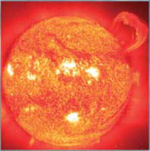
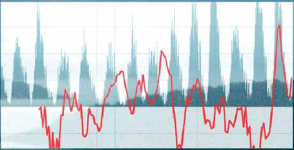
Solar Activity and Earth's Climate

Rasmus E. Benestad



Second Edition



Springer

PRAXIS

Solar Activity and Earth's Climate

Second Edition

Rasmus E. Benestad

Solar Activity and Earth's Climate

Second Edition

Springer

Published in association with
Praxis Publishing
Chichester, UK

PRAXIS

Dr Rasmus E. Benestad
The Norwegian Meteorological Institute
Oslo
Norway

SPRINGER–PRAXIS BOOKS IN ENVIRONMENTAL SCIENCES
SUBJECT *ADVISORY EDITOR*: John Mason B.Sc., M.Sc., Ph.D.

ISBN-10 3-540-30620-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-30620-7 Springer Berlin Heidelberg New York

Springer is part of Springer-Science + Business Media (springeronline.com)

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available from the Internet at http://dnb.ddb.de

Library of Congress Control Number: 2005937945

Second edition 2006
First edition published 2002
Printed in Germany

Cover design: Jim Wilkie
Project management: Originator Publishing Services, Gt Yarmouth, Norfolk, UK

Printed on acid-free paper

Contents

Preface to the second edition

I have in this second addition added a list of exercises and questions at the end of the book for testing whether the readers/students have read and understood a number of central issues in solar–terrestrial relations. Furthermore, the text has been updated with a list of 48 new publications on the solar–climate connection. However, I have not updated the book in terms of new findings on general solar physics, as the progress in the understanding of the solar interior doesn't have any direct consequence for the climate connection. A couple of minor errors have also been corrected, as pointed out in the reviews of the first edition.

Preface to the first edition

This book started as a personal project when Clive Horwood from Praxis Publishing contacted me about writing a book on the connection between solar activity and Earth's climate: Could I do it? The question of whether our climate is somehow affected by changes in the Sun has recently received an increasing amount of attention, after having been neglected by most climate researchers for decades. One reason for the renewed interest relates to the issue about anthropogenic warming, and whether there may be alternative explanations for the recently observed global warming of the Earth's surface. However, the intriguing question about a relationship between solar activity and global warming also puzzled people back in the 19th century and even before that, and can be considered as one of *the* classical conundrums that still remain unsolved.

When I first started learning about the ways in which solar activity may affect our climate, I was surprised by how little most climatologists seemed to know or even care about the issue. It was difficult to find introductory textbooks on the subject that did not assume some expertise in solar physics. The scientific aspects surrounding a link between solar activity and Earth's climate span a wide range of disciplines, which requires good grounding in them in order to understand such a link. When posed with the question of writing this book, I felt I had the advantage of having worked on many different issues that are relevant to this solar–terrestrial problem. Furthermore, I had gained a good deal of insight into this problem and appreciated the need for a textbook that could be used as an introduction to this field.

To Katja, Jacob & Becky

Acknowledgements to the second edition

I would like to than Drs Drew T. Shindell and Joanna D. Haigh for updating me on the latest progress on solar–stratospheric links.

Acknowledgements to the first edition

I am grateful to my wife, Becky, for her support, patience, and assistance with the English language. I would also like to thank Drs John Mason, Jón Egill Kristjánsson and Asgeir Brekke for reviewing the manuscript and giving advice on the text. Thanks to Mike Shardlow for his technical assistance in editing and preparing the book. Dr David B. Stephenson at the University of Reading and Dr Nils Gunnar Kvamstø at the University of Bergen brought my attention to the book "Temperature variations in the North Atlantic ocean and in the atmosphere" written by Bjørn Helland-Hansen and Fridtjof Nansen in 1920, which inspired much of the historical emphasis of this book. Gustav Bjørbæk and Sofus Lystad at the Norwegian Meteorological Institute have supplied me with material of which I otherwise would not have been aware.

Professor Shine, at the University of Reading, kindly sent me a report reviewing the current state of research on the link between solar activity and our climate, and Professor Bas van Geel kindly provided me with a series of articles on proxy data and evidence for solar-terrestrial relationships. Part of the material was taken from the proceedings of the First Solar and Space Weather Euro-conference: The Solar Cycle and Terrestrial Climate, at which I participated on behalf of the Norwegian Meteorological Institute.

The project took longer than anticipated, and I'm grateful to Clive Horwood and Neil Shuttlewood for their patience. Also, this book was only possible because I have a patient wife and well-behaved children, Jacob and Katja, who let me work during the evenings as the book was written in my spare time. Part of the literary research was done while commuting on Oslo's public transportation, and I got plenty of time for reading, thanks to the frequent delays.

This project would never have been completed without the free distribution of *Linux*. the operating system *Debian GNU/Linux*.[1] Thanks to Trond Bø at the Norwegian Meteorological Institute for his assistance setting up *Linux* on my home computer, X-windows environment, the GNU emacs editor, the LaTeX typesetting using a style file written by Stephen Webb, and the GNU *R* data analysis language.[2] The large number of (to me anonymous) programmers who spent their time developing free software paved the way for this book, and deserve an acknowledgement. Finally, the work of the scientists in the USA and some other countries has made observational data available on the Internet for the public (partly governmental policy), and the observations analysed in this book are taken from these sources. Some of the chapters would not have been possible to write without the free availability of these data. The Norwegian climate observations were taken from the Norwegian Meteorological Institute's database.

[1] The GNU project (Free Software Foundation) is an acronym for "GNU is Not Unix". URL: http://www.fsf.org/.

[2] http://cran.r-project.org/

Figures

Tables

1

Introduction

1.1 THE PHILOSOPHY OF THIS BOOK

Since the existence of sunspots was established beyond doubt in the early 17th century, people have wondered about their role in association with various incidents on Earth. This conundrum has sometimes been entangled with a common belief that events on Earth are determined by celestial bodies and perhaps fuelled by a human passion for organising the universe into order. Dr D. B. Stephenson of the University of Reading once described this behaviour:[1] "We have grouped stars that we now know are randomly distributed into stellar constellations, which we even have named". Likewise, there have been numerous attempts throughout history to explain climatic events in terms of sunspots. The question whether sunspots affect our climate and weather is perhaps one of the oldest scientific enigmas. We know of hypotheses on links between sunspots and terrestrial temperatures dating as far back as 1651. This question has yet to be resolved and the endeavour still continues today. In September 2000 the first *Solspa* conference on the solar–terrestrial climate connection was held in Santa Cruz de Tenerife. We will refer to any link between solar activity and the Earth's climate as a "solar–terrestrial link" in the remainder of the book. There are valid reasons to believe that solar activity has *some* impact on Earth's climate, but on the other hand, the history of science can also point to numerous unsupported statements about sunspots and their influence on climatic variations.

This book is a result of my own exploration into the cross-disciplinary subject of solar–terrestrial relationships. After getting engaged in climate research, I quickly learned of various hypotheses on relationships between solar activity and Earth's climate. However, I found that these hypotheses often were fragmented or illuminating only part of the subject, and there were few sources offering a comprehensive discussion on the solar–terrestrial relationships relevant to the climate near Earth's

[1] Statistics course given in Bergen, September 2000.

surface. This text will try to give a comprehensive treatment of the subject, bringing together the different aspects of the science of solar–terrestrial relationships. The book may be regarded as a tour of the solar–terrestrial connection, touching on subjects from archaeology and palaeoclimatology to the aurora borealis and solar physics.

This book is written from a climatologist's point of view, but the first chapter will nevertheless give the reader a basic background on our current knowledge of the Sun in order to give a holistic treatment on the subject of solar activity and Earth's climate. This book will try to bridge the gap between the various subjects that solar–terrestrial science represents and will therefore span a wide range of disciplines, including: astronomy, physics, geography, chemistry, geology and biology. It is not expected that the reader should be familiar with all these subjects. Nor does the author propose to master all these disciplines, and some aspects will therefore be discussed with more weight than others. For more depth, the reader will be referred to other texts on the particular topics.

There are many concepts that are taken for granted in the respective solar science and climate research communities for which it sometimes is hard to find definitions in the literature on solar–terrestrial relationships. With a background in physics, electronics, cloud microphysics, ocean dynamics and statistical analysis, part of the terminology was new to me as I learned about solar science. One goal of this book is to explain the basic concepts and terms commonly used in solar physics and climate research. Hence, some effort is devoted to explaining concepts and terminology that are taken for granted by solar scientists, but with which climatologists may be unfamiliar, and vice versa.

This book will include a brief review of the historical progress of understanding the solar–terrestrial connection but will also objectively scrutinise various hypotheses. A central theme in this book is "How do we know what we know?". The emphasis will therefore be on the "scientific method", discussed by the philosopher Karl Popper (Magee, 1973). The "scientific method" states that it is crucial that the hypotheses in principle can be falsified (i.e. proven wrong), that they are based on objective tests, and that the results must be repeatable. Moreover, it is crucial that these hypotheses are formulated in such a way that they may be falsified when observational evidence really does not support them (Feynman, 1985, pp. 338–346). This can for instance be done by using established relationships between solar activity and Earth's climate for the prediction of future climatic variations. Scientific work must be easily replicated and the methods must be clear and unambiguous.

Since our knowledge on how solar activity may affect Earth's climate is still very limited, most hypotheses are based on empirical studies, and it is crucial to address the issue of how reliable the data are. The data quality is of central importance to empirical work on solar–terrestrial relationships. Can we trust our observations, and can systematic errors affect our analysis? Furthermore, the solar theories, and especially hypotheses about a climatic response to the solar cycle, are to a large extent based on observations, and must therefore be regarded in the context of the observational procedures. The book will therefore give the reader a gentle introduction to

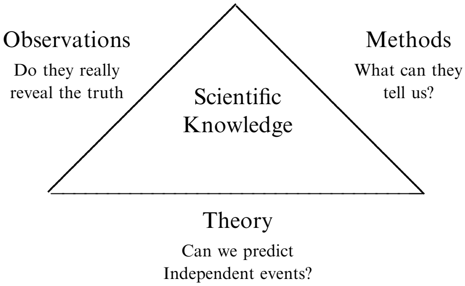


Figure 1.1. A schematic illustrating the "pillars" of scientific knowledge.

the observations, instruments and observational practices. However, an acquaintance with the data and its quality is not enough for judging the statements made about sunspots and the terrestrial climate. The observations and the theories are regarded in connection with the analytical methods used in arriving at the conclusions on which the various theories are based. Thus, the reader will be served a "formula" combining physics with statistics and information about the observations (metadata). One message is that our scientific knowledge is founded on three "pillars" (Figure 1.1): observations (the quality and shortcomings of these), the scientific tests (methods), and theory (physically based hypotheses). The reader should appreciate that there are limitations to what the various statistical and analytical methods can teach us and it is crucial to understand how the various conclusions are reached. Therefore, a number of statistical concepts will be explained in conjunction with the discussion of hypothesised solar–terrestrial links.

Before the various hypotheses can be examined, however, the reader must also have *some* elementary background knowledge of mathematics, physics and chemistry. Perhaps too often, physics is presented in a clean, simple, theoretical and idealistic fashion with too many fictitious imaginary point masses and charges that allow simple and elegant solutions. The real world, however, as many professional physicists have experienced, is not clean and simple. Data are usually "noisy". Real-life physics often deals with the messy world by finding a signal embedded in noise, often through a "best-fit" of some idealised model to "imperfect" data influenced by many factors. The Sun is a remarkable physics laboratory, and it is available for the study of physical processes in extreme conditions. The Earth is also a willing subject of scientific studies, each day displaying awesome, as well as catastrophic, phenomena, such as hurricanes, tornadoes, thunderstorms, formation of precipitation, lightning, and phenomena like the El Niño Southern Oscillation, just to mention a few examples. Hopefully, this text will open the eyes of the reader to the richness of physics: atmospheric microphysics, fluid dynamics, energy and momentum conservation, cosmic rays, geomagnetism and electrification, sunspots, plasma physics, hydromagnetism, and fusion.

There is an abundance of publications on associations between sunspots and climate, and many of these were published in the 19th century. It is not possible within the scope of this book to cover everything. Here only the most popular hypotheses will be discussed. The history of solar–terrestrial studies can point to several studies which do not qualify as science according to Popper's criterion. Here, some of these will be used as case studies since one may learn from the mistakes made as well as from the progress made in the past. This review may also serve as a discourse on analysis, research methodology, and the application of the scientific method. In this respect, part of the book may be of interest to a wider group, such as physics students.

The Sun does affect many processes in the Earth's atmosphere, including some which are not usually regarded as important for our climate. Some of these will nevertheless briefly be discussed here. As the knowledge of solar–terrestrial relations is still incomplete, there is no guarantee that aspects not taken into account are unimportant.

It is important that the reader is aware of the rapid progress continuously made in understanding the Sun and Earth's climate. There have been jumps and spurts in the progress of understanding the Sun in the past which have been driven by advancements in observational capabilities, and new discoveries of solar features are being made at the time of writing, so this book will only attempt to summarise and review the knowledge about our Sun acquired so far. Many of the most recent discoveries are due to the space missions launched in the 1980s and 1990s to observe the Sun, and the picture is changing so fast that many sides of the Sun conveyed in 1953 have almost been "forgotten" in 1992 in favour of "spicy" new details.

The scope of this book will embrace a wider part of the solar system and will not be limited just to discussing the Earth and the Sun. We will glean some information from two other earth-like planets, Venus and Mars, as well as the Moon, and compare these to our own Earth. Both the planets have an atmosphere which exhibit similar features as well as different attributes to the Earth's atmosphere. If variations in the Sun produce changes in the Earth's climate, one may expect to see similar fluctuations in the brightness temperature on Venus and perhaps Mars. Long time-series of the planetary mean temperatures are necessary for empirical studies of the relationship between the atmosphere of these planets and the Sun, and such records do not exist yet to my knowledge. It is important that the measurements of these planets are not made through Earth's atmosphere, if variations in the transparency of the air (at different wavelengths) can affect the analysis. Therefore, the observation of the other planets ought to be made from space. Coherent variability may give strong circumstantial evidence for solar-induced climate variations. In addition to comparing the Earth to similar planets, the Sun will be viewed in the context of general stellar properties.

Since we do not have a firm picture of how, if at all, solar activities influence Earth's lower atmosphere, the reader cannot expect this text to come up with the true and final answer. The Intergovernmental Panel on Climate Change (IPCC) Third Assessment Report (TAR) (Houghton *et al.*, 2001) quotes the level of knowledge on links between solar activity and the climate as "very low". The scope of this book will merely be to review and discuss some popular hypotheses and familiarise the reader with the basic concepts.

I must admit that I have encountered entrenched positions on the subject and some prejudice, as the issue of whether solar activity may affect terrestrial temperatures appears to be laden with political and personal agendas. One important message that this book tries to convey is that solar influence on Earth's climate, be it on decadal or longer timescales, does not represent an antithesis to the so-called "enhanced greenhouse effect". Various hypotheses of solar regulation of our climate have sometimes been presented as a scientific challenge to the established view that human emissions of greenhouse gases represent a disrupting influence on our climate. Scientifically speaking, such a challenge would be sound and the best thing that could happen for further progress. However, this challenge may, from an environmental and practical point of view, be unfortunate if we indeed are perturbing the natural energy balance. The reason is that any delay to act on anthropogenic climate disruption can result in more severe climate change if real, and an "alternative" explanation for the recent climate changes can be used as an argument for not acting. It is important to note that a falsification of solar–terrestrial links would not necessarily prove that the enhanced greenhouse effect is true, and vice versa. Putting all these environmental, ethical, political, and economical issues aside, this book will focus on just the scientific aspects of possible solar forcing of Earth's climate.

1.2 THE LAYOUT

All chapters will start with a general synopsis of the subjects. This overview will also serve as a general introduction to the subject, avoiding heavy mathematical equations and detailed physics. An elaboration of the topic follows the overview.

Chapter 2 discusses issues concerning solar observations, and the third chapter deals with aspects of solar physics describing the various solar features. A discussion of solar activity and the effect of the solar cycle on the total irradiance ("solar constant") is given in Chapter 4. The basic aspects of Earth's climate system are outlined in Chapter 5. These first chapters introduce some statistical concepts that are useful for the description and discussion of solar activity and its relation to the Earth's climate. Some discussion is also given here on statistical inference, in connection with correlation and regression studies on the sunspots and terrestrial temperature. Hence, these first chapters provide background knowledge for later studies on the connection between solar activity and variations in Earth's climate.

It is not easy to arrange the material concerning the "solar–terrestrial" link into separate categories, as there are overlaps between the various hypotheses. The remaining chapters of the book cover solar–terrestrial links mediated through stratospheric processes, the solar influence on geomagnetism or electric fields and how these may affect our climate, and empirical studies on various solar–terrestrial links. A review is given of studies of how variations in ultra-violet (UV) light associated with the solar cycle can affect the chemistry in the atmosphere and hence the dynamics. The Quasi-Biennial Oscillation (QBO) will also be discussed here. The chapter on solar activity and magnetism gives a discussion of recently proposed hypotheses about the solar magnetic field, galactic cosmic rays, and their influence on the Earth. There will also be a discussion of the relationship between solar activity, the northern lights, and our climate. The last of these chapters gives a brief historical account of proposed links between solar activity and regional climate variations. A final chapter will address issues such as links between the solar cycle and El Niño Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), and the south Asian monsoon. The last chapter will also address other possible sources of climate forcing, such as volcanoes, anthropogenic forcing, and the Moon.

2

Solar observations

2.1 SYNOPSIS

The Sun drives our climate, which is often defined as the *average weather* for a given location. Furthermore, photosynthesis and chemical reactions on Earth are principally driven by the energy coming from our Sun. In order to understand Earth's climate, it is therefore important to have some knowledge of how the energy that Earth receives is generated in the first place. By knowing which physical processes give rise to the luminance, stream of particles, and magnetism of the Sun, and hence most of the energy reaching Earth, it may be possible to identify how these behave, and ultimately affect Earth's climate. Here, the most basic features about the Sun and its energy production will be outlined.

Our knowledge about the Sun and solar activity is derived from our observations, which again give a basis for physical models. The observations of the Sun and the solar activity are also crucial for the study of the link between solar activity and Earth's climate. In this respect, it is crucial to know how the observational data have been derived and what their qualities are. Moreover, this information enhances our understanding of how well we know the Sun.

2.2 INSTRUMENTS FOR OBSERVING THE SUN

The only way to study the Sun from the Earth is by examining the solar quantities that reach us, which includes particles, magnetic fields, and electromagnetic radiation. In addition, the action of the Sun's gravity on other bodies in our solar systems may be used to infer the solar mass. There is a great deal to learn from the electromagnetic radiation. The visible light and its directional variation is the foremost source of information on what happens on the Sun. It is through the visible light that we can see the Sun directly with our own eyes[1] or by photographs.

[1] It is harmful to stare directly at the Sun!

Telescopes were therefore the first instrument that were used for solar studies. During total solar eclipses, it is possible to observe the coronal structures directly. Past observations also have indicated that the solar diameter is not quite constant, but may vary as a result of solar activity. Heliography involves photographic studies of the Sun, and has been used to record sunspots. This observational technique reveals the spatial structures on the Sun.

2.2.1 Measuring the total solar irradiance

One instrument for measuring total radiative energy flux is the *bolometer* (Fleagle and Businger, 1980). A water-cooled black target has often been used to estimate the total solar irradiance (TSI) (also known as the "Solar constant"[2]), which is estimated from the energy budget of the black plate and the cooling system. Satellites have been recording the TSI since the late 1970s when a solar irradiance monitoring system finally was put into orbit.[3] There have been a series of different satellites measuring TSI[4] starting from 1978. Fröhlich and Lean (1998a) attempted to merge the various readings from the different satellites. The differences between the different satellites suggest that the TSI estimates from different satellites are not exactly the same and that these may be in the range 1365–1373 W/m^2 (see Figure 2.1). There is some debate as to whether the TSI level is the same during the two sunspot minima or whether the TSI at solar minimum has increased. Willson (1997) has suggested a trend of 0.036% (0.09 W/m^2) increase per decade.

2.2.2 There is more than total irradiance

There is more information about the Sun carried away by the electromagnetic radiation than the total energy. The radiation has several properties which are dependent on the solar conditions during the emission of light. The electromagnetic waves have an amplitude (or photon density associated with the light intensity), a wavelength (which in terms of photons reflect the energy of each particle), and polarisation. The effect of the Sun's gravitation on the objects in our solar system can also be used to estimate the solar mass.

Temperature is a bulk measure of the vibrational energy of the atoms. The wavelength of light's so-called continuous spectrum, is associated with the *black body* radiation which is dependent on the Sun's temperature. The simplest way to picture thermal radiation is to consider electric fields extending from electric charges. Electric charge is one of the four elementary physical quantities that we know, the others being to mass, time, and distance. All physical concepts may be explained in

[2] This label is a misnomer.

[3] There was a lot of debate about whether this project really was necessary, as everyone "knew" that the "solar constant" was constant. It is only in hindsight that we know the value of these observations.

[4] HF/ERB on NIMBUS-7, ACRIM I & II on SMM and UARS, the solar monitor on ERBS, SOVA2 on EURECA, and VIRGO on SOHO.

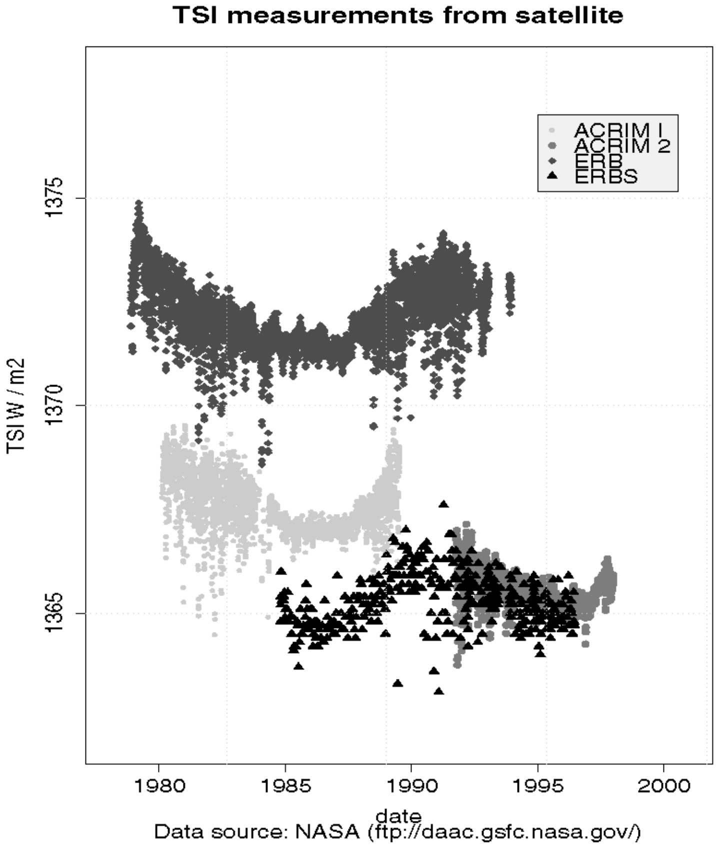


Figure 2.1. Comparison between four different satellite-based TSI measurements. Data from NASA's Climatology Interdisciplinary Data Collection.

terms of these entities. For instance, Einstein showed in his theory of special relativity that mass and energy are related by $E = mc^2$, and are in fact two sides of the same thing. When charge accelerates (oscillates or vibrates) then the electric field is disturbed, and moving charges produce magnetic fields (a relativistic effect). In a "classical physics interpretation", the perturbation of the fields is not instant everywhere, but the disturbances propagate at the speed of light, c, and hence take the form of a wave. Materials with absolute temperature above zero are made up of atoms and molecules that vibrate, and the charged protons and electrons follow the motion of the atoms. Thus, the electromagnetic radiation associated with these oscillating charges reflects the temperature. In Section 3.2.5.1 the total solar irradiation is used to infer the temperature of the solar surface.

The particles do not all vibrate with the same energy and frequency, but some oscillate more vigorously than others. There is a spread in individual energies which follows a common statistical law describing random motion. By assuming a

Gaussian distribution[5] in the vibrational energies of the atoms, it is possible to derive a spectral distribution law (Planck's law) of emitted thermal emission as a function of temperature.

On top of the continuous spectrum are spectral lines associated with narrow wavelength bands, which are a result of atomic transition from an unstable energetic state to a more stable state. The latter type of radiation is known as line emission (Fraunhofer lines), and is a central topic in atomic and quantum physics. Each element is associated with line emissions of different wavelength, and by examining the line spectra it is possible to learn about which elements are present in the Sun. It is possible to decompose light into its spectral components by passing it through a prism. Light with a short wavelength bends more than light with a longer wavelength. Other spectral techniques include diffraction gratings. The study of the different wavelengths is known as *spectroscopy*.

Because of the overlap between the two types of emission, one may get a wrong temperature estimate unless both types are taken into account. There are two ways of inferring the black body continuum: (i) by measuring the continuum intensities between the spectral lines ("Continuum windows") or (ii) broad spectral band measurements with line spectrum corrections (Maltby, 1992)

2.2.3 Spectrography and polarisation

Instruments that observe the solar spectrum are based on spectrography, which measures the intensity of light as a function of its wavelength. When electrons change energy states (ΔE), they may emit or absorb light of a wavelength (λ) according to $\Delta E = h\nu = hc/\lambda$ (line emission). In addition to depending on the atomic energy levels, the line spectra are also affected by the presence of magnetic fields during the line emission. A quantum physical theory, known as the *Zeeman effect*, says that the electrons bound by an atom can only have a spin with the axis oriented in certain directions. The presence of the magnetic field thereby modifies the atomic energy levels, usually splitting these into three levels: a *Zeeman triplet*. There is, however, also more complicated Zeeman splitting, or in some cases, the magnetic field has no effect on the atomic energy levels. There is furthermore the inverse Zeeman effect, where light is absorbed by matter in the presence of a magnetic field, rather than being emitted. The Zeeman effect and the inverse Zeeman effect have different effects on the line spectra. Instruments for measuring the Zeeman split include the magnetograph.

By studying the Zeeman splitting of Fraunhofer lines[6] it is in principle possible to learn about the solar magnetic field. The Zeeman splitting is illustrated schematically in Figure 2.2.

The Zeeman triplet consists of two or three components, depending on the direction of the magnetic field, $\vec{H}$, compared to the observer. When looking straight into the field (Figure 2.2a) there are two components which have the wave-

[5] Also called normal distribution.

[6] Named after the discoverer.

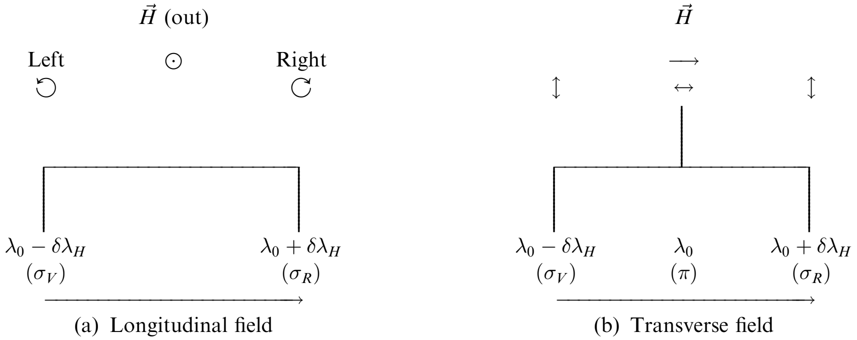


Figure 2.2. A schematic showing the three components in a Zeeman triplet when viewed into the beam (a) and perpendicularly to the beam (b). The horizontal axis is wavelength.

lengths $\lambda_0 - \delta\lambda_H$ and $\lambda_0 + \delta\lambda_H$, often referred to as σ_V and σ_R. By looking in a direction normal to the field (transverse field), three components can be seen, with the wavelengths $\lambda_0 - \delta\lambda_H$, λ_0, and $\lambda_0 + \delta\lambda_H$. The two symmetric components are σ_V and σ_R, and the central line is called the π component.

The electromagnetic wavelengths may also be affected by the relative motion between the emitting material and the observer due to the Doppler effect. One example of the *Doppler effect* that probably most readers are familiar with is the change of pitch heard when an approaching train passes an observer. Likewise, the frequency of light is shifted when the object moves with respect to the observer. The apparent frequency, f', can be expressed in terms of the true frequency f_0, the phase speed of the waves (often the speed of light or sound), and the velocity of the observer and object. If the observer is moving away from the wave source at a speed v_o, then this apparent frequency $f' = f_0(c - v_o)/c$ (v_o is negative if the observer is approaching the source), but if the source is moving away from the observer (v_s) then the appropriate expression is $f' = f_0 c/(c + v_s)$. The Doppler effect will affect all the wavelengths equally and is not influenced by $\delta\lambda_H$. Through the study of the Doppler shifted light, it is possible to estimate the velocities of the solar material. One example is the discovery of the motion in the light sunspot regions, known as the *Evershed effect*.

The polarisation of the light emitted from the Sun may hold some information about the Sun. For instance, the components of the Zeeman triplets have different polarisation. In the longitudinal field, the σ_V and σ_R are left- and right-handed circularly polarised respectively. The σ_V and σ_R components in the transverse case are transversely polarised with respect to the magnetic field, whereas the π component is polarised parallel to the magnetic field.

2.2.3.1 The magnetic field associated with sunspots

The strong magnetic fields in the sunspots give rise to Zeeman splitting of the line spectra. It is usually assumed that the lines examined are Zeeman triplets, however, more complicated forms of Zeeman splitting may also take place. In the sunspots, it is the *inverse* Zeeman effect, caused by the absorption of light under the influence of a magnetic field, which is believed to dominate. The measurable quantities from light emitted from the sunspot regions are the wavelength, the line intensity, and the polarisation of light. The relation between the intensities of the three lines of a transverse field Zeeman triplet is $I_{\sigma_R} : I_\pi : I_{\sigma_V} = \frac{1}{4}(1 - \cos^2\gamma) : \frac{1}{4}\sin^2\gamma : \frac{1}{4}(1 - \cos^2\gamma)$ (Bray and Loughhead, 1964). The magnetic field in the sunspot is assumed to be uniform. The shift in the wavelength from its real value, λ_0, due to the Zeeman effect is:

$$\delta\lambda_H = \frac{eH}{4\pi m_e c^2}\lambda_0^2 \tag{2.1}$$

Plate analysers may be used to examine the polarisation of the light and Figure 2.2 shows how each line in the triplet is polarised. It is in principle possible to infer the strength and direction of the magnetic field from these quantities, given an idealised situation where there is only an inverse Zeeman effect (only absorption under the influence of a magnetic field, and no emission), and the triplet is dominating over all other forms. The magnetic field must also be uniform and not change with depth.

2.2.3.2 Potential problems associated with spectrography

In practice, spectrographs have in the past been limited by low spatial resolution and intensities, and long exposure time is often required. Because the observations are influenced by fluctuating atmospheric conditions, there may be need for "seeing correction". Contaminating stray light from scattering in the atmosphere may also introduce errors in the observations. Light may furthermore be subject to undesired partial polarisation by the mirror arrangement, and it is therefore important to carefully design the instruments to minimise these effects. It is usually assumed that there is no interference from light absorbed or emitted in the path of the light between the Sun and the Earth, but if this assumption is false then the observations may be distorted. Recent space-borne instrumentation and technological advancements have improved the situation.

2.3 THE HISTORY OF SOLAR OBSERVATIONS

Many aspects of the Sun remain a mystery even today, although recent observations have given us important insights into how the Sun works. Historical observations of the Sun have been the foundation for many hypotheses on solar–terrestrial relations. Some of the most important observations of the Sun in the early days were made of dark small regions that were later known as *sunspots*. Some of the earliest known references made to such solar features in the western hemisphere can be dated as far back as the ancient Greeks, to Theophrastus of Athens around 350 BC, but there are also records of sunspot observations in China from 28 BC. Sunspots, however, were not in good accord with the beliefs of most ancient Greeks, which was that the Sun was perfect and without blemishes. Nor were the sunspot concepts particularly popular with the Catholic Church during the Middle Ages. In the early Middle Ages, sunspots were sometimes mistaken as the passing of Mercury in front of the solar disk. Sunspots did therefore not receive much attention until the invention of the telescope by Galileo in the early 17th century, when their existence was demonstrated beyond doubt. The first telescopes consisted of an arrangement of two lenses, but more recent telescopes also use various arrangements of mirrors.

The invention of the telescope allowed better solar observations, and the existence of sunspots became too evident to ignore. According to Bray and Loughhead (1964), the first telescopic studies of the sunspots were started in 1611[7] by four astronomers: Fabricius, Galileo, Scheiner and Harriot. Fabricius deduced from sunspot observations that the Sun must rotate with a period of around 27 days. The solar rotation was of course calculated from a terrestrial frame of reference, and the solar rotation rate is slightly higher viewed from a galactic frame of reference. Galileo deduced that the dark regions were part of the Sun because their shape and size did not behave according to expectations, had they been planets. But these sunspots have ever since been an enigma. There is still no definite answer as to how and why they appear, although several hypotheses have been proposed.

Most of the historical solar observations were made through telescopes, and the first observations were intermittent because the Sun could not be seen during night-time or overcast conditions. The first telescopes were situated at various locations, often chosen to be in the proximity of the enthusiast astronomer and not necessarily where visibility (“atmospheric seeing”) was best. Some observatories were also moved, or the telescopes improved. In 1858, some of the first operational photoheliographs (photographs of the Sun) were made at the Kew observatory near London, but the telescope was moved to Spain in 1860. The telescope was re-erected at Cranford[8] in 1861 and moved to Greenwich in 1873, where the observations commenced in 1874 and have continued to the present day.

The Royal Observatory installed a telescope at the Cape of Good Hope in 1875. A new enlarger was fitted on this telescope in 1889, giving images of 8 inches as opposed to 4 inches with the previous telescope configuration. The telescope’s 4-inch lens was replaced with new ones in 1910 and 1926, presumably to obtain improved solar images. In 1949, the telescope was moved to Herstmonceux Castle in Sussex. Other observatories operated over a short period, such as the Durham observatory which was in operation from 1853 to 1861. The Zurich observatory has been in operation from 1855 to the present day.

The Mount Wilson observatory was established in 1904 when the Snow horizontal telescope at Yerkes Observatory was moved to the 1700-m high summit in California. In 1907 a 60-foot solar tower telescope was erected, and in 1912, a 150-foot tower was built. Hale and Adams were the first to make high-dispersion

[7] An earlier date has also been given: December 10, 1610 according to Helland-Hansen and Nansen (1920), p. 147.

[8] Middlesex, U.K. (Bray and Loughhead, 1964, p. 6).

photographs of the sunspot spectra at Mount Wilson (1906), and established that sunspots were cooler than the surrounding photosphere as opposed to being regions with higher absorptivity. They found certain line spectra of metals with stronger intensity than elsewhere from the solar surface and other spectral lines were weakened, consistent with laboratory experiments demonstrating that cooler gases are associated with more intense metallic lines than hotter gases (Kuiper, 1953, pp. 8–9).

One of the problems with telescopic observations from Earth's surface is the effect of atmospheric conditions on the image. A relocation of an observatory to a location which has less (more) clouds may result in a change in the quality of the observations. For instance, longer cloudy periods may result in faint and small-scale solar features going undetected. Winds, turbulence and haze can degrade the solar image quality, and clouds block the Sun and reduce the observing time. Poor observational conditions can result in undetected small sunspots near the Sun's limb. Small and short-lived sunspots may also go unnoticed due to extended periods of cloudiness. Many observatories are needed to track the evolution of solar events, and more recently a network of observatories around the world has joined in a collaborative effort to observe the Sun. With the recent solar satellites, it is possible to obtain uninterrupted observations of the Sun from just one platform.

2.3.1 The importance of good observations

Most of our knowledge about the Sun and our climate is derived from data of some form, be it actual observations or model data. These furthermore provide the framework for analytical analysis and physical models. It is therefore important to look at the data pool and assess its quality. The data quality is of utmost importance in science, as biases can have profound effects on analytical tests.

2.3.2 Criteria for good observations

A Norwegian meteorologist, Godske (1956), once proposed five criteria for making measurements: (i) the measurements must be unambiguous; (ii) repeated measurements of the same condition must give the same answer; (iii) one must know exactly what is measured; (iv) the instruments must be adapted to the conditions they measure; and (v) the observation must not alter the system. These conditions must be fulfilled if the observations are to be used in empirical studies of solar–terrestrial relations. The first two criteria reflect the quality of the measurements, and if these cannot be quantified in an objective way, there is no way that the data can be used to derive objective conclusions. The third point may seem obvious, but there may also be subtle aspects to this criterion. For instance, if the temperature is measured in direct sunlight, the observation is *not* of the air temperature, but that of the sun-exposed thermometer itself. If this criterion is not fulfilled, then it will be impossible to determine whether there is a real relationship between the two objects being studied. The fourth criterion is related to the third, and if (v) is not satisfied, then this implies a violation of (iii) for subsequent observations.

2.3.2.1 The quality of the sunspot record

Because of the limitations of the atmospheric visibility, observations were later made from high-altitude balloons, aircraft, and spacecraft. The quality of sunspot observations before 1849 has been questioned[9] (R. M. Wilson, 1998). When using solar data in connection with climate studies, it is extremely important to ensure that the sunspot record is not ''contaminated'' by the climate itself. For instance, if observations are made from a few observatories, as they were in the early record, then long overcast periods may result in undetected sunspots. Any atmospheric contamination can lead to a circular argumentation in solar–terrestrial studies. It may nevertheless be possible to support the direct sunspot observations by independent isotopic data from tree rings and ice cores as long as the climate does not affect these too. Therefore, the data quality places some limitations on empirical studies of solar–terrestrial links based on long data records. The quality of measurements of terrestrial variables also has a tendency to deteriorate with age.

It is not a trivial task to obtain long-term records of solar activity from historical observations. One difficulty is associated with the solar rotation, which causes periodic disappearance of the features on the Sun's surface. It is also known that the same feature may have different appearance for different observers, which may lead to some confusion.[10]

There exist two ''official'' sunspot records, the American and Zurich observations since 1950. A comparison between these is shown in Figure 2.3. The curves are very similar, but there are also some differences, such as a tendency for the American sunspot number tending to be higher in the first cycle shown in this plot, and mostly lower values in the second cycle. These discrepancies indicate that the sunspot number is associated with some degree of uncertainty. The sunspot numbers are usually determined from visual observations with a refractor of modest size and using a fairly low magnification. Wolf used a Fraunhofer refractor with an 8-cm aperture, a focal length of 110 cm, and a magnification coefficient of 64. A similar set-up is still used today at Zurich.

It is important to make certain that any changes in observation practices over time do not affect the results and hence give misleading impressions about long-term trends. The systematic improvements made to the telescopes ever since 1611 may, for instance, result in instruments that are capable of getting higher resolution and can capture more small sunspots. Furthermore, the gradual extension of the observational network may imply an improvement over time of sunspot observations. Thus, there is a risk that the historical sunspot record may suggest that the total number of sunspots has increased over time when many spots in the early record could have been missed due to the record being inhomogeneous. There is an east–west asymmetry in the number of sunspots observed discussed by Kuiper (1953). A zonal symmetry in the sunspot occurrence is expected unless the sunspots themselves are reclining with respect to the radial axes. The east–west asymmetry in the sunspot

[9] ''Poor'' before 1818, ''fair'' between 1818 and 1848, and ''good'' from 1849 to present.

[10] K. O. Kiepenhauer in Kuiper (1953), p. 322.

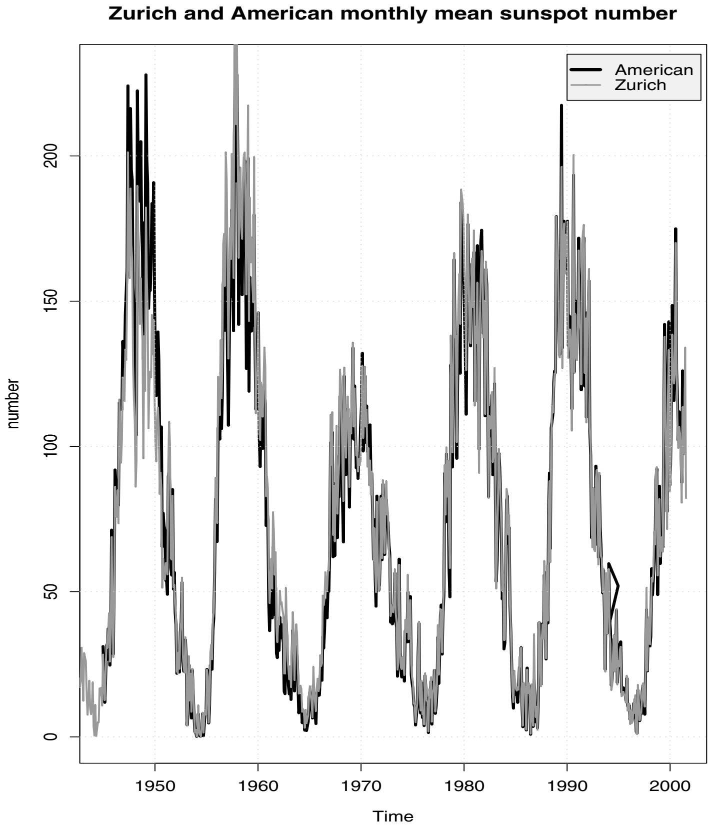


Figure 2.3. A comparison between two different estimates of the sunspot number. Both from ftp.ngdc.noaa.gov.

statistics (see Section 4.7.2) may be interpreted as an indication of not seeing all sunspots. Historically, we have only been able to observe one side of the Sun at any instant of time, and therefore approximately only half the number of sunspots at any time. As the Sun rotates, the hidden sunspots come into view whereas the sunspots in the west disappear behind the limb (see Figure 4.14).

Another source of inhomogeneity may be long-term changes to atmospheric transparency. Haze in the atmosphere or clouds may hide some sunspots, and if the amount of haze or clouds have increased, for instance, as a result of a warmer climate, then this can result in a systematic "under-count" of sunspots. According to Orlove *et al.* (2000), the Incas in the Andes may have used the visibility of Pleiades for centuries to make predictions about the next season's crop. In some Andean villages, the star *Pleiades* is celebrated in the month of June, and some forecasts were based on the size of Pleiades. A large apparent size was associated with a good harvest. There are indications of the upper stratospheric clouds being affected by El Niño events, which again affect the visibility of the stars so that they are dimmer around the onset of El Niño. This observation may have implications for the study between sunspots and climate. It may nevertheless be possible to correct for such atmospheric interference by using a number of independent stars as a baseline, the differences between the brightness of the subject and the baseline can be used to remove much of the atmospheric bias.

High-quality astronomical observations are not easy to obtain, and there are various factors that may degrade the observations. For instance, scintillation from stellar studies (10 Hz) have suggested that the amplitude of the twinkling intensity depends on the telescope aperture and high-altitude atmospheric winds. There may also be an image degradation and rapid image motion caused by local conditions near the telescope, such as surface winds. Furthermore, there may be slow image motions caused be atmospheric disturbance and telescope heating. Kuiper (1953) "guesstimated" that usually there are excellent observing conditions only 1% of the time. Presently, some of these effects may be corrected numerically through computer post-processing of the data.

Parasitic light is scattered light from Earth's atmosphere and optical instruments. Usually one assumes that the light scattering is independent of time and position, so that this effect can thus be corrected for by looking at the sky where there are no objects. Such corrections may be appropriate for usual astronomical studies, but it is important to ask whether they are good enough for deducing slow and small long-term changes that may be related to changes in the terrestrial climate.

The various concerns regarding the sunspot record, in addition to similar problems with climatic observations, may suggest that the sunspot number is not sufficiently reliable for studying long-term trends in solar–terrestrial relationships. It is therefore important to use caution when analysing these data records. Empirical studies may also be backed up by independent evidence, for example from so-called proxy data such as palaeo records.

