The dashed lines correspond to the thermal maturity contours for the top of the oil window (long and short dashes) and the top of the gas window (short dashes) (Figs. 2A and 2B).

3.1.3 Geographic Extent

The geographic extent of the petroleum system at the critical moment is described by a line that circumscribes the pod of active source rock and includes all the discovered petroleum shows, seeps, and accumulations that originated from that pod. The critical moment is a snapshot of the petroleum system at the time when most of the hydrocarbons were generated and migrating. A map of the Deer-Boar(.) petroleum system, drawn at the end of the Paleozoic time (250 Ma), includes a line that circumscribes the pod of active source rock and all related, discovered hydrocarbons. This area represents the geographic extent or known extent of the petroleum system (Fig. 2A).

To draw this petroleum system map, we assume that all the petroleum occurrences emanate from the same pod of active source rock. The pod of active source rock is shown as a shaded area, and the depth to the source rock is assumed to be close to the thickness of overburden rock which is contoured in thousands of meters. Based on the locations of the accumulations, drill-stem tests, and exploratory wells with and without oil shows the geographic extent of the petroleum system is drawn. Using the guidelines discussed earlier, the location of the cross section and the burial chart are indicated. This petroleum system map emphasizes the short migration path of the discovered oil fields and that the long migration path to the northwest lacks an oil field. This map strongly suggests that the most likely place to find undiscovered oil is above or close to the pod of active source rock and along the preferential

migration paths as expressed by the geographic extent of the petroleum system.

3.1.4 Location of Cross Section

The cross section is placed on the map so that it passes through the largest accumulations, thickest overburden rock, and extends over the entire geographic extent of the petroleum system (Figure 2B). If possible, start with an available present-day cross section unless the petroleum system is so old or structurally altered that a restored cross section representing a previous time is required to depict the time when most of the hydrocarbons migrated and accumulated. Here, a cross section at the end of the Paleozoic (250 Ma) was used (Fig. 2B). The largest accumulations are included because they are usually located on the simplest, most efficient migration path from the pod of active source rock. A transect through the thickest overburden rock shows the most likely area where the source rock is thermally mature and, therefore, the provenance of the hydrocarbons. The cross section should transect the entire the petroleum system so that the basis for the geographic extent can be demonstrated.

3.1.5 Location of Burial History Chart

The location of the burial history chart is along the cross section line within the pod of active source rock (Fig. 2B, E). At this location, the source rock must be thermally mature (active or spent) otherwise petroleum would be absent in the conduits or migration paths. The reconstruction of the burial history provides the basis for the times of the onset (O) of generation-migration-accumulation, the depletion (D) of the source rock, and the critical moment (CM).

3.2 Table of Fields

The table showing all the oil and gas accumulations included in the folio sheet provides important information about the petroleum system (Fig. 2C; Table I). First, the discovery dates and sizes of the fields are useful for field-size distributions and discovery-rate modeling. Second, the complexity of the hydrocarbon plumbing system is suggested by the number of reservoir rocks. One reservoir rock for all fields indicates a simple plumbing system, whereas many reservoir rocks indicates a more complicated system. Third, the size of the petroleum system and the generation and expulsion efficiency can be determined by using the total volume of recoverable oil and gas for all fields. Finally, the reservoir rock with the highest percentage of oil or gas reserves is to

be used in the petroleum system name. For example, if all the oil is in the Boar Sandstone, it is included in the name (Fig. 2D).

For example, the Deer-Boar(.) is a 1.2 billion barrel petroleum system with a simple plumbing system. This size designation using recoverable petroleum is most useful to the explorationist who is interested in (1) comparing the sizes of different petroleum systems to rank or plan an exploration program and (2) comparing the field-sizes in a petroleum system to determine the most likely prospect size and what volumes can be produced. However, the size (volume of recoverable petroleum) of the petroleum system needed for material balance equations is quite different. Here, three additional types of information are estimated: the in-place petroleum for each field, what is left behind along the migration path, and what was lost in surface seeps and exhumed accumulations. These estimates are made by the investigator. This volume of petroleum can then be compared to the estimated volume of petroleum generated in the active source rock.