2.4 PALAEO RECORDS OF SOLAR ACTIVITY

2.4.1 Isotopic records

Isotopic ratios in air bubbles trapped in ice-cores, calcite shells, and limestone, as well as carbon atoms in coal and wood may be used as proxy data (Bowler, 1999; Wagner *et al.*, 1999) for past solar activity or climatic variations. Energetic cosmic rays (mostly protons, *p*) enter the top of the atmosphere and collide with atoms so that subatomic particles may be "knocked out" of air atoms, and hence produce rare and unstable isotopes. Common types of isotopes produced by cosmic rays (cosmogenic isotopes) are carbon-14 and beryllium-10 (Table 2.1). As these isotopes are unstable, they decay into other elements with time, and

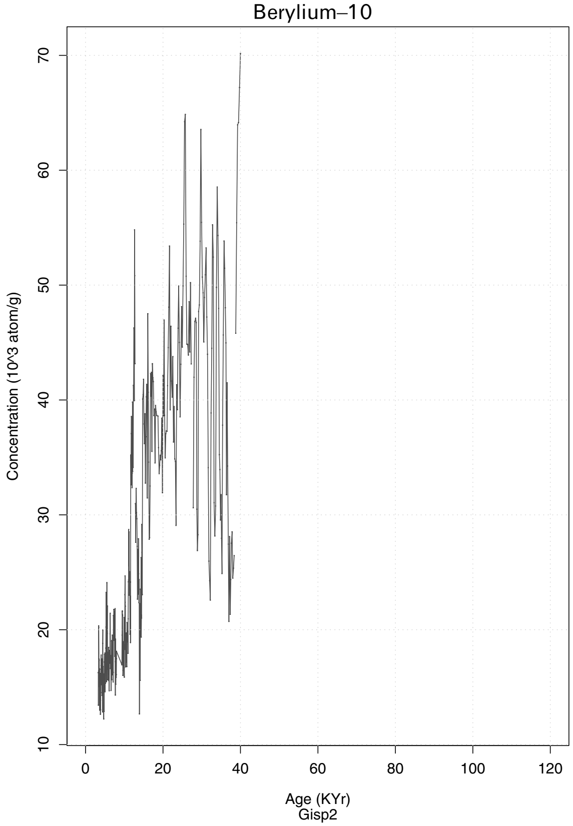


Figure 2.4. The relationship between (a) the ^{10}Be and (b) (opposite) the δ^{18}O cosmogenic isotope ratio from the GISP2 core. The y-axis indicates the isotope values and the x-axis gives the age in kilo (1,000) years with the most recent values shown to the left.

Table 2.1. Summary of some cosmogenic isotopes. Note, carbon-14 and chlorine-36 are also produced by nuclear bomb testing which began in 1954.

Isotope	Name	Lifetime (years)	Production	Decay
$^{3}_{2}$He	helium-3	12.3		
$^{10}_{4}$Be	beryllium-10	1,600,000	$^{14}_{7}\mathrm{N} + n \rightarrow {}^{10}_{4}\mathrm{Be} + 3p + 2n$	
			$^{14}_{7}\mathrm{N} + p \rightarrow {}^{10}_{4}\mathrm{Be} + 4p + n$	
			$^{16}_{8}\mathrm{O} + n \rightarrow {}^{10}_{4}\mathrm{Be} + 4p + 3n$	
			$^{16}_{8}\mathrm{O} + p \rightarrow {}^{10}_{4}\mathrm{Be} + 5p + 2n$	
$^{14}_{6}$C	carbon-14	5730	$^{14}_{7}N + p \rightarrow {}^{14}_{6}C + n$	$^{14}_{6}C \rightarrow {}^{14}_{7}N + \beta^{-}$
$^{36}_{17}$Cl	chlorine-36	30,000		

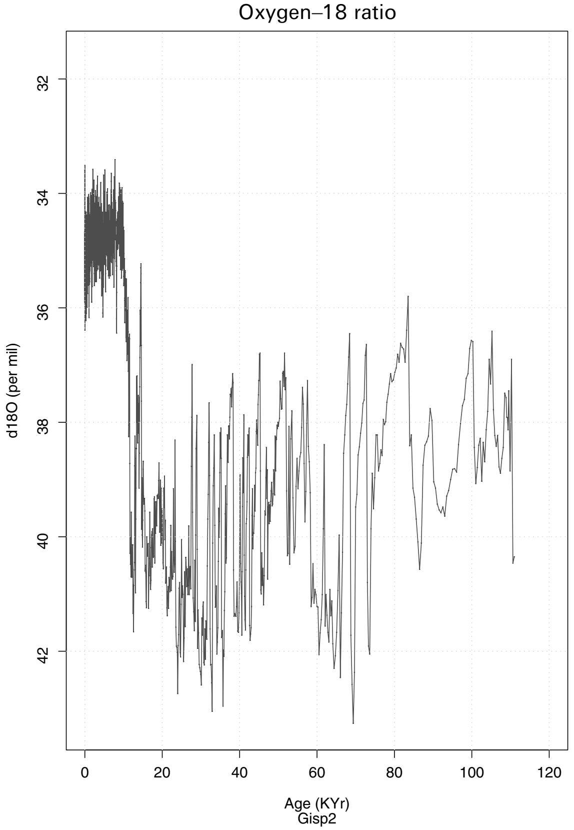


Figure 2.4. (b)

therefore high numbers of these isotopes is a sign of recent exposure to cosmic rays.

As several types of isotope concentrations depend on the exposure to cosmic rays, among other things, it may be possible to get a measure of how intense the high-energy particle bombardment has been in the past. Such footprints can be found in ^{14}C and ^{10}Be records. It is crucial to ensure that variations in the climatic conditions do not affect these isotope ratios, as seen in the δ^{13}C and ^{18}O records, if these proxies are used for studies between the past solar activity and the terrestrial climate.

It is possible to measure the isotope ratio directly with particle accelerators, a technique known as *mass spectroscopy* where a particle with mass m and charge q is accelerated from rest by a potential difference (V_0, which is the electric field E_0 multiplied by distance). The mass spectrometer consists of several stages with different arrangements of electric and magnetic fields.

A particle's charge (q) to mass (m) ratio can easily be inferred since the kinetic energy gained by the particle equals the potential energy lost in the electric field: $q/m = v^2/(2V_0)$. This ratio can be found given a value for the speed v. A moving charge feels magnetic forces ($\vec{F}_B = q\vec{v} \times \vec{B}_1$), but is also affected by electric fields ($\vec{F}_E = q\vec{E}_1$). The magnetic B_1 and electric E_1 fields can be arranged so that their associated forces are opposite to each other (which means that the fields are perpendicular to each other) as well as being perpendicular to the particle's trajectory. When these balance and the particle is not deflected from its original course, then $v = E_1/B_1$. Thus, the particle's speed can be determined by adjusting the fields until they hit a fluorescent screen in the same spot as they would when there are no electric or magnetic fields. The mass spectrometer technique uses the electric and magnetic field arrangement to determine the particle's charge to mass ratio in conjunction with a secondary magnetic field, separated by slits. The magnetic force felt by the particle in the second arrangement is always perpendicular to its velocity and hence alters the particle's direction in the same way as a centripetal force keeping an object in a circular orbit with a radius R. Hence, the magnetic force works as a centripetal force $qvB_2 = mv^2/R$. The radius can be measured, the magnetic field B and the charge to mass ratio q/m are known, and the mass can be estimated according to:

$$\frac{m}{q} = \frac{RB_2}{v} \tag{2.2}$$

The detection of cosmogenic isotopes can be done using cyclotrons, or tandem van de Graaff accelerators, where electrons are attached to isotopes such as ^{14}C. Simple "mass spectrometers" consisting of magnets and collimating slits are used once before and three times after the acceleration to select particles with the right charge to mass ratio. The particle beam is then passed through a detector that measures the energy that the particles lose as they pass through matter, and the larger the atomic number Z the larger is the loss. This type of apparatus gives a clean separation of ^{14}C from isotopes such as ^{14}N, ^{13}C, and ^{12}C. The carbon-13 isotope occurs naturally with a ratio of 1 to every 99 carbon-12 atoms.

2.4.1.1 Carbon-14

During photosynthesis, plants take up air with similar δ^{14}C-isotope ratios[11] as found

[11] The ratio is defined as δ^{14}C-isotope ratios and the expression for this ratio is:

$$\delta^{14}\text{C} = \frac{\left(\frac{^{14}\text{C}}{^{12}\text{C}}\right)_{\text{sample}} - \left(\frac{^{14}\text{C}}{^{12}\text{C}}\right)_{\text{standard}}}{\left(\frac{^{14}\text{C}}{^{12}\text{C}}\right)_{\text{standard}}} \times 10^3$$

in the atmosphere. The carbon-14 isotope is unstable and decays into nitrogen with a half-life of 5730 years. A living organism takes up carbon from the atmosphere with similar δ^{14}C ratio as the atmosphere, but when it dies the δ^{14}C ratio slowly drops as carbon-14 decays. Thus, if the original atmospheric δ^{14}C ratio is known, one can estimate the age of the organism. This kind of dating is known as *carbon dating*. But it is also possible to estimate the original concentrations, n_0, of these unstable isotopes, if there is some information about how long ago these elements were exposed to the cosmic rays, because the decay rate is known. The chronology can be found in wood by counting tree rings, but can also be inferred from the depth (number of layers) in ice. One way to take into account the variable production of cosmogenic isotopes is to apply a "wiggle match" between a profile and a calibration curve determined by wood samples. The wood samples can be dated though ring counting.

The number of ^{14}C atoms formed at any time depends on the flux of cosmic rays, but also on the concentration of ^{14}N and its cross-section (which may possibly vary with cosmic ray energy). The rate of production of carbon-14 isotopes, dn_{C14}/dt, can be expressed mathematically as the product of the galactic cosmic ray flux $F_{gcr}(z, E, \vec{B})$, the concentration of $n_{N14}(z)$ at altitude z, and the cross-section (reaction efficiency) $\sigma_{N14}(E)$ of the nitrogen. The galactic cosmic ray flux can be assumed to be a function of z, the energy E, and the magnetic field $\vec{B}$. This rate can be expressed as:

$$\frac{dn_{C14}}{dt} = F_{gcr}(z, E, \vec{B}) n_{N14}(z) \sigma_{N14}(E) \tag{2.3}$$

Assuming that the cross-section and the ^{14}N concentrations are constant, it is possible to get a first-order estimate of the past cosmic ray intensity, given the decay rate and the time since the exposure of ^{14}N to the cosmic rays. The accuracy of such calculations may be degraded by variations in the atmospheric ^{14}N concentrations and uncertainties associated with the dating of the isotopes. The ^{14}N concentration is not the limiting factor in equation (2.3) as the atmospheric nitrogen concentration is 78% by volume, and it is also often assumed that the cosmic ray energy spectrum is constant. However, the entire chain of events from the cosmic ray flux intensity, the nitrogen concentrations, the cross-sections, the transport of the cosmogenic isotopes down to the surface, the uptake in the biosphere or entrapment in ice, decay and migration during storage, and the representativity, accuracy and reliability of the measurements must be considered.

It has been recognised that ^{14}C concentrations may be influenced by climatic factors (van Geel and Mook, 1989) as well as variations in the geomagnetic field and solar activity (Wagner *et al.*, 2001). Therefore, any inference about past climatic anomalies must be supplemented by other evidence, such as from bog and lake deposits. It is important to keep in mind that there are factors other than solar activity that may influence the δ^{14}C ratio in the palaeo records. Geomagnetic variations may affect ^{14}C production in a similar way as the variations in the interplanetary magnetic field (IMF). Furthermore, the biological uptake of ^{14}C may be affected by climatological conditions. It is also important to be aware of the dangers of circular argument when using these proxies to study the relationship between solar activity and climate.

The δ^{14}C ratio in the deep ocean is often low as this water mass has not been exposed to the atmosphere recently. When there is substantial upwelling activity whereby the atmosphere comes into contact with deep water, an equilibrium process may take place lowering the atmospheric isotope ratio while raising the δ^{14}C in the upwelled water.

Spectral analysis of ^{14}C isotope ratios, according to Lean and Rind (1998), suggests variations with timescales of 88 years (Gleissberg) and approximately 210- and 2300-year cycles. Cycles with a 2500-year periodicity have also been identified in glacier proxies, marine records and ice core readings. Lean and Rind (1998) argue that these periodic variations are unlikely to be caused by internal variability, but suggest that increased ^{14}C production is related to lower temperature. The low isotope δ^{14}C-ratio (Spörer 1450 to 1550 and Maunder 1645 to 1715 minima), they remark, coincide with the "Little Ice Age" 1450 to 1850. However, the events known as the "Little Ice Age" and the "Medieval Warm Period" are not well defined, nor do they stand out prominently in all of the climatic records. A recent study by Solanki *et al.* (2004) based on ^{14}C levels and Beryllium-10 suggested that the 10-year averaged sunspot number has been higher in the last 60 years than any other period during the last eight millennia. This observation had also been reported by Usoskin *et al.* (2003). However, since the past partition of ^{14}C between deep ocean, ocean mixed layer, biosphere, and atmosphere really is unknown, Solanki *et al.* had to base their conclusions on the assumption of no major change in the state of these ^{14}C reservoirs. Therefore, these proxy data may not be appropriate for inferring major climatic fluctuations in the past, as this could involve circular logic. The contention that the Sun has been more active in the last 60 years than previously appears to inconsistent with the finding by Benestad (2005a) based on SCL that the sun appeared to be more settled after 1900.

2.4.1.2 Carbon-13

For ^{13}C, on the other hand, the photosynthesis process favours carbon-12 to the stable carbon-13 isotope, resulting in a lower ^{13}C/^{12}C ratio in biological compounds. Hence more negative δ^{13}C for organic carbon compounds implies a higher rate of photosynthesis.

2.4.1.3 Beryllium-10

Galactic Cosmic Ray (GCR) proxy records can be obtained from beryllium-10 (^{10}Be) trapped in polar ice caps and carbon-14 (^{14}C) in trees. Isotope records of ^{10}Be can be obtained from ice core data (Beer, 2000). For proxy data of ^{10}Be, there are two different quantities: ^{10}Be concentration and ^{10}Be flux. The ^{10}Be *concentration* is the amount per volume unit of ice whereas *flux* refers to the deposited amount per surface unit, per year. The ^{10}Be records depend on the cosmic ray flux as well as the concentrations of ^{14}N or ^{16}O. The palaeo records usually involve ^{10}Be trapped in ice, which implies that the ^{10}Be concentrations depend on precipitation. These proxies therefore can be used to study both climatic changes as well as solar activity levels. Once again, care must be taken when using ^{10}Be for solar–terrestrial studies, as the influence of both solar and internal climatic origins can easily lead to circular argumentation. Another source of ^{10}Be may be from the solar wind (Nishiizumi and Caffee, 2001).

$$^{14}\mathrm{N} + n(p) \rightarrow {}^{10}\mathrm{Be} + 3p(4p) + 2n(1n) \tag{2.4}$$

$$^{16}\mathrm{O} + n(p) \rightarrow {}^{10}\mathrm{Be} + 4p(5p) + 3n(2n) \tag{2.5}$$

Lean and Rind (1998) compared the solar irradiance reconstructions from Hoyt and Schatten (1993) and reconstructions inferred from ^{10}Be and ^{14}C and remarked on their differences. They observed that these differences reflect the large uncertainties in reconstructing historical solar irradiances from a limited solar monitoring database. Furthermore, they thought that because of the rudimentary knowledge of the physical processes involved and the incomplete understanding of the solar origins of these variations, these differences cannot be resolved. Usoskin *et al.* (2003) suggested that discrepancies between observed and sunspot number estimated from ^{10}Be isotope record could be due to local climatic effects. Furthermore, the ^{10}Be record can also be contaminated by variations in the geomagnetic field. They argued that the ^{14}C record was not sensitive to climatic fluctuations because CO_2, which contains carbon-14, is a globally well-mixed gas. They also claimed that the reconstructed sunspot number derived from the ^{14}C record exhibited a correlation of 0.83±0.07 with the observed ones when a 20-year lag was accounted for.

2.4.1.4 Oxygen-18

There are other isotopes, such as oxygen-18 (^{18}O), that are stable and occur naturally. Harold Urey suggested in 1947 that oxygen isotopes in fossilised sea shells may provide information on past temperatures. The ^{18}O/^{16}O-ratio in biologically produced calcium carbonate in the ocean (corals) is often used to infer the ^{18}O/^{16}O-ratio in the sea-water. The difference between the sea-water ratio and the carbonate ratio is a well-defined function of temperature, and records of these isotopes can be used to estimate past sea temperatures. High ratio is associated with low temperature. The conversion from the isotopic ratios in calcium carbonate assumes that the oxygen-isotope composition is known for the sea-water.

During evaporation, the lighter ^{16}O is evaporated faster than the heavier isotope. As a result, the sea-water tends to have high ^{18}O concentrations, but snow and ice formed from the water evaporated from the oceans have low ^{18}O content. When a substantial part of the water is trapped in ice-sheets, the oceans have high ^{18}O/^{16}O-ratios, and this proxy has therefore been used as an indicator of ice ages. One source of error is that the fluids in sediments can shift the oxygen isotopic compositions away from the original value during and after fossilisation.

Around the mid-1980s a hypothesis based on ^{18}O was proposed stating that despite the warm Cretaceous to late-Eocene (67 to 35 million years ago), the atmospheric CO_2 concentrations (1000 ppm[12]) were higher than now, the polar regions were warm, but the tropics were about 10°C cooler than now (Schwarzschild, 2001). The ^{18}O from fossils of planktonic foraminifera from this warm period, living near the ocean surface, indicate high ^{18}O levels, and hence cool temperatures. This is known as the "cool-tropics paradox". It has recently been shown by Paul Pearson and co-workers that the ^{18}O-level may be misleading due to subsequent re-crystallisation at the bottom of the sea long after the plankton had died, fallen down to the bottom, and fossilised.

2.4.2 Geomagnetic field measurements

Geomagnetic measurements[13] may capture magnetic field variations which arise from the coupling between the terrestrial magnetic field and interplanetary field lines carried with the solar wind. The measurements of the magnetic fields are used to derive various indices which are related to specific physical phenomena. For instance, the *K-index* is designed to measure the solar particle flux through its effect on the geomagnetic field. The K-index may be contaminated by "noise" from interactions between electromagnetic processes in the magnetosphere and the ionosphere. Currents induced in the Earth's surface, magnetosphere and field-aligned currents may also interfere with the observations. The geomagnetic field itself undergoes changes, and care must be taken to filter out this contribution when these measurements are used in studies of slow long-term changes in solar magnetism.

Another entity is the *aa-index* (Mayaud, 1972) which has been used as a proxy data for solar magnetism by Lockwood *et al.* (1999) (see Section 7.10.5.1). This index suffers from similar weaknesses as the K-index, although some of the noise may be eliminated by subtracting one measurement from another at the opposite side (antipodal) of the Earth. The aa-index can be derived from measurements from the close-to-antipodal Greenwich and Melbourne observatories which have been in operation since 1867. The data from Abinger-Hartland and Toolangi observatories have superseded the data from Greenwich and Melbourne. The aa-index is defined as the 3-hour average of K-indices from antipodal observatories after the K-amplitudes have been transformed into gammas. The K-indices show prominent diurnal variations with a local night-time maximum, but there are also annual variations with a maximum in each hemisphere at the summer solstice.

The aa-index may nevertheless be sensitive to small-scale changes to the local geomagnetic field. Figure 2.5 shows the temporal evolution of the aa-index, and although the curve appears noisy, the aa-index has an 11-year signal. There is furthermore a long-term trend in the record, as documented by Lockwood *et al.* (1999). An increase in the magnetic activity, as shown in Figure 2.5, may have

[12] Parts per million in terms of volume.

[13] For instance, see Mayaud (1972).

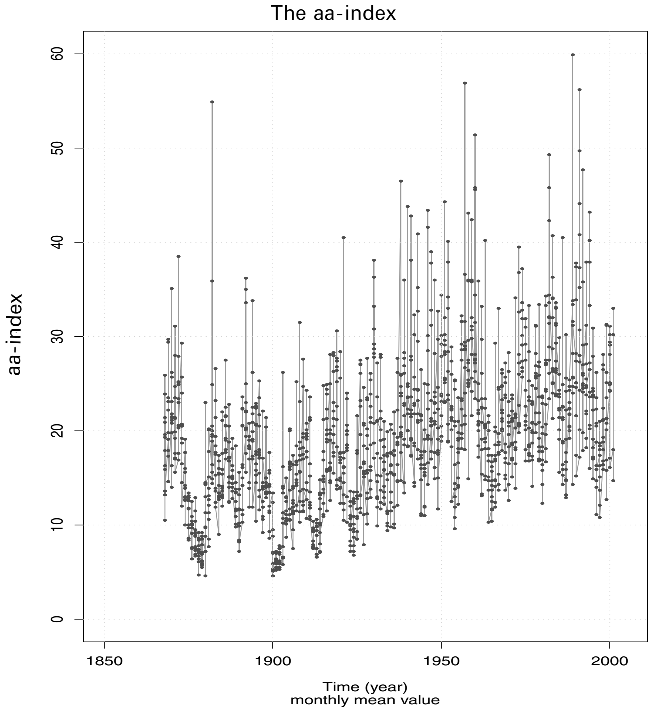


Figure 2.5. A time-series showing the aa-index. This index is derived from the difference between magnetograph measurements made at two locations located at opposite sides of the planet. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

implications for Earth's climate (see Section 7.10.5.1). It is important to ask why and how there is a trend in the aa-index. Any long-term trend in solar magnetism must have a physical explanation, such as a long-term change to the solar dynamo or the intensity of the solar wind. Figure 2.6 shows that the sunspot number may explain the 11-year variations in the aa-index, but not the long-term increase. Hence, a trend in the magnetic activity also calls for an explanation as to why a similar trend is not seen in the sunspot record. Any hypothesis stating that the long-term trend in the aa-index is related to a long-term trend in Earth's climate must demonstrate that the trend in the aa-index reflects real changes in solar magnetism.

2.5 SPACE-BORNE SOLAR OBSERVATIONS

Modern observations of the Sun are also based on satellite-borne instruments and space probes which eliminate the problems associated with atmospheric disturbances as the observations are no longer made through the atmosphere. Magnetism may still affect instruments if they are not thoroughly shielded, as the electronics may be subject to a Hall effect.

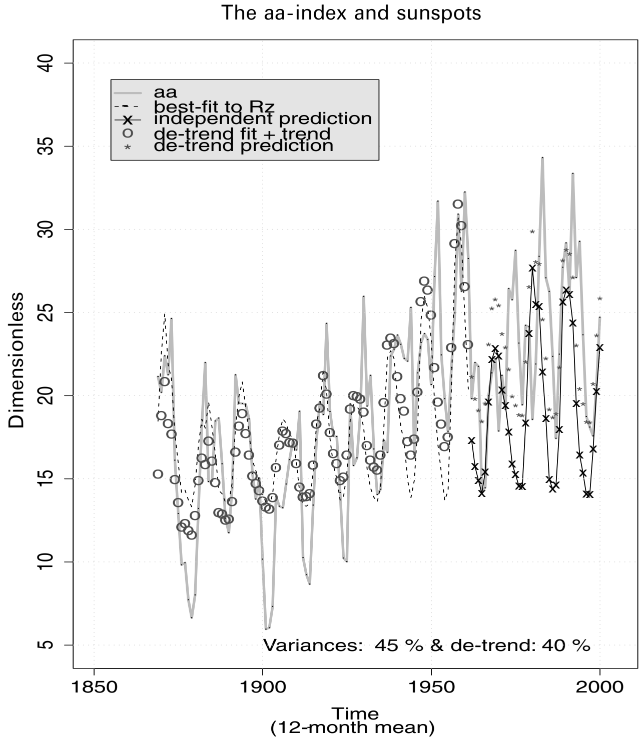


Figure 2.6. Predicting the annual mean aa-index using the sunspot number. The low aa-values in the early part of the record and the most recent high numbers cannot be predicted by the sunspot number. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

The *Wind* probe is a NASA mission launched in November 1994. Its objectives are to (i) explore the ejected solar plasma,[14] energetic particles, and magnetic fields for (terrestrial) magnetospheric and ionospheric studies and (ii) to study the interaction between the solar wind and Earth's geomagnetic field.

The *Ulysses* spacecraft is another NASA/ESA mission to explore the Sun's polar regions. It was launched from the Space Shuttle in October 1990, and carries instruments designed to study the solar wind. Measurements are made of the particles, magnetic field, and electromagnetic radiation at a range of wavelengths ranging from radio waves to gamma radiation.

The *TRACE* mission (acronym for "Transition Region And Coronal Explorer") is a solar research satellite designed to study the connections between the fine-scale magnetic fields on the Sun and associated solar plasma structures. The observations focus on the photosphere, the transition region, and the corona, and involve temperature measurements with high spatial (one second of arc) and temporal

[14] Plasma is ionised matter, in the Sun usually a cloud of free protons and electrons.

(one second between different wavelengths) resolution. The goal of the project is to obtain coronal and transition region thermal topography and to estimate the 3-D magnetic field structure, temporal evolution of photospheric flow, and time-dependent coronal fine structures. The craft was launched in April 1988.

The *SOHO* mission, which is a NASA/ESA project, stands for "SOlar Heliospheric Observatory". The spacecraft was launched in December 1995. This spacecraft carries an array of instruments: scanning ultraviolet spectrometer (SUMER), extreme ultraviolet spectrometer (CDS), extreme ultraviolet imaging telescope (EIT), ultraviolet chronograph spectrometer (UVCS), large-angle chronograph for visible light (LASCO), whole-sky Lyman alpha mapper (SWAN), solar wind composition and extreme UV flux probe (CELIAS), low-range energetic particles (COSTEP), high-range energetic particles (ERNE), global low-degree velocity oscillation (GOLF), solar irradiance and luminosity oscillation (VIRGO), and Michelson Doppler Imaging of solar oscillation (MDI/SOI). SOHO uses new ways of probing the solar interior by helioseismology, where acoustic waves seen on the Sun's surface are used to derive conditions about the solar interior as these are caused by pressure fluctuations in the Sun.

3

The physical properties of the Sun

3.1 SYNOPSIS

The Sun is believed to be around 4.6 billion years old, and about half of the original amount of hydrogen (the fuel used in the energy production) in its core is thought to have been used up by now. It is estimated that the Sun is mid-way in its life. Since its formation, the solar energy output is believed to have increased by 30 to 40%. According to modern solar theory, when the hydrogen in the Sun's core burns up, the Sun shrinks and produces an increasingly denser core. Higher temperatures accompany the higher core density, and as a result, the thermonuclear energy production is boosted. This theory therefore implies that the Sun becomes brighter as it evolves.

In order to understand how solar activity may affect Earth's climate, it is important to know about the most basic features of the Sun.

3.2 THE SUN AS A STAR

3.2.1 Solar size and mass

Our Sun is one of many so-called G2 stars.[1] The Sun has a diameter that is measured to be 1.39×10^9 m and a mass estimated to be 1.989×10^{30} kg. The solar mass and parallax have traditionally been derived indirectly from studies of the orbit of the asteroid *Eros*. The parallax is the mean angle subtended (in arc-seconds) by Earth's

[1] Stars are classified by a letter and a number describing the nature of their spectral lines and their surface temperature. The types are denoted by the letter: O, B, A, F, G, K, and M, where O stars are the hottest and M the coolest. The numbers are simply subdivisions of the major classes. The Sun comes under the G2 category, to which there are estimated to be 100 billion other stars belonging.

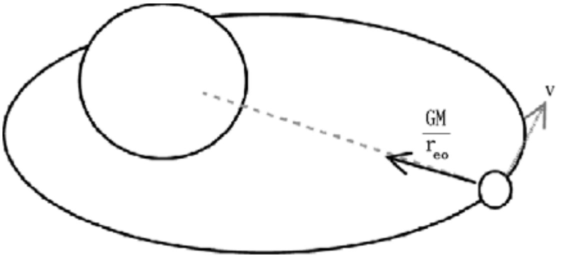


Figure 3.1. A schematic figure illustrating the balance of forces for an orbiting object. The gravitational pull is responsible for the centripetal force.

equatorial radius at the axis between the Sun and the Earth (Kuiper, 1953, p. 17), and is found from the perturbation of Earth's gravitational force on nearby objects.[2] The solar mass is found as a by-product of the parallax studies. An approximate estimate of the solar mass can also be derived considering the centripetal and gravity forces: these must be equal if a satellite is to stay in a circular orbit.

$$\frac{GM_s m_e}{r^2} \approx m_e \frac{v^2}{r} \tag{3.1}$$

After cancellations and rearrangements, the solar mass M_s can be expressed in terms of Earth's mean velocity v, distance to the Sun ($r = 1.50 \times 10^{11}$ m) and the universal constant of gravity ($G = 6.67 \times 10^{-11}\ \mathrm{N\,m^2\,kg^{-2}}$). A simple way of estimating an approximate value for the mean distance between the Sun and the Earth is to assume that Earth's orbit is circular and using the relationship between the circumference of a circle and the radius: $s = 2\pi r$. The distance along the orbit is the product of the Earth's speed and the length of the year: $v\tau = 2\pi a_0$, and the mean orbital radius is $a_0 = v\tau/(2\pi)$. The velocity of the Earth can be measured using the stars as a reference, for instance by measuring the annual variations in the Doppler shift in the light from the stars. A mean orbital speed of $2.98 \times 10^4\ \mathrm{m\,s^{-1}}$ gives a mean orbital radius of 1.50×10^{11} m. When the mean orbital speed and radius are known, the solar mass can be estimated according to:

$$M_s \approx \frac{v^2 r}{G} = \frac{4\pi^2 r^3}{G\tau^2} = 2.00 \times 10^{30}\ \mathrm{kg} \tag{3.2}$$

The first-order estimate is in good agreement with the value derived from the parallax studies. The fact that there are several objects affecting the gravitational field and that Earth's orbit is elliptical (not circular) complicates the precise estimation of the solar mass.

3.2.2 The solar rotation

The outer part of the Sun rotates with different angular velocity near the equator and

[2] See Webb (1999) for detailed discussion on astronomical measurements.

the poles, and not as a solid body. The equatorial rotation rate is 25.4 days per cycle, whereas the angular velocity near the poles is slower: 36 days per rotation. From a terrestrial frame of reference, the bulk of the Sun appears to rotate once about every 27 days.

3.2.3 The solar material

The solar material is thought to be mostly ionised matter, also known as plasma. The plasma is a conductor, and electromagnetic theory states that there cannot be a steady electrical field inside a conductor because the charges will move to cancel the field. A classical example is Faraday's cage, which shields its interior from electrical disturbances outside. An electric current is nevertheless believed to be present in the solar interior to explain the Sun's general magnetic field. This paradox is still problematic to explain, however; the most recent theory of the Sun's magnetic field is based on the Earth's geomagnetic model, which will be discussed in more depth in Section 3.3.

3.2.4 The Sun's core

The Sun may be thought of as being made up of several regions. The innermost part ($r = 0$ to 175,000 km) of the Sun is called the *core*. The core is believed to extend out to around 25% of the Sun's total radius, and is the region where most of the energy production takes places. The temperatures in the core are thought to be as high as 1.56×10^7 K. The pressure in the Sun's core is calculated to be 2.5×10^{11} atmospheres (1 atm = 1013 hPa).

3.2.4.1 The mass–energy relation

The energy production of nuclear reactions is mainly through a transfer from mass to energy. Albert Einstein became world-famous for the discovery of the equivalence between mass and energy, and the relationship between these forms: $E = mc^2$. As the products of the nuclear reaction have smaller mass in total than the initial elements, the mass-discrepancy equals the amount of energy released by the reaction.

3.2.4.2 The nuclear reactions in the Sun

The Sun's energy is produced through thermonuclear processes in the Sun's core, where hydrogen atoms (H) fuse to create helium (He). Nuclear reactions are believed to take place between two colliding protons[3] to form deuterium (hydrogen atom consisting of one proton and one neutron (n)[4]):

$$^1\mathrm{H} + {}^1\mathrm{H} \Rightarrow {}^2\mathrm{H} + \beta^+ + n + \text{energy} \tag{3.3}$$

The rate of reaction and energy production in equation (3.3) depends upon the

[3] Protons are one basic type of atomic particles with charge $+1e$. A hydrogen atom normally consists of one proton only.

[4] A neutron is an atomic particle with no charge.

cross-sections (probability of collision) and the potential barrier (repulsive force between the two protons) through which they have to penetrate (forces to overcome) and the particles' kinetic energy (speed). The reaction produces β^+ radiation (positron emission[5]) and a neutrino[6] in addition to deuterium. One of the products of this reaction is deuterium ($^2\mathrm{H} \equiv \mathrm{D}$). The deuterium will react quickly with other protons in the central regions of the Sun to form helium. This process releases energy in the form of gamma[7] rays. The final reaction that results in helium atoms with atomic weight 4 (^{4}He) is where two helium-3 nuclei combine: $^3\mathrm{He} + {}^3\mathrm{He} \Rightarrow {}^4\mathrm{He} + 2{}^1\mathrm{H}$. The total energy production of such a cycle, i.e. for the formation of each ^{4}He atom, is estimated to be 4.48×10^{-12} J (4.48×10^{-5} ergs[8] or 28 MeV). The gamma radiation is absorbed locally where the solar material is opaque and transformed into local thermal energy.

Helium atoms cannot form directly from the combination of two ^{1}H atoms since ^{2}He nuclei do not exist. The process of creating helium involves two more reactions. One of these is between two deuterium atoms to form tritium ($^3\mathrm{H} \equiv \mathrm{T}$): $\mathrm{D} + \mathrm{D} \rightarrow \mathrm{T} + p + 4\,\mathrm{MeV}$ kinetic energy. One deuterium and one tritium atom can combine to form helium: $\mathrm{T} + \mathrm{D} \rightarrow {}^4\mathrm{He} + n + 17.6\,\mathrm{MeV}$ kinetic energy.

There are also other nuclear reactions in the Sun between protons and trace elements, such as beryllium (^{8}Be), carbon-12 and carbon-13 (^{12}C and ^{13}C), nitrogen (^{13}N, ^{14}N, and ^{15}N), and oxygen (^{15}O):

$$\begin{aligned}
&{}^{12}\mathrm{C} + {}^1\mathrm{H} \Rightarrow {}^{13}\mathrm{N} + \gamma \\
&\qquad {}^{13}\mathrm{N} \Rightarrow {}^{13}\mathrm{C} + \beta^+ + \text{neutrino} \\
&{}^{13}\mathrm{C} + {}^1\mathrm{H} \Rightarrow {}^{14}\mathrm{N} + \gamma \\
&{}^{14}\mathrm{N} + {}^1\mathrm{H} \Rightarrow {}^{15}\mathrm{O} + \gamma \\
&\qquad {}^{15}\mathrm{O} \Rightarrow {}^{15}\mathrm{N} + \gamma^+ + \text{neutrino} \\
&{}^{15}\mathrm{N} + {}^1\mathrm{H} \Rightarrow {}^{12}\mathrm{C} + {}^4\mathrm{He} + \gamma
\end{aligned} \tag{3.4}$$

3.2.4.3 *The neutrino mass-deficiency paradox*

Neutrinos escape from the Sun without contributing to the Sun's luminosity (electromagnetic radiation), and neutrinos do not interact readily with matter. They were long believed to be massless, but discoveries made by a team of US and Japanese physicists in 1998 (Wolfstein, 1998) suggest that they have a mass of about $0.1\,\mathrm{eV}/c^2$. By comparison, the electron mass is $0.5\,\mathrm{MeV}/c^2$, or about a million times as heavy. The neutrinos have no charge, but they have a spin. Since the neutrinos (responsible

[5] A positron may be thought of as an electron, but with a positive charge.

[6] A neutrino is a small chargeless particle.

[7] Gamma rays are electromagnetic radiation of very high frequency in the range $\nu = 10^{19}$–$10^{21}\,s^{-1}$

[8] 1 joule = 10^7 ergs.

for around 5% of the total energy) escape from the Sun without interaction, they are not regarded as a part of the solar energy production.

There has until recently been an unsolved problem in solar physics regarding the solar neutrinos. Less than half of the neutrinos expected from the solar models were detected.[9] This puzzle was reported to have been solved in the July 2001 issue of *Physics World*: the neutrinos produced inside the Sun change "flavour"[10] on their journey to Earth. According to the standard solar model, the Sun produces electron neutrinos when boron-8 undergoes beta decay, but the two other flavours are not created. It has been shown that the disappearance of neutrinos of one flavour is accompanied by the appearance of another. All the neutrino flavours undergo neutral-current reactions, and all are susceptible to "elastic scattering" off electrons in the detectors. But only the electron neutrino can take part in charged-current reactions. The previous detection has only been sensitive to the electron neutrinos (1.75 million neutrinos $\text{cm}^{-2}\,\text{s}^{-1}$).

3.2.4.4 The energy production

The energy production associated with the reactions in equation (3.4) is estimated to be 4.48×10^{-12} J for every ^{4}He nucleus formed. The total energy production (ϵ) associated with these two cycles depends on their respective reaction rates and the total mass of the active region in the Sun. The reaction rate is calculated by the number of collisions per unit volume, determined by experimentation and calculations. The total energy production, which is the sum of the energy production (ϵ) of the two cycles is a function of temperatures and ranges between 2.4×10^{-10} and 5.3×10^{-7} J ($\epsilon = 0.0024 - 5.3\,\text{erg g}^{-1}\,\text{s}^{-1}$) for temperatures of 5×10^6 K to 30×10^6 K respectively. It is thought that the proton–proton process is the dominant energy source in our Sun, where the central temperature is in the range 13 to 16×10^6 K and the central density is 8.6×10^3 to $94 \times 10^3\,\text{kg/m}^3$.

If the energy is radiated isotropically, then it is possible to estimate how much energy the Sun produces by making energy flux (total solar irradiance) measurements along Earth's orbit (the "solar constant" is $1370\,\text{W/m}^2$) at a mean distance a_0 from the Sun. The mean orbital radius is estimated to be 1.50×10^{11} m.

The total energy production rate can be estimated using the divergence theorem,[11] which in this case is equivalent to multiplying the flux density observed at the mean distance between the Earth and the Sun with the total surface area of a sphere with a radius with similar length as this distance. Thus the total solar energy production can be estimated through the expression for the energy density:

$$e_{\text{tot}} = 1370\,\text{W/m}^2 \times 4\pi a_o^2 \tag{3.5}$$

[9] The Sudbury Neutrino Observatory, SNO, detects of the order of 10 neutrinos per day.

[10] Elementary particles may have several properties, which can be categorised as charge, mass, spin, and flavour. A neutrino may have three different flavours: electron neutrinos, muon and tau.

[11] Also known as Gauss theorem: $\int\int F \cdot n\,dS = \int\int\int \nabla \cdot F\,dV$ where the former term describes a flux through a surface and the latter is a source term.

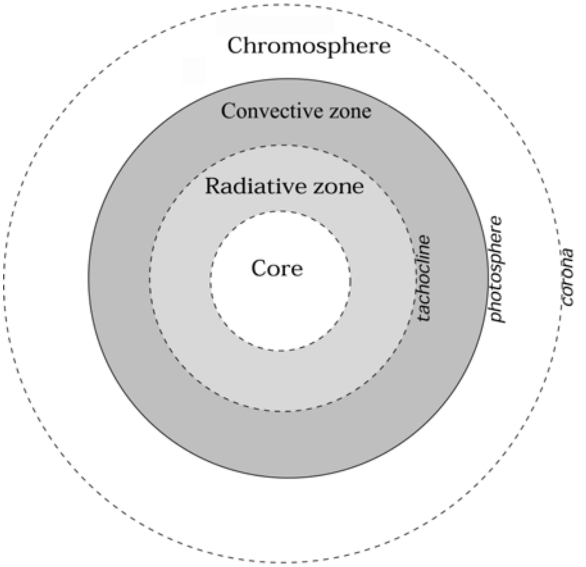


Figure 3.2. A schematic showing the basic features of the Sun: the core; radiative zone; convective zone; the chromosphere; and the corona.

The energy production is, according to equation (3.5), 3.87 × 10^{26} W. In other words, the solar energy production is about 400 billion billion megawatts, and each second, 7 × 10^8 tons of H is transformed into 6.95 × 10^8 tons He per second and 5 × 10^6 tons/s (1 ton = 1016 kg) energy production equivalent. The Sun's total radiative output, or luminosity, is approximately constant.

3.2.4.5 Energy transport

Outside the core is the radiative zone (region: $r = 0.25R_s \rightarrow \approx 0.8R_s$). The energy transfer within the radiative zone is by photons (gamma rays). In the interface zone between the radiative zone and the outer convection zone, the primary energy transfer changes gradually from photon flux to convection by vertical motion. The convection zone is the outer part of the opaque region of the Sun, and few photons manage to penetrate through this region. The basic parts that make up the Sun are illustrated in Figure 3.2.

3.2.5 The photosphere

The photosphere is the surface of the Sun that most of us are familiar with, and this is what we normally see if we look at the Sun. The photosphere is thought to be about 100 km deep. Features such as dark sunspot pores, sunspots, bright faculae and granules are observed on the photosphere.

3.2.5.1 Temperature

The temperature of the photosphere determines the properties of the solar black body[12] thermal continuum radiation (Section 2.2.2). A good discussion on the black body radiation is given by Fleagle and Businger (1980). The radiance can be expressed in terms of the wavelength (λ):

$$S(\lambda) = \frac{2hc^2}{\lambda^5} \frac{1}{e^{\frac{hc}{k\lambda T}} - 1} \tag{3.6}$$

where $h = 6.63 \times 10^{-34}$ J s is the Planck constant and $c = 3 \times 10^8$ m/s is the speed of light in vacuum. The integral over all wavelengths gives the total energy flux, or the Stefan–Boltzmann law:

$$Q = \sigma T^4 \tag{3.7}$$

The wavelength associated with maximum radiance can be found using Wien's displacement law:

$$\lambda_m = \frac{\alpha}{T} \tag{3.8}$$

where $\alpha = 2897.8 \times 10^{-6}$ m K. This expression is useful for estimating the temperature of remote black bodies, and can be used to infer the temperature at the Sun's surface. The spectrum of the solar irradiance is shown in Figure 3.3. The wavelength corresponding to the peak[13] is $\lambda_m = 450.4$ nm (4.504×10^{-7} m) (vertical dashed line), which according to equation (3.8) would imply a solar black body temperature of:

$$T = \frac{2897.8 \times 10^{-6}\,\text{m K}}{4.504 \times 10^{-7}\,\text{m}} = 6434\,\text{K} \tag{3.9}$$

The solar surface temperature can also be inferred from the energy production (E) and equation (3.7) if the solar radius ($R_\odot$) and hence the Sun's surface area (A_S) are known: $E = QA_S$. This expression can be written out in the form: $3.87 \times 10^{26}\,\text{W} = \sigma T^4 4\pi R_\odot^2$. Hence, the solar temperature can be estimated from:

$$T = \left(\frac{3.87 \times 10^{26}\ \text{W}}{4\sigma\pi R_\odot^2}\right)^{\frac{1}{4}} = 5770\,\text{K} \tag{3.10}$$

The differences in these two estimates illustrates how difficult it is to obtain accurate values for quantities such as the solar temperature (which is probably closer to 5780 K).

Figure 3.4 compares the energy distribution as a function of wavelength of the solar radiation and the long-wave infrared radiation emitted from Earth's surface. Since the Sun and the Earth radiate in two entirely different parts of the wavelength

[12] A "black body" is an object that absorbs all incident radiation and none of the radiation is reflected. Black bodies have unit emissivity because they emit energy at the same rate as it is absorbed.

[13] This estimate derived from the spectrum may be inaccurate due to the presence of line spectra on top of the continuous spectrum; however, this estimate is a good approximation.

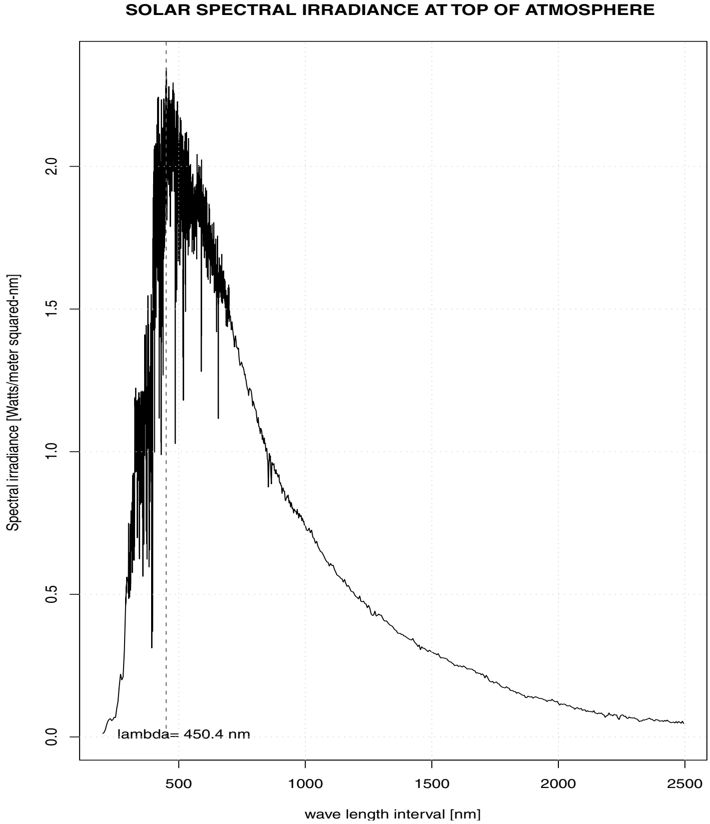


Figure 3.3. Extraterrestrial solar spectral irradiance at the Earth's mean solar distance UV-to-visible. Spectro-radiometric measurements were performed during eleven research flights on board a NASA CV-990 aircraft at altitudes between 11.6 km and 12.5 km. Precision of the measurements was better than ±1%. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA (see also Arvesen *et al.*, 1969).

spectrum, the radiation of either may be treated separately from each other. The difference in the wavelength also has an implication for the second law of thermodynamics: the total entropy[14] on Earth must increase as the energy leaving the Earth is of "lower grade" (associated with lower temperature) than the incoming energy.

3.2.5.2 Photospheric granules, bright rings and convection cells

There have in the past been alleged observations of bright rings around most of the medium-sized sunspots (Bray and Loughhead, 1964, p. 67). These rings were

[14] Entropy is defined as $S = Q/T$, where Q denotes the energy flow and T the temperature.

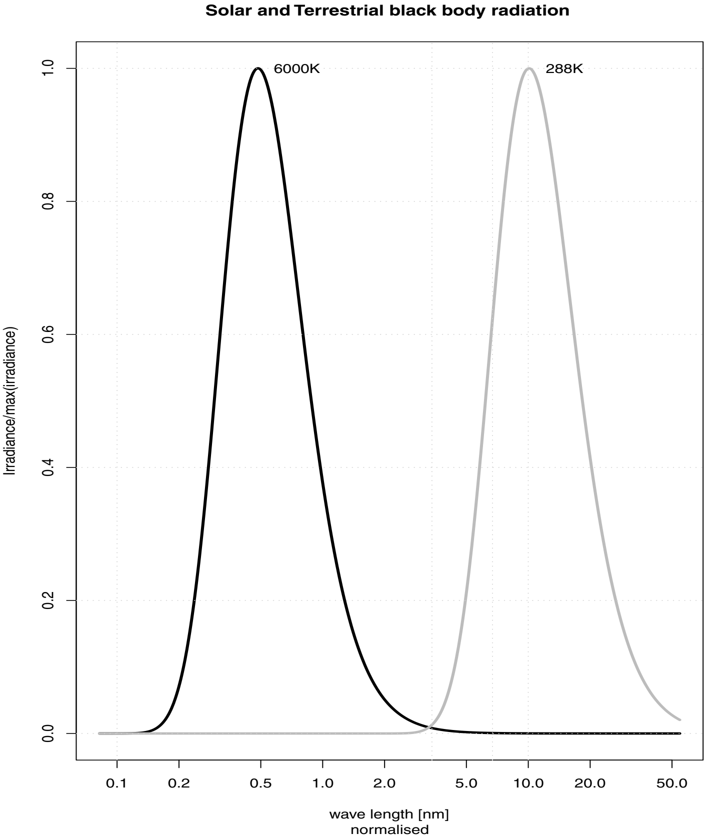


Figure 3.4. Standardised theoretical black body emission spectra for two bodies with respective temperatures of 6000 K and 288 K. The overlap of these two curves is minimal, implying that the short-wave radiation is almost entirely from the Sun and the long-wave radiation comes from the Earth.

supposedly about 2 to 4% brighter than the rest of the photosphere, but the atomic emission lines from Fe (3820.5 Å) and Ca II (K_{1R} 3937 Å and H_{1R} 3931 Å) are about 10% brighter and the K_2 portion of the K-line is up to 50% brighter than the surroundings. The width of these regions were supposed to be about the same as the sunspot radius, but the extra emission from the bright rings was believed to be insufficient to compensate for the reduced black body emission, "flux deficit", caused by the sunspots. Therefore, it has in the past been postulated that the Sun radiates slightly less energy during sunspot maximum, contrary to conclusions from more recent observations. The existence of such bright rings around the sunspots "damming up" energy beneath the sunspots has now been questioned (Thomas and Weiss, 1992).

Studies of the photosphere have revealed the presence of small granules which are interpreted as a sign of convection. Narrow bright rings in the violet and blue-violet spectral bands have also been observed (Hale and Ellerman, 1906) on the Sun's surface partly or completely encircling the sunspots, but any explanation for why and how these form has been elusive. These bright features are rarely seen in the green-orange region of the spectrum. Waldmeier suggested in 1939 that a sunspot group normally is surrounded by a bright zone, which can take the form of a bright ring when the group is reduced to a sunspot. Maltby proposed in 1960 that the bright rings are unusually intense around sunspots with high Evershed velocities, but Waldmeier argued that the emission comes from a deeper layer of the photosphere (Bray and Loughhead, 1964, p. 67). Such bright rings, however, are not mentioned by Thomas and Weiss (1992), and the presence of these do not seem to have been established for sure as it is hard to find any reference to these in the most recent literature. One old hypothesis is that they were a product of diverted energy flux from the base of the sunspot. There is nevertheless little doubt about the presence of bright areas, known as *facular brightening* (Lean, 2000; Fröhlich and Lean, 1998a) and these are associated with high sunspot activity, although it is not entirely clear in the literature whether these are related to such "bright rings".

The facular brightening is estimated using ground-based measurements of the He 1083 nm line as a proxy and using the correlation with the ACRIM I data from which the sunspot blocking has been removed.

Dark lanes have been seen between granules, and it is understood that these may be a sign of rising loops of magnetic field lines. Bray and Loughhead (1964) were among the first to witness the process of sunspot pores forming, and reported that the granules aligned before the dark lanes appeared (Bray and Loughhead, 1964, p. 68).

3.2.6 The chromosphere

Outside the photosphere lies the chromosphere, which is thought to have extremely high temperatures (about 20,000 K). Because of the high temperatures in the chromosphere, hydrogen can emit light with a reddish colour (H α emissions). Energetic light at wavelengths shorter than 160 nm have their source in the chromosphere or corona (White, 2000). The extreme conditions also ionise calcium (Ca II) in the violet spectrum (many electrons are stripped off the Ca atoms).

Outside the chromosphere is the narrow transit region separating the chromosphere from the corona. The light emitted from this region is dominated by ionised carbon (C IV), oxygen (O V) and silicon (Si I) with three electrons stripped off. The latest missions studying the transition region include TRACE and SOHO.

Chromospheric holes are regions in the chromosphere with low densities. Bright regions in the chromosphere are known as *plage*. "Spicules" is a term used to describe a grass-like pattern observed in the solar atmosphere.

3.2.7 The corona

The corona is regarded as the outermost atmospheric region of the Sun, and can be seen during total solar eclipses. It has an extremely high temperature ($\approx 10^6$ K), and it

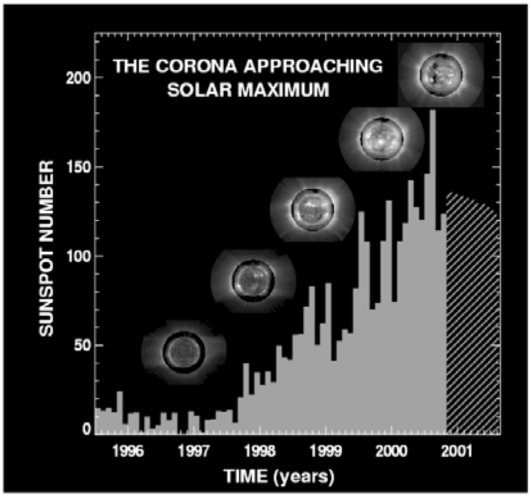


Figure 3.5. The evolution of the corona shape from near sunspot minimum in 1997 to the sunspot maximum in 2001. Courtesy: SOHO/MDI (ESA & NASA).

is still not entirely established why it is warmer than the photosphere. Hypotheses about magnetic warming have been put forward. A bright emission line emanates from the corona, stemming from completely ionised substances such as hydrogen, carbon, nitrogen, and oxygen. Iron and calcium may still have some of their innermost electrons left. The corona also emits X-rays. The appearance of the corona changes with the sunspot cycle, exhibiting streamer features near the solar equator during sunspot minimum and a more isotropic structure during maximum (Figure 3.5). The strongest and weakest flattening of the corona precede the minimum and maximum sunspot number by about two years. The differences in the coronal appearance are illustrated nicely by Diego (1999), but Figure 3.5 also illustrates the main differences between solar maximum and solar minimum.

3.2.8 The solar wind

The Sun emits a stream of charged particles called the solar wind. These particles mainly consist of protons and electrons with a mean speed of 400 km/s. The solar wind velocity varies from 300 km/s in the presence of streamers and 800 km/s in coronal holes, regions of very low green corona brightness. Sýkora *et al.* (2000) have proposed that the coronal holes may play a role in solar–terrestrial links. The source of the solar wind is the Sun's corona, where the temperature is so high that the Sun's gravitational force is insufficient to trap these particles. The kinetic energy of these particles is $1/2 \times m_e v^2 = 0.5 \times 1.67 \times 10^{-27}\,\text{kg} \times (4 \times 10^5\,\text{m/s})^2 = 1.34 \times 10^{-16}\,\text{J} = 835\,\text{eV}$ for protons and $7.28 \times 10^{-20}\,\text{J} = 0.455\,\text{eV}$ for electrons, where one electronvolt is $1\,\text{eV} = 1.60 \times 10^{-19}\,\text{J}$. The solar wind carries with it the solar magnetic field lines. The interaction between Earth's geomagnetic field and the solar wind produces storms in Earth's magnetosphere, also seen as northern or southern lights (aurora borealis and aurora australis, see Section 7.2). The spacecraft Ulysses has made a complete orbit of the Sun and mapped the speed of the solar wind, the magnetic field strength and direction and its particle composition. This solar mission has established that the solar wind is not uniform in all directions. Streamers in the corona can often be used to identify the solar poles.

3.3 THE GENERAL SOLAR MAGNETIC FIELD

For a long time, it was uncertain whether the Sun had a general magnetic field beside the strong magnetic field lines associated with the sunspots. The existence of a solar general magnetic field is now acknowledged, but it is much weaker (10–100 gauss) than those of the sunspots (2000–3000 gauss), yet stronger than the geomagnetic field on Earth (0.3–0.6 gauss, or in SI units, about 10^{-4} T (tesla)). The magnetic field of the sunspots will be discussed later. The existence of such a large scale field is not easy to reconcile with classical electromagnetic theory (Tobias and Weiss, 1999; Kuiper, 1953), and is still regarded as an unresolved topic. Circular and symmetric conductors cannot maintain steady magnetic fields (Kuiper, 1953).

The Sun's general magnetic field looks like the geomagnetic field when the sunspot activity is low, displaying a dipole structure. This general magnetic field flips during each sunspot maximum. At the sunspot maximum of 2001, the solar magnetic north pole reversed from being in the northern solar hemisphere to the southern hemisphere. One explanation for the solar magnetic reversals is that the surface flow on the Sun carries magnetic field lines from the mid-latitude sunspots to the Sun's poles and that the southward-pointing magnetic flux is transported towards the north magnetic pole and the northward-pointing magnetic flux towards the south pole. But this hypothesis has not yet been verified. The reversal of the general field should not be confused with the arrangement of the magnetic polarities in bipolar sunspot groups discovered by Hale in 1925. These sunspot group field polarities are opposite in the northern and southern hemispheres and flip at the beginning of each 11-year cycle (Bray and Loughhead, 1964, p. 242). Thus, a 'complete' magnetic cycle lasts for about 22 years and is twice as long as the 11-year Schwabe cycle. Richardson found in 1948 that only 3.1% of the bipolar groups monitored over the 1917–1946 period (5814 cases) failed to obey this rule. Recently, attention has been focused on double-peak features in the 11-year cycle with an apparent emergence of a secondary peak in the sunspot number since 1980 (T. Phillips, Science@NASA). The secondary peak appears approximately 18 months after the primary peak.

3.3.1 The decay of magnetic fields

The solar magnetic field decays because of motions within the Sun (Kuiper, 1953, pp. 544–557). High conductivity restricts velocities relative to the force lines, but also inhibits rotational symmetric magnetic field lines. The decay time of solar magnetism can be estimated according to the formula $t_0 \approx 4\pi\sigma l^2$, where l is the spatial scale of the magnetic region and σ is the conductivity. For the general solar magnetic field $t_0 \approx 1.5 \times 10^{10}$ years ($\sigma = 5.67 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$). For sunspots, the spatial dimensions are smaller, and the timescale is around 1000 years. In either case, the decay times are much longer than the observed timescale of the general solar magnetic field reversals and the sunspots' life-time. Thus, other mechanisms than just decay are responsible for destroying the sunspots.

The current theory assumes an azimuthal magnetic field deep in the solar convective zone. The field can, according to this theory, bulge outward and extend outside Sun's surface (Parker, 1997), giving rise to phenomena known as *omega loops* (Ω) of magnetic field lines.[15] The result of all these Ω-loops produces a net bipolar magnetic surface field.

3.3.2 The geomagnetic-based solar model

The most recent sunspot models are based on geomagnetic field-like theories, explaining the presence of the magnetic fields as a consequence of a dynamo-like process, like on the Earth. On Earth, it was proposed by Elsasser in 1945 that convective motions within Earth's liquid-metal core may generate magnetic fields. The observation that the magnetic field lines move around by as much as 0.18°/year may suggest that the outer regions of the liquid core rotate more slowly, and thereby stretch out the magnetic field lines and give rise to both an azimuthal and a dipole field orientation (Parker, 1997).

Steady and long-lived magnetic fields with rotational symmetry cannot be maintained by fluid motions alone (Cowling, 1934), but the discovery of 12 magnetic features in Earth's surface magnetic field lines points to the presence of 12 convective cells. Each convective cell has a central ascending region surrounded by descending fluid. The ascending fluid at the bottom of the cell rotates more rapidly than the rest of the cell, and this cyclonic motion may lift and rotate the azimuthal field and generate vertical loops of magnetic field lines. The same mechanism is believed to be responsible for the solar magnetic field. However, the presence of the general solar magnetic field is still problematic in terms of classical electromagnetical theory, and the mathematical equations describing the physical processes cannot be fully solved (Tobias and Weiss, 1999). The solar dynamo-action is located in the vicinity of the Sun's outer convective region, where small-scale convective cells produce disordered magnetic fields. Helioseismic observations have revealed strong rotational gradients that may partially explain the dynamo process (Howe *et al.*, 2000).

At the present, all physically based solar-dynamo models simulating the large-scale magnetic fields represent the small-scale turbulence by bulk parameterisation schemes, as it is still not possible to analyse both large-scale and small-scale processes in sufficient detail in such computer models, due to the high demand of computer resources. This problem also restricts general circulation models describing

[15] Ω-loops are magnetic lines which emerge from the solar interior, named according to their shape.

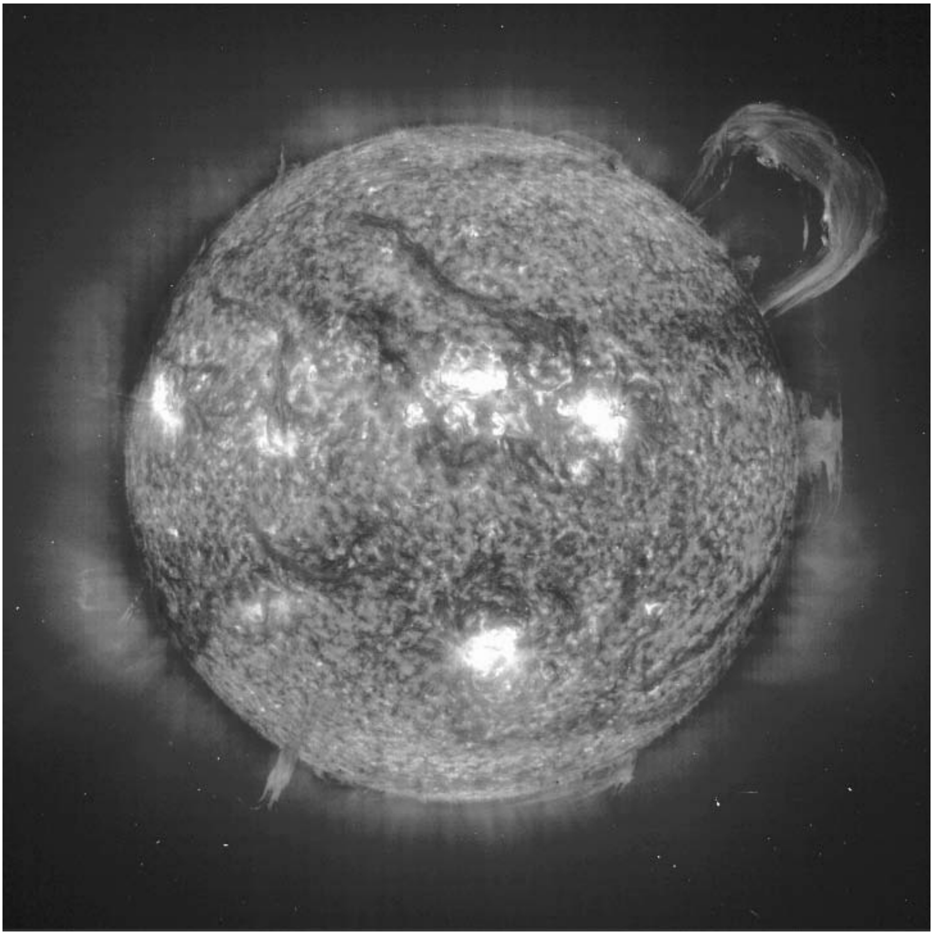


Figure 3.6. A solar prominence where the solar material is slung out of the photosphere and bright regions of the photosphere with flare activity. Courtesy: SOHO/MDI (ESA & NASA).

the Earth's atmosphere and oceans. The solar mean-field models nevertheless capture the important large-scale features, and describe both the basic solar cycle, the modulation of the amplitude, and periods of low solar activity such as the "Maunder minimum". These numerical models thus give increased confidence in the dynamo theory.[16] Some of the difficulties include a description of the turbulent processes which control the angular momentum and the magnetic fields. It is still not entirely clear how the turbulent motion produces differential rotation and such sharp gradients at the base of the convective zone (Tobias and Weiss, 1999).

[16] Such simulations, however, can strictly never prove that these theories are true.

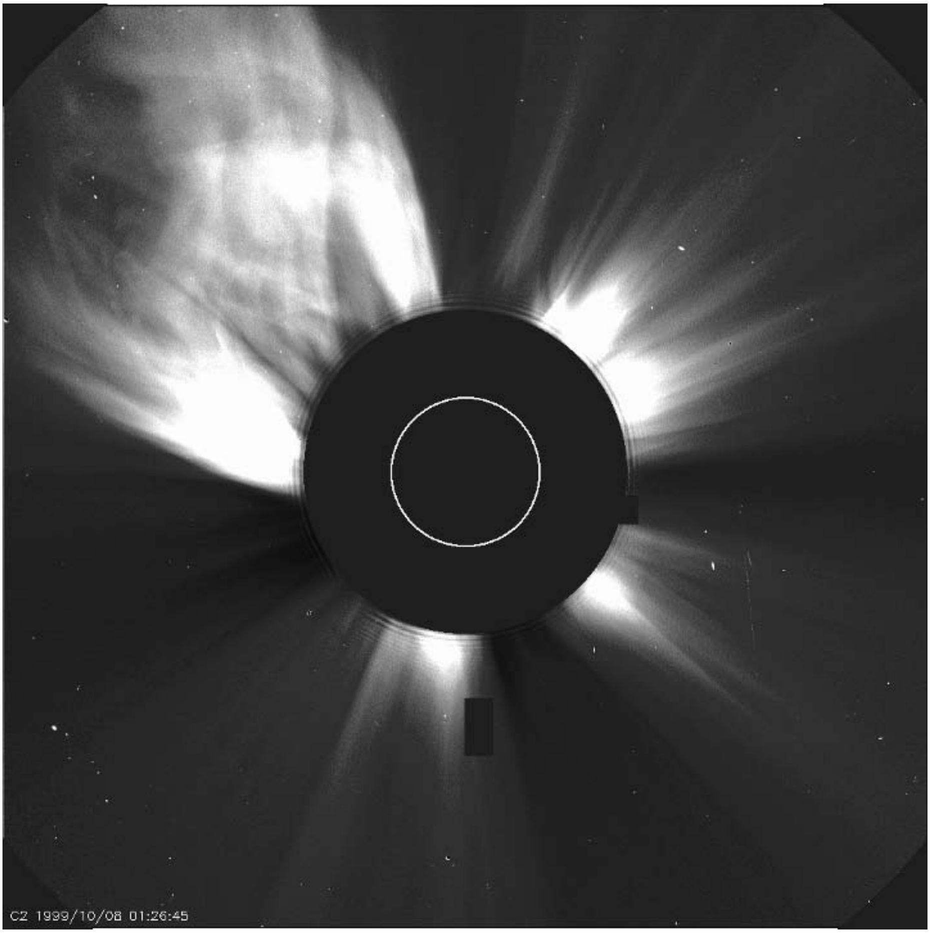


Figure 3.7. Coronal mass ejection (CME) observed from SOHO. Courtesy: SOHO/MDI (ESA & NASA).

If the magnetic field is embedded in a conducting material (plasma), then the field lines move with the material. They are in other words “frozen” in the plasma, according to a theorem called the Alfvén theorem (Alfvén, 1942). The magnetic field lines behave in a similar fashion as light, thin rubber bands in a thick viscous fluid. Transverse waves, known as Alfvén waves, form as a result of the field lines being frozen in the fluid motion and a restoring force that opposes the stretching of these lines.

The solar magnetic field varies with timescales from hours to centuries. The magnetic field is influenced by the sunspot activity, prominences,[17] and eruptions

[17] Prominences are filaments on the Sun’s limb seen in emission against the dark sky.

that lead to flares[18] (Figure 3.6) and coronal mass ejection (Figure 3.7). In the case of coronal mass ejection, clouds of ionised matter are blown out of the Sun at a speed of nearly 1000 km/s and may carry with them the solar magnetic field lines. The solar activity is closely related to the Sun's magnetic "weather" and "climate".

3.3.2.1 Other stars

By comparing the Sun with other stars, one can gain further understanding of the solar magnetism. Young stars are observed to rotate more rapidly than the Sun, but slow down as they age. The stellar activity[19] of these young stars is believed to be higher and more erratic, and the level of activity decreases with the stellar age. The older stars also exhibit a less erratic behaviour with a more cyclical character, just as is seen in the Sun. The magnetic fields of the remote stars are measured from their line spectra.

Recent discoveries from the Michelson Doppler Imager data on SOHO (Howe *et al.*, 2000) suggests a 16-month oscillation close to the regions believed to be in the vicinity of the solar dynamo action, and where the turbulence in the outer region meets the more "orderly" interior (tachocline). The speed difference in the layers above and below the tachocline may change by as much as 20% in 6 months. The lower gas accelerates as the upper layer decelerates, and later vice versa. Observations suggest fluctuations in these speed differences over 3 periods over 4.5 years. These fast changes may play a role in the generation of the 11-year solar cycle and in sunspot generation.

[18] A flare is a sudden eruption of energy on the solar disk lasting from minutes to hours emitting radiation and particles.

[19] Stellar activity is analogous to solar activity that takes place on the stars.

4

Solar activity

4.1 SYNOPSIS

The term *solar activity* comprises photospheric and chromospheric phenomena such as sunspots (Figure 4.1), prominences (see Figure 3.6, on p. 42), and coronal disturbances. Solar activity also refers to the level of solar magnetism, often giving rise to sunspots (Thomas and Weiss, 1992). Prominences and coronal disturbances are often associated with sunspots, but there are also types of solar activity which do not involve sunspots directly, such as the appearance of magnetic flux tubes and variations in the global solar magnetic field.

Sunspots are dark regions on the photosphere associated with strong magnetic fields and lower temperature than the rest of the photosphere. Thus they may be regarded as surface regions where the magnetic energy is concentrated. The sunspots are made up of dark central regions known as the *umbra* and outer lighter regions called *penumbra* as illustrated in Figures 4.2(a) and 4.2(b). The word "penumbra" literally means "almost shadow".

The sunspots start their life as small *sunspot pores*, which is often defined as a small, dark, long-lived region with no penumbral structure. These pores sometimes develop into sunspots, but most of them do not. The sunspot pores tend to be associated with weaker magnetic fields than the larger sunspots. Most sunspot pores dissolve after a short time (typical lifetimes shorter than one month). According to Thomas and Weiss (1992), the sunspots can be regarded as the advanced stage of pores that have acquired a penumbra.

In 1843, Schwabe brought attention to the periodic nature of the sunspots. After lengthy observations he discovered that the number of sunspots fluctuates between minima and maxima with a periodicity of approximately 10 years. The maximum number of sunspots may be as high as 250 for a month (the highest number of sunspots was 254, recorded in October 1957), and sometimes no sunspots are seen during minimum solar activity. The variation in the sunspot activity is often referred to as the *Schwabe cycle* in honour of the discoverer, or just the "solar cycle". In 1852,

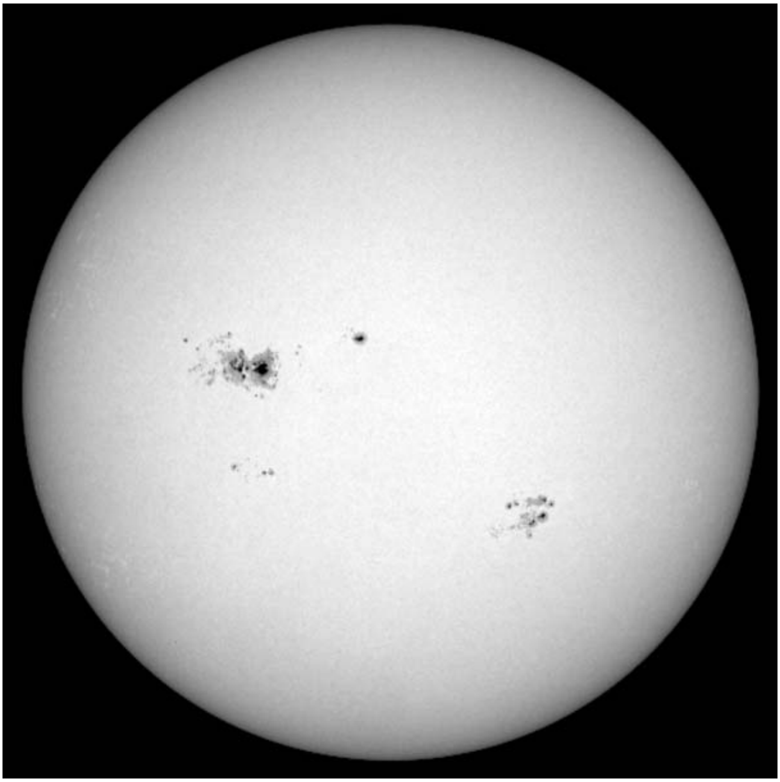


Figure 4.1. Sunspots observed on 20 September 2000. The sunspot area within the group spanned 2140 millionths of the visible solar surface. Courtesy: SOHO/MDI (ESA & NASA).

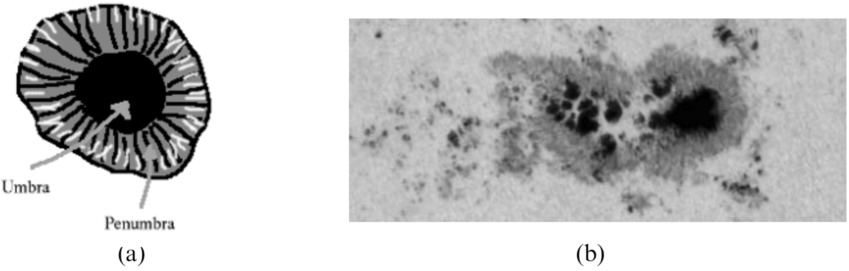


Figure 4.2. A schematic diagram showing the structure of a large sunspot with umbra and penumbra (a). Close-up photograph of sunspots showing the umbral and penumbral regions (b). Courtesy: SOHO/MDI (ESA & NASA) (b).

Wolf found that the length of the solar cycle was closer to 11.1 years. More recently, it has been acknowledged that the length of the individual cycles is not constant, but may vary between 9 and 12 years. Usually the rise of sunspot activity (the time between a minimum and subsequent maximum) is of shorter duration than the fall-time (time between maximum and the following minimum), with a few exceptions.[1]

The solar magnetic fields associated with the solar cycle exhibit certain systematic characteristics, such as preferred scales, orientations and polarity. The observations suggest that the magnetism associated with the sunspots involves coherent bundles with a toroidal shape. One view is that this magnetic field is a product of processes deep in the convective zone, but with stray loops that periodically penetrate out through the photosphere and thus provide seeds for sunpores and spots. These bundles of magnetic field lines are shredded as they extend out of the photosphere. The magnetic bundles are initially responsible for only small field intensities, but the flux lines then combine to produce sufficiently strong fields so that pores form at the junctions. The pores have almost uniform field strength of approximately 1900 gauss, and their radius may exceed 2000 km. Big pores may produce penumbra within a couple of hours, but several pores may also combine to produce a sunspot. In the latter case, the interface of each individual pore is often preserved, giving rise to seam-like structures in the spots which may develop into light bridges.

The granules are smaller in active regions where sunpores are formed, probably due to the presence of strong magnetic fields (Thomas and Weiss, 1992). Pores form where the magnetic field is converging. Sunspots are often surrounded by annular *moat cells*, which are large, long-lived granules stabilised by the sunspot occupying its core. These have diameters which are around twice those of the sunspots.

The sunspots are associated with strong magnetic fields, and they tend to form in groups with other sunspots that are associated with the same magnetic field structure. Observations by Evershed suggested that a transverse magnetic field is present in the umbra, but the field is longitudinal in the penumbra. Evershed's observations are in contrast to the classical picture of a symmetric field with maximum value over the central parts of the umbra directed away from the umbra. One possible explanation for Evershed's findings is that some magnetic field loops emerge from the umbra in such a way that both "ends" are anchored in the umbra.

Some solar observations have revealed intriguing facts about the sunspots. For instance, when the sunspots approach the solar limb, there is a tendency for a foreshortening of the penumbral region furthest away from the limb. This foreshortening is known as the Wilson effect, after the discoverer, and there has been some dispute about the explanation for this. It was originally thought that the sunspots were indentations in the solar surface, like saucers with a centre at a lower height level than the rest of the surface. It is now believed that the Wilson effect is due to the differences in the transparency of the sunspot and the photosphere, and that the sunspots are more transparent than their surroundings.

The solar cycle is associated with changes in the coronal structure, and it is believed that the sunspots themselves influence the corona. Streamers flowing out

[1] Min-max-min: 1755.2-1761.5-1766.5 and 1798.3-1805.2-1810.6.

from the equator are visible in the corona during solar eclipses when there are few sunspots,[2] whereas during sunspot maxima, the corona has a more isotropic appearance and is roughly circular. The extension of the corona near the equator during sunspot minima had already been noted in 1878 (Kuiper, 1953, p. 7). Mitchell noted in 1936 that the strongest and weakest flattening of the corona took place two years before the sunspot minima and maxima respectively. One may explain the differences in the coronal appearance between the sunspot maximum and minimum in terms of the solar activity during maximum blowing the corona out in all directions. Coronal mass ejections are usually seen during sunspot maximum (Diego, 1999).

4.2 THE UMBRA

The umbral intensity has been measured for about 125 years, but only the last 25 years of data are believed to be reliable (Thomas and Weiss, 1992, p. 10). The umbral intensity varies with the solar cycle (the penumbral intensity varies also, but not by as much as the umbra), which may suggest that the energy source of the sunspots is transported from beneath.

Larger umbra tend to be darker than smaller umbra, but there is no systematic relationship between the size and the umbral temperature.[3] Michard argued in 1953 that the umbra is more transparent than the photosphere, but Mattig (1958) believed that the photosphere was more transparent.

The gas pressure inside the sunspot is lower than outside, and the solar surface has slight depressions in the sunspot regions. Magnetic forces are required to maintain a dynamical balance and avoid an inflow of material that destroys the low pressure and the surface indentation. The umbral surface is depressed by 500–700 km (Thomas and Weiss, 1992), and the number of dark umbral regions is about the same as the total number of sunspots: $R_{\text{umbra}} \approx R_{\text{spot}}$. In a steady state, the pressure difference inside and outside a sunspot (umbra) must balance the magnetic force:

$$\Delta p = \frac{|\vec{B}|^2}{8\pi} \tag{4.1}$$

The umbra gas pressure is lower than in the photosphere by a factor of 2–3. The magnetic field lines have tube-like shapes, and electric currents must be present in these flux tubes due to induction if the magnetic field is changing over time.

Electrical currents in the sunspots can be dissipated by Joule heating, and can play a key role in the formation and dissolution of the sunspot magnetic fields. The conductivity in the sunspots has a magnitude of the order 1–2 lower than in the photosphere (Bray and Loughhead, 1964, p. 122), which is consistent with Michard's

[2] Diego (1999) shows photographs of the corona during solar eclipses at sunspot minimum in 1995 and maximum in 1980.

[3] Bray and Loughhead (1964, p. 112): There is a substantial statistical scatter: $\Delta\theta \approx 1000^{\circ}$C.

explanation that the sunspots are more transparent than the photosphere. A foreshortening of the penumbral regions near the Sun's limb (*Wilson effect*) also suggests that the sunspots are more transparent than their surroundings.

Granulations (cell-like structures) are also seen in the umbra, and these granules are the only direct evidence of umbral convection. The convection is most effective in the upper 1500 km of the photosphere, and a horizontal temperature gradient is required for these convective cells to exist. Strong magnetic field lines impede the convection in electrically conducting fluids, but the magnetic fields in the Sun probably do not shut down the convective processes completely.

4.3 THE PENUMBRA

The penumbra has filamentary structures (Thomas and Weiss, 1992), and the early models of the penumbra assumed a shallow structure. More recently, however, it has been deduced that the penumbra must be deep. The penumbral striations have an average lifetime of around 2 hours (Table 4.1), a length of 5000 km, and width of around 300 km. They seem to consist of elongated grains which migrate towards the umbra. Magnetic field lines have been used to explain the thin filaments, where lateral distortions are restricted by the fields. The magnetic field lines in the darker regions are taken to be horizontal whereas the brighter filaments are thought to be associated with a magnetic field that is inclined with respect to the vertical (35° near the outer penumbral border, and 45° near the umbra).

Radial motions have been measured in the penumbra (Bray and Loughhead, 1964, p. 146), and this is referred to as the *Evershed effect*. It was long believed that the motion was restricted to radial motion, but irregular tangential motion has also been observed by Abetti (1957). Although the observed motions in the sunspots are predominantly radial, vertical and tangential flow may also be present. Matter emitting weak spectral lines associated with the photosphere appears to move out of the sunspot centre, whereas strong chromospheric spectral lines are associated with an inflow. The Evershed effect stops abruptly at the outer penumbral border because the gas flowing along the magnetic field lines becomes transparent outside this region. The flow does, however, not cease beyond this border. The radial flow may have velocities of the order 1–10 km/s, and the motion extends well into the photosphere. The maximum flow rate is estimated to be approximately proportional

Table 4.1. Umbral and penumbral timescales (Bray and Loughhead, 1964).

Penumbral filaments	≈2 hr
Penumbral bright regions	30 min
Umbral granules	15 min–2 hr
Facular granules	2 hr
Photospheric granules	7–10 min

to the umbral radius, and the maximum values are found in the penumbra for large spots and in the photosphere for small spots.

The measurement of Doppler shifts also suggests a sinking motion in the sunspot region and radial inflow at the top and outflow at the bottom similar to a convection cell. Thus, the Evershed outflow observed in the penumbra decreases with height until a critical level where the flow is reversed and explains an outflow in the photosphere (lower levels) and inflow in the chromosphere (higher levels).

Penumbral temperatures, inferred from intensity measurements, are not uniform and there appears to be a gradual increase from the interface between the umbra and penumbra to the edge toward the photosphere. The penumbral intensity tends to be independent of sunspot size, but may differ between the individual spots.

The Wilson effect can also be seen as a foreshortening of the penumbral regions furthest away from the limb. The fact that this effect is seen for the penumbra suggests that the umbra is more transparent than the photosphere.

4.4 SUNSPOT GROUPS

The sunspots often appear in groups in which the individual spots are associated with the same magnetic field structures. Some sunspots are referred to as "leading" and "following" spots, meaning spots in a group leading or following in terms of the Sun's rotation.

Hale and Nicholson (1938) devised a magnetic classification scheme for the sunspot groups (Table 4.2) with three different types: unipolar, bipolar and complex. Bipolar groups are sunspot groups which consist of two members associated with opposite magnetic polarity. Most groups with lifetimes longer than one day are initially bipolar. In the period 1917–1924, 46% of the observed sunspots were unipolar, 53% were bipolar, and 1% was complex.

Bray and Loughhead (1964) provide some interesting statistical summaries of the sunspots and sunspot groups. Groups tend to start as one or more spots and the

Table 4.2. Hale and Nicholson magnetic classification of sunspot groups (Bray and Loughhead, 1964).

Class	Description
α	Unipolar: symmetrical faculae preceding and following the group
αp	Unipolar: followed by elongated faculae
αf	Unipolar: preceded by elongated faculae
β	Bipolar: preceding and following members almost equal area
βp	Bipolar: preceding member is the principal feature
βf	Bipolar: following member is the principal feature
$\beta\gamma$	Bipolar characteristic but lacks well-marked division between regions of opposite polarity
γ	Complex

Table 4.3. Zurich classification of sunspot groups. Note, there is no "I" category (Bray and Loughhead, 1964).

Class	Description
A	Single pore or group of pores showing no bipolar configuration
B	Group of pores showing a bipolar configuration
C	Bipolar group with one spot possessing a penumbra
D	Bipolar group whose main spots have penumbra and at least one spot has simple structure length of group $>10^\circ$
E	Large bipolar group with two main spots having penumbra length of group $>10^\circ$
F	Very large bipolar or complex group: length $>10^\circ$
G	Large bipolar group containing no small spots between main spots length of group $>10^\circ$
H	Unipolar spot with penumbra and diameter $>2.5^\circ$
J	Unipolar spot with penumbra and diameter $<2.5^\circ$

initial group may encompass a handful of pores concentrated within an area of 5–10 square heliographic degrees but they may also start as a single spot at the preceding (westward in terms of the solar rotation) end of the subsequent group region. The area of the sunspot group typically grows in the first 5–12 days, and then the small spots between the preceding and following components tend to disappear.

The sunspot groups can be classified according to a Zurich classification scheme (Table 4.3) or according to a classification by Mount Wilson (Table 4.6, p. 73). The sunspot groups typically start as A and evolve through the various categories through their life cycle. The evolution commonly passes quickly through classes A to E and spends most of the time in G to J when the region slowly disappears. The highest flare activities tend to be associated with classes D, E and F. Small groups tend to follow more symmetrical development curves like A-B-C-B-A, whereas large groups are less systematical in their development (e.g. A-C-C-D-C-H-J-A).

The sunspots grow rapidly, but decay slowly so that most of their life is spent in their decaying phase. Hoyt and Schatten (1993) established a relationship between the sunspot decay rate and the ratio of umbra to whole spot area (U/W), so that given U/W one can estimate the decay rates (equation (4.2)):

$$D_{\text{spot}} = 63.3 - \frac{(6.0 \pm 0.4)}{\bar{U}/\bar{W}} \tag{4.2}$$

The decay rate of umbra and penumbra are different, with the umbral decay rate being constant and the penumbral rate variable. In many cases, the umbral regions disappear before the penumbra so that the last stage of sunspots often involves just penumbral structure. Spots with only penumbral structure are referred to as penumbral spots. Hoyt and Schatten (1993) give the ratio of umbral decay rate

Table 4.4. Sunspot groups by area and occurrence (Bray and Loughhead, 1964).

Mean area during disk passage (millionths of visible hemisphere)	Percentage of all groups
1–250	85.6
250–500	9.2
500–750	3.0
750–1000	1.0
1000–1500	0.61
1500–2000	0.38
>2000	0.23

(D_{umbra}) to that of all sunspot area (D_{spot}) lasting N days as:

$$\bar{U}/\bar{W} = \frac{2U_0 - (1+N)D_{\text{umbra}}}{2W_0 - (1+N)D_{\text{spot}}} \tag{4.3}$$

The sunspot decay rate has been observed to vary proportionally with the sunspot perimeter and linearly with time for 95% of the sunspots, and the decay rate is independent of the maximum sunspot area.

It is well-known that sunspots tend to appear in the *spot zone*, a narrow belt between ±35° from the equator, but there have been occasions when sunspots have been observed at 75°. The average latitude depends on the phase of the solar cycle with occurrence at the highest latitudes at the beginning of the cycle (beginning of rise in sunspot activity) and lower latitudes towards the end. There is nevertheless a significant spread in latitudinal distribution within a cycle (Table 4.5).

The number of sunspots varies between a minimum of almost no sunspot activity to a maximum of around 254 sunspots per month (1957). Schwabe counted the sunspots over a long period and discovered that they repeatedly appeared with maximum numbers at intervals of around 10 years, and this variation in the sunspot activity is often referred to as the solar cycle or the

Table 4.5. Latitude drift of the sunspot belt over a sunspot cycle (Bray and Loughhead, 1964).

Time from minimum (years)	Average latitude (±°)
0	28
2	21
4	16
6	12
8	10
11	7

Schwabe cycle. The sunspot cycle usually starts with a small number of sunspots that appear near 30° from the equator. As more and more sunspots appear, they are formed closer and closer to the equator and at sunspot maximum the sunspots often form near ±15° latitude. At the end of a cycle, before the minimum, the sunspots form closest to the equator ($\approx \pm 8^\circ$ latitude), and the sunspots associated with the southern solar hemisphere have even been known to penetrate into the opposite hemisphere. The equatorward migration is not regular and periodic, nor is the length of the individual solar cycle. The nature of the solar cycle in the two hemispheres may be slightly different, for instance the maximum in the northern hemisphere takes place before that in the southern hemisphere and more sunspots are seen in the southern hemisphere (Kuiper, 1953, pp. 336–337). If the beginning of each solar cycle may be defined as the time when the leading spots appear at maximum latitude, then solar cycles seem to overlap each other by around 3 years.

Gnevishev studied the sunspot group lifetimes of some 3000 groups between 1912 and 1934 and found that more than half of the groups have lifetimes less than 2 days. Such short-lived sunspots may easily go undetected unless an uninterrupted observing system is in place. Less than 10% of the groups last more than 11 days. In other words, there are few long-lived sunspot groups (Table 4.4). Furthermore, groups at high latitudes but within the spot zones may have shorter lifetimes than those nearer to the equator. The lifetime is also approximately related to the maximum group area according to the "rule-of-thumb" $T \approx 0.1A_{\max}$ ($A_{\max}$ is measured in millionths of the visible hemisphere).

Sunspot groups often have a roughly oval region whose major axis is slightly inclined with respect to the parallax of latitude. The preceding spots (p-spot, west) (also referred to as leading spots) tend to lie at a lower latitude than the following (f-spot, east) spots. Differential motion may therefore move the p- and f-spots further apart. These motions reflect the dynamical behaviour of the magnetic flux loops attached to the spots.

Whereas monthly mean sunspot numbers give a reasonably smooth curve, daily sunspot numbers exhibit very large variations that may last for a few days to a few months. In fact, the short-term variation in R_z appears to be completely random with a weak 27-day signal due to the solar rotation. The 27-day cycle tends to stand out most prominently when the sunspot number is low and one or more long-lived groups are present.

4.4.1 Sunspot brightness and temperature

One of the legendary solar scientists, Hale, found that the sunspots were cooler than the photosphere. His finding was based on a photographic study of solar spectra, where light either from the sunspots or from the photosphere had been blocked out. The intensity of the line spectra from the sunspots were lower than the rest of the photosphere, and laboratory studies had revealed that the line intensities were related to the temperatures. Sunspots are cooler (3800 K) than the photosphere (5800 K), and their size (diameter) may be as large as 50,000 km. The temperature of the sunspots can also be derived from the total energy flux (I and I^*) integrated over all wavelengths using the following expression:

$$T^*_{\text{eff}} = T_{\text{eff}} \left(\frac{I^*}{I} \right)^{\frac{1}{4}} \tag{4.4}$$

It therefore suffices to use the measurements of I/I^*, or the ratio of the radiative energy from the sunspot region and the surrounding photosphere. Correction must be made for ‘‘parasitic light’’ (Wander’s method) and selective attenuation in the telescope and atmosphere. The energy–wavelength ratios differ for the sunspots and the photosphere, measuring $I/I^* \approx 0.27$, and taking $T_{\text{eff}} = 5785\,\text{K}$ gives $T^*_{\text{eff}} = 4160\,\text{K}$. The difference between this estimate and 3800 K may be due to different temperatures in different sunspots, but it may also give an indication of the uncertainty in the sunspot temperature estimates.