3.3 Cross Section

The petroleum system cross section (Fig. 2B; Table I), drawn at the critical moment, or time when most of the hydrocarbons were generated, shows the geographic and stratigraphic extent of the petroleum system and how each rock unit functions within the system to distribute the oil and gas. Stratigraphically, the petroleum system includes a petroleum source rock, reservoir rock, seal rock, and overburden rock. This cross section is in contrast to structural or stratigraphic cross sections.

The presence of adequate overburden rock in the correct geometry provides (1) the burial needed to thermally mature the source rock, (2) the up dip vector needed for oil and gas to migrate to shallower depths, (3) the burial depth variations needed to form traps for petroleum accumulations, and (4) the burial depth of accumulations that allow for the preservation or biodegradation of oil. If the history of the petroleum system is to be correctly modeled, the age, thickness, and erosional history of the overburden rock is required. The cross section, drawn to represent the end of the Paleozoic (250 Ma), shows the geometry or structural style of the essential elements at the time of hydrocarbon accumulation, or critical moment, and best depicts the stratigraphic extent of the system (Fig. 2B).

3.4 Petroleum System Name

The name of a petroleum system labels the hydrocarbon-fluid system or distribution network (Fig. 2D) in the same way the name Colorado River designates an aqueous distribution system, the river and its tributaries. The name of the petroleum system includes the geological formation names of the source rock followed by the major reservoir rock (Fig. 2C) and then the symbol expressing the level of certainty. For example, the Deer-Boar(.) is the name of a hydrocarbon fluid system whose source rock, the Deer Shale, most likely generated the petroleum that charged one or more reservoir rocks, which in this case is the Boar Sandstone. It is the major reservoir rock because it contains the highest percentage by volume of hydrocarbons in the petroleum system.

A petroleum system can be identified at three levels of certainty: known, hypothetical, and speculative. At the end of the system's name, the level of certainty is indicated by (!) for known, (.) for hypothetical, and (?) for speculative. The symbol indicates the level of certainty that a particular pod of active source rock has generated the hydrocarbons on the table of accumulations (Fig. 2C). In a known petroleum system, a well-defined geochemical correlation exists between the source rock and the oil accumulations. In a hypothetical petroleum system, geochemical information identifies a source rock, but no geochemical match exists between the source rock and the petroleum accumulation. In a speculative petroleum system, the link between a source rock and petroleum accumulations is postulated entirely on the basis of geological or geophysical evidence.

In certain frontier areas of the world, especially in the offshore, the stratigraphic units are poorly understood and frequently undesignated. Here, the judgment of the investigator is required. The geologist should avoid using ages, such as Jurassic, in the name because it fails to uniquely identify the petroleum system. Other situations arise where it is difficult or confusing to follow the naming rules. For example, when a rock unit that includes both the source and reservoir forms more than one petroleum system, the same name might be used more than once, such as Monterey(!) petroleum systems. Here, a geographic modifier can be used to differentiate the systems. Another naming problem arises in frontier areas where formal rock units are lacking, so only ages or geophysically mapable units are used. A geographic name or the name of an accumulation in the petroleum system may be used. If it is impossible to follow the formal designation of a petroleum

system, the investigator should select a unique name that identifies the fluid system, not the rock units.

3.5 Burial History Chart

The purpose of the burial history chart is to show the essential elements, and three important hydrocarbon events, which are (1) the onset (O) of generation-migration-accumulation, (2) the partially spent or depleted (D) active source rock, and (3) the critical moment (CM) of the petroleum system (Fig. 2E; Table I). The top of the oil and gas windows, and the lithology and name of the rock units involved should also be shown. This chart uses sedimentologic and paleontologic evidence in the overburden rock to reconstruct the burial or thermal history of the source rock. The onset of generation-migration-accumulation usually occurs when the source rock reaches a thermal maturity at a vitrinite reflectance equivalence of 0.6% R_o and ends when the source rock is either uplifted before all the hydrocarbons can be expelled or is depleted as the source rock is more deeply buried. The location of the burial history chart is shown on the petroleum system map and cross section.