The sunspots appear to suppress the granules, which are interpreted as a sign of convection near the photosphere, but the convection is not completely suppressed. Weaker granules have been observed inside the sunspots in both the penumbra and umbra. It is believed that the suppression of the convection is the reason why the sunspots are cooler than the surrounding photosphere. The solar surface is assumed to be in energetic equilibrium, which means that it does not gain or lose energy over time. The energy required to keep the Sun’s surface at a constant temperature must come from the interior, as the energy that the surface emits through radiation must be balanced by energy produced in the core. The photosphere is opaque, which means that the photons do not bring much energy to the surface. Instead, the main energy transfer from the solar interior takes place through fluid convection near the surface. The convection process involves cells of fluid with rising and sinking motions. The rising fluid brings heat to the surface whereas the sinking fluid returns colder fluid to the base of the convection zone. These convective cells resemble a type of convection studied in laboratories, known as Bérnard cells. The stratocumulus clouds found on Earth also exhibit similar convective characteristics.

4.5 SUNSPOT MODELS

Readers interested in matters on solar–terrestrial links should appreciate the complexity behind the theories of solar activity and sunspots. It may be beneficial to have an idea of where the solar activity comes from in order to get a holistic view on the relationship between solar activity and Earth’s climate. The appreciation of such matters requires some physical insight into the processes taking place on the Sun.

4.5.1 Dynamo action and magnetism

What drives the solar activity and gives rise to sunspots? This is a key question which is still difficult to answer, although there are some hypotheses about various mechanisms. The answer is most likely to involve magnetic interaction with turbulent flow and electric conduction.

Petrovay (2000) provides a short review on the matter, basing the discussion around an equation describing magnetic induction in a turbulent conductive medium:

$$\partial_t \vec{B} = \nabla(\vec{U} \times \vec{B} + \vec{\epsilon}) - \nabla \times \eta \nabla \times \vec{B} \tag{4.5}$$

In this equation the term $\partial_t \vec{B}$ denotes the rate of change in the magnetic field $\vec{B}$, and the parameter $\vec{\epsilon}$ describes the flow structure: $\vec{\epsilon} = \alpha \vec{B} - \beta \nabla \times \vec{B}$, where the parameters α and β are functions describing the nature of the turbulent velocity fields.

According to Petrovay there are three different basic models attempting to explain the solar activity:

a Overshoot models
b Interface dynamos
c Flux transport circulation conveyor belt

These models must be able to account for the observed solar features: (i) the pole-to-equator diffusion in the convective zone with an 11/22 year periodicity; (ii) the characteristic migration patterns in the butterfly diagram shown in Figure 4.3; (iii) the confinement of large active regions to $\pm \leq 35^\circ$; (iv) the radial field at low latitudes appears to be π radians out of phase with the toroidal field at the same latitude; and (v) the phases of the two branches of the butterfly diagrams tend to have a difference in phase by $\pi/2$ radians so that the polar field reversal occurs slightly after the sunspot maximum.

Large-scale solar active regions in Figure 4.3, according to Petrovay, can be interpreted as tracers of a subsurface toroidal magnetic field. The latitudinal migration pattern can be partly due to the toroidal and partly due to a poloidal field structure. A toroidal field gives a zonal displacement whereas a poloidal structure can explain meridional displacements.

The mean field dynamo theory assumes that most of the shear (Ω) is concentrated in a thin layer near the bottom of the convergence zone also known as the *tachocline*. The tachocline thickness is $\approx 0.1 R_\odot$ ($R_\odot$ is the solar radius). The mean field dynamo theory has been the framework for a number of different models. The most recent models can be classified under four different categories: (a) overshoot layer models (OL dynamos), (b) distributed wave models (IF dynamos), (c) co-spatial transport models (CP dynamos), and (d) distributed transport models (BL dynamos). A short summary of these different model types is given below. A main difference between the various models can be regarded as different ways of interpreting and treating the α-effect (see box): cyclic convection (α is positive in the unstable

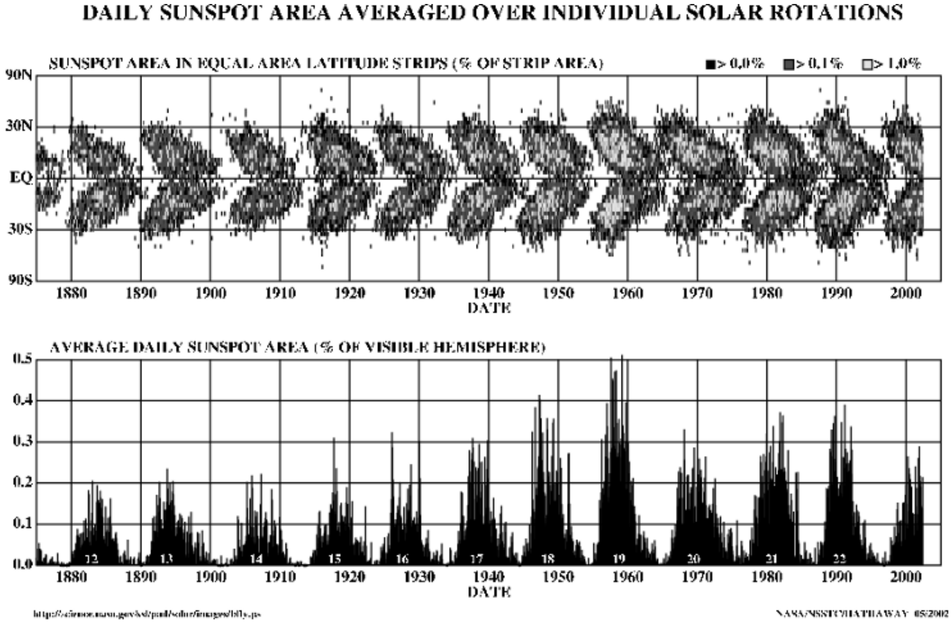


Figure 4.3. Butterfly diagram of solar activity showing time–latitude distribution of sunspot groups. Courtesy of David Hathaway/NASA/NSSTC.

layer and negative in the overshoot layer below); magnostrophic waves (α is negative); flux loop α (positive); unstable global-scale Rossby waves on the tachocline with non-zero mean helicity that may produce an α-effect.

Models with positive so-called "α-effect"[4] (in the solar convective zone) that is consistent with an inwards increasing rotational rate can reproduce the 11-year cycle period. The equatorward migration of the sunspots (butterfly diagram branches) can be explained in terms of a dynamo wave and longitude–latitude phase relationship. The problem is that the radial rotation profile that this model assumes seems to be wrong.

It has been observed that short-lived active regions in the high latitude tend to lie on the backward extension of the low latitude butterfly wings similar to polar faculae. There have also been suggestions that about half of the polar faculae are associated with east–west oriented magnetic dipoles which have been interpreted as parts of the toroidal field.

It was discovered in the 1970s that the magnetic field in the solar photosphere is present in strong intermittent form and concentrated in long flux tubes. This observation is problematic in terms of our physical interpretation. The flux tube theory,

[4] Interaction between the flow structure and the magnetic field on which $\vec{\epsilon}$ depends in equation (4.5).

on which the flux emergence models are based, states that the tubes are stored near the convection zone base or lower overshoot layer, but this means that the field strength is difficult to explain ($|\vec{B}| \approx 10^5$ G).

Overshoot layer models, also known as "co-spatial wave models", require $\alpha < 0$ in order to get the right migration direction. These models do a good job predicting the butterfly diagrams, but tend to produce cycle periods that are too short.

Distributed wave models assume an abrupt spatial change in diffusivity that results in the excitation of dynamo waves. Strong toroidal fields can be explained by these models, but evaluation of these are hindered by numerical difficulties.

Co-spatial transport models explain the field migration as a result of density pumping or advection of the magnetic field. These models do not address the question regarding the origin of the deep toroidal field.

Distributed transport models do not interpret the butterfly diagram as a manifestation of a dynamo wave, but a consequence of a conveyor belt action. Low latitudinal confinement of strong activity needs another dynamo mechanism kick and requires unrealistically low turbulent diffusivity. These models account for the confinement of the strong activity at low latitudes and predict migration characteristics similar to those in Figure 4.3.

In summary, none of the models can fully account for the observed features of the solar cycle and all these have to make some assumption about the Sun. There have been some severe difficulties, even with the classical theory of the dynamo wave origin of the butterfly diagram. Thus, the jury is still out, according to Petrovay, on what makes the Sun "tick".

Pores and sunspots are exclusively formed within regions of enhanced magnetic fluxes recently emerged from the solar surface, referred to as bipolar active regions. The emergence time is defined as the time span from the first appearance of the bipolar region[5] to maximum development.

The origin of the solar cycle was briefly discussed above, and now attention will be given to the individual sunspots. How do these arise? There are various hypothetical sunspot models that come under two types: (i) convective or hydrodynamical or (ii) magnetic cooling. The first category explains the origin of the sunspots as being due to an initial cooling and subsequent intensification of the magnetic fields.

The second category, which is more generally accepted nowadays, assumes that the sunspots are formed as a result of regional intensification of the magnetic fields. The magnetic fields are thought to suppress the convection in the sunspots, and thereby produce a region of colder and darker material. The presence of a solar magnetic field is explained by a dynamo theory similar to the geomagnetic field model, and will be elaborated in Section 4.5.3.

[5] On Kitt Peak magnetogram.

4.5.2 Convective and hydrodynamical sunspot models

Some of the first convective sunspot models were proposed by Russel in 1921 and Rosseland in 1926. These models assumed the presence of a stable layer under the surface, where the vertical temperature gradient is smaller than adiabatic[6] temperature change of a rising volume of gas. Rising motion in such a stable layer cools the fluid due to volume expansion and lower pressure, and the gas becomes cooler than its surroundings. The models suggest that the sunspots form when a cyclic convection cell transports the material at the top of the convective cell horizontally away from the convective centre and then the cool gas sinks to its original depth. The darkness of the spot is postulated to be due to the adiabatic cooling as the material rises. The bright faculae around the spots are conversely thought to be due to adiabatic heating of the descending motion. The problem with these models is that their motion opposes the force of gravity, which would rapidly destroy the motion.

Villhelm Bjerknes proposed in 1926 a second model analogous to tornadoes on Earth. Bjerknes' explanation focused on the solar cycle and was more vague on the details of the individual sunspot. He assumed that the magnetic fields of a spot were related to the rotation of ionised material and proposed that the sunspots were vortex-like phenomena spinning around a vertical axis. The sense of the rotation is determined by the direction of the magnetic fields of the sunspots, according to this theory, and the leader spots in one hemisphere in any solar cycle are related to the vortex motion of the other sunspots of the same cycle (the sense of rotation). The direction of rotation is postulated to be opposite in the two hemispheres, and the spin of the sunspots reverses between each solar cycle. Bjerknes' model assumes the presence of vortex rings around the Sun underneath the surface, and sunspots form from a pair of vortex rings that break away and extend up to the surface.

The vortex model can explain the latitudinal preference of the sunspot appearance, and the polarity was explained by two vortex rings in each hemisphere. The difficulty with Bjerknes' model is that only a few sunspots exhibit spinning motion, and these show no preferred sense of rotation associated with the sense of the magnetic field. Furthermore, the theory is regarded as inadequate at explaining the lower temperatures in the sunspots.

4.5.3 Magnetic cooling models

The formation of sunspots can be explained by the magnetic cooling theory, in which strong magnetic fields inhibit the convection and slow down the radial energy transfer. Thus, the strong magnetic fields are responsible for the cooling of the sunspots. As the convection is suppressed in the sunspots, the energy transfer from the solar interior is reduced, and the sunspot cools. The effective temperature

[6] Adiabatic describes a process within a closed volume from which no energy enters or leaves. This volume may change in time.

of the sunspots can be measured from the spectral continuum of the emitted light according to the Stefan–Boltzmann law.

The cooling associated with sunspots is only believed to take place at the depth of a few thousand kilometres, or otherwise the sunspots would have a blurred appearance as opposed to sharp edges. Biermann suggested in 1941 that the cooling takes place in the convective zone, and not the thin radiative layer which is considered to be too thin to obstruct the light. The strong magnetic fields in the sunspots inhibit the convection and restrict the flow from following the lines of force. As the convection is the primary agent for energy transfer in the outer solar region, the reduction in the convection intensity results in less energy arriving at the surface of the sunspot. The sunspots appear dark because of the contrast with the convective heating in the surrounding area, not because the reduced convection itself is responsible for cooling. Thus, the magnetic fields influence both the mechanics of the fluid as well as the thermodynamics. Hoyle proposed in 1941 that the convection below a sunspot is not destroyed, but merely confined to the direction of the magnetic field lines. The field lines funnel out near the surface and obtain greater cross-sections than below the sunspots. This divergence dilutes the energy transported from the interior and therefore can explain the cooler temperatures.

The magnetic field is thought to be concentrated in sheets and tube-like structures where the field strength is limited by diffusion and the action of convection. The magnetic field is dissolved by the convective collapse of flux tubes, a process known as “flux expulsion” (Thomas and Weiss, 1992), where strong fields are formed in narrow rapidly sinking regions which become less dense (“evacuated”). Hence, the magnetic flux is concentrated into the cool lanes of sinking material between the granules, and this concentration inhibits heat transport into the flux tube, which cools and collapses further because of radiative losses at the solar surface.

The magnetic field arises as an instability in a self-sustaining dynamo, and these instabilities are caused by differential motion or turbulent convection near the solar surface. A key ingredient to this model is a rotating fluid confined between a spherical annulus and where convection is driven by heating from below. The Lorentz force limits the growth of the magnetic field by altering the motion. One of the problems with these models is explaining how the activity zones migrate (Thomas and Weiss, 1992), but the model shortcomings may be related to the simplified representation of the turbulent convection by parameterisation. The simulations with the dynamo models nevertheless demonstrate that the mechanism actually works.

Helio–seismological studies indicate that there is a strong gradient in the rotation rate just below the convection zone. Such gradients may explain the presence of strong toroidal magnetic fields and may be a key factor in the oscillator theories. In these models, the underlying poloidal field is wound up through differential rotation rates and thus result in a toroidal field shape. The toroidal magnetic field grows until it is sufficiently strong that it reverses the motion. This field must be present near the interface between the radiation interior and convection zone in a dynamo model, because a sufficient gradient in the rotation rate is found in this region (Zwaan, 1992).

The differential rotation rates may be another explanation for the presence of sunspots: the magnetic field lines are moved with the ionised material (plasma) and may be bundled closely together in the shear regions. The strength of the magnetic field intensifies with closer bundling.

4.6 SOLAR ACTIVITY AND PREFERRED TIMESCALES

4.6.1 Solar activity and the sunspot cycle

There are various methods for quantifying the solar activity. The classical method is using the Wolf sunspot number, also known as the Zurich sunspot number. Wolf, who was the director of the Zurich observatory, adopted in 1849 the following equation for estimating the *sunspot number* R_z (which is an index):

$$R_z = k(10g + f) \tag{4.6}$$

The symbol f is the total number of (physical) sunspots, g is the number of sunspot groups, and k is a scaling factor depending on the observing method. The Wolf sunspot record was extended back to 1749, and there is a gap between 1611 and 1749 in which no sunspot counts exist (Kuiper, 1953), but only the time of maximum and minimum sunspot activity.

In 1882, Wolf's successors at the Zurich observatory modified the counting method, and the value of k was changed from 1 to 0.60 (its present-day value) to accommodate this change.

4.6.1.1 Statistical description of the sunspots

Records of one or several observations made a number of times are referred to as "time-series" or just "series". An example of a time-series is shown in Figure 4.4 which shows the (Wolf) monthly mean sunspot number (represented by the symbol R_z). The number of sunspots is plotted along the (vertical) y-axis, whereas the (horizontal) x-axis indicates the time when this observation was made. We will refer to entries in an arbitrary time-series, x, as x_n, where x means the value and the subscript n refers to the time of observation. The number of observations is N, so that n counts from 1 to N. Thus, $x = [x_1, x_2, \ldots, x_n, \ldots, x_N]$. In the case of Figure 4.4, x is the monthly mean sunspot number, R_z, between 1749 and 2001. Additional important data is the *metadata*, describing the times of observation, location, instruments, observational practices, holes of missing data, and the local environment. The times of observations is another time-series t of same length as x. The data are also associated with a physical unit, although some data records represent indices or ratios which are dimensionless.

4.6.1.2 The mean

The mean value (or average) of a time-series describes a value about which the quantity fluctuates, or a "typical" value describing the data. The mean is therefore

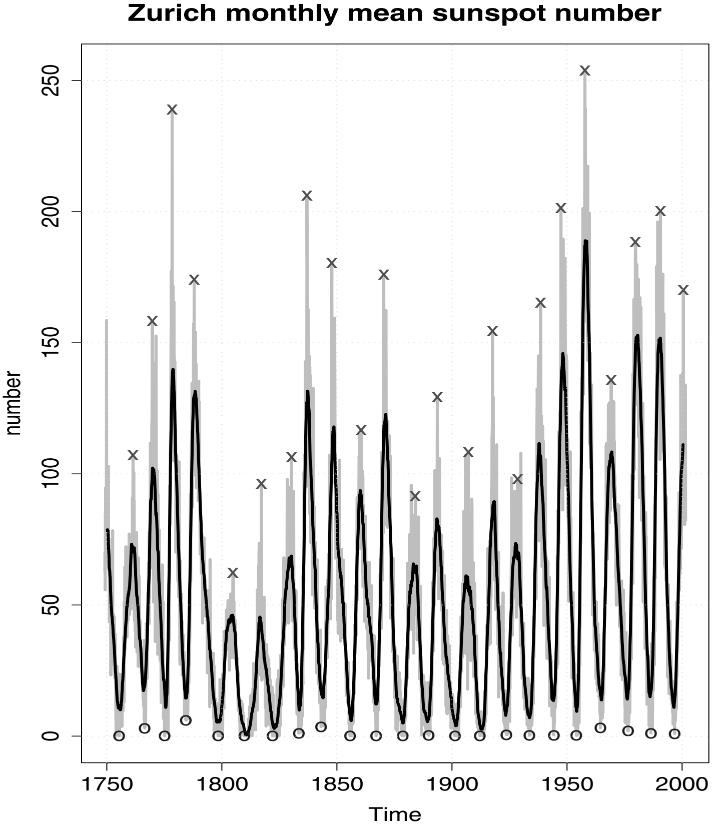


Figure 4.4. The unfiltered (grey) and 27-month low-pass-filtered (black) temporal variation of the monthly mean Zurich sunspot number (R_z). The low-pass-filtered (smoothed) curve was used for finding the maxima and minima (marked as "x" and "o"). Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

called a *location* descriptor. One way to define the mean value is to count the sum of all the observations and divide this evenly among each observation.

Mathematically, the mean can be defined as the sum over time divided by the number of observations:

$$\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n \tag{4.7}$$

where $\sum_{n=1}^{N}$ means the sum over all values of x_n from $n = 1$ to $n = N$.

This estimate belongs to a class of statistics called *moments estimators* or just M-estimators. We will henceforth reserve the notation $\bar{x}$ to the temporal mean (average over time), whereas $\langle x \rangle$ means spatial average or the average over a number of stations. The mean sunspot number for the period January 1749 to July 2001 is $\overline{R_z} = 52.52$. Another type of location descriptor is the median value which describes the number x_n which is ranked in the middle of the data values and the median of R_z for the same period is 42.55. The median is less sensitive to extreme values (and outliers) than the mean value.

4.6.2 Spectral analysis

The term "spectral analysis" may have two different meanings, which may cause some confusion, although the two are related. Figure 3.3 in Section 3.2.5.1 shows the spectral analysis obtained through spectroscopy. One example of spectroscopy is a prism that separates the different wavelengths of light by bending the rays at different angles. Spectral analysis can also be carried out numerically through a Fourier transform (FT). In this section, the latter type, i.e. FT-based spectral analysis, will be considered.

4.6.2.1 A spectral study of the solar cycle based on sunspots

So far, we have considered time-series as an array of observations given in a chronological order and with a fixed time interval between each entry. This type of representation is called "time-domain". In addition to the timing of the time-series fluctuation, it is also possible to characterise them by their timescales. The analysis of timescales can be illustrated with a spectral analysis of temperatures measured in Oslo (near 60°N) and it is evident that the temperature is dominated by the daily variations (the diurnal cycle of 24 hours, not shown) and the seasons (the annual cycle of 365.25 days), and a spectral analysis such as a Fourier transform or a wavelet transform will therefore indicate strong variations with frequencies of 1/(24 hours) and 1/(365.25 days).

Timescale analysis is often done in the frequency-domain, in which the data is analysed according to the strength of a range of timescales, or frequencies. For readers who are interested in spectral analysis, more details about these techniques are given by Box and Jenkins (1976), Wilks (1995), von Storch and Zwiers (1999), Press *et al.* (1989), and Torrence and Compo (1998). There are several different methods, and all of these ought to give similar results if the features are robust. One common mathematical technique is the Fourier transform[7] (FT).

Some important aspects of spectral analysis are the fundamental limitations set by the *Nyquist frequency*: the highest frequency for which we can hope to get information, the Nyquist frequency, is dictated by how frequently the measurements or observations are made. The Nyquist frequency is defined as $f_L = 1/(2\Delta t)$. If fluctuations with higher frequency are present, then these will cause an *aliasing*

[7] $F(\nu) = \int_{-\infty}^{\infty} f(t) e^{-2\pi i t \nu}\, dt$.

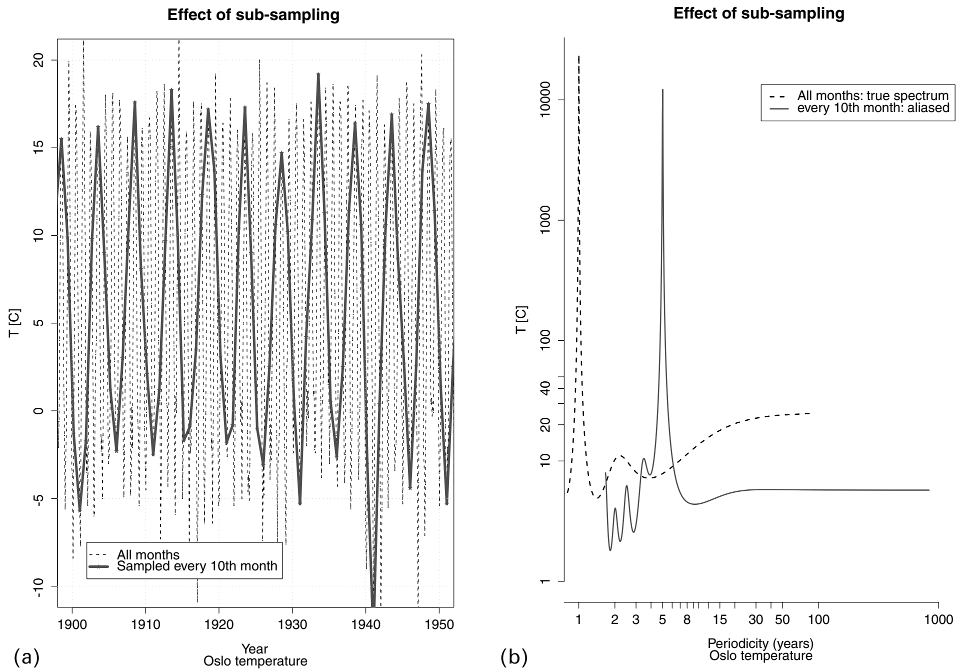


Figure 4.5. Example of aliasing due to sub-sampling. The sub-sampled time-series is shown in grey in panel (a) and the corresponding spectrogram is given in (b). The whole monthly mean temperature series is shown in black for comparison. Data from: the Norwegian Meteorological Institute.

problem that can produce spurious peaks at lower frequencies (see Figure 4.5). Any variation which varies at a faster rate will affect the estimates of the frequencies lower than f_L, and thus result in an erroneous spectral estimate. A common indication for this effect taking place is when the highest frequency estimates near f_L have a significant amplitude. It is therefore desirable to have spectral estimates that tail off at the high frequency end. Aliasing is not believed to be a problem for the monthly mean sunspot record with the prominent 11-year cycle, but may be a problem for climate observations if sampled at every solar maximum or minimum. Aliasing is not just a problem in spectral analysis, as Figure 4.5(a) illustrates, but may also affect correlation and regression analyses.

In short, the FT tries to describe the time-series as the sum of a range of sinusoids.[8] The spectrograms in Figure 4.5 show that temperature series may have a pronounced annual cycle, but that there are also indications of a longer, albeit weaker, cycle with a timescale somewhere near 2–3 years and longer.

One common strategy for inferring a relationship between various quantities is to compare the power spectra: do they have power peaks at the same timescales? By

[8] A sinusoid consists of regular and periodic oscillations and has the mathematical notation: $a\sin(\omega t) + b\cos(\omega t)$.

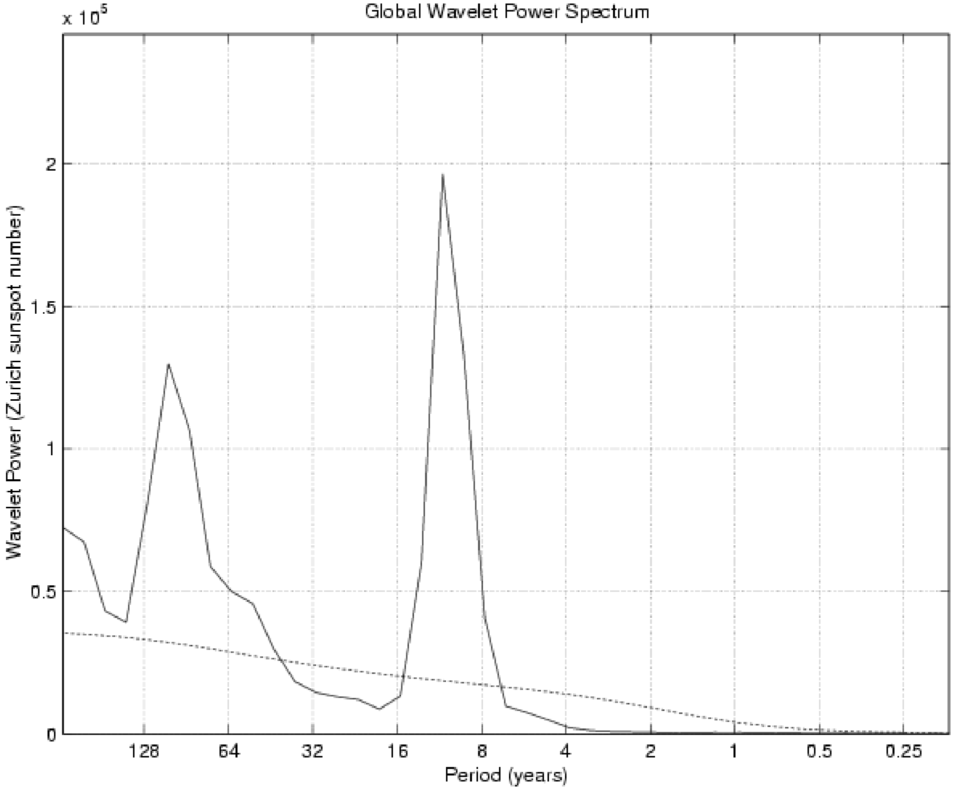


Figure 4.6. Spectrogram (estimate of the power spectrum) of the Zurich sunspot number indicating the 11-year cycle. In addition to the 11-year Schwabe cycle, the spectrogram exhibits a prominent spectral peak for timescales around 100 years. The dashed line shows an estimate of the 5% confidence level assuming a red noise process. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

comparing Figures 4.6, 4.7, and 4.8 we can draw some conclusions about the linear relationships between the sunspot number, the aa-index, and the solar radio emission. If there is a nonlinear relationship, then the spectral peaks may not be associated with the same timescales. Furthermore, a collocation in terms of frequency or timescale does not prove that these are related. However, the similarities provide “circumstantial evidence” for a relationship between these quantities.

4.6.3 Wavelet analysis

It is also possible to compute how the various frequencies contribute over time through a wavelet analysis (Torrence and Compo, 1998). The wavelet analysis is a method that estimates how important the various timescales are at various times (Figure 4.9, see colour plate section). The result of a wavelet analysis applied to a

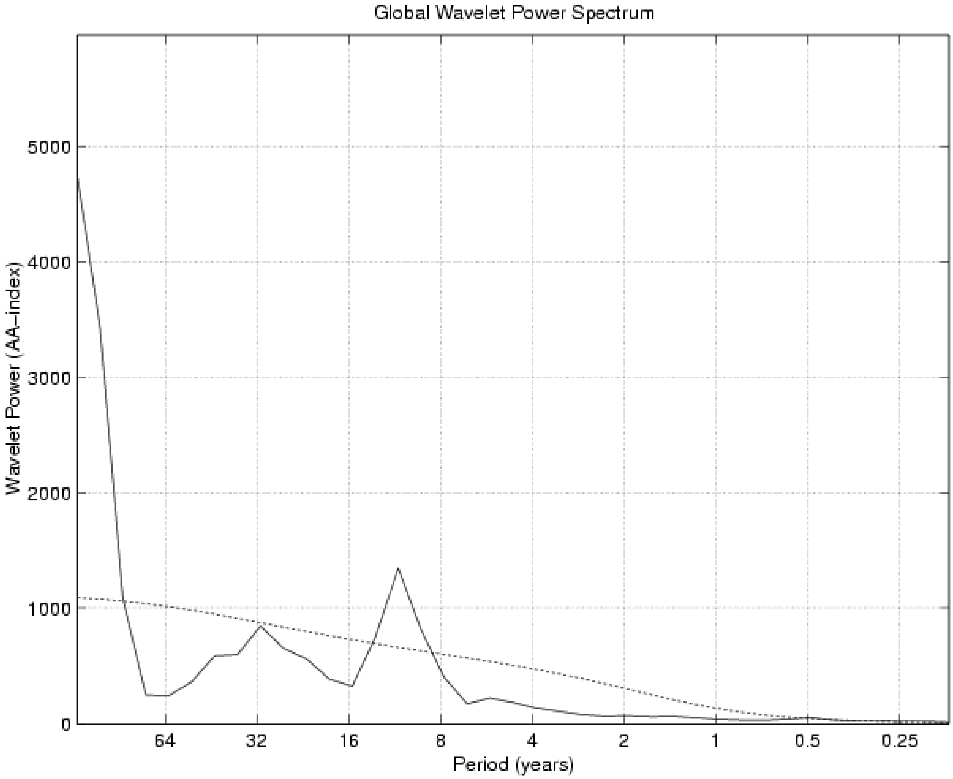


Figure 4.7. The power spectrum of the aa-index indicates the presence of an 11-year cycle in the magnetic field. There are also hints of a cycle with 32-year periodicity, but this is too weak to be significant. The estimation of the longer periodicities becomes extremely uncertain due to the short length of the aa-record. The dashed line shows an estimate of the 5% confidence level assuming a red noise process. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

time-series is a two-dimensional complex field. The results of such an analysis on the sunspot number is shown in Figure 4.9. The colours indicate how prominent the variations associated with timescales (read along the y-axis) are at a given time (the x-axis). This plot may be thought of as a series of spectrograms similar to that in Figure 4.6, but oriented along the y-axis, where each "spectrogram" is representative for a brief local time interval (centred on the time given by the x-axis). The wavelet analysis can furthermore be used to compute the solar cycle length (SCL), which is shown as a thick light grey curve in Figure 4.9. The black contours indicate the variations that have a statistically significant signal and the "cone" at either end shows the regions which are too close to the beginning or end of the record to give reliable estimates of the spectral power.

The wavelet analysis of the sunspot record tells us that the Schwabe cycle has been the most prominent feature throughout the record. The solar cycle lengths fluctuated relatively strongly before 1850. A "100-year signal" is also visible, but

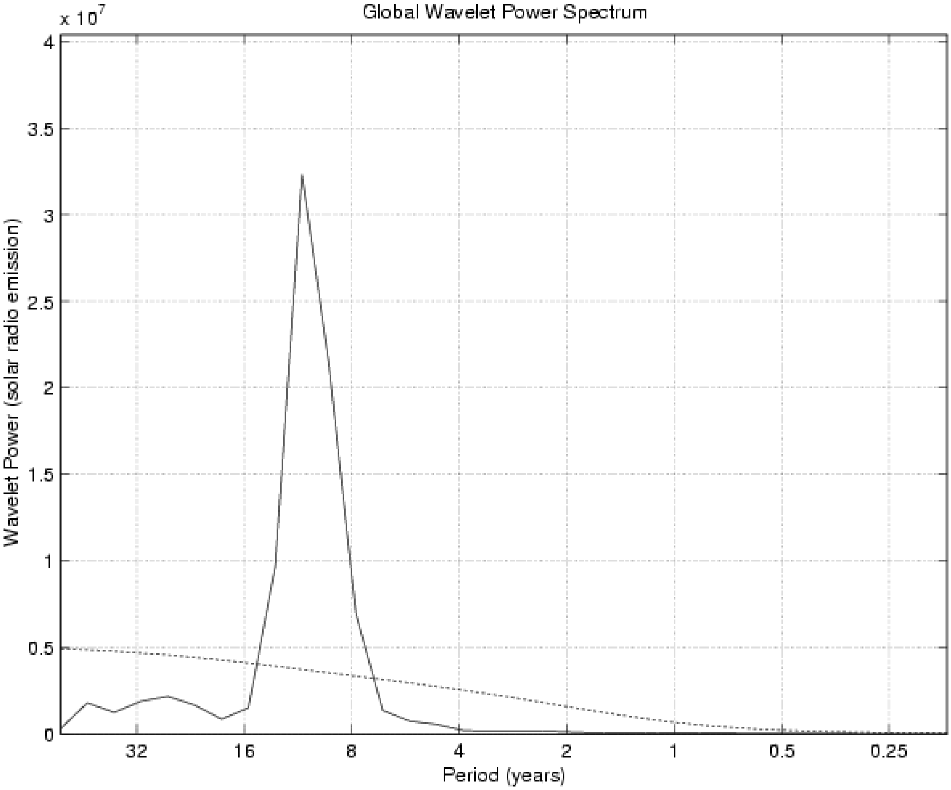


Figure 4.8. Spectrogram of the solar radio emission gives a clear indication of an 11-year cycle, suggesting that there is a close association between the solar radio emission and the sunspot cycle. The dashed line shows an estimate of the 5% confidence level assuming a red noise process. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

the timescale of this cycle has increased after 1850. There are occasions when weak 4-year cycles may be present in the sunspot record (around 1775, 1850 and 1960). There is little evidence of an annual cycle in the early sunspot record, which may suggest that the first observations were not severely hampered by seasonal variations in the atmospheric transparency (for instance cloud cover or haze).

The wavelet technique allows a decomposition of the sunspot record in terms of frequency bands, and Figure 4.10 shows the power in the 7–15-year frequency band. The variance of the Schwabe cycle increased up until 1950, and has since decreased. There were periods with weak 11-year variations before 1750 and around 1800.

The solar cycle affects the solar radio emission as well as the sunspot number. The emission is high during high sunspot activity (Figure 4.11). The spectral analysis of the solar radio emission does not give any indication of slow cycles; however, the solar radio record is much shorter (1947–2001) than the sunspot

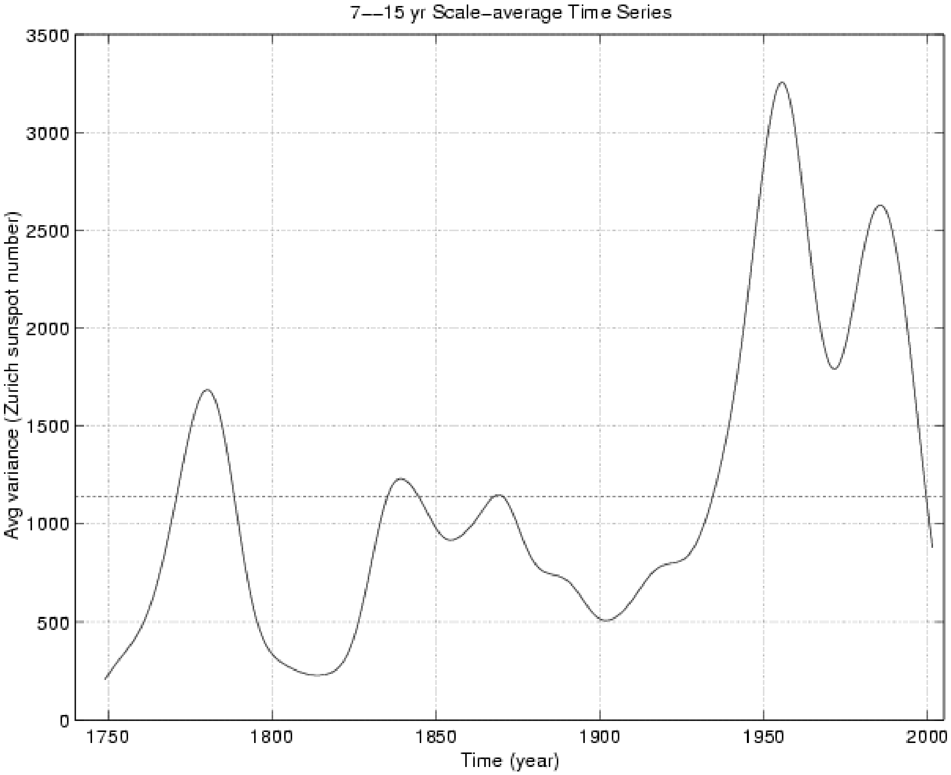


Figure 4.10. The evolution of the variance in the sunspot record accounted for the 7–15-year band, spanning the Schwabe cycle. There has been an intensification of the sunspot variations within this frequency band since 1900. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

number, and therefore a 100-year cycle cannot be identified as easily in the radio emission. The wavelet analysis shown in Figure 4.12 (see colour plate section) shows hints of 0.5–1.0-year cycles modulated about every 10–12 years. It is also evident from Figure 4.11 that this high frequency feature is enhanced during high sunspot activity.

The 11-year sunspot cycle (the 7–15-year band) is modulated about every 30 years, with peaks in the late 1950s and late 1980s. These peaks correspond with the two last local maxima seen in the sunspot curve (Figure 4.10), and the strength of the 11-year cycle in the solar radio emission follows that of the sunspot record. There is a clear indication of some relationship between these two features.

Although the spectral methods give estimates for the low-frequencies, these must also be treated with caution in a statistical sense. The reason is that a random process may by chance produce similar undulations over a few cycles. The statistical confidence of the spectral estimates increases with the number of cycles in the time-series segment.

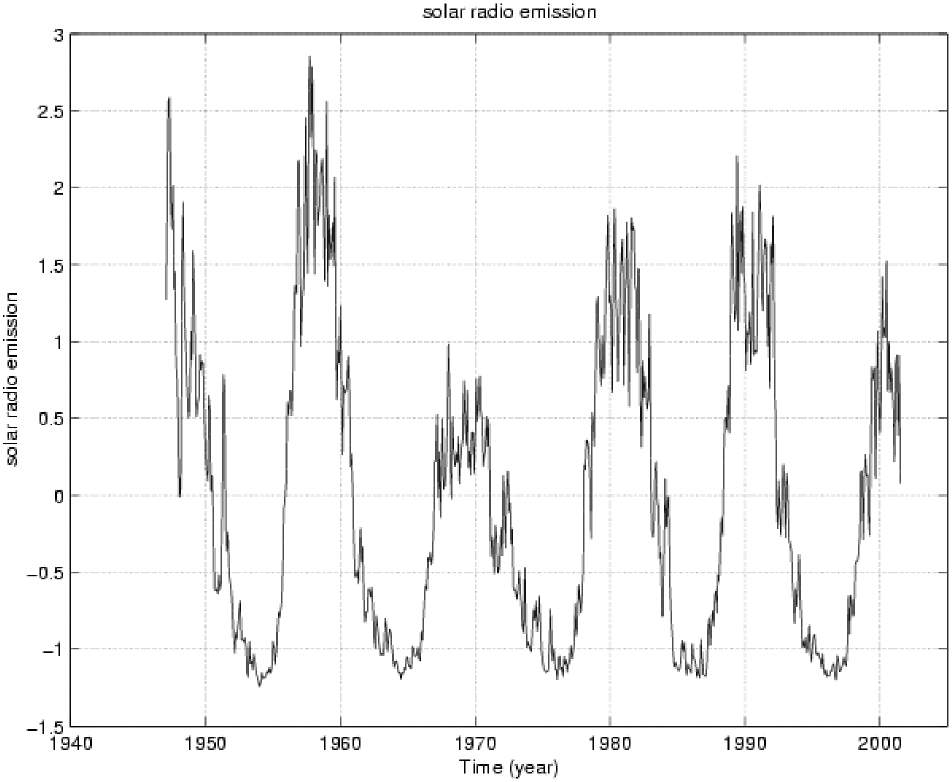


Figure 4.11. The measured solar radio emission, showing high frequency variations superimposed on top of the 11-year cycle. The high-frequency signal is most prominent during sunspot max. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

4.6.4 Comparison between preferred timescales

Monin (1972) considered extraterrestrial influences, such as meteor showers and solar activity, on climate. He noted that there may be an increase in the amount of precipitation and number of downpours in the month following the meteor showers. As regards the solar activity, Monin acknowledges the influence on geomagnetic phenomena (magnetic storms, auroras, etc.), but argues that the hypotheses concerning a physical mechanism whereby the solar activity can influence the weather lack convincing substantiation. The energy variations associated with the corpuscular beam (see Section 4.9.4) that is intercepted by Earth is small compared to the energy of cyclones. Szocs and Kosa-Kiss (2001) have recently argued that the cyclones may, on the other hand, be influenced by the solar activity and that the solar-particle flux plays an important role. They considered the possibility for these corpuscles acting as a triggering mechanism, and searched for statistical evidence to support this hypothesis.

By comparing the periodograms for the air temperature in Moscow (130 years) and Leningrad (St. Petersburg, 198 years) and the Wolf sunspot number, Monin found that the temperature variations are associated with a significantly shorter timescale than the 11-year cycle. By considering the harmonics of the 11-year cycle and the fact that the rise of sunspot numbers is faster (4 years) than the fall (7 years), the results indicate some coherency around the 2–5-year timescale although this is no evidence for a link between the solar activity and the temperatures. Monin cited a study by Brier who summed eight station temperatures and found no periodicity for the 22–23 year, 11–12 year or the 5–6 year timescales. Furthermore, the recent warming could not be a result of increasing solar activity as the growth in the Wolf number from 1720 to 1790 was not accompanied by a similar warming. Hence, Monin argues that the solar activity plays no major role in Earth's weather or climate. Bond *et al.* (2001) inferred from ice drift tracers at the sea floor an 'enigmatic-but-pervasive quasi-periodic 1500-year cycle' in 70–1800 year band-pass filtered geological proxies over the last 12,000 years. From isotope proxy records (see Section 2.4.1.1), Solanki *et al.* (2004) proposed that there have been 31 periods with 10-year average sunspot number counts exceeding a threshold value of 50 during the past 11,400 years. They noted that the these high-sunspot-count episodes lasted on average for about 30 years, and that the longest of them was 90 years. Furthermore, the current state of solar activity appears to be both unusually high as well as having lasted unusually long. Isotopic records may also suggest a presence of $\sim$2400, 208, in addition to the 90-year periodicities (Lean, 2005).

4.7 SUNSPOT GROUPS AND THEIR MAGNETIC FIELD

Hale and co-workers discovered in 1913 that in each hemisphere, the magnetic fields associated with the sunspots changed direction approximately every 22 years. The magnetic fields associated with the sunspots are aligned almost in parallel with the equator (inclination). The sunspot magnetic fields in the northern hemisphere are in opposite directions to the corresponding fields in the southern hemisphere. The 22-year variability is known as the *Hale cycle*. The toroidally shaped magnetic field reverses direction at each sunspot minimum. This cycle is further believed to be modulated by a 200-year envelope (Thomas and Weiss, 1992). There are two types of model that have been proposed to explain the Hale cycle: (i) oscillator models and (ii) dynamo theories. So far, there is no widely accepted explanation of how an oscillator mechanism can give rise to the solar cycle, and the dynamo theories are more widely accepted.

Wolf suggested in 1862 that there is also a 78-year cycle associated with the sunspots, which is composed of seven 11-year cycles, and has maxima in 1778, 1860 and 1947. The 78-year cycle is seen in the sunspot distribution on the northern and southern hemispheres. Bray and Loughhead (1964, p. 239) argue that the 78-year and 200-year cycles are weak and, despite the numerous studies of various timescales, it is only the 11-year periodicity in the sunspot cycle that is established beyond doubt. A prominent secondary peak that may correspond to the 78-year cycle is seen in Figure 4.6, but only three complete cycles of this periodicity fit in a 250-year-long record.