In this example, the Deer Shale (rock unit 3) is the source rock, the Boar Sandstone (4) is the reservoir rock, the George Shale (5) is the seal rock, and all the rock units above the Deer Shale (4, 5, 6) comprise the overburden rock. The burial history chart is located where the overburden rock is thickest, and indicates that the onset of generation-migration-accumulation started 260 Ma in Permian time and was at maximum burial depth 255 Ma. Oil generation ended about 240 Ma because the source rock is depleted. The critical moment as judged by the investigator is 250 Ma because modeling indicates most (>50%) of the hydrocarbons have been expelled and are accumulating in their primary traps. However, the investigator would be correct to have chosen anytime between 250 and 240 Ma, but 250 Ma was chosen because the best geologic information is available to reconstruct the map and cross section. The time of generation-migration-accumulation ranges from 260 to 240 Ma and is the age of the petroleum system.

As mentioned in the cross section discussion, knowing the age and thickness for each increment of overburden rock is crucial for any modeling exercise. Each increment needs to be bounded by time lines whose dates are supported by paleontologic dates, isotopic age dates, or other means of dating strata. As the number of increments in the overburden rock

increases, the details of the burial history of the source rock will be better understood. For instance, in the previous example, the overburden rock is Permian, undifferentiated. Suppose, however, that paleontologic dates indicate that 95% of the overburden rock is Early Permian and that the rest is Late Permian. This increase in increments in the overburden rocks will change the time when the underlying source rock becomes thermally mature. In this example, the change in time that the source rock becomes mature might be considered insignificant, but in other examples the difference could be significant.

3.5.1 Location of Burial History Chart

The burial history chart chosen to show the three hydrocarbon events for a petroleum system should be located in the pod of active source rock where, in the judgment of the investigator, much of the oil and gas originated. Usually this location is down-dip from a major migration path to the largest fields.

Petroleum systems are seldom so simple that only one burial history chart adequately describes the same three hydrocarbon events for every location in the pod of active source rock. The investigator chooses the burial history curve that best suits the purpose. If the investigator is presenting (oral or written) a petroleum system investigation, he or she would use the burial history curve down-dip from a major migration path to the largest fields. However, if the subject is a play or prospect, the burial history curve would be down-dip on a suspected migration path to the area of the play or prospect.

3.5.2 Critical Moment

The generation, migration, and accumulation of oil and gas in a petroleum system never starts when the source rock is being deposited and seldom extends into the present day. If a source rock is deposited in the Paleozoic, it may be the Mesozoic before it becomes thermally mature and charges adjacent traps, and by the Cenozoic, this source rock is probably depleted. The time over which the process of generation-migration-accumulation takes place could be tens of millions of years. This is a long period of time to choose from if an investigator needs to select the most appropriate moment during this process to make a map and cross section that shows the petroleum system when most (>50%) of the hydrocarbons were migrating and accumulating. To help the investigator with this important exercise, the critical moment was introduced and incorporated into the petroleum system folio sheet.

Geologists use the concept of the critical moment for other exercises. Whenever a map, such as an isopach map, is constructed, it is frequently reconstructed to its original thickness at the moment of deposition. The kinematic development of a fold-and-thrust belt occurs over many millions of years, but it is frequently represented by one cross section, or a snapshot in time. A structural cross section of a fold-and-thrust belt reconstructed at the end of the Cretaceous, for example, utilizes the critical moment concept.

The critical moment is the time that best depicts the generation-migration-accumulation of hydrocarbons in a petroleum system. This definition needs an explanation and an example to be better understood. A moment is a brief, indefinite interval of time that is of particular importance. For a camera to take a picture, a moment is less than a second. In geology, the further one goes back in time, the interval becomes thousands or even millions of years. For the petroleum system, moment relates to the shortest measurable time. Critical refers to the moment that best shows, in the judgment of the investigator, the process of generation-migration-accumulation.