4.7.1 Alternative measures of solar activity

Another method of estimating the sunspot activity is by the *area* occupied by the sunspots. This method became more reliable after heliographs came into routine observations. The sunspot area and the Wolf number tend to give similar results, although there are some differences. An approximate relationship between monthly mean sunspot number and area is $A_{\text{spot}} \approx 16.7R_z$ (in millionths of the visible hemisphere).

The Sun's diameter, it has been argued, varies slightly with time and the changes are found to be proportional to the inverse of the sunspot activity ("Secchi-Rosa law") (Kuiper, 1953, p. 18). The solar magnetism is believed to affect the radius as well as the solar output by regulating the heat transport in the convective layer. In other words, the Sun's luminosity varies with the solar cycle as seen in Figure 4.13, and solar physicists have tried to measure the ratio between the relative radius and luminosity changes (W) because it is hypothesised that this ratio gives an indication of the depth of the underlying mechanism causing the changes in the luminosity and radius. Calculations suggest that such a regulating mechanism taking place near the solar surface yields $W \approx 0.002$, near the base of the convective source gives $W \approx 0.2$, and in the core $W \approx 0.5$. A recent estimate of W is around 0.04, suggesting that the regulating mechanism responsible for the variations in the luminosity and radius is not likely to be located deeply inside the Sun. On the other hand, this estimate also suggests that the cause of luminosity change is not simply a change in the effectiveness of the convection at the solar surface.

Wittmann and Bianda (2000) have recently made measurements of the solar diameter. The solar semi-diameter was measured from two locations which gave the mean values of $R = 960.63 \pm 0.02$ (Izaña, Tenerife, 1990–2000) and $R = 960.66 \pm 0.03$ (Locarno, Switzerland, 1990–1998) respectively. They find, in contrast to other findings, that the measured solar semi-diameter does not show long-term variations outside the range ± 0.0003/year nor solar cycle-related variations exceeding ± 0.05.

4.7.2 Observed east–west asymmetry in sunspot statistics

Statistical studies by Mrs Maunder in 1907 have revealed a small west–east asymmetry of sunspot appearance (east–west ratio of around 0.98 for sunspots and 0.93 for groups), suggesting that there are more spots seen on the east side than on the west side of the Sun. Sunspot formation is also more frequently found on the east side than in the west hemisphere, and more sunspots enter the east limb by solar rotation than disappear on the west limb (Figure 4.14). This fact may suggest that not all sunspots are discovered before disappearing behind the limb,

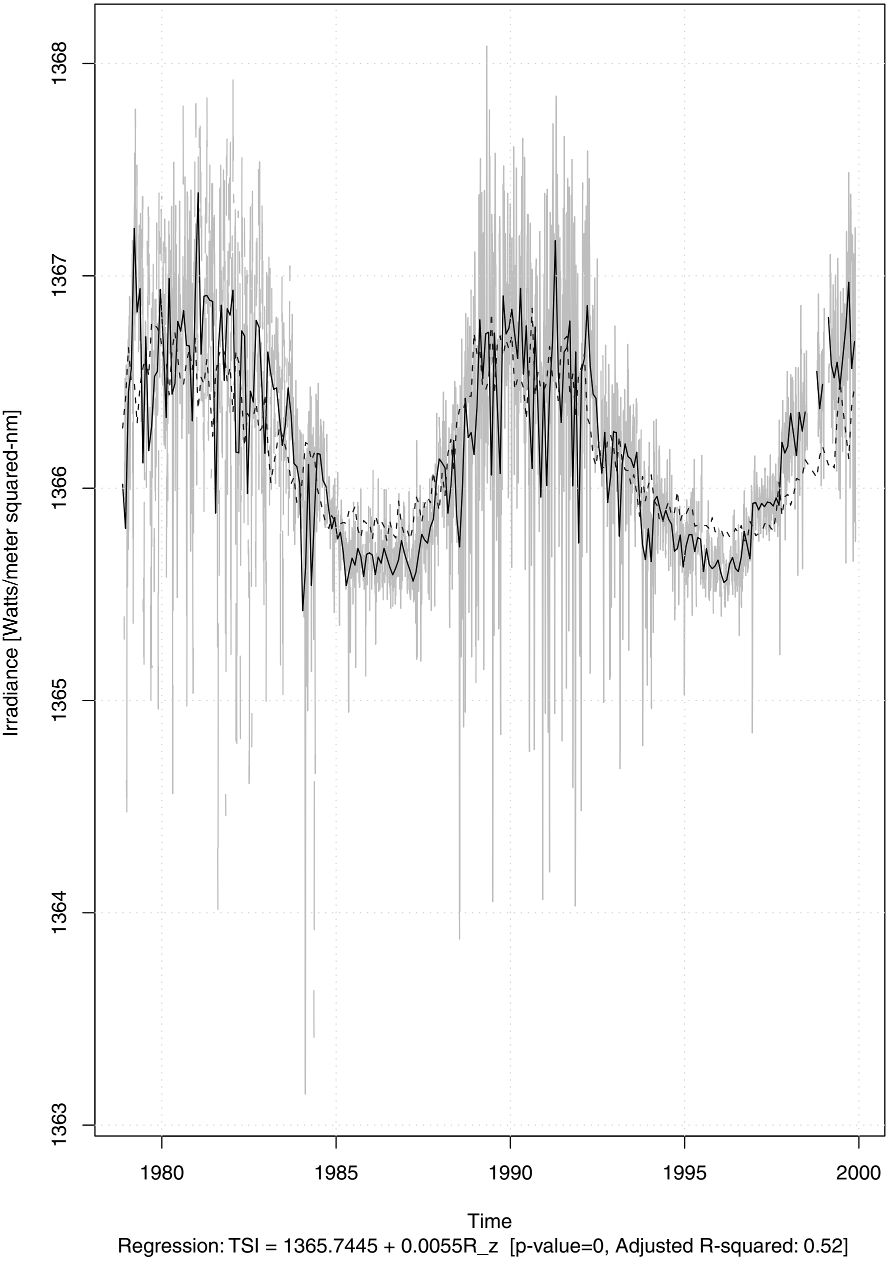


Figure 4.13. A composite of total solar irradiance (TSI) measurements from satellites (grey). The black line shows the monthly mean TSI and the dashed line represents the reconstruction of the TSI using a model based on the Zurich sunspot number. 52% of the variance in TSI can be associated with the sunspot cycle. The *p*-value (the probability that the null-hypothesis is true) is 0. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

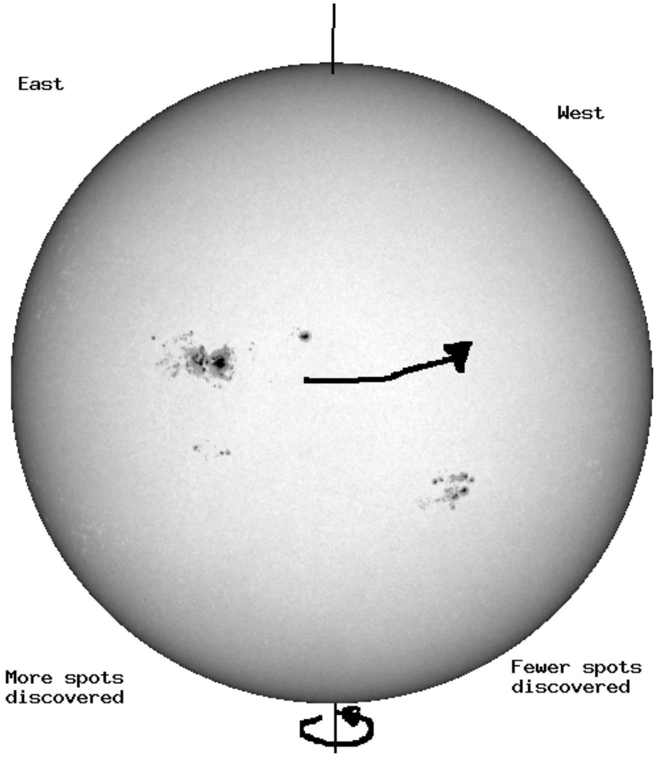


Figure 4.14. A schematic diagram showing solar rotation from east to west. Courtesy: SOHO/MDI (ESA & NASA).

as there is no known physical reason why the number of sunspots should vary with longitude. The Earth, and hence the longitudinal angle from where we observe the Sun, changes continuously as the Sun itself spins. It is argued that the statistical asymmetry is small enough for its effect on the visibility function to be ignored (Kuiper, 1953, p. 328).

Mrs Maunder also suggested that the east–west asymmetry may be interpreted as the sunspot having axes that lean toward the west, and that such "tilted" sunspots become more invisible on the west due to an additional foreshortening effect by the Sun's curvature. The degree of westward tilt depends on the sunspot age, with the oldest spots having greatest tilt. Preceding spots (p-spots, or spots leading the rest of the group in terms of solar rotational motion) are usually more compact than following spots (f-spots, or spots following the p-spots in terms of solar rotation) and are therefore more visible.

4.8 THE SUNSPOT CYCLE AND TOTAL RADIANCE

Space-borne measurements of the total spectral irradiance have been made since the late 1970s, and there is a clear relationship between the solar cycle and the total

Table 4.6. Mount Wilson sunspot magnetic classification (Bray and Loughhead, 1964).

Class	Description
Alpha	A unipolar sunspot group
Beta	Sunspot groups having both bipolar magnetic fields
Gamma	Sunspot groups complicated and irregular positive and negative polar structures
Beta-gamma	Sunspots with bipolar structure, but with sufficient complexity that one continuous line can be drawn between spots of opposite polarity
Delta	Umbra separated by less than 2 degrees from penumbra with opposite polarity
Beta-delta	Groups with beta magnetic classification and one or more delta spots
Beta-gamma-delta	A beta-gamma group with one or more delta spots
Gamma-delta	A sunspot group under gamma category, but with one or more delta spots

electromagnetic energy output from the Sun. The solar cycle is associated with small variations of the total solar luminance (radiative energy summed over all frequencies) of the order 0.1%. Analysis by Judith Lean and others (Lean and Rind, 1998; Fröhlich and Lean, 1998a,b; Bertrand and van Ypersele, 1999) indicate variations between 1363 and 1368 W/m^2, with peak energy in the late 1970s to early 1980s and early 1990s (Figure 4.13). The brief dips in the TSI are interpreted as being due to the passage (blocking) of sunspots across the solar disk on a rotational timescale (Solanki and Fligge, 2000). The changes to the incoming solar radiation are estimated to be too small to produce the large global mean temperature fluctuations seen on Earth (see Sections 5.4 and 8.6.1.1), unless there is an amplification mechanism involved. Willson (1997), however, proposed that the TSI was 0.5W/m^2 higher during the minimum of solar cycle 22 (1996) compared with cycle 21 minimum (1986). Other TSI reconstructions (Frölich and Lean, 1998a; referred to as the 'PMOD' composite) show almost identical TSI levels at these two minima, which is also obtained through TSI reconstructions from the sunspot number (Figure 4.13). The analysis by Frölich and Lean (1998a) suggested an insignificant 'trend'[9] as well as lower TSI at solar maxima than corresponding estimates in the 'ACRIM' TSI composite of Willson and Mordvinov (2003). Willson and Mordvinov argued that the lower 'trend-estimate' and lower TSI at solar maxima are both due to artefacts of ERBS degradation (note: it is easy to misunderstand the abstract of their paper, as they really argued that TSI at solar maxima was lower in the PMOD composite than corresponding estimates from their own reconstruction; not that the TSI in the Frölich and Lean (1998) analysis was lower during solar maxima than for solar minima). If there really has been a trend in TSI over the most recent solar cycles, then this would have taken place despite no systematic changes

[9] A difference between mean states of two brief periods hardly constitutes as a 'trend'.

in other solar indices (e.g., sunspot number,[10] GCR, 10.7 cm flux[11]). Moreover, the 'trend-estimate' is sensitive to how the various data sources are pre-processed and combined. Whereas the PMOD data has been obtained after using the ERBS to adjust downward the estimates and is in better agreement with solar proxies, the ACRIM is closer to the original data (Willson and Mordvinov, 2003). Willson and Mordvinov (2003) claimed that a degradation in the ACRIM1-part of the PMOD data was due to an incorrect assumption regarding the cumulative solar exposure of its spin-mode SMM sensor. The jury is still out on this issue, as the PMOD's degradation corrections cannot be validated since the NIMBUS/ERB did not have a degradation self-correction capability. Willson and Mordvinov (2003) suggested that the most likely explanation for why the ERBS measurements differ from those of ACRIM1 is an uncorrected ERBS degradation, however, this cannot be proved to be the case. Furthermore, their argument hinges on the assumption that the (three-fold redundant) degradation self-calibration works flawlessly. Another problem is that there is no explanation for how the TSI can change while other solar proxies indicate the same level of activity. It is the TSI during solar minima that is purported to have changed from 1986 to 1996, but the minimum monthly sunspot is the same for the two minima ($R_z = 1$ both in 1986 and 1996; Table 8.3). Furthermore, the monthly sunspot number was lower in year 2000 ($R_z = 170$) than in year 1990 ($R_z = 200$).

4.8.1 Sunspots and the solar irradiation

Although the individual dark sunspots reduce Sun's total radiative output, the Sun tends to be brighter when there are many sunspots because other factors such as plages and faculae also affect the solar brightness. Lean and Rind (1998) show that these variations are evident as 27-day variations with magnitudes as much as "a few tenths of percent" in the total solar irradiance in addition to the prominent 11-year cycle (0.1%). The short-term variations are measured accurately but space-borne observations of the long-term trends are associated with great uncertainties due to the short lifetimes of the space-based solar monitor systems. Instrumental uncertainties may result in spurious readings in individual solar radiometers (often most serious during the first year of deployment). Instrumental changes could have produced an artificial upward trend in 1992. Discrepancies between the readings by the ERB and the ERBS data during 1990–1992 may be due to systematic errors related to temperature or aspect drift. de Toma and White (2000) observed that the rise phase of cycle 23 (2000–) has fewer sunspots and faculae than the rise of cycle 22 (1990–2000), but the TSI measurement indicates similar radiative output as during the rise of cycle 22. This observation brings up the question of whether the solar radiative variability is affected by other factors than sunspots, faculae and enhanced network.

[10] Figures 2.3 and 4.13.

[11] Figure 4.11.

On January 29 1993 fewer sunspots were visible than on February 10 1993, and the total solar irradiance measurement had dropped from around 1367.8 to 1366.2. Thus, over the short timescale of solar rotations (27 days), the darkening effect of sunspots tends to dominate the total solar irradiance, but over longer timescales, such as the 11-year timescale, the facular brightening is on average twice as strong as the sunspot depletion. The ultraviolet (UV) light with 200-nm wavelength is mainly associated with faculae increase with the number of sunspots and is not affected much by the sunspots themselves.

The facular brightening tracks the monthly mean sunspot number throughout the Schwabe cycle. On timescales longer than 11 years, however, it is assumed that the network facular emission varies with the overall solar activity.

4.8.2 Sunspot-irradiance models

4.8.2.1 Regression

One of the best ways of testing a scientific hypothesis is to make predictions. In order to make predictions, one needs a model. In this case, we can develop empirical models based on past statistics. There are several ways of making such models, and one of the most common methods is to apply a linear regression to estimate the best-fit linear relationship between two or more quantities.

Linear regression is a method for finding coefficients, m and c, that give the best fit between two variables, and the "best-fit" criterion is often made by minimising the root-mean-square errors (RMSE) by doing a least-squares-fit. Studies of solar–terrestrial links often involve regression between the sunspots (x) and temperature (y) by assuming a linear relationship $y = mx + c$. By transforming the quantities (for instance by taking the square of a quantity: $x \rightarrow x^2$), any linear method may be used to describe any nonlinear relationship. Regression analysis is discussed in most books on statistics (e.g. Wilks, 1995 and Strang, 1995) and is available in most statistical tools (e.g. R, S-plus, SAS, Minitab). The reason for reviewing the regression analysis is that it is commonly used in the study of relationships between two quantities, and it is important to know the strengths and weaknesses of this type of analysis. It is necessary to have a detailed knowledge of the regression for understanding the strong and weak points of the method.

Let $\hat{y}$ be the best-fit solution of the linear relation $y = mx + c$.
The root-mean-squared error is defined as:

$$\begin{aligned} e_{\text{rms}} &= \sqrt{\frac{\sum_{n=1}^{N}(y_n - \hat{y}_n)^2}{N}} \\ &= \sqrt{\frac{\sum_{n=1}^{N}(y_n^2 - 2mx_n y_n - 2cy_n + m^2 x_n^2 + 2mx_n c + c^2)}{N}} \end{aligned} \tag{4.8}$$

In order to find the optimal solution that minimises the error, one applies a variational method, which implies differentiating e_{rms} with respect to m and c respectively and equating these expressions to zero:

$$\frac{\partial e_{\text{rms}}}{\partial m} = \left[\frac{\sum_{n=1}^{N} 2(x_n y_n + m x_n^2 + x_n c)}{\sum_{n=1}^{N} (y_n^2 - 2m x_n y_n - 2c y_n + m^2 x_n^2 + 2m x_n c + c^2) N}\right]^{\frac{1}{2}} = 0$$

$$\frac{\partial e_{\text{rms}}}{\partial c} = \left[\frac{\sum_{n=1}^{N} 2(m x_n + c - y_n)}{\sum_{n=1}^{N} (y_n^2 - 2m x_n y_n - 2c y_n + m^2 x_n^2 + 2m x_n c + c^2) N}\right]^{\frac{1}{2}} = 0$$

Which implies that

$$\sum_{n=1}^{N} (x_n y_n + m x_n^2 + x_n c) = 0$$

$$\sum_{n=1}^{N} (m x_n + c - y_n) = 0 \tag{4.9}$$

It is straightforward to solve equation (4.9) by calculating the sums of the individual terms, given x and y. The constant and the slope can be found solving (assuming that the uncertainty σ_i is similar for all observations):

$$\Delta = N \sum_{n=1}^{N} x_n^2 - \left(\sum_{n=1}^{N} x_n\right)^2$$

$$c = \frac{(\sum_{n=1}^{N} x_n^2)(\sum_{n=1}^{N} y_n) + (\sum_{n=1}^{N} x_n)(\sum_{n=1}^{N} y_n x_n)}{\Delta}$$

$$m = \frac{N(\sum_{n=1}^{N} x_n^2) + (\sum_{n=1}^{N} x_n)(\sum_{n=1}^{N} y_n)}{\Delta} \tag{4.10}$$

The next step is to estimate the uncertainties associated with these coefficients. The error estimates can be calculated using:

$$\sigma_c^2 = \frac{\sum_{n=1}^{N} x_n^2}{\Delta}$$

$$\sigma_m^2 = \frac{N}{\Delta} \tag{4.11}$$

The probabilities that the best-fit estimates in equation (4.10) are different from zero can be estimated using the standard deviations from equation (4.11) and assuming that the true values belong to a population that is normally distributed around the estimates from equation (4.10).

An inverse regression can be written as: $x = m^{-1} y + c'$. In principle, the slope m estimated using ordinary regression should be the inverse of the slope of the latter type. However, this is usually not so in real analysis using traditional least-squares where errors in y are minimised, as the optimal root-mean-squared error in terms of x may differ from that in y. It is therefore a good practice to do the regression both ways, and compare the two solutions. If the two solutions are similar, then the fit is good. However, if the two quantities are unrelated so that the slope is small, then one of the cases is ill-defined and the two lines will have very different slopes.

Regression may also be carried out on more than one variable just as well as on one variable (univariate regression) as described above. *Multiple regression* is appropriate when a single time-series y is a function of more than one factor, and may be solved by finding the optimal fit to the equation $\hat{y} = m_1 x_1 + m_2 x_2 + c$. *Multivariate regression* refers to a regression method that describes several time-series $(y_1, y_2, \ldots, y_n)$ in terms of a number of predictors (i.e. $m_{11} x_1 + m_{21} x_2 + c_{11}$) and is most easily done by adopting linear algebraic notations, and by writing the time-series as a vector in data space: $y \rightarrow \vec{y}$ and $x \rightarrow \vec{x}$.

Strang (1995) shows that regression can be done by applying a projection transform of $\vec{y}$ onto $\vec{x}$ to get $\vec{\hat{y}}$. The projection of one vector onto the other implies that the difference between $\vec{y}$ and $\vec{\hat{y}}$, which is defined as the error, is normal to $\vec{x}$, and hence involves the shortest distance (smallest error) from $\vec{y}$ and any point on the line $\vec{x}$.

$$\vec{\hat{y}} = \vec{x}\mathbf{A} \tag{4.12}$$

$$\mathbf{A} = \vec{x}(\vec{x}^T \vec{x})^{-1} \vec{x}^T \vec{y} \tag{4.13}$$

A least-squares-fit regression will by definition always return a solution that minimises the RMSE, which means that the outcome may be sensitive to outliers and long-term trends (see Section 8.4.4). When the different observations are of different accuracy (reading errors), weights may be applied to the data so that those of the highest quality have a stronger effect on the outcome than those of low quality. Such weighting must always be justified. The regression results are often summarised in an analysis-of-variance (ANOVA) table. In addition to the estimated variables (equation (4.10)) and the associated uncertainties (equation (4.11)), the ANOVA typically gives: (i) the degrees of freedom (e.g. the number of data points minus the number of estimated parameters); (ii) the "r-squared" (R^2) which is a measure of how much of the variance the solution can account for; (iii) *f-statistics*;[12] and (iv) the probability estimate (p-value) for the rejection of the null-hypothesis saying that there is no relationship between the two quantities. The ANOVA therefore gives an indication of the goodness-of-fit and the likelihood that there is no

[12] The *f-statistics*, also referred to as the f-ratio, is a ratio of the variance of the regression fit to the variance of the residual, and provides a measure for the "strength" of the regression.

relationship between the quantities. An R^2 close to 1 and a zero p-value give a strong indication of a real link, whereas $R^2 \approx 0$ and p-value > 0.10 suggest that there is no connection between the two series. Ideally, the test of an empirical model ($\hat{y} = mx + c$) obtained through regression should involve prediction of independent (future) data.

Regression analysis is ideal for the study of physical problems where there is an *a priori* reason to believe that the quantities are related to each other. The regression coefficients may, for instance, give an indication of amplification of a signal.

Figure 4.13 gives an example of a regression analysis. Here we have applied the regression to find the relationship between the monthly mean sunspot number and the TSI. The fit between these two quantities is very good ($R^2 = 52\%$, p-value $= 0$), and the linear regression of the sunspots onto the composite of space-borne measurements TSI from various satellites, shown in Figure 4.13, yields the following relationship between the two quantities:

$$\mathrm{TSI} = S = (1365.7445 + 0.0055R_z)\ \mathrm{W/m^2} \tag{4.14}$$

Wilson and Hudson (1988) arrived at a similar equation for the same relationship:

$$S = (1366 + 7.71 \times 10^{-3} R_z)\ \mathrm{W/m^2} \tag{4.15}$$

Hoffert *et al.* (1980) also proposed a relationship between the TSI and the sunspot number, that can be described by the equation

$$S = S_0(1 + \beta[R_z(t) - R_{z0}])\ \mathrm{W/m^2} \tag{4.16}$$

The terms S_0 and R_{z0} are reference values (the ANOVA is unfortunately not available). Reid (1987) assumed $\beta = 1.08 \times 10^{-4}$ and calculated a change in TSI of $\approx 0.6\%$ for the increase in the sunspot number between 1910 and 1960 $[\Delta R_z(t) = 57]$.

Hoyt and Schatten (1993) proposed a simple relationship between the solar activity component of the irradiance and sunspot number as:

$$\Delta S = 0.01R_z$$

They compiled a table of reconstructed solar irradiance variations using different solar activity proxies, reproduced in Table 8.4, and noted that there is a relatively good phase agreement between the various solar indices, but the solar rotation reconstruction appeared to be approximately 11 years out of phase with the other curves (their figure 8).

Lean and Rind (1998) used a climate sensitivity of $\Delta S = 0.1233\Delta R$ which together with a TSI reconstruction can explain a 0.27°C warming since 1900 and 0.12°C since 1970.

4.8.2.2 Residuals

The difference between the real data and the best-fit is often referred to as the "residual". The residual may sometimes hold information about the variable that could not be explained by the time-series against which the data was regressed. Such information may come from additional factors influencing the quantity. For a perfect regression where y only depends on x, then the residual is expected to have zero mean, hold random values, and be normally distributed. Often it is assumed that the residual of a perfect-fit is a white noise process (a series with random values that have zero autocorrelation). In earth science, however, there are many processes more similar to red noise processes,[13] and one must therefore expect to find residuals with red noise character as well.

4.8.2.3 Solar activity and UV radiation

Most of the power is in the wavelength range of 400–700 nm, with very little power at wavelengths shorter than 300 nm (Figure 3.3). However, the proportional variation is highest for the shorter wavelengths. The solar energy output varies with the solar cycle, (a) because faculae, which radiate more energy, are more frequent during sunspot maximum and (b) because the sunspot areas are associated with a lower energy output. The faculae brightening is more dominant than the sunspot darkening and the Sun emits more energy at sunspot maximum than at sunspot minimum. Although the solar ultraviolet (UV) frequency band represents a small fraction of the radiative energy, the solar cycle variability in the UV band is of the order of 0.7% (Lean, 2000).

4.8.2.4 The rise-time of solar cycles

Empirial studies (Kuiper, 1953, p. 333) suggest that the area under the rising slope of the sunspot cycle is approximately constant, which means that it takes a longer time to reach maximum sunspot activity when the peak sunspot number is small than for high peak values. This observation is illustrated in Figure 4.15, but it seems that this rule of thumb does not fit for the more recent observations. There seems to be no such relationship for the decaying part of the solar cycle. The time between a sunspot minimum and a subsequent maximum tends to be longer than the time between a maximum and the following minimum.

4.8.2.5 Long-term variations in the solar cycle and the solar radiation

The question whether the solar radiation is related to solar activity remains unanswered for timescales longer than one solar cycle. There have been proposals suggesting that the solar output is related to the length of the solar cycle, but this hypothesis has not yet been proved to be true. Hoyt and Schatten (1993) argued that the perturbation of the energy transfer in the convective zone due to active regions may have long-term effects on the solar output, even if these features do not extend deeply into the convection zone. There are four common types of models that relate the solar output to the solar cycle: (a) the constant quiet Sun model, (b) the activity envelope model, (c) the solar diameter model, and (d) the umbra–penumbra ratio model. These types of model will be discussed in more detail later.

[13] A red noise process, also referred to as an AR1 process, produces a series of numbers that are partly random and partly correlated with the adjacent data points.

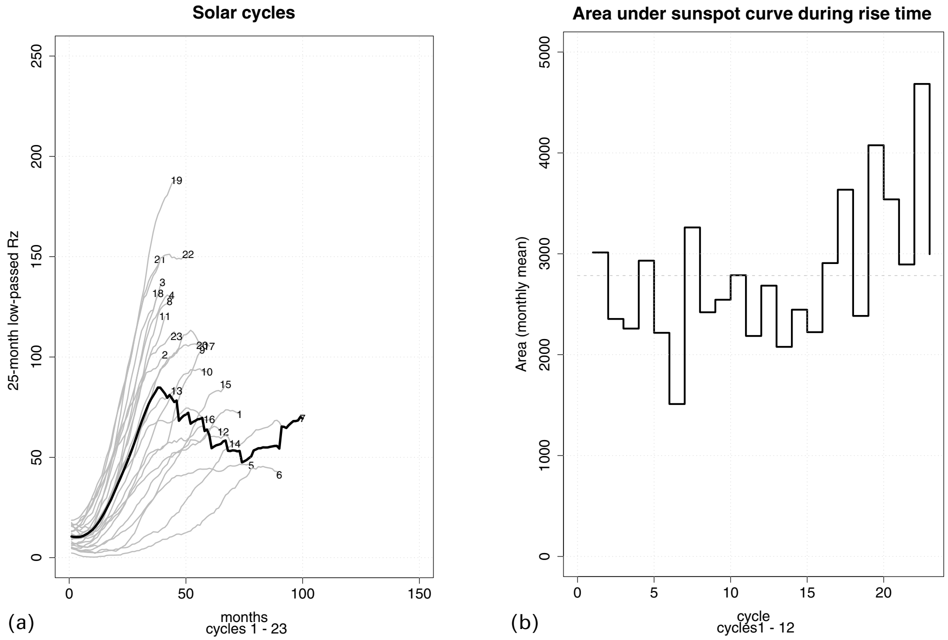


Figure 4.15. (a) The comparison of the rise from minimum to maximum of individual solar cycles (grey) and a composite (black). There is a clear tendency for the peak being earlier when the maximum sunspot number is high. (b) The area under the curves of the monthly mean (25-month low-pass-filtered) R_z over the rise interval. The horizontal dashed grey line marks the mean area over all 23 solar cycles. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

The main argument justifying these models is that the solar irradiance depends on the energy transport from the solar interior though the convective zone. A simple model of the solar irradiance can be expressed in terms of the total energy production in the solar core ($\eta(t)$) and the transport efficiency ($\alpha(t)$) from the solar interior to the solar surface. If the thermonuclear energy production in the solar core varies with time, the solar irradiance will change accordingly, everything else being constant. Furthermore, the energy reaching the solar surface depends on the energy transport, and is therefore sensitive to the zones with the lowest transfer rates in a similar fashion to the water transport of a river depending on the point with lowest transport (rapids with strongest current). A simplified model describing the solar irradiance at the photosphere, I_p, can be written mathematically as $I_p(t) = \alpha(t + \tau)\eta(t)$. Thus, the solar brightness is dependent on the efficiency, $\alpha(t + \tau)$, at which the turbulent cells convect energy outward.

4.8.2.6 "Quiet Sun" models

Model (a) describes a Sun whose irradiance only depends on the Schwabe cycle and postulates that all variations in the solar output can be related to solar activity. Solar mechanisms that may affect the irradiance may include sunspot blocking, enhanced emission due to faculae, and *network emission*. Network emissions are extensive strand-like (like networks) features with enhanced irradiation.

The solar convective energy transfer may be enhanced by the presence of sunspots, which are associated with large-scale eddies. It has been postulated that these eddies are more efficient at convecting heat from the solar interior than other convection cells in the convective zone.

4.8.2.7 Penumbral structure and solar activity

Other proxies may be the rate at which the sunspots decay, the fraction of sunspots that have penumbral structure, the decay rate of the solar cycle, and the mean solar activity level. The solar convection disperses the magnetic field lines associated with the sunspots and hence cause the decay of the spots. It is usually assumed that the rate at which the sunspots dissolve is proportional to the mean convective velocity.

4.8.2.8 Umbra–penumbra ratios

If the umbra is more persistent than the penumbra, this may suggest that the penumbra is more sensitive to the photospheric convection rates. The ratio of umbral to penumbral area may therefore be another measure for the convective intensity and hence the convective energy transfer from the solar interior. The fraction of penumbral sunspots seems to vary with the convective activity.

The umbra–penumbra variation models assume the sunspot structure to be manifested in the ratio of umbral to penumbral area, and adopt this proxy as a measure for the solar irradiance. The umbral–penumbral area ratio is believed to be a function of the convective activity level, and therefore a measure of the convection rate and hence the energy transfer.

4.8.2.9 Solar rotation models

Convection may be closely related to the solar rotation though the Coriolis force, and therefore the solar rotation rate may be taken as a proxy for the solar output. Changes in the rotation rate affect the convection at the deeper levels, which again affects the convective energy transport and thereby the Sun's brightness (Hoyt and Schatten, 1993). It has been argued that the convection intensities may influence the rate at which the sunspots decay and dissolve.

4.8.2.10 The solar diameter and solar activity

The activity envelope models describe the long-term variations in the solar irradiance as a function of long-term changes in the solar activity, and the solar diameter model assumes a relationship between the size of the Sun and its output. There appears to be some controversy around the historical record of the solar diameter variations, which may imply that empirical models derived from these will be subject to errors.

4.8.2.11 The solar cycle length – SCL

The sunspot decay rate may influence the fall-off time of a given Schwabe cycle. An increase in the photospheric convection may result in more sunspots through the outward advection of magnetic flux tubes. Hoyt and Schatten (1993) proposed that this mechanism may also affect the solar cycle length by causing more rapid decay of the individual sunspots and "speeding up" the solar cycle.

Solar cycle lengths (SCL) are highly correlated with the sunspot decay rate (D_{spot}, 85% variance in common). Stewart and Panofsky (1938) and Hoyt and Schatten (1993) proposed that the solar cycle length is related to the sunspot decay time (equations (4.17) and (4.18), respectively). The relationship between the SCL and the decay time according to Stewart and Panofsky (1938) can be expressed as:

$$L_m = (246.5 \pm 0.7) - (4.73 \pm 0.07)D_{\text{spot}} \tag{4.17}$$

whereas Hoyt and Schatten (1993) obtained a slightly different relationship:

$$L_m = (251.1 \pm 5.4) - (3.98 \pm 1.02)D_{\text{spot}} \tag{4.18}$$

It is interesting to note that the model from 1938 (equation (4.17)) suggests lower uncertainty than the more recent model (equation (4.18)). This difference may be interpreted as a non-stationarity in the data series or insufficient data for proper calibration of the earlier model.

The solar cycle growth period (rise-time) is approximately constant[14]: 4.30 ± 1.10 years, and the variance in the solar cycle length that can be described by the rise-time is 12% compared to 36% for the fall-off time (cycles 1–21). For cycles 8–20, for which the data is regarded as more reliable, the solar cycle decay periods can account for 65% of the cycle length. Hoyt and Schatten (1993) give an empirical relationship between the sunspot decay rate and the Schwabe cycle decay rate:

$$D_{\text{spot}} = -(3.93 \pm 1.58) + (192.3 \pm 66.7)\frac{|(dR_z/dt)_{\max}|}{R_{z,\max}} \tag{4.19}$$

The solar cycle length is furthermore thought to be related to the mean cycle sunspot activity. The time interval in months between successive sunspot minima can, according to Granger (1957), be expressed as:

$$L_m = \frac{12}{(0.074086 + 3.47 \times 10^{-4}\overline{R_z})} \tag{4.20}$$

where $\overline{R_z}$ is the mean Wolf sunspot number for the given cycle. This expression can account for 66% of the variance in L_m.

According to Hoyt and Schatten (1993) the solar rotation leads the other solar irradiance indices by approximately 11 years. They speculated whether the solar rotation responds to the conditions near the base of the convective region and whether the other indices are related to near-surface photospheric processes. The

[14] It is usually the area under the curve during the rise of sunspot activity that is considered constant.

idea is that it takes some time for changes in the convection zone base to appear on the solar surface. The umbra-to-penumbra ratio model further leads the mean solar activity model by approximately 20 years.

There are also long-term variations in the solar irradiance which are not correlated with the sunspot number. There have recently been observed trends in the equivalent widths of the spectral lines which are not correlated with the Schwabe cycle. This suggests that there are changes in the photospheric temperatures believed to be related to the convective energy transport that are independent of the sunspots.

The notion of secular changes in the solar convective energy transport is supported by observations of the carbon 5380 Å line which is formed relatively deep in the photosphere. This line has systematically widened from 1978 to 1990, which suggests that the temperature gradient is changing secularly. Changes in the convective energy transport may result in variations in the photospheric temperature gradient, and more intense convection will lead to higher temperatures and slightly more pronounced limb darkening. Such an enhancement of the limb darkening is difficult to detect, but may possibly appear as changes in the solar diameter. However, the evidence for changes in the temperature gradients also points to secular changes in the solar limb darkening which may be unrelated to the level of solar activity.

Hoyt and Schatten (1993) postulated that the low-frequency variations in the solar irradiance involves changes in the convective energy transport that are deeper than those of shorter timescales such as sunspots and faculae. The latter last for a few days and involve only a few thousand kilometres of the upper solar convection zone. Two models have been proposed for explaining the variability in the solar irradiance: (i) statistical fluctuations in the total energy transport by a finite number of stochastic or turbulent convective cells; (ii) pressure fluctuations that may be a result of changes in the solar magnetic activity.

According to Harrison and Shine (1999), the reconstruction of the solar output involves three steps: (i) characterising the variability of the solar irradiance in terms of observable proxies; (ii) estimation of the size of the irradiance variations associated with the changes in the proxy; (iii) the absolute level of solar irradiance is set to recently measured values. In many of the recent studies, the size of irradiance in step (ii) has been obtained using the Maunder minimum as a reference, and measurements of calcium emission lines from "sun-like" stars.

Reid (1997) reconstructed a total solar irradiance record assuming a linear relation to the solar activity. He applied a 15-point Gaussian filter to capture the variations with timescales longer than the 11-year cycle, and the slope of fit was derived by assuming that the cold temperatures in the mid-17th century (Maunder minimum) were purely a result of changes in solar activity, and assuming that it was 1°C colder during this period than at present (Harrison and Shine, 1999). A simple energy balance model was then used to estimate a value for the solar irradiance of $1363.7\,\mathrm{W\,m^{-2}}$ during the Maunder minimum.

4.8.3 Prediction of sunspots

Various statistical methods can be used for sunspot predictions, and most of these rely on identifying recurring patterns. It is probably hopeless to predict the appearance of individual sunspots, but the sunspot statistics, such as monthly sunspot number may nevertheless be predictable to some degree. As the solar cycle involves some nonlinear chaotic processes, such empirical schemes may be limited. The intriguing regularity of the sunspot statistics may nevertheless indicate some predictive skill. All these methods must use historical observations as a basis for the predictions. The simplest techniques may involve lagged regression of sunspot measures and various tendencies. By smoothing the data record, one may focus on the long-term aspects.

Waldmeier suggested in 1955 that the course of the sunspot can be described in terms of a single parameter: $R_{\max}$, the maximum sunspot number. The rise-time, T_R of cycles with high or medium $R_{\max}$ is shorter than the fall-time T_F and this asymmetry becomes more pronounced with higher $R_{\max}$. For cycles with low maximum sunspot numbers, the rise-time is approximately equal to the fall-time. Waldmeier found empirical relationships between the rise-time and maximum sunspot number. For even-numbered cycles: $(0.17 \pm 0.02)T_R = (2.69 \pm 0.09) - \log R_{\max}$; for odd-numbered cycles: $(0.10 \pm 0.02)T_R = (2.48 \pm 0.10) - \log R_{\max}$; and for both even and odd cycles: $T_F = (3.0 \pm 0.6) - (0.030 \pm 0.006) \log R_{\max}$. Furthermore, he developed an empirical formula for predicting the sunspot number 5 years after maximum:

$$R_5 = (0.29 \pm 0.06)R_{\max} - (11.4 \pm 6.7) \tag{4.21}$$

When the analysis of Waldmeier is repeated today with an updated sunspot record (Figure 4.16), a different relationship is obtained:

$$R_5 = (0.13 \pm 0.8)R_{\max} + (3.74 \pm 12.78) \tag{4.22}$$

The scatter of the points in Figure 4.16 suggests that there is a poor relationship between the peak sunspot number and the sunspot number 5 years afterwards. This finding is confirmed by the ANOVA score adjusted R^2, which indicates that this lagged regression model can only account for 7% of the sunspot number after 5 years. The p-value (probability) of 12.45% suggests that if the same prediction is repeated for 100 test series with completely random numbers, i.e. where there is no relationship between "$R_{\max}$" and "R_5", then one would expect to see 12 cases obtaining a better prediction score by chance.

The evolution of individual sunspots is characterised by a brief growth phase (days) and a longer decay phase (up to a month's duration) and is intimately related to magnetically bipolar active regions (Zwaan, 1992).

Schatten[15] predicts the long-term (decadal) evolution of the sunspot activity using a solar dynamo model based on physical principles. But, as the physical understanding of the generation of the solar magnetic field is still incomplete, this model is not sufficiently comprehensive and cannot describe all the physical processes in detail (details may become important if the solar dynamo behaviour is chaotic). It is nevertheless possible that this model, as with weather and climate

[15] NASA/GSFC Greenbelt, MD 20771.

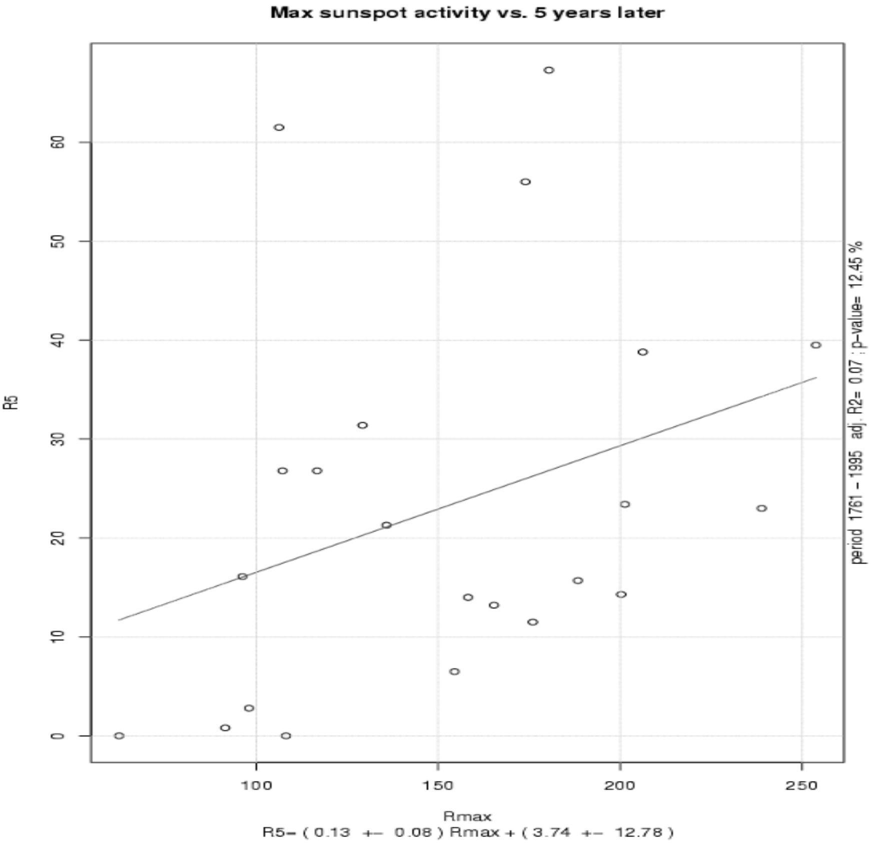


Figure 4.16. The relationship between $R_{\max}$ and R_z five years after sunspot maximum.

models, can give a reasonable description of the main large-scale features important for the evolution of the system.

There has so far been little success at predicting the sunspots with physical models. Small-scale convective cells near the solar surface generate disordered magnetic fields that are believed to emerge from the solar surface and heat the corona. The solar cycle, however, involves an ordered interaction between these turbulent convections and the solar rotation to produce a 22-year magnetic cycle. One of the hurdles of sunspot predictions is the mystery of how turbulent motion in the convective zone can produce differential motion and sharp gradients in the rotation at the base of the convective zone. At best, the present solar models give a realistic reproduction of the primary large-scale conditions when small-scale turbulence is parameterised (i.e. often represented by statistical formulae for the entire grid box volume).

If the solar cycle is truly ‘‘chaotic’’, this implies that it is not predictable at lead-times longer than a couple of cycles (20 years). Thus, the prediction of solar activity suffers from similar problems as weather prediction, where the theoretical limit of predictability is dictated by the characteristic atmospheric timescales.

4.9 FLARES, PROMINENCES, FACULAE, AND CORPUSCULAR CLOUDS

4.9.1 Flares

Sometimes, regions of the Sun become extremely bright for a short time, and these phenomena are called *flares* (see Figure 3.6, on p. 40). The appearance of flares tends to follow a general rule: a rapid increase in the intensity, followed by a brief period (often less than 1 minute) of maximum activity and then a slow decay. There tend to be fewer flares per solar rotation during high sunspot activity (Kuiper, 1953, p. 379). Table 4.7 gives an overview of different flare categories, defined according to their intensity (magnitude of peak burst). The number of flares per solar rotation varies to a lesser extent than the mean number of sunspots ($\bar{R}$). Kuiper gives a simple expression relating the reduced (N_F) number of flares to the mean sunspot number for the solar cycles 17 and 18:

$$N_F = a(\bar{R} - 10)\begin{cases} a = 1.98 \pm 0.02 & \text{rotations} \quad \text{(cycle 17)} \\ a = 1.47 \pm 0.02 & \text{rotations} \quad \text{(cycle 18)} \end{cases}$$

Solar cycle 18 had more sunspots than cycle 17.

4.9.2 Prominences

Sometimes bright strands of relatively cool gas appear in the corona. These phenomena are known as *prominences*, and are best seen at the edge of the Sun against the dark background of empty space (see Figure 3.6, on p. 42).

4.9.3 Faculae

Faculae are bright areas near the sunspots seen in the visible light and chromospheric lines. They appear both on the photosphere and in the chromosphere. Ultraviolet and radio emissions are also connected with the sunspots (Figure 4.11). The faculae appearance precedes the sunspots and they often last for several solar rotations after the sunspots have disappeared.

Table 4.7. Classification of flares (Bray and Loughhead, 1964).

Class	Magnitude of peak burst (W/m^2)
B	$<10^{-6}$
C	$10^{-6} \ldots 10^{-5}$
M	$10^{-5} \ldots 10^{-4}$
X	$>10^{-4}$

4.9.4 Corpuscular clouds

Corpuscular clouds are gases of ionised matter being ejected from the Sun during flares with speeds of the order 1,000,000 m/s and an extension (length) typically of 10^9 m. The clouds get rarefied as they spread out with the distance from the Sun, and their density decreases from typical values in the corona (2×10^{13} protons/m^3) to a typical density of approximately 10^6 protons/m^3 at one astronomical unit.[16] The corpuscular beams are also associated with fast-travelling radio-emission sources. It is thought that the energy associated with these solar eruptions is extracted from the sunspot's magnetic field. Szocs and Kosa-Kiss (2001) suggest that the corpuscular radiation affects the atmosphere over the north pole through an influence on the polar eddies.

Solar "cosmic" rays are often associated with chromospheric flares and consist of high-energy nuclear particles, but not all flares produce cosmic rays (Kuiper, 1953, p. 451). Sometimes, the peak solar cosmic ray counts are measured 1–2 hours after the maximum flare intensities. Through the collision with atmospheric molecules, the cosmic rays produce secondary neutrons with lower energy. The flux of cosmic rays reaching Earth's surface is influenced by the geomagnetic field, but meteorological conditions such as clouds may also possibly affect the flux near sea level.

Sir Norman Lockyer, one of the founders of the magazine *Nature*, investigated the coronal shapes from solar eclipses between 1860 and 1930, and found that the prominences of solar high-latitude zones took place at greatest latitudes (±80°) during sunspot maximum when the corona was most circular.

4.9.5 Solar brightening, sunspots and faculae

According to Lean (2000) an expression for the change in the solar luminance can be expressed as:

$$\Delta S(\lambda, t) = \alpha_s F_q(\lambda) \frac{C_s(\lambda) - 1}{C_s^{\text{bol}} - 1} P_s(t), \tag{4.23}$$

where $C_s(\lambda)$ is the residual contrast between the emission ratio of the faculae brightening and sunspot darkening, the constant $\alpha_s = 0.99$, $C_s^{\text{bol}} = 0.68$ is the bolometric contrast, and $P_s(t)$ is taken as the sunspot proxy.

[16] Mean distance between the Sun and Earth.

5

Earth's climate

5.1 SYNOPSIS

Earth's climate is a complex system which would require several volumes to describe adequately, so we will only cover the most relevant basic details here. Our climate embraces subjects such as solar heating, the atmosphere, oceans, ice, snow, land-processes, ecosystems, and human activities. Here the focus will be on the atmospheric, oceanic, and cryospheric (ice and snow) processes. Although land-processes, ecosystems, volcanism and human activities are outside the scope of this book, this does not mean that these factors are not important.

Earth's weather systems are primarily driven by the energy it receives from the Sun. The latitudinal temperature profile, however, cannot be explained in terms of radiative balance alone: the polar regions are warmer and the equatorial regions colder than a pure radiative balance suggests. The atmospheric and oceanic circulation must play a role through energy transport (heat advection). Earth's climate may be contrasted with those on the Moon and Mars[1] which have virtually no atmosphere and oceans, and hence no heat advection. Much larger spatial variations are seen in temperature on the Moon and on Mars (140 K at the winter pole to 300 K on the day-side during summer).

In order to examine how Earth's climate is affected by variations in the solar processes, it is necessary to have an idea of what "climate" is. It is normal to refer to climate as the average weather over a long period of time. The word "climate" originated from the Greek word for inclination, referring to the Sun's position in the sky.

The weather is highly chaotic and may seem random despite repeating apparently similar patterns. It is generally accepted that weather cannot be predicted with a lead-time longer than about 10 days due to its chaotic behaviour and its sensitivity to small disturbances. Lorenz (1967) was the first to show that there is a limit to the

[1] Mars has a thin atmosphere and a surface pressure of around 1 hPa.

predictability because small changes in the initial state may result in completely different weather types a few days later. Lorenz's theory is popularly known as the "butterfly effect", as the flapping of a butterfly's wing in China may, according to this theory, produce hurricanes over the Atlantic a week later. A well-written introduction to the concept of chaos is given by Gleick (1987).

Although the atmosphere is chaotic and difficult to predict, that does not necessarily imply that the long-term weather statistics (or climate variability) are unpredictable. External forcings, such as solar radiation, may only have a weak effect on the short-term weather development, but may nevertheless be important in the long run. Daily fluctuations are large, and a warm January day may occasionally (rarely in the mid-latitudes) be as warm as a cold day in May in the northern hemisphere. It is nevertheless possible, and easy, to predict that the general January weather in the northern hemisphere tends to be colder than in May. Furthermore, the amplitude of the temperature variations varies geographically, with largest daily differences occurring in high altitudes (spring and autumn) and far inland, and small variations in coastal regions.

There are clearly other factors than the external solar forcing and chaotic behaviour that determine our climate, namely geographical features such as oceans and mountains. It is therefore necessary to know the basic components that make up Earth's climate before looking at how the external forcings, such as variations in solar activity, affect the Earth.

Our knowledge about Earth's climate is mainly inferred from observations as well as from theoretical considerations based on physics and mathematics. Climatology has traditionally embraced different disciplines such as geography and geophysics. The former usually focuses more on observation and description of the climate whereas the latter tends to emphasise theory. The theoretical understanding of the atmosphere is highly advanced, partly because atmospheric models are run several times every day in weather forecasting. However, the weather models may give good weather predictions over a long period, and then the forecasts may suddenly be poor. By studying the reasons why the models suddenly fail, one can gain further insight into the climate system. In many other disciplines, on the other hand, a model may be run a few times, and after some tuning may give a reasonable description of the system it is simulating. However, few models are tested as thoroughly as weather models.

This chapter only aims to convey the most important and relevant climatic features in conjunction with solar–terrestrial relationships. Some excellent introductory books on the atmosphere are Fleagle and Businger (1980) and Houghton (1991), giving a more comprehensive treatment of atmospheric physics. Peixoto and Oort (1992) and Hartmann (1994) discuss not just the atmosphere but also the oceans and ice, and Gill (1982) goes into the detailed dynamics of the atmosphere and oceans. There are also textbooks dedicated to the subject of cloud physics (Rogers and Yau, 1989) and aerosols (Götz *et al.*, 1991). The latest IPCC third assessment report (Houghton *et al.*, 2001) also gives an account of the latest progress in the understanding of our climate.

5.2 THE OBSERVATION OF EARTH'S CLIMATE

5.2.1 Instrumental data

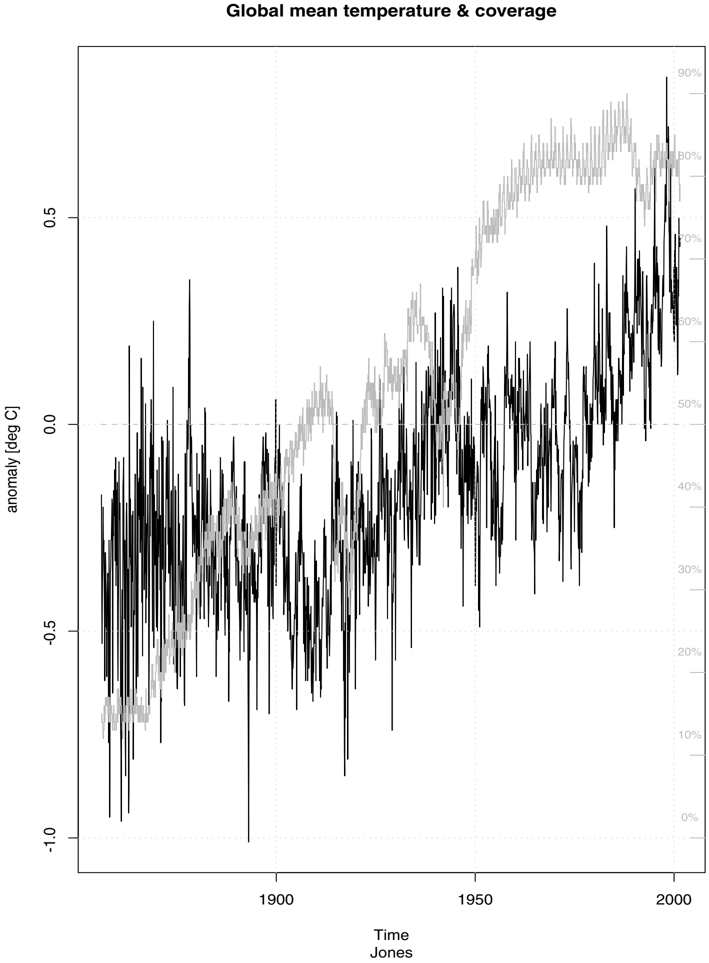
Climatological observations on Earth only date back to the 17th century, when instruments such as the thermometer and barometer were invented. Galileo (1564–1642) invented the thermometer, but there are no temperature records available from his work. His student, Torricelli (1608–1647) invented the barometer, which consisted of a 1.2-m long tube filled with mercury. The Grand Duke of Tuscany, Ferdinand II, founded the *Accademia del Cimento* in Florence in 1657 which made the Florentine thermometer and the condensation hygrometer. In 1654, Ferdinand set up the first meteorological observational network with stations at Florence, Pisa, Parma, Curtigliomo, Vallombrosa, Bologna, Milan, Innsbrück, Osnabruck, Paris and Warsaw. These stations measured the temperature, sea level [air] pressure (SLP), wind direction, humidity and the local weather. The observations ceased when the academy was closed in 1667.

5.2.1.1 Temperature measurements

The temperature readings of air usually refers to the mean temperature of a small volume of air (of the order m^3) for a period of time (minutes). The temperature therefore gives a bulk measure of air molecules' average kinetic energy in a finite volume of air. Daily mean temperature observations are often taken as the mean between two temperature measurements: taken at noon and midnight (maximum and minimum temperatures). The instruments measuring the temperature are called *thermometers*, of which there are various types. The mercury thermometer is usually regarded as the most reliable, and has traditionally been the standard thermometer within meteorology and climatology. The most common atmospheric temperature unit is degrees Celsius (°C) after the Swedish astronomer Anders Celsius (1742), although the kelvin is the proper SI unit (0°C = 273 K and 100°C = 373 K). Meteorological and climatological temperature readings tend to have an accuracy of the order of 0.1°C.

The air temperature must be taken in the shade, as direct sunlight will heat up the thermometer itself and result in a substantial warm bias. The early temperature readings were sometimes made on the north side of buildings (in the northern hemisphere), however, and such readings will be affected by direct light during summer in the high latitudes. The present temperature measurements tend to be made inside small white huts with ventilated thermometers. This arrangement aims to minimise heating by direct solar radiation and errors due to wind chill.

As temperature varies with height, and may be affected by surface processes close to the ground, standard heights (2 m above the ground in Europe and 1.5 m in the USA) have been decided for temperature readings. Various terrain has different properties, and the local environment may affect the readings. For instance, bushes and buildings may shelter the hut against the wind and create turbulence, which may mix the air close to the surface with that at the



(a)

Figure 5.1. (a) Estimates of global mean temperature according to Jones (black) and the spatial coverage of the observations (grey). (b) (opposite) Map showing the geographical distribution of the mean temperature. The data is taken from CRU.

reading level. Readings made on a lawn, on the other hand, may be situated in a more laminar flow, and may not be affected by warmer air just above the ground.

The longest instrumental climate record we have is the *Central England Temperature* (CET) which goes back to 1659. Other long temperature records are available from Berlin in Germany (1700), de Bilt in the Netherlands (1706), Germantown in Pennsylvania (1731), Milan in Italy (1740), and Stockholm in Sweden (1756). In 1781, 14 permanent observing stations had been established. The early climatic records were few and only give a fragmented picture of

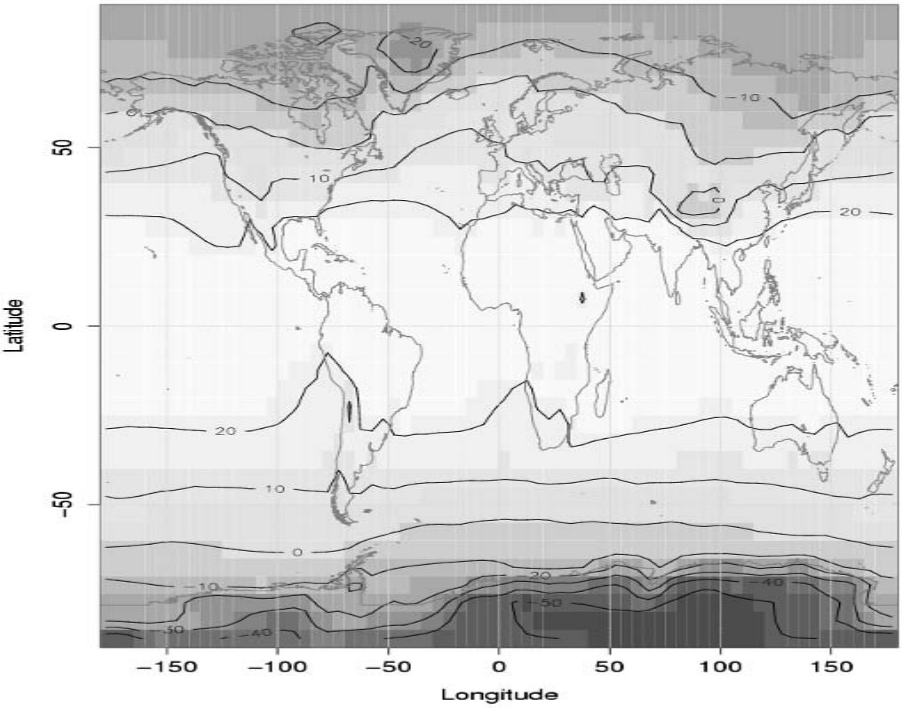


Figure 5.1. (b)

the historical climate. Furthermore, one measurement or a "consensus" made out of only a few observations are more prone to errors than an average of many *independent* contemporary measurements. It is therefore generally acknowledged that data reliability has improved with time as more and more observing stations have come into operation. In order to get an understanding of the geographical climate variations, Jones *et al.* (1998a) and others compiled data sets for the northern hemisphere temperatures, dating back to the 1850s. Figure 5.1(a) shows the estimated global mean temperature ($\langle T \rangle$) and the data coverage (grey), while Figure 5.1(b) shows the geographical distribution of the (temporal) mean temperature ($\bar{T}$). The global mean temperature includes data from more than 7000 stations (Hansen *et al.*, 2001), but the number of stations diminishes when going back in time. Not all of the stations are used, due to questionable quality.

There may be substantial small-scale geographical temperature variations, often dependent on height above the ground. Readings made high up on a south-facing slope in the northern hemisphere may therefore give much higher temperatures than similar measurements at the bottom of the valley or north-facing slopes in the northern hemisphere. Sometimes, on cold and calm winter days, cold air may be trapped in the bottom of the valleys (inversions), whereas the higher surroundings may be substantially warmer (by 10–20°C).

5.2.1.2 Urban heat islands

Temperature readings over a long time may be subject to landscape changes, such as an encroaching urban environment. The urbanisation may lead to warmer temperatures, both because man's energy consumption results in spill heat and because of the changes to the landscape properties themselves (e.g. asphalt). The former is often referred to as urban heat islands,[2] whereby the urban development can explain a gradual local warming (Couzin, 1999). Conversely, increased irrigation in rural regions is thought to produce colder local climates. Such changes to the environment may lead to inhomogeneity in the long-term temperature records. There are, however, ways to correct for some of these effects, usually by comparing with neighbouring stations where no such environmental changes have taken place.

A classic example is the problem of identifying a long-term global warming trend in a data set subject to urban heating. How much of the warming is caused by local heat contamination and how much reflects a global warming? In this case, it is easier to look at the geographical warming trends, as the urban heating effect is expected to be effective only near the major cities (Hansen *et al.*, 1998b). For climate station data, it is a common practice to compare the data with those of its closest neighbour stations. Urban heating effects may be estimated by comparing station data in the city and from the countryside. It is important to keep in mind that observations near fields subject to irrigation may have a cold bias because of the local evaporation. Hansen *et al.* (2001) used satellite measurements of night light intensity in the USA and classified the climate stations according to their proximity to the light pollution. About 210 of the 1200 US stations were in unlit areas and may be assumed to be rural.

Changing the thermometer or the hut type may affect the temperature reading and introduce a systematic shift in the data after (before) the change. Moving a temperature station for instance from a south-facing slope to north-facing slope will also often lead to a jump in the mean values. To search for such inhomogeneities, one may compare the data records with neighbouring stations.

One may furthermore use overlapping periods to estimate correction factors between succeeding instruments. This has been done to the MSU data record, which is made up of measurements from many different satellites. The lifetime of each individual satellite is usually too short for climatological studies.

5.2.1.3 Testing for non-homogeneity

It is always a good idea to test a time-series for non-homogeneity.[3] How does one do this? A simple test for homogeneity can involve looking for changes in the difference between two neighbouring climate stations. Jumps, discontinuities or trends in the time-series of station differences can be a sign of non-homogeneity. To illustrate this,

[2] Changes to the landscape properties as a result of urbanisation also contribute to the urban heat effect.

[3] Changes in observational practices or a relocation or replacement of the instruments may render an observational record non-homogeneous.

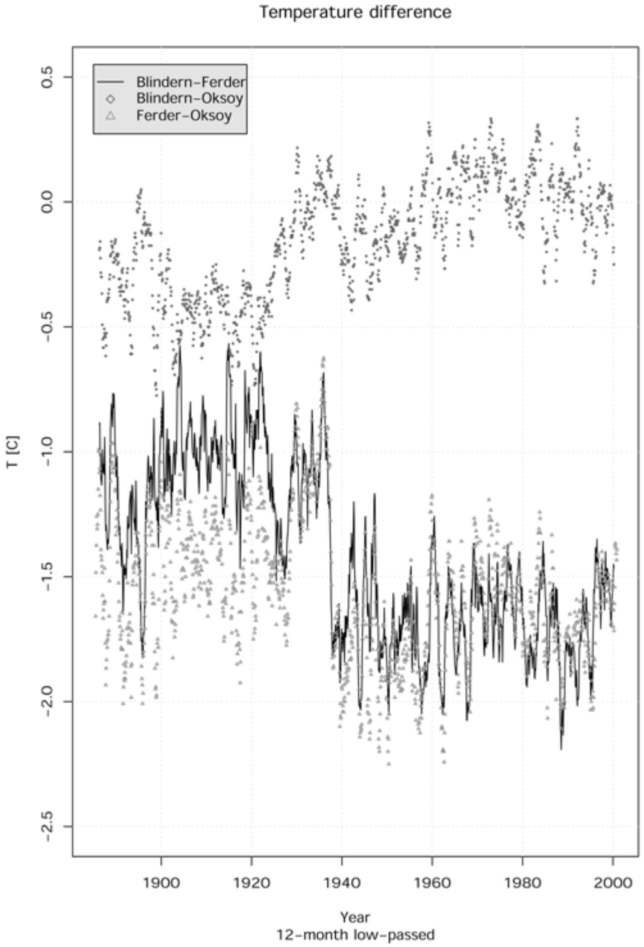


Figure 5.2. The difference between measurements of the annual mean temperature taken from a city environment (Oslo–Blindern) and a rural environment (Ferder lighthouse). Also shown is the difference between two rural stations (Ferder lighthouse–Oksøy lighthouse). The data is from the Norwegian Meteorological Institute.

we can examine neighbouring temperature records from southern Norway: Oslo, Ferder and Oksøy. The so-called *urban heat island* effect is one specific example of the non-homogeneity problem. Urban development and paved areas can result in a local warming because of the spill heat effect from buildings and lower albedo from tarmac. It is possible to identify some of the cases affected by the urban heat island effect by comparing urban and rural stations, but it is important to keep in mind that the local temperature in rural places may have dropped over time due to increasing use of irrigation (wetter soil and more evaporation). In Figure 5.2 the temperature in Oslo and at Oksøy lighthouse are compared. Lighthouse records are ideal for these kinds of studies, as they are not affected by changes in the environment or irrigation.

Figure 5.2 shows the difference between the temperature recorded at Oslo and Ferder lighthouse, Oslo–Oksøy and Ferder–Oksøy. There is a sharp discontinuity in 1937 in the Oslo–Oksøy and Oslo–Ferder difference curves, with an increase in the temperature in Oslo relative to the two lighthouse records. This change may be related to a relocation of the temperature station from St. Hanshaugen to Blindern, and hence the Oslo temperature series is not homogeneous. The temperature at Ferder also gets systematically higher than at Oksøy between 1920 and 1940. There is, on the other hand, little sign of long-term systematic trends in the difference between Oslo and the lighthouse series after 1940. This example illustrates one method for diagnosing urban heat island effects, but there is no evidence pointing to an urban heat island effect in Oslo.

5.2.1.4 Metadata

Observations must be regarded as an incomplete data set without *metadata*, which is any relevant information about the data, instrumentation and observational practices. The metadata may for instance comprise time and place of observation, units, uncertainty estimates, diaries and log books documenting relocation, instrumentation replacements, change of observer, encroachment of buildings and vegetation. One example of how the metadata can be useful is given by Hansen *et al.* (2001) who report changes in the time of observations by cooperative observers in the USA.

It is possible to correct for inhomogeneity if there is information about the observing practice (metadata), or there is a redundancy in the data. However, it is unlikely that such corrections truly remove all of the inhomogeneity. Care must be taken when applying such corrective measures, especially if there is a lack of sufficient metadata as such steps are based on the expected behaviour of the subject.

Some analytical methods require that the data are normally distributed, and there may be need for a test of Gaussian distribution. For quantities expected to be normally distributed, such a test may reveal data discrepancies if it fails to produce the expected results. However, some quantities, such as daily rainfall, are not normally distributed, and the analysis of such data records may require a transformation that produces a new normally distributed data series. Gamma or Weibull distribution is often used in conjunction with daily rainfall.

5.2.1.5 Global mean temperature

A dense network of independent measurements will give a more reliable mean temperature estimate for the whole region than just a few observations. Therefore, regions with sparse observations, such as polar regions, the interior of Africa, and the south Pacific are associated with a greater degree of uncertainty than Europe and the United States. Although such holes may be filled in by various interpolation techniques, they do represent a problem for estimating global mean temperatures.

The record of global mean temperatures is not homogeneous: it is based on a number of stations that varies with time (see grey curve in Figure 5.1(a)). The global mean temperature cannot be measured directly, but must be computed from a large number of measurements made from a global network of climate stations. There are different ways of calculating the global mean value and there are more than one global temperature record. The differences in the various records reflect uncertainties which may be due to analytical method, the selection of station observations included, and strategies to deal with non-stationary records.

Other gridded data sets include the maritime Comprehensive Ocean Atmosphere Data Set (COADS, Slutz *et al.* (1985)) which dates back to 1854. The observations included in COADS are from "ships-of-opportunity", and the observations tend to crowd along the major shipping routes. Other regions where few vessels venture tend to be blank. The observations at the beginning of the record show only a handful along the India route. The COADS data include sea surface temperature (SSTs), surface air temperature, and sea level pressure (SLP).

The incomplete and time-varying geographical data coverage and non-homogeneities degrade the data quality as a global mean and may affect the outcome of analytical tests. One crucial question is whether the observations are sufficiently good for empirical studies of solar–terrestrial links over long timescales. Efforts have been put into so-called re-analysis of gridded data (Gibson *et al.*, 1997; Kalnay *et al.*, 1996), where observations are assimilated into atmospheric models in order to obtain physically consistent data reconstructions, but these data sets tend to be of limited length (they presently do not cover the period before 1940).

5.2.1.6 Sea surface temperature

The measurements of sea surface temperature (SSTs) are done by ships, by drifting and anchored buoys, and by satellites. Whereas the *in situ* measurements make readings of the temperatures at some depth (e.g. 2 m), the satellites only see the skin temperature (the temperatures of the oceans' surface film). The early SST measurements were made by throwing in a bucket and reading the temperature of the water fetched up. However, the type of bucket has changed with time, and corrections must be made for this in order to get homogeneous SST records. More recently, the SSTs are measured automatically at the ship's engine-cooling water intake. This change in observational practice must be taken into account, as must the different ship-dependent depths from which the cooling water is taken. Modern ships tend to have a metallic hull which has a different heat capacity from wood, leading to a higher hull temperature under direct exposure to sunlight. The ship measurements are done by "ships-of-opportunity", and therefore made along the major shipping routes.

Drifting buoys do not have the same homogeneity problems as ships do; however, they tend to live their own life as location is concerned. Moored buoys, such as the *tropical atmosphere ocean* (TAO) array provide temperature readings fixed in space. Measurements may be made from the surface to depths of 500 m or deeper. These temperature records are more homogeneous than most other *in situ* readings, although they only date back as far as the early 1980s. The buoys nevertheless sometimes suffer problems caused by fish nibbling at the sensors, theft or sabotage, and lack of maintenance.

The data coverage changed abruptly after the opening of the Panama and Suez canals.

5.2.1.7 Sea level pressure

The 17th-century scientist Torricelli discovered that the air has a mass, and Otto v. Guericke demonstrated that the air pressure could keep two hemispheres together by strong forces when the space between them is evacuated. As a consequence of the atmospheric mass and Earth's gravity, the atmospheric pressure decreases with height. The legend tells of Pascal dispatching his brother with a barometer to the top of Puy de Dome in France to demonstrate this height dependency of the air pressure. The pressure at sea level (sea level pressure, SLP) is associated with the total air column mass above this level. It was also discovered that the SLP varies with time.

The instrument measuring pressure is called a *barometer*. There are various types of barometers, such as mercury barometers and aneroid barometers. The mercury type is regarded as the most reliable, although less practical to use. The barometers are usually calibrated by measuring the pressure and temperature of water's boiling point.[4] The barometers must be calibrated against height, as the pressure measured in the mountains is usually lower than at sea level. Correction must also be made for the latitude, as the rotation of the Earth produces a centrifugal force that partly counteracts the gravity near the equator (small effect). The SI pressure unit is the pascal (Pa), and the atmospheric pressure is commonly reported in hectopascals (hPa = 100 Pa) or millibars (mbar = hPa). The old units are millimetres of mercury (1000 hPa = 75 mm Hg, and 1 hPa = 0.075 mm Hg).

Blaise Pascal (1623–1662) established that the changes in the sea level pressure are related to changing weather. Sea level pressure observations are important since they give an indication of how the air is moving around. The SLP difference between Tahiti in the central tropical Pacific and Darwin in Australia is used as an index for the strength of the trade winds over the tropical Pacific. The oscillations of this pressure index are associated with the Southern Oscillation (see Figure 5.3) and therefore also El Niño events (El Niño Southern Oscillation, ENSO). At least one instrumental record of the southern oscillation index starts in 1876 (Figure 5.3(a)). Similarly, the pressure difference between Lisbon in Portugal and Stokkisholmyr in Iceland is taken as an index of the North Atlantic Oscillation (NAO) strength (Figure 5.3(b)). Instrumental records of the NAO index start in 1821, but with reconstructions based on proxy data one may go further back in time.

5.2.1.8 Precipitation measurement

Precipitation is usually measured with rain gauges, which are basically cylinders that collect rain or snow. The amount of rainfall is read off as volume of liquid

[4] This requires an accurate thermometer.

water, which implies the melting of snow. Local wind and turbulence may affect the catchment efficiency of the rain gauges, and to remedy some of these effects the instruments are often fitted with a screen. However, the local environment, such as trees, bushes, buildings, and hills may influence the readings. More recently, radars have also been employed in the measurement of the precipitation, although the radar measurements are subject to important limitations, as will be discussed in Section 5.2.3. Inhomogeneities in the precipitation records may be a result of station relocations, changes to the local environment of the instruments, or the replacement of observers accustomed to different practices.

Robert Hooke (1635–1703) gets the honour for inventing the rain gauge. Precipitation measurement goes back almost as far as temperature measurement. The observation of precipitation suffers from the problem of not giving a complete picture of the large-scale conditions. Rain is often a local phenomenon, and a high density of stations is needed to capture the geographical variations. Unfortunately, there are large sparse regions even at the present. One of the greatest challenges is to measure the rainfall over the vast oceans.

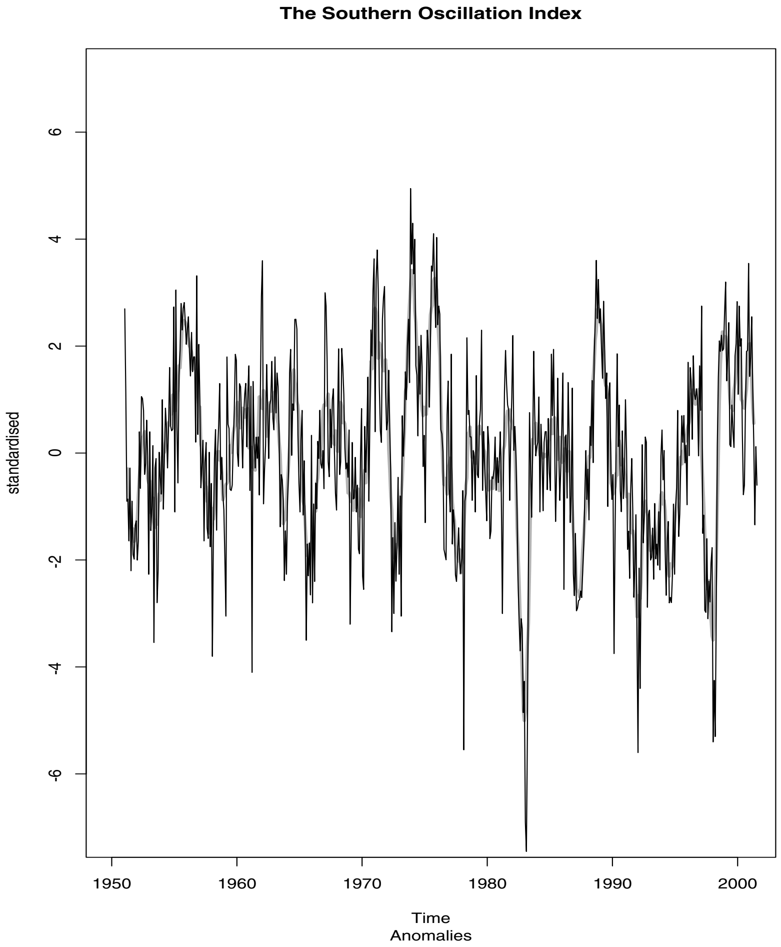
5.2.2 Upper air data

Upper air observations are only recent. Routine radiosonde profiles started after World War I (1930s), and satellite observations of the atmosphere started in the 1960s. The first (polar-orbiting) weather satellite, the Television Infrared Observation Satellite (TIROS), was launched on April 1st 1960. The polar orbiting satellites (e.g. Nimbus or Tiros) orbit the Earth at an altitude of around 1000 km and complete about 14 orbits a day. Geostationary satellites, on the other hand, are at an altitude of approximately 35,000 km and remain directly above a fixed point on the equator.

Radiosondes make measurements of the vertical atmospheric profile of pressure, temperature and humidity. These devices consist of a light balloon-borne instrument called a meteograph that records the vertical profiles and transmits the data by a small radio. The first radiosondes were deployed between the two world wars. However, an extensive network of radiosondes was not in place until the International Geophysical Year in 1958.

5.2.3 Earth-observing satellites and space-borne UV measurements

UV measurements were carried out from balloons and rockets in the 1970s, but these data did not have a high quality. Comparisons of various measurements revealed up to 20% differences above 200-nm wavelength. Later measurements from space platforms such as Space Lab I & II, NOAA (launched 1985 and 1988), UARS, the ATLAS missions, ERS2 (launched 1995) and SOHO yielded data with much higher quality (4% RMS). The measurements of the vertical stratospheric profiles has been measured through the scattered light as a function of wavelength (NIMBUS7 SBUV spectrometer: 1978–1994).



(a)

Figure 5.3. (a) The Southern Oscillation Index (SOI) defined as the standardised difference SLP(Darwin)–SLP(Haiti). (b) (opposite) The North Atlantic Index which is the standardised difference SLP(Lisbon)–SLP(Iceland). The seasonal cycle has been removed from both these series.

Station observations of weather and climate can give a fairly reliable description of the local conditions, but due to the sparseness of stations in certain regions, ground-based observations are far from ideal for climate studies with a global data coverage. For a global view, satellites give a good overall picture.

5.2.3.1 Deriving climate data from visible observations

Satellites have recently been employed to observe Earth's weather and climate. Meteorological satellites are routinely used in weather forecasts and analysis, and

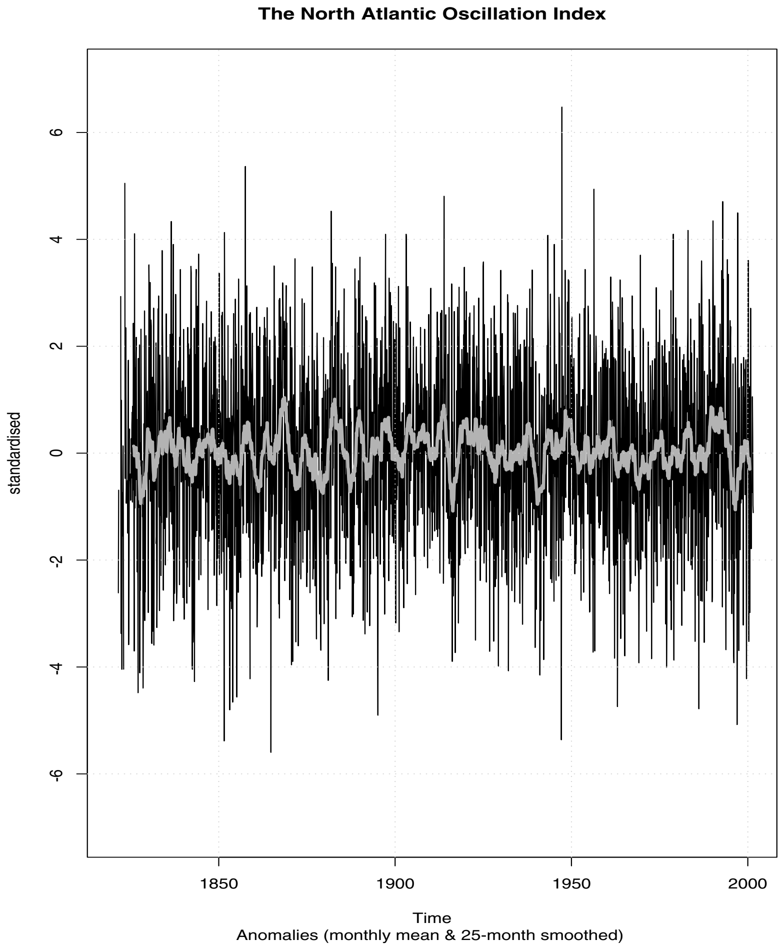


Figure 5.3. (b)

satellite images are becoming increasingly common in the presentation of the weather on TV. The satellites may carry various instruments that measure physical quantities in addition to taking visible pictures of Earth's weather. Weather satellites measure the infrared (IR) radiation, which gives an indication of the temperature of the cloud tops which appear as cold regions, whereas the clear regions where the satellites measure the ground temperature are characterised by warm temperatures. In a similar way, the satellites may measure the sea surface temperatures (Reynolds and Smith, 1994, 1995). The surface temperatures of the ground are more difficult to estimate due to brightness changes of different vegetation and landscape types.

Visible images of the Earth may be used to estimate the winds through cloud tracking. Scatterometers measure the surface winds from the backscatter of radar signals from waves on the ocean surface. Satellite images in the visible light range have also been used in the study of the ice extent of ice-sheets and glaciers. Satellite ice data records date back to 1976. The electronic scanning microwave radiometer was the first satellite instrument designed for recording ice data. More modern satellite-borne ice observation systems, such as the scanning multi-channel microwave radiometer (SMMR), employ a technique called multi-frequency brightness temperature (T_B) measurements in the microwave range, and were deployed in 1978. In 1987, the SSMR was replaced by an instrument called the special sensor microwave imager (SSM/I).

Satellite observations require different data retrieval algorithms for night-time and day-time observations, due to the reflection of the sunlight during the day (Figure 5.4, colour plate). In both cases, the presence of clouds must be examined, and the satellite observations are usually calibrated against *in situ* observations on the ground. Although the satellites in principle should give high precision measurements, there are various error sources that reduce the measured accuracy. Disruptions of the satellite's orbit can produce systematic biases, for instance as the timing of the observations will be shifted with respect to the diurnal cycle. It has recently been established that the atmospheric drag on the satellite is affected by the solar cycle, and that the satellites often fall during sunspot maximum. In order to make inferences about atmospheric temperature from satellite measurements, the optical depth and a set of weighting functions must be calculated according to the mass ratio of absorbing constituents in the atmosphere. Since aerosols may affect the optical depth and may alter the weighting functions, satellite-based temperature measurements may be affected by the presence of atmospheric dust particles, haze, and pollution.

5.2.3.2 Microwave sounding unit

The TIROS-N satellites routinely measure the air's moisture and temperature. These satellites are in sun-synchronous polar orbit (1000-km altitude) and cross the equator at the same local time each day. The satellite crosses the equator at 7 am and 2:30 pm when northbound and 7 pm and 2:30 am southbound. An instrument known as the *microwave sounding unit* (MSU) measures the air temperature as well as humidity between the surface and about 8 km above the surface from information received on a 4-channel Dicke type passive radiometer (50.30, 53.74, 54.96 and 57.95 GHz). J. Christy and R. Spencer have been the pioneers in developing the MSU data set. For the tropospheric measurement, about 10% of the signal over the oceans and 20% over land comes from Earth's surface. It is important to account for this surface contribution when using the data in climate studies, and the surface temperature at radiosonde release is needed for validation against the radiosonde observations.

The satellite may be affected by changes in the performance (e.g. linearity) of the transducers and electronic circuit, such as amplifiers, due to variations in temperature, magnetic fields, and cosmic ray bombardment and instruments are calibrated continuously by taking a measurement of cold space and a black body target. The satellite lifetime is typically short in terms of climatic scales, frequent replacements are required and there is often substantial large scatter between different satellite readings (Fröhlich and Lean, 1998a,b). There has been a drift in various satellite heights due to increased drag during sunspot maximum, and the associated loss of altitude has resulted in a slight shift in the timing of the measurements and hence produced a spurious cold bias in the estimated long-term temperature trend. The recent versions of the MSU data include a correction for this error.

Over a 16-year interval, the correlation between radiosonde measurements and the MSU is 0.97. There has been a discrepancy between the global mean near-surface temperature trends derived from the station networks ($\approx$ 0.2°C/decade between 1979 and 2001) and the MSU data ($\approx$ 0.03°C/decade between 1979 and 2001). The discrepancy may partly be due to trend differences in the free atmosphere and at the surface. Part of this discrepancy may also be related to the fact that the station observations use the mean of the local maximum and minimum values whereas the MSU records temperatures at different times of the day depending on the latitude. The fact that the night-time (minimum) warming has been about twice as strong as the day-time warming (Houghton *et al.*, 2001) and the minimum may carry more weight in the station estimates of the daily mean, may also explain part of the different trends. It is also important that long-term variations in the atmosphere's density profile, chemical composition and aerosol concentrations are taken into account when using the MSU data to study climatic trends. Furthermore, the calculation of the temperatures assume local thermodynamic equilibrium (LTE) and require vertical weighting functions. More recent work shows that the trends converge, with MSU trends adjusted to higher values and in better agreement with GCMs and surface observations.

Independent observations from radiosondes suggest that there has been greater warming at the surface than in the free atmosphere (Gaffen *et al.*, 2000). The effect of anthropogenic factors, stratospheric ozone depletion and volcanic aerosols may explain part of the differences in the lower tropospheric and surface trends (Santer *et al.*, 2000).

The satellite temperature measurements are based on estimates of the atmospheric transparency and the black body radiation (Houghton, 1991, p. 195). The thermal radiation (radiance) associated with a particular frequency (ν) seen by a satellite is the sum of the thermal emission ($B_\nu(T)$) at the different heights (z), but is also affected by the atmospheric transparency (optical depth, τ_ν).

$$I_\nu = \int_0^\infty B_\nu(T)\frac{d\tau_\nu(z,\infty)}{dz}dz + B_\nu(T_s)\tau_\nu(0,\infty) \tag{5.1}$$

The transparency is related to the atmospheric density, aerosol concentrations, and the chemical constituency.

The MSU has a superior spatial coverage, and more than 30,000 readings are made each day, sampling more than 75,000 km^3 of air. The estimation of

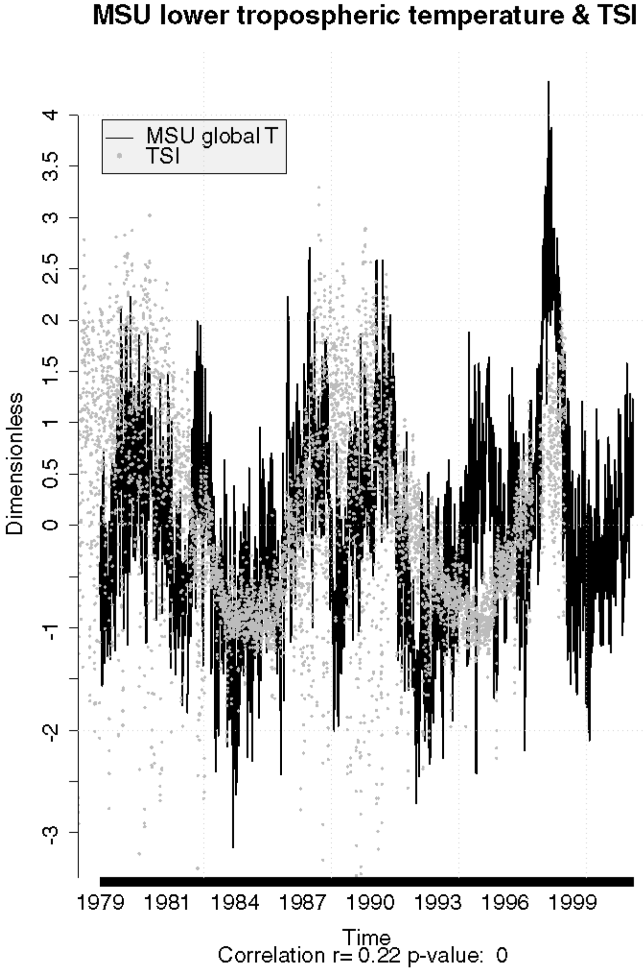


Figure 5.5. A comparison between the daily mean TSI and MSU lower tropospheric temperature measurement. There is a weak but highly significant correlation between the two curves. Data from http://daac.gsfc.nasa.gov/CAMPAIGN_DOCS/atmospheric_dynamics/ad_data/msu.html and ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

the temperatures does not use distinct algorithms for night and day measurements, but a correction is applied to compensate for diurnal drift. The radiation in the microwave band is well-separated from the solar radiation (Houghton, 1991, p. 192).