Best is a keyword in this definition. Best contains the criteria the investigator should use to select the appropriate moment. The best time needs to fulfill several criteria: (1) it must be within the age of the petroleum system; (2) it must be when most, or more than half, of the hydrocarbons are migrating and accumulating; and (3) it must be shown as an arrow, not an interval, on the burial history and events charts.

The critical moment of a petroleum system can vary depending on its purpose. If the purpose is a petroleum system case study, then the critical moment should be representative of the entire system. However, if the purpose is related to an exploration play or prospect, then the critical moment should be for that part of the pod of active source rock most likely to charge the traps in the play or prospect. Depending on the size, thickness, and variation of the thermal maturity of the pod of active source rock and the objective of the investigator, these could be different best times, none of which are incorrect. In fact, the investigator may need to make numerous burial history charts of a large, thick pod of active source rock that has a wide range of thermal maturity to determine which best moment properly depicts generation-migration-accumulation for the intended audience.

The burial history chart omits important information available in most modeling software packages that explains the critical moment from a different perspective (Fig. 3). This graph shows cumulative volumes of generated or expelled oil and gas.

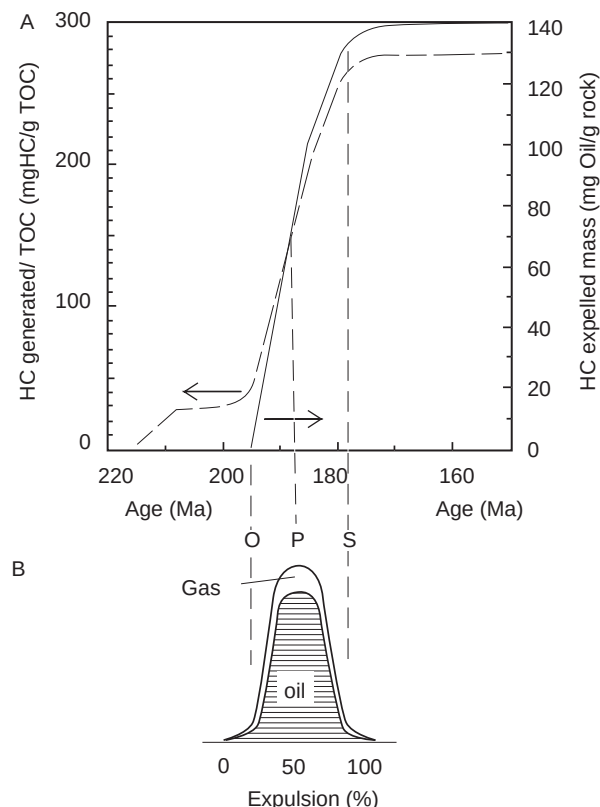


FIGURE 3 (A) Cumulative curves showing the time over which hydrocarbons (HC) are generated and expelled. (B) Distribution curve for oil and gas (HC) expulsion using the same information and showing the onset (O) of expulsion, peak (P) expulsion, and depletion (D) of the source rock. The critical moment (CM) is selected to be any time between P and D. From Magoon, L. B., and Dow, W. G. (2000). Mapping the petroleum system—An investigative technique to explore the hydrocarbon fluid system. Memoir 73, p. 64 (as Figure 7). AAPG © 2000. Reprinted by permission of the AAPG whose permission is required for further use.

Wherever the curves are horizontal, no volume is being added. This graph shows the onset of generation (dashed line) precedes expulsion (solid line) by almost 20 million years. According to this graph, most expulsion occurs more than 15 million years from 195 to 179 Ma. When this cumulative expulsion curve is transformed to a curve showing the distribution of expulsion, it shows peak expulsion at 187.5 Ma (Fig. 3). At this time, at least half of the petroleum is migrating so the critical moment can be selected by the investigator as any time between 187.5 to 179 Ma.