A comparison between the MSU-derived lower tropospheric temperature and the TSI measurement suggest a statistically significant positive correlation (Figure 5.5). Hence, the MSU data indicate that the variations in the TSI do affect the atmosphere. Temperature records also contain variations due to ENSO variability and volcanic eruptions, which are not seen in the solar activity proxies, and these other factors reduce the signal-to-noise ratio and affect the analysis. There

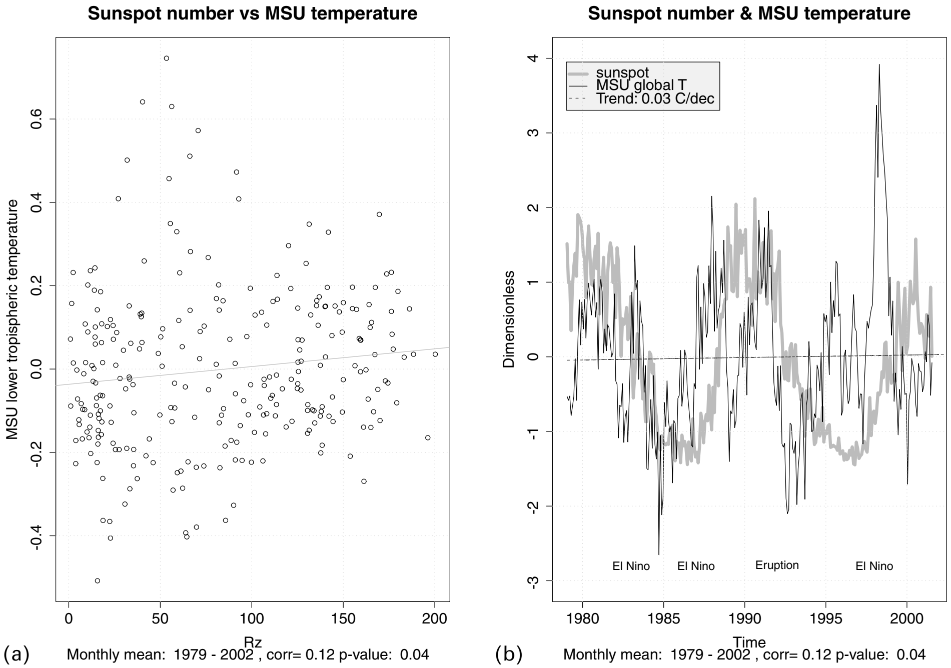


Figure 5.6. A comparison between the monthly mean sunspot number and MSU lower tropospheric temperature measurement. There is a weak correlation between the sunspot number and the MSU temperature measurement, statistically significant at the 5% level. The linear trend in the MSU temperature is estimated to be +0.03°C/decade ± 0.004°C/decade and is associated with a *t*-value (a standard statistics parameter used for the "strength" of the fit) of 9.291 and a *p*-value $<2 \times 10^{-16}$.

is, however, a weaker correlation between the (monthly mean) sunspot number and the MSU data (Figure 5.6), even though this correlation qualifies as 5% statistically significant. Figure 5.6(b) indicates that the influence of solar activity on the lower tropospheric temperature is weak. A high correlation between the TSI and the sunspot cycle and a weak non-zero correlation between the sunspots and the MSU record may indicate that the TSI variations influence the temperature rather than the solar activity.

5.2.3.3 *Satellite altimetry*

Satellite altimetry measures surface heights. The ERS-1 and Topex-Poseidon satellites observe variations in the sea level height to an accuracy of around 5 cm. The Topex-Poseidon programme started in 1992, and gridded altimeter data may be constructed with a temporal resolution of 15 days and a high spatial resolution. The sea level height anomalies (SLA) are important quantities since they reflect the geostrophically[5] driven currents. Furthermore, the sea level height may also give an indication of the heat content of the upper ocean layers.

5.2.4 Observation of planetary atmospheres

Space probes have been used for solar and planetary exploration since the 1970s. Furthermore, lunar rocks which have been exposed to a continual bombardment of cosmic rays may provide clues about the history of the solar activity.

One use of space platforms has been in occultation index studies of remote stars by planets to infer the refractive indices of the other planets' atmosphere. *In situ* measurements were made on Mars with the Viking probes in 1976–1977. The Venera and Pioneer Venus missions in the 1970s and 1980s brought *in situ* sensors to Venus. The Galileo fly-by and the Magellian radar instruments during the 1990s explored Venus under its thick cloud cover.

5.2.5 Palaeo data – "proxy data"

5.2.5.1 Biological proxies

It is sometimes possible to learn about past climate by examining elements that one way or another have been exposed to the climate and bear an imprint of the past climatic conditions. Such data records, which themselves are not direct measurements of the climate elements, but are witnesses of the past, are called *proxy data*. One common type of proxy data is tree rings from old trees and fossilised wood. Tree rings provide two types of information: by counting the rings that show faster growth during summer and slower growth during winter, it is possible to date events in chronological order. The width of the rings gives an indication of how warm and wet the summer has been. The tree ring chronology can be used for calibrating carbon dating methods and vice versa. The proxy data do not give a direct measure of the conditions, but one must use empirical models to relate the tree ring characteristics to the climate elements. It is possible to develop statistical relations (empirical models) between the tree ring width and, for instance, summer mean temperature through regression analysis for a period when both the temperature is known and there is tree ring information. Various factors, such as solar energy, moisture, pests, competition among different species, as well as temperature may affect the growth rate, so the tree ring data usually do not give a perfect record of the past climate. This is the general rule for proxy data, and because these often are based on a number of assumptions, such as stationarity,[6] these records are less reliable than instrumental records.

[5] A geostrophic flow is a situation where the pressure force is balanced by the Coriolis term (an effect of Earth's rotation). The net effect is that frictionless flow is along the isobars (contours of constant pressure) with lower pressure to the left in the northern hemisphere. The flow around a low-pressure region in the northern hemisphere is counterclockwise (cyclonic flow).

[6] The statistical properties do not change over time, see Section 5.2.6.

Various plants thrive under different conditions, and the population of particular species can hold some information about the local climate conditions. For instance, fossils of fish in the Sahara suggest that the desert once was wet. Organisms, such as plankton, have been found in fossils. Stomatal frequency has been found to decrease linearly with increasing CO_2 concentrations, and the stomatal frequency has been estimated by the stomatal (cell) density in tree and plant leaves. Many of these land-based biological proxies are only sensitive to the warm growing season, and therefore can only provide a clue about the summer temperature or precipitation (Shindell *et al.* 2004). Coral remains can also be used to reconstruct past climate variations. The species, the leaf structure, and the population give indications of Earth's past climate. The population of a species rarely depends on just one factor. Plankton population, for instance, may depend on the temperature, salinity, availability of nutrients, solar energy (UV), and the population of herbivores (fish).

5.2.5.2 Geological proxies

Geological records may also be used as proxy data records for climate conditions. Glaciers and ice-sheets form under cold conditions and retreat during warm periods, and moraines from glaciers, ice-rafted debris, polished or weathered rocks, and past sea levels may indicate past climate variations. The dating of such geological features tends to be highly uncertain, but approximate dates can be obtained from isotopic ratios, counting geological layers, and chemical composition. Lake sediments have also been used for past climate reconstructions. Records of oxygen isotopes and gypsum ($CaSO_4$) concentrations have to be used to infer relative changes in the ratio of evaporation to precipitation. The lakes are assumed to be near the saturation point for gypsum and the mineral only precipitates in shallow water during normal climatic conditions. The water volume, and hence the water level, is reduced during dry periods and the saturation point is exceeded, favouring gypsum precipitation throughout the lake. Isotopic proxies can also be used to infer past climate variations (see Section 2.4.1).

Some geological studies from southeast Africa have suggested a link between the African climate and the solar cycle (Verschuren, 2000). Sediments on the bottom of lake Naivasha in the Rift Valley (Kenya) can be used to reconstruct climate a thousand years back in time. High concentrations of organic carbon assumed to be remnants of algae are assumed to be indications of high water levels, and the findings suggest that there was a dry period between 1000 and 1270 AD, in addition to shorter intense droughts in the periods 1380–1420, 1560–1620, and 1760–1840 AD.

Bond *et al.* (2001) used ice drift tracers to study past climate. These tracers comprised of ice rafted debris such as detrital carbonate, fresh volcanic glass from Iceland and hematite-stained grains (lithic in e.g., 63–150 μm size range), and radiocarbon dated with a mass spectrometer. Excursions southward were interpreted as an indication of cold periods.

5.2.5.3 Archaeological data

It is sometimes possible to use archaeological records as proxy data for more recent climate records. Diaries on the time of harvest and the yield can be used as a measure, albeit highly subjective and uncertain, of what the summers were like. Viking settlements on Greenland during the 11th to 15th centuries suggest that there were warmer conditions in the northwestern Atlantic during that period. Fortunately, people in the past relied on the winds when sailing the seas, and the mean wind and ice conditions must have been favourable for the Vikings to traverse the Arctic and North Atlantic seas. Likewise, whalers logged sea-ice conditions, and their records can be used to partially reconstruct the historical ice extents.

In Latin America and the tropical Pacific, El Niño Southern Oscillation (ENSO) has strong effects on the environment. The Peruvians have long known about the phenomenon, and one of the first accounts that we have of these climatic fluctuations are from these fishermen. They named the warm current off the coast of Peru "El Niño" ("The Child" after Christ), as this appeared around Christmas time with an irregular periodicity in terms of number of years. El Niño had a devastating effect on the fishing, as well as the rest of the ecosystem in the region. There are also references to El Niño from the conquistadors, who had to cross a desert that normally is dry, but is in full bloom during El Niño years. Some historians believe that the conquistadors would not have managed to cross the desert under normal conditions. There have also been suggestions that Indians may have felt the wrath of El Niño.

Sometimes, the slightest information about past weather has been used to infer past climate conditions, such as studying climatic features in landscape paintings. For example, the scenes of oil paintings by Pieter Brueghel the Elder changed from sunny summer weather to wintry scenes around 1560. Joseph Turner painted a large number of pictures featuring the light of the sky, clouds, and sea. Glorious red skies may reflect the atmospheric effects of volcanic eruptions in the Azores (1811) and Tambora (1815). Of course, the use of fine art and landscape painting for climate reconstruction is highly uncertain and may be tinted by the artist's eye.

5.2.5.4 Synthesised data – model simulations

In the last decade, computers have become a major tool of research. The increased computational capacity and the advancement of computer science have enabled increasingly more realistic climatic simulations in a virtual realm. Computer modelling has been used in a wide range of applications, from simple empirical modelling that may relate temperature and isotopic records to complicated general circulation models (GCMs) describing the full 3D wind-fields and energy transfer in the atmosphere. Computers are powerful tools for both prediction (e.g. weather and climate models) and testing various hypotheses. An ordinary personal computer may also be used as a "test universe" where various analytical methods may be tested against constructed data, such as stochastic series (random numbers) and artificial data where *a priori* relationships or conditions have been imposed.

General circulation models can be used for the reconstruction of past climates;

however, there is as yet no guarantee of a true description of the past climates as the climate models themselves are limited in terms of model shortcomings. They always need to be able to demonstrate the ability to make good and reliable predictions. Furthermore, they are no replacement for the real world observations even when they can reproduce the observed climatic features; there is always the need for good observations and for continuous model evaluation. In climate science, it is important to reconcile the model world with the actual historical observations and proxy data records. Since Earth's climate is driven by solar energy, it is also necessary to get confidence in solar models,[7] which must be able to reconstruct the solar cycle.

Sometimes, the geographical dependence of quantities, such as temperature, precipitation, or sea level pressure, may be reconstructed by employing empirical models based on regression methods. Such reconstructions relate a small number of key observations (predictands) to large-scale patterns from the recent past when the observational network is the most extensive. These statistical relationships are subsequently used to predict the evolution of past large-scale features, given reliable long data records of the key observations. There are various methods for constructing geographical distributions of climate variables, and these are based on the fact that the spatial structures are coherent over some distance (Kaplan *et al.*, 1998). The simplest way to reconstruct such maps is by regression or projection of so-called modal structures (EOFs or eigenvectors) that represent a common regime. The projection method does not always produce reliable data in regions where the observations are sparse. More advanced methods, such as *optimal interpolation* and *kriging*, improve the estimates of the data in the voids of missing data, and further refined methods, such as *Kalman filtering* and *optimal smoothing*, also take into account the relation between the successive data maps (autocorrelation). Haigh (2003) summarised work done by Stott *et al.* (2002) and proposed that GCMs seem to underestimate the response to solar forcing by a factor of 3 and that important amplification mechanisms are not accounted for in the GCMs.

5.2.6 Climate observations

Climate studies require long data records, which at the present unfortunately rules out the modern satellite data. The longest ground-based data records may be used for some types of study; however, several hundreds of years, or even thousands of years, are needed for proper examination of causal relationships. Only proxy data have such lengths, but these have a very low temporal and spatial resolution and are often not appropriate for particular studies of climatic processes.

One problem associated with the longest instrumental data records is the lack of stationarity and homogeneity, such as systematic changes in amplitude, mean value (non-zero trend), or relationship with other stationary entities. Analytical methods may be sensitive to such non-stationary properties, and thus non-stationarities may introduce misleading biases into the analysis.

[7] Which are even more uncertain than the climate models.

Another major concern is that changing observational practices, changes to the instruments themselves or their environment, can affect the observations and hence the analyses. Changes to observational practices, movement of observation sites or replacement of old instruments with new imported ones may render the data record inhomogeneous. In such cases, there may be a systematic change in the data with time which does not reflect the real climate. In other circumstances, where an instrument has been employed for a long time, there may be an "instrumental drift" where the readings are systematically biased over time, for instance due to a grounding problem.

5.3 BASIC CLIMATE PHYSICS

In order to understand our climate, it is important to have an understanding of how the climate processes work. There is a small number of entities in the climate systems which are known as *conserved quantities*, and these are *mass*, *energy*, *charge*, *momentum*, and *angular momentum*.

5.3.1 Mass conservation

Under normal conditions on Earth, mass is conserved because it is neither created nor destroyed in the physical processes taking place in Earth's climate.[8] This means that when it rains, the water does not disappear as soon as it hits the ground, but is absorbed by the ground, fills the natural water reservoirs, and dribbles out to the rivers. The water eventually reaches the oceans from where it may evaporate again, producing more rain, or the water may evaporate directly from the ground. Thus, the water is moved around in cycles, or stands still, but there is the same amount of water on Earth all the time as long as no water enters from or escapes out to space. Similarly, the air is not created out of nothing and is not annihilated. When the winds move the air masses around, fresh air must fill up the void that otherwise would have been left behind. Thus, the movement of the air implies a circulation pattern of a cyclic nature. The conservation of mass leads on to the *equation of continuity*.

The equation of continuity is expressed mathematically as:

$$\frac{\partial \rho}{\partial t} + \frac{dx}{dt}\frac{\partial \rho}{\partial x} + \frac{dy}{dt}\frac{\partial \rho}{\partial y} + \frac{dz}{dt}\frac{\partial \rho}{\partial z} = 0 \tag{5.2}$$

In equation (5.2) ρ represents density, t time, and x, y, and z the three spatial dimensions. A short-hand for this equation is: $\rho_t + \vec{u} \cdot \nabla \rho = 0$, where $\vec{u}$ represents the fluid velocity.

[8] Mass is only "destroyed" when being converted to energy by nuclear reactions, for example, as when unstable isotopes, such as ^{14}C, are created; but these mass losses are so small that they can usually be ignored.

5.3.2 Energy conservation

The energy on Earth usually determines the temperature, movements in the air and sea, the evaporation of water, the brightness, and the chemical reactions. The energy is conserved in a similar way as the mass, but with the difference that there is a "steady"[9] supply of energy received from the Sun which is balanced by the energy radiated by Earth's surface. As long as there is a balance between the incoming and outgoing energy transport, the Earth will not be subject to net energy change. Moreover, in an energy equilibrium, the temperatures will not increase substantially, and the mean circulation will be constant if the temperature contrast between the equator and the poles does not change. The meridional temperature gradient (geographical difference) drives the climatological "heat engine" responsible for much of the heat transfer to the high latitudes.

The fact that energy is manifested in both the temperature and circulation, suggests that it is present on the Earth in various forms. There is thermal energy associated with the vibration (kinetic energy) of atoms and molecules. The motion of the various particles tends to be random and disordered, and the thermal energy is often referred to as a low-grade energy. The large-scale motion is due to *ordered* kinetic energy, where large-volumes of fluid move coherently. There is a transfer from highly ordered large-scale motion, through turbulence to lower grades. This energy cascade follows a certain rule, in which there will always be a fixed energy cascade rate.

5.3.2.1 Entropy

The climate system obeys the laws of thermodynamics. The *first law of thermodynamics* states that when heat is added to a gas (Q), the gas may do work (W), and the change in the internal energy is the difference between the heat added and the work done: $\Delta U = Q - W$. The internal energy change is independent of how the change was achieved.

The *second law of thermodynamics* states that the total energy of a closed system is always transformed into more disorderly forms until a state of maximum disorder is reached. This observation can be explained in terms of statistical considerations, where the interaction (collision) between the particles results in random motion due to the large number of particles involved. Thus, the second law is also a statement about probability, where a system moves to a more probable state and highly ordered states are less probable than disordered states. The concept of randomness can to some extent be justified in terms of quantum mechanics. One manifestation of the second law of thermodynamics is found in the difference in the wavelength of the incoming and outgoing electromagnetic energy on Earth: short-wave electro magnetic energy entering the terrestrial system drives the atmospheric circulation and is ultimately converted to heat, which in turn radiates the energy to space as long-wave radiation (see Figure 3.4).

[9] "The solar constant" varies by about 0.1%, and therefore is not completely invariable.

It is possible to define a quantity called *entropy* ($\Delta S = \Delta Q/T$), which describes the rate production of disorder. The entropy tends to increase in irreversible processes where a system cannot return to its original state without a net energy exchange with its surroundings.

Entropy is represented by the symbol S, and is related to the heat transfer and temperature according to $\Delta S = \frac{\Delta Q}{T}$.

5.3.2.2 Potential energy and latent heat

Potential energy is also present in the climate systems, and tends to be closely associated with the latent heat of evaporation. Water is evaporated and carried aloft, where it is condensed into cloud droplets while giving off heat. Thus, the latent energy is converted to thermal energy as well as potential energy during the condensation, as no energy can be created or destroyed. The potential energy of the hydrological cycle is harnessed by humans when lakes and rivers are developed into hydroelectric dams. The production of large amounts of hydroelectric energy for human consumption, which only involves a tiny fraction of the total potential energy of water in nature, gives an indication of how much energy is involved. Similarly, the melting of ice-sheets and glaciers involves an enormous amount of energy, and while at melting point, the temperatures will not change.[10] In summary, the melting and evaporation involve substantial changes to the energy available on Earth's surface; however, the temperature may not change much while these processes take place.

5.3.2.3 Electromagnetic energy

Electromagnetic energy is present on Earth in the form of light, infrared radiation, and charge separation between Earth's surface and the ionosphere. During lightning discharges, impressive amounts of energy are released within a small volume of air. Clouds, snow-cover, and ice modify the radiative forcing through reflection and absorption of light and infrared radiation (Figure 5.7). The electromagnetic energy received by the Earth and returned to space may also be associated with the distortion and disturbances of the geomagnetic field lines, such as those responsible for the auroras (northern lights).

5.3.2.4 Partitioning of energy

One fundamental geophysical question is whether the total ratio between the amount of energy in the various forms is approximately constant, or whether the energy is transferred between the various forms regardless of their state. If

[10] This is only true for equilibrium conditions.

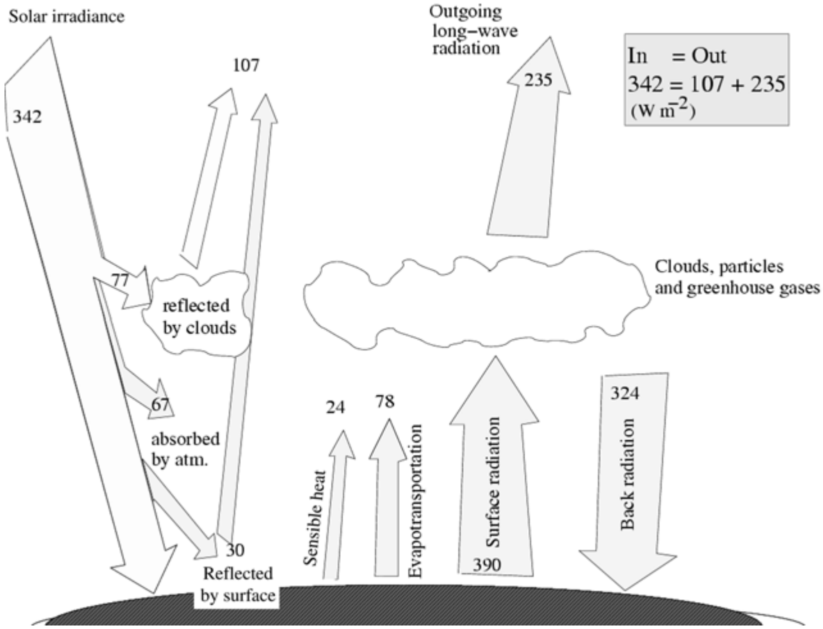


Figure 5.7. A schematic showing the balance and conversion between different forms for energy.

the energy is divided equally between the possible modes, then the process is *equi-partitioned.* One example of an equi-partitioned phenomenon is the second law of thermodynamics implying that heat flows from high temperature towards low temperature. Objects of different temperature but in thermal contact will eventually reach common temperature through a redistribution of the molecular kinetic energy, and hence there will be an equi-partitioning of the heat. Another example is the equatorial Kelvin wave, for which the kinetic energy equals the potential energy. Thermal and kinetic energy in the oceans and the atmosphere are related by the ‘‘thermal wind equation’’, as temperature gradients give rise to pressure differences, and hence atmospheric circulation. Systems which are not equi-partitioned, on the other hand, include the classical pendulum, where the instant energy is transformed back and forth between potential and kinetic energy. On average, however, the level of potential energy equals that of kinetic: $\overline{\mathrm{PE}} = \overline{\mathrm{KE}}$.

If the energy on Earth is equi-partitioned between the various forms of energy, then an increase in temperature due to an imbalance between the incoming and outgoing energy fluxes will also imply a strengthening in the circulation. On the other hand, if the Earth is in radiative equilibrium, then it is expected that the global temperatures are constant as well as the planetary kinetic energy being constant.

5.3.3 Momentum conservation

The fluid in Earth's oceans and atmosphere follows Newton's laws: (i) a body in motion will continue in a straight trajectory unless it interacts with a force, (ii) that the product between the acceleration and mass equals the force acting on the body, and (iii) any body acting on another will feel an opposite force with an equal strength. Newton's laws imply that the *momentum*, which is the product between the mass and velocity of a body, is conserved. The definition of acceleration is *the rate of change of velocity*, which according to Newton's second law requires that a force is exerted on the object.

Pressure is defined as force per unit area, and when air is under a pressure gradient (the pressure varies with distance), then it will accelerate. Pressure gradients may arise in connection with atmospheric waves as well as heating and cooling of the fluid. Differences in the moisture content in the atmosphere and salinity in the sea may also produce pressure gradients, and thus initiate motion.

The force of gravity is an important constraint on the large-scale geophysical fluid motion. The vertical temperature and pressure profiles of the troposphere are largely determined by a balance between the gravity and pressure forces, and this balance is often referred to as the hydrostatic equilibrium. The vertical velocities in the atmosphere are usually small, except for in small-scale weather systems such as extreme weather events. Therefore, the large-scale atmospheric flow is often assumed to be approximately 2-dimensional with only horizontal components, and the net vertical forces on the atmosphere are taken to be zero. There are certain regions, such as near the equator (the Hadley and Walker circulations), where the vertical motion cannot be ignored. The vertical density profiles are hydrostatically stable or unstable according to whether a vertical displacement will be restored to its original height or whether the motion accelerates.

> A statically stable atmosphere has no vertical acceleration, and the gravitational force is balanced by pressure force due to the weight of the air above. The hydrostatic equation describing the balance between gravity and pressure (buoyancy) is $\partial p/\partial z = -g\rho(z)$.

The motions of the oceans and the atmospheres are decelerated[11] by friction and drag force. The friction between the air and the ground converts the momentum from the atmosphere to the solid Earth, and the drag on the winds from the oceans transfers momentum from the atmosphere to the oceans. The winds are the primary driving source for the ocean currents.

[11] Negative acceleration or slowing down.

5.3.4 Effects of Earth's rotation

5.3.4.1 Angular momentum conservation

To a lay-person, one of the more obscure conserved physical quantities on Earth is the *angular momentum*. The fact that this is conserved means that a spinning object continues to spin as long as nothing interferes with it.

The angular momentum is usually represented mathematically by the symbol $\vec{L}$ and is defined as the cross-product between the particle momentum ($\vec{p}$) and a position vector from the axis of rotation ($\vec{r}$):

$$\vec{L} = \vec{r} \times \vec{p} \tag{5.3}$$

The conservation of angular momentum is analogous to the conservation of (linear) momentum, but whereas a force is needed to alter the linear momentum, changes in the angular momentum are produced by torques.

The rate of change in the angular momentum is the product between the force ($\vec{F}$) and the shortest distance to the axis of rotation:

$$\delta\vec{L} = \vec{r} \times \vec{F} \tag{5.4}$$

A common quantity used in meteorology is the *vorticity* which is defined as: $\vec{\zeta} = \nabla \times \vec{v}$. The vorticity is related to a conserved quantity also commonly used in meteorology (fluid mechanics) known as the *potential vorticity*, and is defined as

$$\mathrm{PV} = \frac{\zeta + f}{\delta p} \tag{5.5}$$

Usually, climate processes are not expressed in terms of angular momentum directly, but in the quantities of *vorticity* or *potential vorticity*. All these terms are related to Earth's rotation and the vorticity quantities are conserved, as these are various forms of the angular momentum.

5.3.4.2 The Coriolis force

Moving objects in a rotating coordinate system will maintain their momentum in an inertial reference frame, and their motion will appear to curve, viewed from the reference frame of the rotating system. This force is of the same type as one feels when trying to walk across a spinning platform at a funfair. From the rotating observer's point of view, there appears to be a force, a "fictitious" force, that bends the object's trajectory and is known as the *Coriolis force*. Low-pressure systems are associated with cyclonic flow[12] because there is a balance between the pressure gradient and the Coriolis force.

The Coriolis force is described by the cross-product between the Earth's angular velocity ($\vec{\omega}$) and the motion of the air ($\vec{v}$). For a finite volume (dV) of air with uniform density (ρ) in coherent motion, it is $\vec{F}_c - 2\rho\vec{\omega} \times \vec{v}$, and for the horizontal flow, it introduces an extra term perpendicular to the flow: fu and fv, where f is the Coriolis parameter [$f = 2\omega \sin \phi$], and where Φ is the angle of latitude.

Similarly to the Coriolis force, the fluid is subject to another fictitious force, the centrifugal force, which results in the Earth bulging out near the equator. The centrifugal force is opposite to the centripetal force that is responsible for the acceleration towards Earth's axis. The centripetal force is "provided" by Earth's gravity. The gravity, of course, is stronger than the centripetal force and hence stops things from flying off out to space. This subject is discussed further by Kleppner and Kolenkow (1978).

When air is lifted, for instance over a mountain range, it is vertically compressed and deflected towards the equator in order to conserve angular momentum.

The Coriolis force acts on any object moving in the rotating reference frame, but the moving object must always conserve its angular momentum. By moving further away from the axis of rotation, the angular velocity is reduced to conserve the angular momentum. The Coriolis force, on the other hand, is a fictitious force on an object which tries to conserve its *linear* momentum in an *inertial* frame of reference. Thus the Coriolis force only acts in the plane normal to the axis of rotation.

5.3.5 Charge conservation

Another physical quantity that is conserved is electrical charge. There is an electric potential between the upper atmosphere and Earth's surface, with the ionosphere being at 400,000 volts with respect to the ground. The electric charge plays a role in lightning and thunderstorms, but may also be important for the initiation of rain as well as formation of ice-crystals and hail in cloud processes. A brief review on the effect of an electric field is given by Harrison and Shine (1999). The electrification may increase the collection efficiency of charged drops and particles: small droplets collide and coalesce forming larger drops. This collision and coalescence mechanism

[12] Anticlockwise motion in the northern and clockwise in the southern hemisphere.

is believed to be an important mechanism by which small cloud drops can grow large and produce rain without the need of freezing.

5.4 EARTH'S ENERGY BUDGET

5.4.1 Variations in solar output and terrestrial temperature

The total electromagnetic energy flux density from the Sun at Earth's average distance from the Sun is measured to be 1370 W/m^2. This energy flux enters the top of Earth's atmosphere, and the total energy received by Earth equals the product between Earth's cross-section and the energy flux density. Some of this energy (30%) is reflected back to space and this fraction is referred to as the Earth's *albedo* (A). The energy is, to a first approximation, distributed evenly over Earth's surface,[13] and the Earth is in radiative equilibrium, which means that it gains no energy over time. The energy radiated from Earth is in the form of long-wave radiation (infrared), and is described by the black body radiation equation (see Section 2.2.2).

The black body radiation is $B(T) = \sigma T^4$ (the Stefan–Boltzmann constant $\sigma = 5.670 \times 10^{-8}\,\mathrm{J\,m^{-2}\,K^{-4}\,s^{-1}}$),[a] and Earth's radius r_e. By assuming an energy balance between incoming solar radiation and re-radiated thermal emission, it is possible to compute a theoretical value for Earth's mean emission temperature:

$$(1 - A)\pi r_e^2 S = 4\pi r_e^2 B(T) \tag{5.6}$$

By re-arranging equation (5.6) and cancelling equal terms, Earth's emission temperature can be expressed as:

$$\langle T_e \rangle = \left[\frac{(1 - A)S}{4\sigma}\right]^{\frac{1}{4}} \tag{5.7}$$

[a] For more detail about black body radiation, the reader is recommended the book by Fleagle and Businger (1980).

Earth's (emission) surface temperature is related to the Sun's energy production, and one would therefore expect variations in the solar energy production to affect the Earth.

[13] Tropical regions are in reality systematically warmer than polar regions.

The fluctuations in the mean temperature accompanying variations in the total solar irradiance can be estimated according to:

$$\frac{\delta T_e}{T_e} = \frac{1}{4}\frac{\delta S}{S} = \frac{1}{4}\frac{\delta F}{F} \tag{5.8}$$

It is, however, evident from equation (5.8) that changes in Earth's mean surface temperature, T_e, are not very sensitive to small fluctuations in the values of S. This means that variations in solar output must be large in order to induce significant changes in Earth's surface temperatures.

A 0.1% change in the solar constant will, according to this simple model, result in approximately 0.025% change near the tropics (where the Sun is at zenith angle) and less at higher latitudes. If the annual mean temperature is taken to be 15°C (288 K), then the amplitude of the temperature response will, according to equation (5.8), be 0.07°C and would probably be too small to be noticeable. Hoyt and Schatten (1993) have proposed that the global climate during the Maunder minimum was about one degree cooler than at the present ($\delta\langle T_s\rangle \approx 1°\text{C}$), which according to our simple radiation balance model implies $\delta F \approx 1.4\%$ (mean temperature assumed to be 288 K). Lean and Rind (1998) review past estimates of the solar forcing during the Maunder minimum, and quote values between 0.2% and 0.6%. Extrapolations based on calcium II brightness from sun-like stars give a reduction of 0.24%.

According to the temperature data compiled by Jones *et al.* (1998a), the global mean temperature has since the industrial revolution increased by approximately $\delta\langle T_s\rangle \approx 0.5°\text{C}$, which would require an increase in the total solar irradiance by $\delta S \approx 9.5\ \text{W/m}^2$ according to equation (5.8). The net change in radiative forcing on Earth would be $S(1 - A)/4 = 1.7\ \text{W/m}^2$.

There are obvious discrepancies between the simple energy balance models and the observed state of Earth's climate, suggesting that there are other factors playing a role in the climate system. The energy balance model does not take into account heat transport due to atmospheric and oceanic circulation (see Section 8.6.1.1), heat stored by the oceans and land, and the radiative effect of atmospheric gases and clouds.

There are variations in the global mean temperature on Earth which may not be a result of variations in the external forcing, but may be influenced by the variations in the climate system itself. Such variations are due to *internal variability* of the climate system. In order to demonstrate how the temperature may vary with time without being affected by external factors, one can use computer models to simulate the climate under a constant forcing. The global mean temperature is influenced by the heat distribution of the ocean surface. One example is that the global mean temperature tends to be high during El Niño events (when the eastern tropical Pacific sea surface is warmer than normal). The extent of sea-ice cover and snow affects the planetary albedo and may also influence the global temperature. Sea-ice is affected by the ocean circulation.

5.4.2 Variation in insolation

5.4.2.1 The diurnal cycle

The fact that the insolation varies diurnally (with the day and night) is a "trivial" example of how the atmosphere responds to changes in the available solar energy. Every 24 hours, the Earth makes a complete rotation about its own axis and the sunlight varies as the Sun is nearly above the equator. The atmosphere tends to cool during night, with greatest differences between day and night-time in the low-latitude, mid-altitude regions with continental climate (spring and autumn) and the smallest diurnal variations near the sea. The diurnal cycle also has more subtle effects on the atmosphere, such as the *nocturnal jet* which is a result of strong radiative cooling near the ground. During the day, the ground heats due to the effective absorption of sunlight, and since warm air tends to be lighter than cold air, surface air rises and mixes with the air above, resulting in a more uniform vertical temperature profile. The mixing gives rise to a frictional stress acting on the free atmosphere above the boundary layer. During night, the ground cools off more quickly than the atmosphere, and the coldest air is accumulated near the ground, hence producing a much stronger vertical temperature gradient. The vertical temperature profile inhibits mixing, and the free atmosphere no longer feels the friction as it does during daytime. Thus, a different flow structure is set up (Gill, 1982).

The diurnal cycle varies with latitude and seasons. The days and nights are of approximately equal duration near the equator, but in the polar regions the sun does not set in the summer and the polar nights (the Sun is below the horizon) during winter last for a couple of months. The reason for these seasonal variations in the diurnal cycle is that the Earth's axis is tilted with respect to its orbital plane and this axis points in the same direction throughout the year. June 21st is the longest day in the northern hemisphere and the daylight hours are shortest on December 21st.

5.4.2.2 The annual cycle

As the Earth rotates around the Sun, the direction of its axis changes with respect to the Sun. Earth's orbit around the Sun is slightly elliptical with an *eccentricity* (e) of $e = 0.017$. *Perihelion* is the point where the distance between the Sun and the Earth is at minimum ($r = r_{\text{min}}$, presently at January 5th).

Earth's distance from the Sun as a function of the day of the year:

$$r(j) = \left[1 + 0.0334 \cos\left(2\pi \frac{(j-2)}{(365.25)}\right)\right]\bar{a} \tag{5.9}$$

The term *aphelion* is the point where a planet is farthest from the Sun. For objects orbiting the Earth, the term *apogee* is used, and *apoapsis* is used for objects orbiting other bodies.

There may be an annual variation in Earth's global temperature due to the eccentricity of Earth's orbit. The global mean temperature annual cycle may also be affected by the different geographical features between the southern and northern hemispheres, which may affect the albedo of the different hemispheres.

The annual variation in the temperature tends to be small in the tropics and large in the high latitudes. Thus, the seasonal variations are another "trivial" example of how the atmosphere responds to variations in the solar energy. There is also a pronounced seasonality in the tropical rainfall patterns, with wet and dry seasons. The south Asian monsoon is a weather system which tends to bring rain to southern India in early June, and is driven by the solar-induced temperature contrast between land and sea. The rains tend to follow the Sun with the wet season taking place during the summer season.

5.4.3 The natural greenhouse effect

One of the first persons to propose the so-called "greenhouse effect" was the Swedish scientist Svante Arrhenius (1896). This theory states that trace gases in the atmosphere, such as CO_2, water vapour (H_2O), and methane (CH_4), absorb the outgoing long-wave radiation that is emitted from Earth. When these gases re-emit this energy, it is radiated equally in all directions (isotropically), and part of the energy is therefore sent back to Earth's surface. According to the simple energy balance in equation (5.8), Earth's mean *emission*[14] temperature is $\langle T_e \rangle = 256\,\text{K} = -17^\circ\text{C}$. The estimated global mean *surface*[15] temperature derived from observations is $T_s = 288$ K $= 15^\circ$C. The natural greenhouse effect increases the global surface mean temperature by approximately 32°C.

The mean surface temperature of Venus (about 740 K) is higher than for Mercury (up to 670 K on the day-side and down to 70 K on the night-side) despite being further away from the Sun. The global mean terrestrial temperature is furthermore substantially higher than can be explained by a radiative equilibrium with the Sun. This observation can be explained in terms of a natural greenhouse effect. (Mars has a thin atmosphere and is just barely warmer than the prediction of the radiative equilibrium model.) The wavelengths at which the radiation is absorbed depends on the molecules' absorption lines and energy bands, which are explained by quantum physical considerations (French and Taylor, 1989; Houghton, 1991). Some of these gases, such as water vapour, may play an important role in feedback processes due to their spectral properties and phase sensitivity to temperature.

[14] Derived from theoretical energy balance.

[15] Derived from observations.

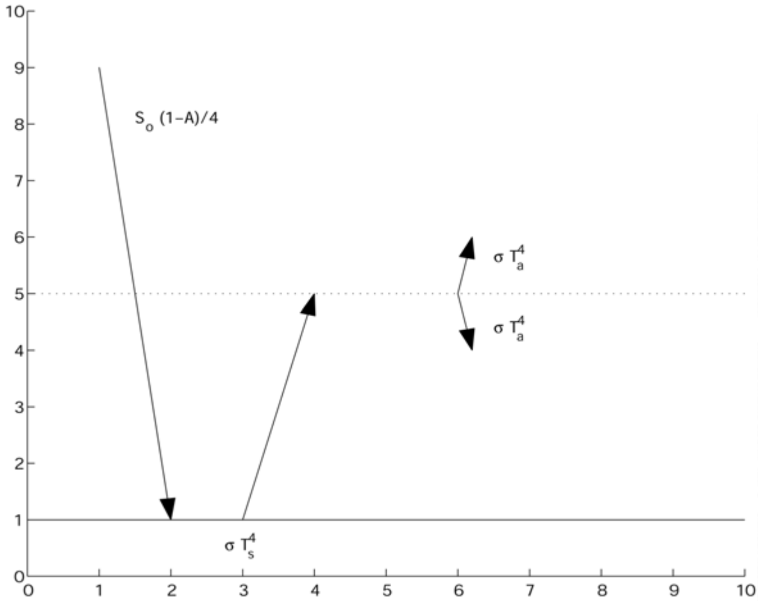


Figure 5.8. A diagram illustrating the radiation balance for a one-layer grey atmosphere. T_a is the temperature of the atmosphere and T_s is the surface temperature.

5.4.3.1 A simple greenhouse effect model

A simple greenhouse gas model may illustrate how the atmosphere can warm the surface. Suppose the atmosphere is completely opaque to the short-wave sunlight, but absorbs the long-wave radiation (infrared). Such an atmosphere is often referred to as a "grey atmosphere", and the model is illustrated in Figure 5.8.

The Earth can only lose energy to space from the top of the atmosphere since all long-wave radiation emitted from the surface is absorbed in the atmosphere. If an equilibrium situation is assumed, so that the Earth does not gain or lose energy over time (warms up or cools off), then the radiative heat loss must equal the solar energy coming into the terrestrial system.

The temperature of the atmosphere must equal the emission temperature (equation (5.7)):

$$\frac{S(1-A)}{4} = \sigma T_a^4 = \sigma T_e^4$$

Since there is no absorption of the sunlight (short-wave) in the atmosphere, the atmosphere must receive energy from the surface. This energy flux is divided into upward and downward radiation, and hence a factor of 2. Assuming that the energy exchange between the ground and the air is through radiation, and assuming the values $S = 1387\,\mathrm{W/m^2}$ and $A = 0.3$ the surface temperature T_s can be calculated:

$$\sigma T_s^4 = 2\sigma T_a^4 = 2\frac{S(1-A)}{4} \rightarrow T_s = 304\,\mathrm{K}$$

In other words, the greenhouse effect gives rise to warmer surface temperatures. This picture is complicated by the fact that the atmosphere also receives a significant amount of energy from the ground through evaporation (latent heat) and convection (Figure 5.7). Furthermore, clouds play an important role by increasing the albedo but also enhancing the greenhouse effect. A rule-of-thumb is that high clouds tend to warm the surface while low clouds tend to result in cooling. Furthermore, the atmosphere is not (yet) completely transparent to the long-wave radiation, as there are spectral intervals where the absorption is low (the wavelength interval 8–12 μm), and a small fraction of the long-wave radiation emitted from the surface escapes to space. The greenhouse effect increases with increased greenhouse gas concentrations even if the atmosphere is completely opaque in terms of infrared emission from Earth's surface (the optical depth continues to increase), and the simple model above can describe this situation by adding more "layers" to the atmosphere: The surface temperature of an n-layer atmosphere can be expressed as $T_s = \sqrt[4]{n+1}T_e$ (Hartmann, 1994).

5.5 THE BASIC COMPONENTS OF EARTH'S CLIMATE

5.5.1 The atmosphere

5.5.1.1 The troposphere

The *troposphere* is the lowest layer of air in the atmosphere (Figure 5.9) and is where the weather processes take place and is the stage for our climate.[16] In other words, it is in the troposphere that virtually all weather processes take place. The troposphere contains the bulk of the atmospheric mass and extends up from the surface to the *tropopause*, which is approximately 10 km a.s.l. The height of the troposphere varies in space and time, and the tropopause[17] is highest near the equator. There is a break in the troposphere somewhere between 30° and 60°, associated with the polar front as well as between the tropics and the subtropics. The polar front is a sharp boundary between cold and warm air masses, often associated with strong jet winds and wind shear. Another semi-permanent (for the winter season) frontal system is the arctic front, usually located polarward of 60°. The sharp meridional temperature contrasts near these fronts also drive jets of strong zonal (westerly) winds.

[16] By "climate" we usually mean the surface climate.

[17] The top of the troposphere.

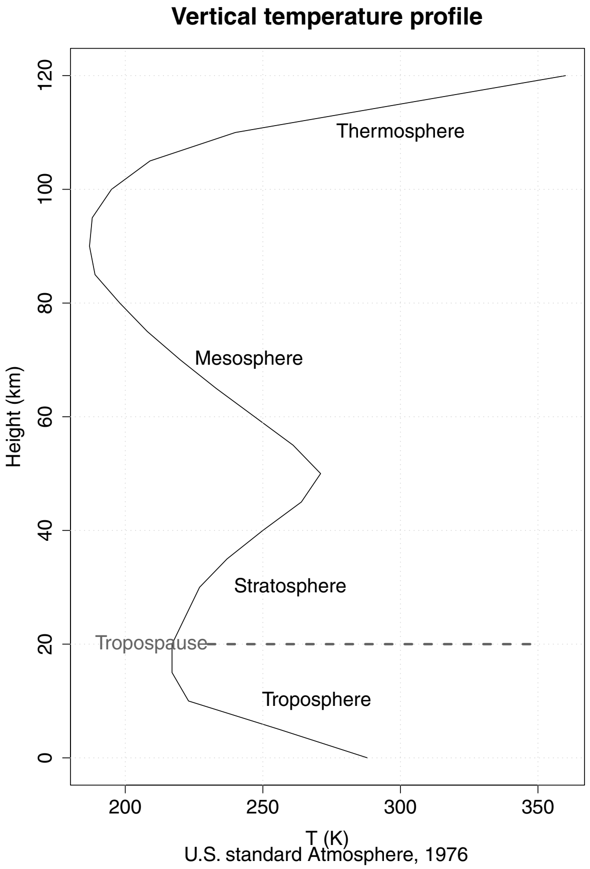


Figure 5.9. A typical vertical profile of temperature in the atmosphere. The different regions are labelled. The boundary between the troposphere and stratosphere is called the tropopause and is marked by the horizontal grey dashed line.

The vertical temperature distribution normally follows the so-called lapse rate of 6–7°C/km. The lapse rate varies with the atmospheric water content, as the water molecules (H_2O: atomic mass 18) are lighter than dry air (mainly N_2, O_2, and Ar with a mean atomic mass 28.97), and according to *Avogadro's* law, there are an equal number of gas molecules per unit volume for equal partial pressure and temperature.

As the polar regions on average receive less energy from the Sun than the tropics, the polar regions tend to be colder than the equator (equation (5.11)). Figure 5.10 illustrates how the radiative energy balance predicts relatively low high-latitude temperatures. However, the observed meridional temperature profile on Earth indicates a smaller gradient than that expected from a pure local radiative equilibrium, and it is clear that factors other than the solar radiation also influence the temperatures. The energy balance model given in equation (5.6) is too simple and does not include other types of energy transfer, such as *advection*, which is physical transport of heat due to macroscopic motions such as warm air replacing cold air. A prerequisite of advection is that there must be some form of atmospheric circulation that brings warm air poleward from the tropics.

5.5.1.2 Local temperature balance

The incidence of sunlight is nearly directly overhead (at zenith, see Figure 5.11) in the tropics, whereas the polar regions are subject to polar nights all winter and receive the sunlight at an angle during summer. The intensity of the short-wave solar radiation, S, is less for a non-zero zenith angle than if the Sun is at zenith, because a beam of light with a unit cross-sectional area hits a larger area when it is not normally incident to the plane. The daily mean radiation varies with latitude (ϕ) and is the main reason for the meridional temperature profile of cold poles and warm tropics. This temperature profile is shown in Figure 5.10.