3.6 Events Chart

The petroleum-system events chart shows the temporal relationship of the rock units, essential

elements, processes, preservation, and critical moment for each petroleum system in bar graph form (Fig. 2F; Table I). The events chart concept is flexible and is used as a risk chart to evaluate plays and prospects.

The events chart shows the following items (Fig. 2F). The rock units include those that are within the stratigraphic extent of the petroleum system. On a certainty scale, the ages of the four essential elements (source, reservoir, seal, and overburden rocks) are usually better established from paleontology or radiometric age dates than those associated with the two processes (trap formation and generation-migration-accumulation). Determining the times over which the traps form and hydrocarbons accumulate (generation-migration-accumulation) is more interpretive because there is less precise temporal information about these processes. Therefore, risk or uncertainty with regard to the times over which the two processes takes place is higher or less certain than for the better established times of development of the four essential elements. This certainty relationship is important if a similar chart is constructed for a complementary play/prospect or assessment unit.

When an events chart is constructed for a complementary prospect, it becomes a risk chart. The risk chart is derived from the petroleum-system events chart which, in turn, is derived from the summation of the events chart for each oil and gas field in the petroleum system. These oil or gas fields are successful prospects. Unsuccessful prospects are dry holes. For example, if a risk chart for a prospect is similar to the petroleum-system events chart, then it can be concluded that this prospect is more likely to contain petroleum than one that has a dissimilar risk chart. Conversely, if an events chart is constructed for each dry hole within a petroleum system, they should be dissimilar from producing fields. This dissimilarity indicates where the greater uncertainty lies. Used this way, the events chart is a useful analytical tool to deal with uncertainty or risk.

One important issue this simple bar graph addresses is as follows: For an evolving petroleum system to effectively trap migrating hydrocarbon fluids, the trap forming process must occur before or during the generation-migration-accumulation process in order for petroleum to accumulate.

When constructing an events chart, these rules should be followed. First, there is only one pod of active source rock for each petroleum system. Second, every effective reservoir rock needs a seal, no matter how thin. Third, show only reservoir rocks

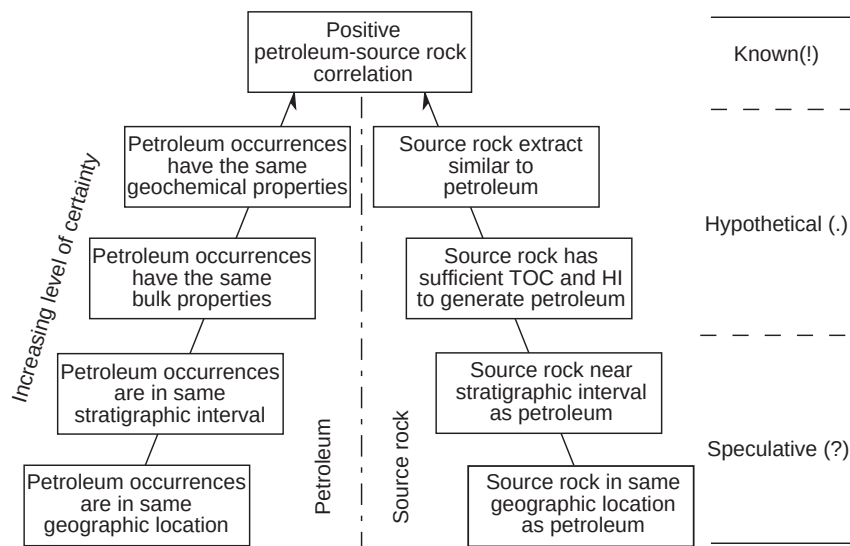


FIGURE 4 The logic sequence used for the levels of certainty from guessing to knowing. From Magoon, L. B., and Dow, W. G. (2000). Mapping the petroleum system—An investigative technique to explore the hydrocarbon fluid system. Memoir 73, p. 56 (as Figure 2). AAPG © 2000. Reprinted by permission of the AAPG whose permission is required for further use.

that contain petroleum accumulations, shows, or seeps. Fourth, show eroded overburden rock with hatcher lines so that it can be incorporated in the modeling exercise. Fifth, the best information for timing of trap formation comes from oil and gas fields. Sixth, the best information for generation-migration-accumulation comes from geological and geochemical information about the source rock that are then incorporated into the burial modeling and kinetics. This information indicates the onset, peak, and end of generation-migration-accumulation or when the active source rock is depleted (spent) or uplifted (inactive source rock). This process takes place over a relatively short period of time. Seventh, preservation time, by definition, starts when generation-migration-accumulation ends and continues to the present. Some petroleum systems have no preservation time. Last, when the critical moment occurs is, as discussed earlier, in the judgment of investigator, but modeling software packages are useful tools as they show the time over which expulsion from the pod of active source rock occurs.

4. THE PETROLEUM SYSTEM AS A WORKING HYPOTHESIS

A petroleum system investigation starts with a working hypothesis for generation, migration, and entrapment of petroleum in a province, based on

available geological and geochemical data, which evolves as more data becomes available (Fig. 2A, B). The investigator starts with an oil and gas field map and related field data for the petroleum province of interest. The geographic location of the accumulations is important because accumulations located close together are more likely to have originated from the same pod of active source rock (Fig. 4). Accumulations that occur in the same or nearly the same stratigraphic interval are also likely to be from the same active source rock. In contrast, accumulations separated by barren rock sections are presumed to have originated from different pods of active source rock. Accumulations of widely differing bulk properties, such as gas versus oil, API oil gravity, gas-to-oil ratios, and sulfur contents, may also be presumed to originate from different pods of active source rock. Detailed geochemical data on oil and gas samples provide the next level of evidence for determining whether a series of hydrocarbon accumulations originated from one or more pods of active source rock. Last, comparing the geochemistry of oils and gases to possible source rocks provides the highest level of certainty as to which active source rock generated which oil or gas type.

By acquiring and organizing information that addresses these issues of location, stratigraphic position, and geochemistry, an investigator can take a working hypothesis of how a particular petroleum system formed to increasing levels of certainty (Fig. 2; Table I). The investigator organizes the

information on the oil and gas accumulations into groups of like petroleum types on the oil and gas field map, cross section and table (Figs. 2A, 2B, and 2C).

With this step completed, the investigator then locates all surface seeps on the oil and gas field map, which now becomes the petroleum system map. The seeps with the same geochemical composition as the subsurface accumulation provides geographic evidence for the end point of a migration path. The stratigraphic unit from which the fluid emanates can be compared to the stratigraphic unit in which oil and gas accumulations are found to determine the complexity of their migration paths. If the stratigraphic units are the same, then the migration paths are simple. If they are different, migration may be more complex. Geochemical information from seeps can be compared with that of discovered accumulations in order to link the seep fluid to the proper petroleum system.

Oil and gas shows in exploratory wells are added to the petroleum system map and cross section to better define the migration paths. As this map and cross section evolve, the investigator is encouraged to anticipate how the final map will look based on the framework geology and petroleum fluid information. Intuitively, exploration risk is high if the petroleum system is complicated and hence less predictable; risk is lower if the petroleum system is simple and thus more predictable.

After similar hydrocarbon fluids in the petroleum system have been mapped individual oil and gas accumulations are tabulated to better understand the size (by volume) and complexity of the petroleum system. The petroleum system table is organized by stratigraphic interval in each field (Fig. 2C). These stratigraphic intervals are zones, members, and formations that produce or contain measurable amounts of oil and gas. The table should include age of the stratigraphic interval, API gravity and sulfur content of the oil, gas-to-oil ratio (GOR), cumulative amount of oil and gas produced, and the remaining amount of oil and gas that can be produced. Other information the investigator may chose to include are lithology, gross and net thickness, porosity and permeability of the reservoir rock, geometry, closure, and area of the trap, and detailed geochemistry of the oil and gas. The information included in the table will depend on what is available and the objectives of the investigation. The required information is used to determine the size (by volume) of the petroleum system, which reservoir rock name is to be used in the name of the petroleum system, and to evaluate the complexity of the migration path.

Now the provenance or origin of the petroleum is mapped as the pod of active source rock. Only one pod of active source rock occurs in each petroleum system. A pod is a contiguous body of source rock that has or is expelling oil and gas. Because this pod has thickness and area, it can be mapped using well, outcrop, or geophysical data. When an organic-rich rock is in close, or reasonably close, proximity, both stratigraphically and geographically, to oil and gas accumulations, shows, or seeps, it is tentatively correlated with those fluids. Based on seismic, well, or outcrop data, the likelihood of correlation increases when the source rock's burial depth is known to reach 3 km, which in the experience of the authors is a reasonable minimum burial depth for thermal maturity or when burial depth modeling indicates a source rock is in or below the oil window. The correlation gains certainty if the source rock, by vitrinite reflectance or some other analytical technique, is established as being thermally mature. If the kerogen type of the source rock is consistent with that of the oil and gas, then confidence increases that the source rock is correctly correlated. If the geochemical composition of the organic matter in the source rock compares favorably with the migrated petroleum, the oil-source rock correlation is considered a match. Using seismic, well, and outcrop data, the suspected or confirmed active source rock is mapped as a contiguous, three-dimensional body, or pod, on the petroleum system map and cross section.

In this manner, the petroleum system map and cross section evolve, as the working hypothesis is taken to successive level of certainty. To further refine this work, a burial history chart and events chart are constructed and the petroleum system is named. This article discussed each of these petroleum system components in sequence, but they are frequently developed in parallel, and their relationship to each other is considered so that the petroleum system can be properly mapped. To organize these components, the petroleum system folio sheet is used.

SEE ALSO THE FOLLOWING ARTICLES

Natural Gas Resources, Global Distribution of • Oil and Natural Gas Drilling • Oil and Natural Gas Exploration • Oil and Natural Gas Liquids: Global Magnitude and Distribution • Oil Recovery • Oil Refining and Products

Further Reading

- Dow, W. G. (1974). Application of oil-correlation and source-rock data to exploration in Williston basin. *American Assn. Petroleum Geol. Bull.* 58(7), 1253–1262.
- Law, B. E., and Curtis, J. B. (2002). Unconventional petroleum systems. *American Assn. Petroleum Geol. Bull.* 86(11), 1851–1999.
- Magoon, L. B. (1995). The play that complements the petroleum system—A new exploration equation. *Oil Gas J.* 93(no. 40), 85–87.
- Magoon, L. B., and Beaumont, E. A. (1999). Petroleum system. In “Exploring for Oil and Gas Traps” (E. A. Beaumont and N. H. Foster, Eds.), pp. 3.1–3.34. American Association of Petroleum Geologists Treatise of Petroleum Geology, Tulsa, OK.
- Magoon, L. B., and Dow, W. G. (1994). The petroleum system—From source to trap. *American Assn. Petroleum Geol. Memoir* 60, 655.
- Magoon, L. B., and Dow, W. G. (2000). Mapping the petroleum system—An investigative technique to explore the hydrocarbon fluid system. In “Petroleum Systems of South Atlantic Margins” (M. R. Mello and B. J. Katz, Eds.). American Association of Petroleum Geologists Memoir 73, pp. 53–68.
- Magoon, L. B., and Schmoker, J. W. (2000). The total petroleum system—The natural fluid network that constrains the assessment unit. In “World Petroleum Assessment 2000” (U.S. Geological Survey World Energy Assessment Team, eds.). U.S. Geological Survey DDS-60, 2 CD-ROMs.
- Magoon, L. B., and Valin, Z. C. (1994). Overview of petroleum system case studies. *American Association of Petroleum Geologists Memoir* 60, 332.