Equation (5.10) gives an expression for the daily mean radiation received at a given latitude.

$$\overline{S_{\text{day}}} = \frac{S}{\pi}\left(\frac{\bar{r}}{r}\right)^2 [h_0 \sin(\phi)\sin(\delta) + \cos(\phi)\cos(\delta)\cos(h)] \tag{5.10}$$

If h_0 is the hour angle (in radians) at sunrise and sunset, then $\cos h_0 = -\tan\phi\tan\delta$, where δ is the declination angle.[a] The Sun's zenith angle, z, is related to the latitude (Φ), solar declination and the local time in hours according to the relation $\cos(z) = \sin(\Phi)\sin(\delta) + \cos(\Phi)\cos(\delta)\cos(h)$. r is the Sun–Earth distance (equation (5.9)). The intensity (energy density) of the light arriving at the surface of the Earth is $S\cos(z)$ for any instant.

On a local scale, an equilibrium temperature can be estimated by balancing the incoming short-wave radiation S and the outgoing long-wave radiation. The angle of latitude is ϕ, with $\phi = 0^\circ$ at the equator and $\phi = \pm 90^\circ$ at the poles. The insolation is latitudinally dependent, and a simple model of the energy budget (assuming no meridional heat transport) can be expressed as:

$$\sigma T^4_{\text{theor}} = S\cos(\phi) \tag{5.11}$$

Figure 5.11(a) shows the theoretical meridional temperature profile computed using this simple expression together with estimations based on real observations. The discrepancy between these two profiles can be explained in terms of meridional energy transport[b] (Q) assuming a steady-state solution:

$$\sigma T_{\text{real}}^4 + \frac{dQ}{dy} = S\cos(\phi) \tag{5.12}$$

By substituting the theoretical temperature profile into equation (5.12) the meridional heat transport can be derived from the two temperature profiles according to $\frac{dQ}{dy} = \sigma(T_{\text{theor}}^4 - T_{\text{real}}^4)$. Figure 5.11(b) shows the solution for the heat transport Q.

[a] The latitude where the mid-day Sun is at zenith. At present, $\delta = 23.5°$.

[b] dQ/dy is known as the divergence of the heat transport, i.e. the difference between the heat coming in from the equator side and that leaving at the pole side.

The latitudinal variation in the solar radiation implies that changes in the solar intensity are most noticeable near the equator, everything else being equal. However, since the real climate system is complex with a number of feedback mechanisms taking part, the strongest signal may be seen at higher latitudes.

5.5.1.3 The atmospheric circulation

Several types of circulation features have been identified in Earth's atmosphere and meridional heat transport may arise as a result of several processes. Near the equator, the Hadley cell (Figure 5.12) is important, but at higher latitudes so-called "eddy-transport" (mediated through cyclones and planetary waves) becomes important. The transport of energy can be through heat advection (warm air moving poleward), latent heat (water vapour), or potential energy. The climate system, through its energy transport, can be regarded as a kind of heat engine (Carnot engine). One important aspect of this poleward energy transport is that it drives the atmospheric circulation. A controversial hypothesis has been put forward (Lopez, 2001) stating that the poleward heat transport adjusts itself in such a fashion that it maximises the mechanical work caused by the fluid. If the heat transport is too great, then the temperature gradient between the tropics and the poles vanishes, whereas too weak a transport results in much warmer tropics and much cooler poles. With very weak flow, there is little energy to convert into work and the amount of work carried out by the fluid initially increases with the flow rate. But at some given flow rate, the mechanical work diminishes because of reduced efficiency which is proportional to the temperature gradient.

An important circulation feature is the Hadley cell, which comprises rising warm moist air at the equator and sinking (subduction) in the sub-tropics. The Hadley circulation is not responsible for much poleward energy transfer, but brings warm air aloft near the equator. It is important for the poleward angular momentum transport, tropical hydrological cycle and tropical cloud formation. Closely connected with the Hadley circulation is the Walker cell, for which the flow is

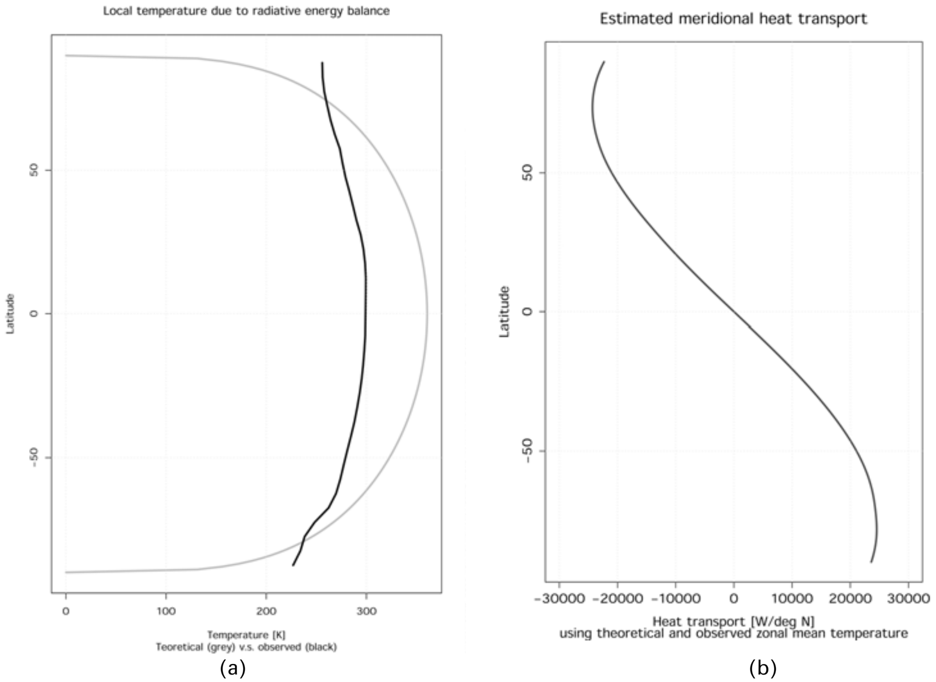


Figure 5.10. Comparison between the theoretical local emission and observed (Climate Research Unit, University of East Anglia) temperatures as functions of latitude (a) and meridional heat transport derived from the difference between the two quantities (b). The spherical geometry is not accounted for in this illustration.

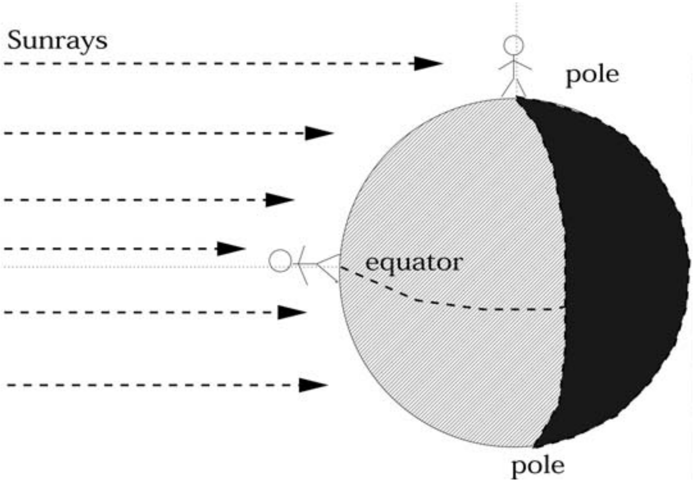


Figure 5.11. An illustration of the difference in the angle of sunlight incidence at the equator and the poles.

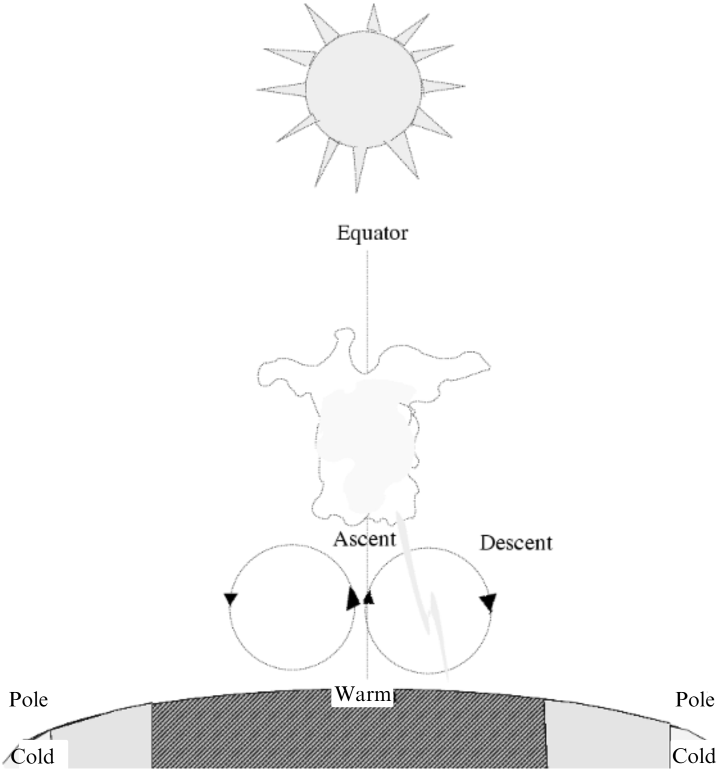


Figure 5.12. A schematic diagram showing the basic features of the Hadley cell.

parallel to the equator as opposed to meridional. The Walker circulation and Hadley cell are related to the trade wind system.

The atmospheric flow in the sub-tropics and the mid-latitudes is responsible for a significant part of the poleward heat advection, but the oceans also play a role. In the mean large-scale circulation, relatively cold air in the high latitudes is forced aloft and rises (known as thermally indirect circulation), whereas relatively warm air sinks in the mid-latitudes (the *Ferrel cell*), resulting in a meridional exchange of air-masses where warm air is brought poleward and cold air moves towards the equator. The mid-latitudinal meridional circulation is part of the Ferrel cell, but much of the energy transfer is facilitated though large-scale eddies, as opposed to a steady flow.

An important part of our climate involves clouds (Figure 5.13). A thick cloud cover usually implies cooler weather during summer, but higher temperatures during winter at high latitudes. The clouds play a dual role: reflecting the solar radiation (short-wave radiation) and trapping outgoing long-wave radiation (infrared). Cloud formation requires atmospheric moisture, but the cloud drops tend not to form spontaneously. Usually, the drops form on particles (cloud condensation nuclei,

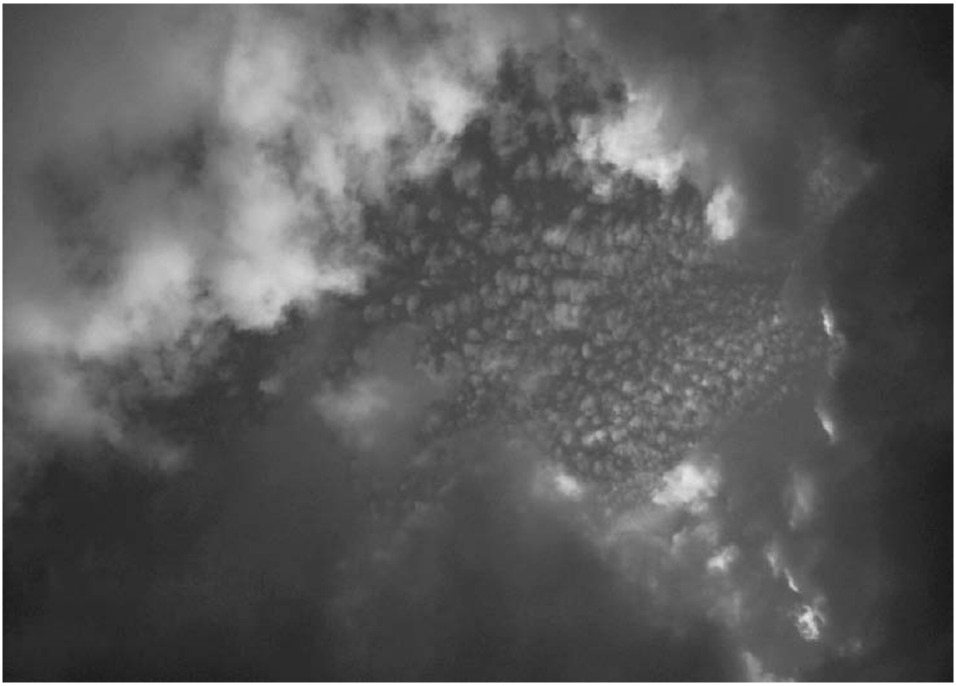


Figure 5.13. Different cloud types have different effects on the surface climate.

CCNs). The size of the cloud drops is determined by the number of cloud condensation nuclei, the availability of moisture (vapour pressure) and temperature. Because each cloud drop competes for water, the number of cloud drops tends to vary inversely with the size ($N_r \sim r^{-3}$). When the droplets reach a critical size, they will grow spontaneously as larger drops require lower humidity.

One of the unsolved problems in atmospheric physics is to explain how rain is formed within a few hours. There are two proposed mechanisms for rain: warm[18] and cold[19] initiation. Current theory predicts a droplet growth rate in warm initiation that diminishes with time in terms of the droplet radius. Other factors than just diffusion must play a role in accelerating the droplet growth, and a collision-and-coalescence process where drops join to produce even bigger drops has been proposed as a likely mechanism. The remaining difficulty is to explain how the droplets grow sufficiently large that the collision efficiency reaches a threshold value where the raindrop formation takes off through a cascading process (avalanche). Only a few large drops are necessary, as they break up after reaching a critical size. Explanations for accelerated growth before collision and coalescence may include giant condensation nuclei and various mixing processes that favour the

[18] Only the liquid water phase is involved – no freezing.

[19] A process that involves freezing.

growth of some drops to others. Electrostatic forces may play a role, as charged droplets may attract dipole molecules such as H_2O. The cold initiation of rain is less problematic than the warm initiation, as the ice crystal growth rate is higher than for liquid water. A more detailed discussion on cloud physics is given by Rogers and Yau (1989).

Clouds with many small droplets are more effective at reflecting the short-wave solar radiation, and these are therefore associated with higher albedo (A) than clouds with fewer and larger drops. For a given amount of water, many small drops give larger total surface area than a few large drops. The cloud characteristics vary geographically, and in time. Continental clouds tend to have more and smaller cloud drops than maritime clouds.

High clouds are capable of trapping more outgoing long-wave radiation than lower clouds (see Section 7.10.4), but the cloud's modifying effect on the radiative balance also depends on the cloud thickness.

Liquid drops have different scattering properties from ice crystals. The geometrical shapes of the ice crystals vary with their history, and the temperatures and humidity tend to decide which form they take. The crucial cloud parameters in a climatological context are therefore cloud base height, cloud top height, cloud drop population, cloud drop or ice crystal shape and size spectrum.

5.5.1.4 The stratosphere

Above the troposphere there is a hydrostatically stable region which restricts vertical motion, and the base of this region is known as the *tropopause* (Figure 5.9). Above this region, the temperature increases with height, which explains the hydrostatic stability. This vertical temperature gradient is caused by photochemical reactions,[20] where solar ultraviolet (UV) light is absorbed by ozone (O_3). The absorption implies energy convergence (accumulation) unless an equal amount of energy escapes by thermal emission.

Atmospheric flow over surface irregularities, such as mountain ranges, sets up wave motions known as *gravity waves*. These waves have a vertical group velocity and may propagate into the stratosphere. Their amplitude depends on the hydrostatic stability (the restoring forces), the atmospheric flow and density. When these waves reach sufficiently high altitudes, the air becomes increasingly rarefied and these waves eventually break and dissipate. The gravity waves thus contribute to the energy transfer from the Earth's surface to the troposphere. These waves may also influence the stratospheric equatorial mean flow which tends to reverse approximately each year. The oscillation in the stratospheric mean flow is known as the *quasi-biennial oscillation* (QBO). The high-altitude winds blow from west for approximately one year before they switch and are easterly for another year. The phase change tends to start at high altitudes and the phase then tends to propagate downwards.

[20] Sunlight (energy) is absorbed through chemical reaction, but the temperature is also affected by excess energy.

Because the stratosphere represents only a fraction of the atmospheric mass and heat content, it is traditionally believed to play a minor role for Earth's surface climate. Moreover, the fact that this region is dynamically isolated from the troposphere suggests that the stratospheric motions have little impact nearer the ground. However, a recent study suggests that there may be a downward signal propagation from the stratosphere (Baldwin and Dunkerton, 2001). Furthermore, recent theories about the Arctic oscillation (AO) propose that the stratosphere may play a role for the tropospheric circulation (Thompson and Wallace, 1998). The existence of the AO is still a disputed topic, as some scientists have argued that this pattern is an artefact of the mathematical analysis, and does not really have a physical meaning (Ambaum *et al.*, 2001).

5.5.2 The oceans

The oceans play an important role in the climate system. They represent an immense heat reservoir, and the upper mixed layer of the world oceans has about 30 times higher heat capacity than the entire atmosphere. The high heat capacity of the seas gives the oceans an enormous thermal inertia, and the persistence of the oceans may act as a memory in the climate system. The fact that the oceans have greater heat effective heat capacity than the continents, is partly due to the heat being more easily transferred to greater depths by currents and mixing than over land as well as being due to the higher heat capacity of water compared to land (Hartmann, 1994).

5.5.2.1 Heat capacity

The net energy required to heat the upper 1 m of the world's oceans by an average of 1.0°C is to a first-order approximation 1.5×10^{21} J (or 4.12×10^6 J/m^2). If such a change was driven by changes in the Sun it would require a linear increase in the solar irradiance of 1.83 W m^2 decade^{-1} over 10 years if it were solely due to a brightening of the Sun, assuming that none of this additional energy is lost due to changes in the long-wave radiation (assuming the average depth of the upper ocean layer being affected ~70 m). To change the entire mixed layer by the observed magnitude of 0.5°C on a global scale would therefore require at least 0.91 W m^2 decade^{-1} increase over 10 years, which is a 0.38%/decade increase in the annual mean short-wave radiative flux of 240 W/m^2 over 10 years. Because warmer sea surface loses more energy in the form of long-wave radiation and the planetary albedo reflects part of the sunlight back to space, this value is an underestimate of the required increase in the solar energy. Using the expression $\Delta S = \frac{4(4\sigma T^3 + 0.7.c.\rho)\Delta T}{(1-A)} \cdot \frac{1}{2t}$, a more accurate estimate of the change in S is equivalent to 3.65 W m^2 decade^{-1}. For comparison, satellite observations from the late 1970s have noted changes in the solar luminosity of around 0.1% (1.37 W/m^2) over a solar cycle (11 years) at the top of the atmosphere which would correspond to

0.24 W/m^2 evenly distributed on the surface. Variations at the surface of the order 0.24 W/m^2 is too low to account for the observed 30-year temperature change in the world ocean mixed layer by 0.5°C.

The land–sea contrast in heat capacity is evident in the differences in the surface heating over the sea and the land. The land tends to be warmer than the sea during hot days, and the warm continental air becomes warmer than its surroundings and starts to ascend. Maritime air flows towards the land to fill in and replace the ascending air, and a land–sea breeze is born. The land–sea temperature contrast has profound implications for large-scale seasonal timescales, and is responsible for monsoon systems.

The sea surface has lower friction than the land surface, and the land–sea friction contrast may affect the prevailing circulation patterns. Therefore, surface winds tend to be stronger at sea than over land.

5.5.2.2 Heat transport

The ocean currents play an important role in the poleward energy transfer, and are estimated to be responsible for up to one-third of the meridional heat transfer (Trenberth and Stepaniak, 2004). The current systems in the Atlantic Ocean are particularly important, as it is the only ocean which does not have closed northern or southern boundaries.[21]

The ocean currents are sometimes steered by the bottom topography, and are therefore an important medium through which geological features may affect the climate. The Denmark Strait is one such feature through which overflowing bottom water from a shallower Greenland-Iceland-Norwegian Sea (GIN Sea) may mix with water found at a much deeper level.

The heat transferred by current systems has a profound influence on local climates. An illustration of the importance of oceanic heat transport is the Norwegian climate where the mean temperatures are 5–10 degrees higher than places in eastern Greenland and Alaska at the same latitude. This mild climate is primarily a result of the warm water transported northward by the Gulf Stream, its extension into the North Atlantic (the North Atlantic drift) and the North Sea, and the Norwegian current. This heat transfer is a part of the "global conveyor belt", which is a global current system driven by density differences (thermohaline circulation due to salinity and temperature differences) and prevailing winds. The prevailing winds over the oceans include easterly trade winds associated with the Hadley and Walker circulations over the tropics and Westerlies in the mid-latitudes. The *Doldrums* are the regions with little wind and grey skies associated with ascending air motion located between these two wind belts.

The wind forcing produces basin-wide circular current systems, known as *gyres*. In the Atlantic, there are the sub-tropical gyre associated with the Gulf Stream

[21] The Arctic Sea to the north (meets the Polar Sea) and the Southern Sea to the south (connected to the Southern Ocean). The Pacific is virtually closed in the north with the narrow Bering Strait, and the Indian Ocean meets the Asian mainland in the north.

bringing warm water from the Gulf of Mexico northward, and the sub-polar gyre in the North Atlantic transporting cold water from the Labrador Sea to the south. The boundary surface currents (coastal) in these gyres are usually much narrower and faster-flowing in the west than the returning currents in the east. The *Coriolis* force acts perpendicular to the motion (to the right in the northern and the left in the southern hemisphere), and wind-forced currents are subject to a balance between the Coriolis force and friction resulting in a surface flow in the direction veering to the right of the winds. The motion may be regarded as the sum of a component along the direction of the wind and one perpendicular to the winds, and the latter is referred to as the *Ekman drift*.

5.5.2.3 The oceans and the hydrological cycle

The oceans are virtually limitless water reservoirs and provide the air with moisture. Hence, the oceans are the source of water in the hydrological cycle, and are influential in cloud formation and precipitation. It is well known that maritime climates tend to be wetter than far inland over the continents. Because of the high thermal momentum (persistence) of the oceans, they tend to dampen and reduce the climatic variations. Coastal temperatures tend to vary to a lesser extent than temperatures further inland. The cloud properties also differ between maritime types and continental types. Maritime clouds often have a lower cloud base than the continental ones. The sea spray caused by wave breaking introduces sea salt into the atmosphere. Sea salt and sulphate represent important cloud condensation nuclei (CCN) for clouds over the oceans, whereas for continental cloud, soil (mineral) particles and aerosols from forest fires are important sources for CCN (Pruppacher and Klett, 1978; Götz *et al.*, 1991; Rogers and Yau, 1989). The presence of oceans influences the cloud drop population, and maritime clouds tend to have fewer but larger cloud drops.

Warm sea surfaces may pump energy into weather systems such as hurricanes and storms. Energy is taken from the oceans by dry surface winds over warm ocean surfaces. This air then becomes moist, and is brought up to higher levels where the water condenses and releases latent heat. In addition to energy transfer by evaporation and condensation (latent heat transfer), the oceans may provide the air with energy through thermal conduction (sensible heat) from the physical contact between the air and the water surface.

The west coasts of Africa, Latin America and the USA are arid compared to the east coasts of the USA and Asia. Arid coasts are usually found along the eastern ocean boundaries in the sub-tropics, and are related to the prevailing winds and cold coastal waters. The coastal sea surface temperatures are cold because of a process known as Ekman pumping where equatorward winds force the surface water westward. Cold sub-surface water from underneath is upwelled in order to replace the original surface water.

5.5.2.4 Oceanic waves

Wind forcing tends to produce ocean waves on the surface, such as those familiar to sailors, but also internal waves under the sea surface with much larger horizontal scales. Such internal waves arise from the restoring action of rotation and gravity, and have various properties. Waves are believed to play a central role in ENSO. They may be regarded as adjustments to disturbances in the oceans and the atmosphere that propagate away from the source of perturbation like ripples on a pond. However, whereas the sound waves, ripples (surface tension waves) and surface gravity waves travel with the same speed in all directions, various oceanic and atmospheric waves travel along preferred directions.

There are various types of waves, and the wave character depends on the restoring force, their frequency and wavelength. The gravity waves, including the surface waves, are primarily subject to gravitational restoring forces, and are found deep in the oceanic interior. The restoring force of the sound waves, gravity waves and ripples are pressure differences, gravity and surface tension respectively.

Slower waves, such as *Kelvin waves*, are also subject to gravitational restoring forces, but are influenced by Earth's rotation (the Coriolis force). The Kelvin waves are so-called boundary waves that cannot propagate in the ocean interior except for along the equator. Kelvin waves can only travel from west to east along the equator and with a boundary to the right (left) in the northern (southern) hemisphere.

Rossby waves, on the other hand, arise from the restoring force of Earth's rotation (Coriolis force), and may be found away from the boundaries. The Rossby waves can travel both eastward and westward, depending on the wavelength. The waves with the shortest wavelength tend to dissipate faster than the long Rossby waves. There are also *mixed gravity–Rossby waves*, subject to both gravitational and rotational restoring forces.

Additionally, there are tidal waves, produced by the drag force of the Moon, as well as sound waves. Because of the tidal effect, the oceans may act as a medium through which the Moon in principle may influence our climate. Tidal currents may produce more mixing (tidal mixing) that may influence the sea surface temperatures and salinity profiles, for instance through various straits such as the Indonesian through-flow region. Some oceanographers, such as Egbert and Ray (2000) and Wunch (2000), have suggested that the effects of the Moon on our climate may be profound although subtle.

The oceans tend to be more sluggish than the atmosphere, due to the larger oceanic thermal inertia and momentum (density). The oceanic processes are furthermore associated with different timescales depending on the depth. The response to forcing is quick near the ocean surface compared to that in the deep ocean. The near-surface currents tend to be faster than deep ocean currents. The vast oceanic inertia provides a memory where deep ocean anomalies are signs of past events.

The equatorial Pacific tends to be cooler in the east than the west, where the warmest sea surface waters on Earth are found (the "warm pool" with temperatures exceeding 28°C). In the Indian ocean, the highest temperatures are found in the eastern and northern basin (Figure 5.14).

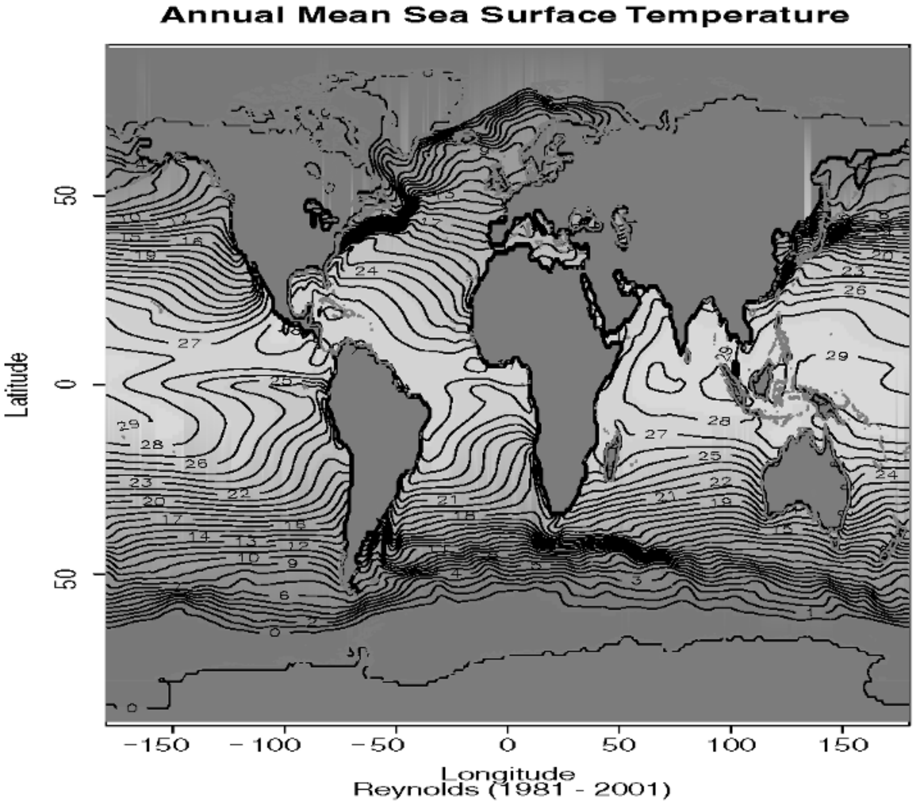


Figure 5.14. Geographical distribution of annual mean SST.

5.5.2.5 Oceans and the carbon cycle

The atmosphere and the oceans constantly exchange CO_2 and both play important roles in the carbon cycle. The oceans may act as a sink for the atmospheric CO_2, and their ability to absorb CO_2 increases as the water cools, everything else being constant. The near-surface oceanic CO_2 uptake would slow down as the ocean surface gets saturated if it were not for the fact that ocean currents renew the surface water. Thus the ocean currents pump the CO_2-rich water to deeper depths (deep water formation), where the water mass may be resident for thousands of years. Furthermore, living organisms, such as plankton, also take up carbon, and when organisms die and fall to the bottom of the sea, carbon may be removed from the oceans.

Weathering can trap CO_2 by depositing the gas on rocks and the deposition may hence act as an atmospheric carbon sink. Such weathering processes cannot take place over oceans, since there are no rock surfaces exposed to the air, and therefore the weathering process only takes place over land (not covered by ice).

The state of the oceans is important for the existence of sea-ice. Warm currents and high-salinity water keep the waters ice-free, as in the Norwegian and Barents Seas. The stress from surface currents and winds stretches and compresses the ice in addition to being the primary cause for ice-drift.

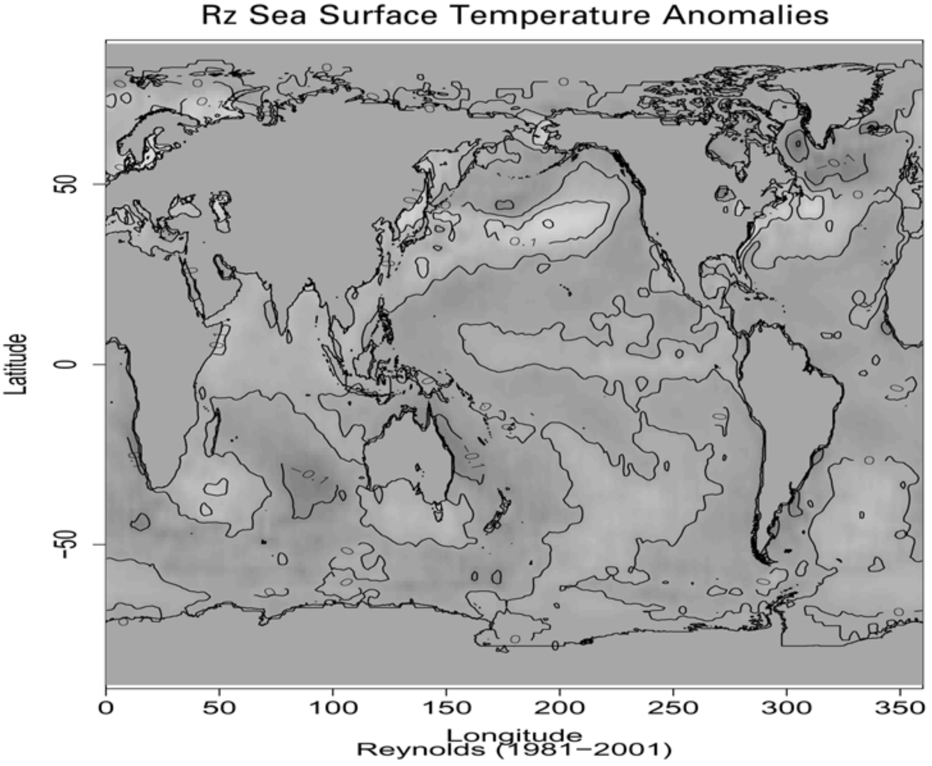
5.5.2.6 Solar activity and the oceans

The oceans affect the climate, and there is a question whether the oceans are influenced by solar activity. Reid (1987) found a correlation between the global SST and the sunspot record. Empirical evidence based on SST from satellite observations suggest that there is indeed a statistical relationship between the sunspot number and the SST (Figure 5.15). The ANOVA results suggests that the link is highly significant, despite the short time span of the SST record. There are two ways that solar activity may affect the oceans: either directly or indirectly through the atmosphere. For instance, if the solar activity affects the cloud formation, there may be a knock-on effect on the oceans through changes in the geographical heat distribution and the hydrological cycle.

5.5.3 The cryosphere

5.5.3.1 The role of ice in the climate system

On long timescales, the interaction between ice and the atmosphere plays a central role in the climate. Ice has a high albedo, which means that it reflects a large portion of the incoming solar radiation. New and old ice have different properties in terms of reflecting and emitting electromagnetic radiation. Fresh ice reflects the sunlight more efficiently (has higher albedo). Sea-ice is also thought to play a key role for the thermohaline circulation, the ocean current system driven by ocean density differences. The sea-water density is a complicated nonlinear function of both salinity and temperature. The density of cold sea-water is for instance more sensitive to salinity than warm water, which implies that when highly saline (salty) water from the warm Mediterranean is transported northwards and then cools, it becomes very heavy. This cooled salty water tends to sink in the Greenland–Iceland–Norwegian Sea (thermohaline circulation). Ice-formation increases the salinity by producing brine (salty water), and *deep-convection* may take place as a result. Deep-convection brings surface water to the bottom of the ocean (deep water formation: the deep water flows southward and is part of the meridional overturning), and is a recycling mechanism for surface water. This process is known to take place only at a few locations: in the Greenland–Iceland–Norwegian Sea and the Labrador Sea, and around Antarctica. The sea-ice may also affect the poleward heat transport in the atmospheric as well as in the ocean, because the climate over ice-covered regions tends to be colder than over open sea. The presence of sea-ice therefore influences the equator-to-pole temperature difference, which is an important factor controlling the poleward heat transport.



(*a*)

Figure 5.15. Variations in the anomalous SST correlated with the sunspots: (a) shows the "loads" derived from a regression analysis, with greater absolute values indicating stronger sunspot signal; (b) (opposite) shows the reconstruction of the sunspot number using the weights in (a). The linear variance R^2 accounted for is 53% and the F-statistic is 12.69 on 20 and 221 degrees of freedom suggesting a p-value of 0.00. The SST data are from Reynolds and Smith (1994).

5.5.3.2 Sea-ice

Sea-ice is an effective insulator that decouples the frigid air from the milder ocean water. The ice also has lower heat capacity than the ocean, and does not modify the air as much as the oceans do. Thus, surface temperatures tend to fluctuate more (during winter) over ice-covered areas and have more of a continental character than over the open sea. The contrast between the open sea and ice-sheets near the ice edge can produce some interesting phenomena, such as (atmospheric) cold-outbreaks and eddy formation in the sea. Furthermore, the presence of ice restrains the (equilibrium) temperature in the near vicinity so that it does not greatly exceed the melting point.

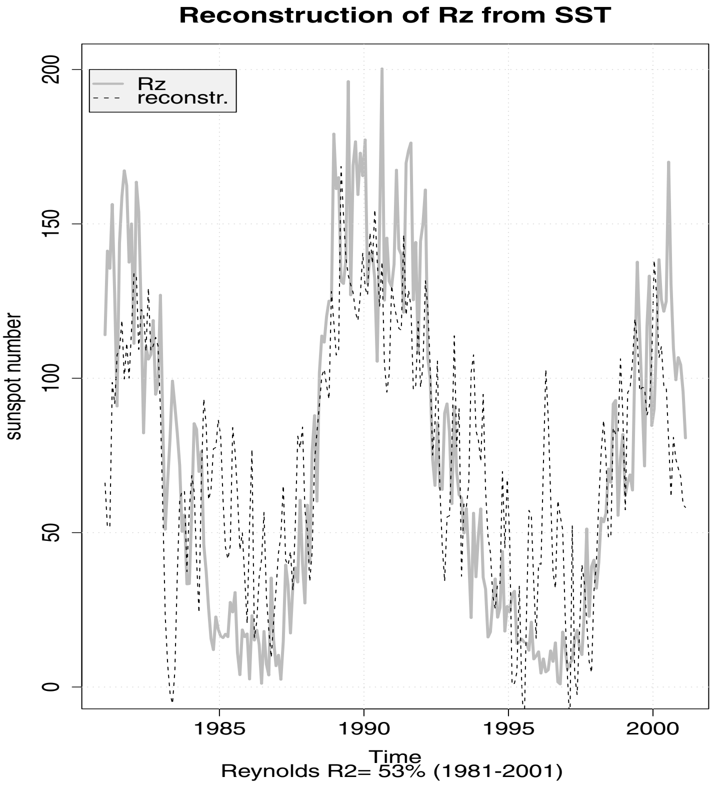


Figure 5.15. (b)

5.5.3.3 Snow

Snow-cover may affect the seasonal climates over the continents. As with ice, snow also has high albedo. However, the albedo may depend on the Sun's zenith angle over forested regions, as the trees tend to obstruct the Sun's view at low inclinations. Snow may affect temperatures by insulating the air from the ground, and may be a warming influence if the ground is freezing cold. The snow-cover is closely linked with the hydrological cycle, both because snow is a result of precipitation processes and because, during spring-time, most of the snow melts and produces high run-off rates.

5.5.3.4 Solar activity and sea-ice

Figure 5.16 shows the results of a regression analysis that identifies the geographical distribution of sea-ice variations related to the sunspot record. High absolute values indicate a strong solar signal in the sea-ice, and are seen near the sea-ice border. A

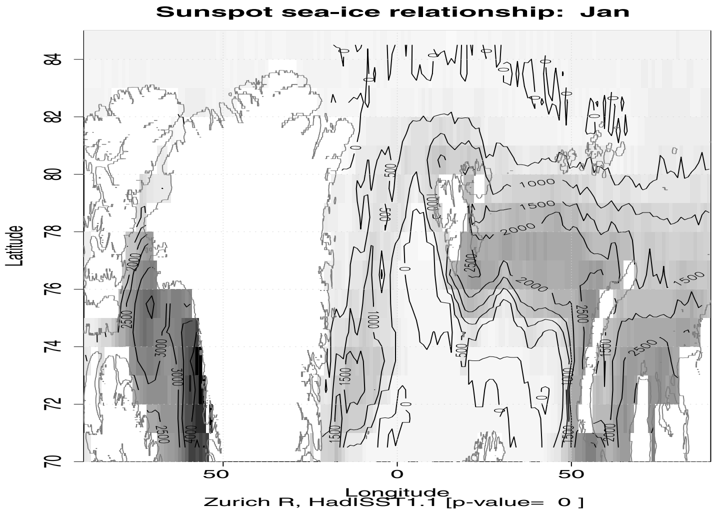

Sunspot sea-ice relationship: Jan
Latitude
70
72
74
76
78
80
82
84
50
0
50
Longitude
Zurich R, HadISST1.1 [p-value= 0]
1000
1500
2000
2500
3000
500
0

Figure 5.16. Variations in the sea-ice cover over Greenland and the Greenland–Iceland–Norwegian Sea correlated with the sunspots for January months. The figure shows the "loads" derived from a regression analysis, with greater absolute values indicating stronger sunspot signal in the sea-ice. The sea-ice data are from the HadISST1.1 product from the UK Meteorological Office and the sunspot record is from ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

"reversed" analysis, taking the time variations of the spatial patterns in Figure 5.16, can be used to make a "reconstruction" of the sunspot record. There are periods of bad sea-ice data, but despite this fact there is a close agreement between the sunspot reconstruction based on the sea-ice data and the actual record (Figure 5.17). This agreement gives a strong indication of a solar signal in the sea-ice data from the HadISST1.1 product from the UK Meteorological Office. Although there is empirical evidence suggesting that the sea-ice may play a role in the solar–terrestrial link, there is no obvious physical explanation. It is possible that the sea-ice is involved by acting as a feedback process (see Section 5.6.3).

There have been reported links between the NAO and sea-ice, and a strong lag-correlation is found when variations in the wind lead the variations in the sea-ice, suggesting that the winds drive changes to the sea-ice. It is also possible that the sea-ice subsequently affects the atmospheric circulation[22] in a "chicken-and-egg" kind of situation. The sea-ice and the NAO are furthermore thought to affect the so-called thermohaline circulation, which is a current system driven by density (temperature and salinity) differences. The thermohaline circulation may play a role in the poleward heat transport which in turn may influence the sea-ice.

[22] For instance, in a nonlinear fashion or over a longer timescale.

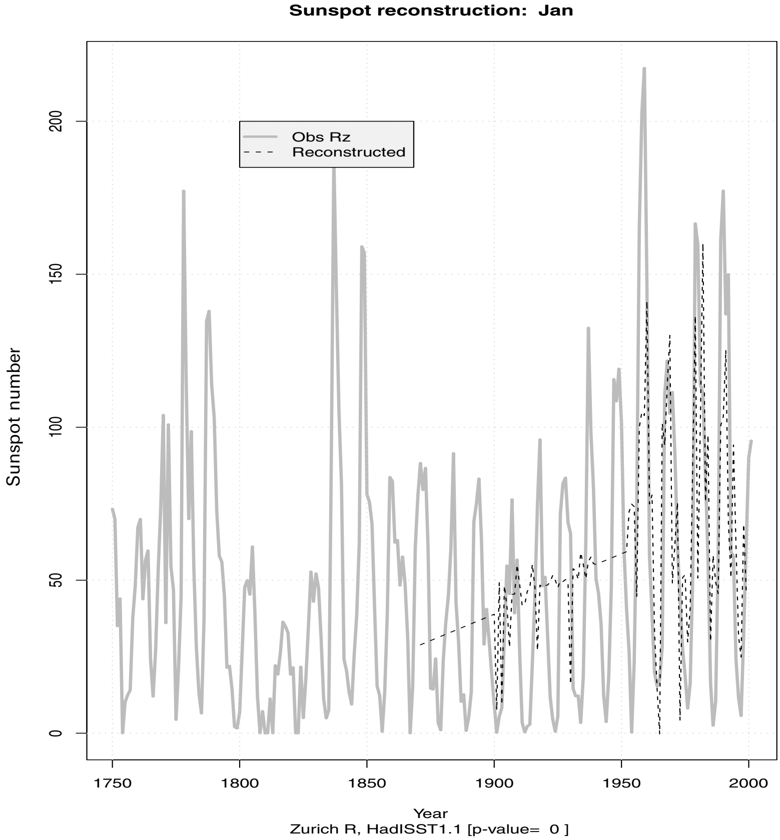


Figure 5.17. A reconstruction of the sunspot number based on the sea-ice cover patterns in Figure 5.16 for the January months. The close agreement between the reconstruction and the observed sunspot record suggests that there is a solar signal in sea-ice data. The sea-ice data are from the HadISST1.1 product from the UK Meteorological Office and the sunspot record is from ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

5.5.4 The biosphere

Trace gases such as carbon dioxide (CO_2) in the atmosphere play a subtle but important role for the climate. Some of these gases are transparent (e.g. CO_2, CH_4, and water vapour) to the short-wave solar irradiation, but opaque in the infrared spectrum. These may affect the energy balance for Earth's surface by trapping heat near the surface (Section 5.4.3). The effect of such "greenhouse gases" is similar to wrapping a blanket around the Earth, warming the surface and cooling the top of the atmosphere. The production of the naturally occurring CO_2 and free oxygen is primarily by living organisms, and there is a marked seasonality in the CO_2 concentrations. The concentrations are smaller during northern summer–autumn, because there is more land area in the northern hemisphere.

Photosynthesis associated with plants absorbs more CO_2 during the northern spring and summer which leads to minimum concentration in the autumn. A release of CO_2 during late autumn and winter explains maximum concentration in the spring (Peixoto and Oort, 1992, p. 435).

Plankton also plays an important role in producing cloud condensation nuclei. There have been reported changes to the CO_2 during strong El Niño events, which indicate that plankton also influences the atmospheric carbon dioxide concentration. Chavez *et al.* (1999) found a biological and chemical response, such as a decrease in the atmospheric CO_2 concentrations due to the 1997–98 El Niño, accompanied by impoverished plant nutrients in the surface layer and low chlorophyll concentrations. There has also been speculation about whether plankton has played a key role in terminating warm periods in the past (Schmitz, 2000). Increased biological activity in the sea leads to increased sequestering of carbon in the oceans.

Vegetation affects the albedo and hence may play a role in the energy balance on Earth. Plants are also sources of moisture through evapotranspiration. The climates in rainforests such as in the Amazon are known to be more humid than in less lush locations. Hypotheses have been proposed about the Sahara once having been wet and lush, due to the discovery of fossils with imprints of fish and other aqua-creatures (Simpson, 1999). The hypothesis states that the Sahara gradually became more arid, and the disappearance of the vegetation produced a positive feedback mechanism where the soil was less able to hold moisture. The clouds are a central part of the hydrological cycle and the precipitation plays a crucial role in the biosphere.

Vegetation may be extremely sensitive to the climate in several ways. The fact that trees grow more slowly in cooler climates can be inferred from tree ring structures, and thus, tree rings can hold information about past climates. Different plant and animal species thrive under different conditions, and variations in the local climate may result in variation in the vegetation. Moreover, changes in the plant and animal species may be one sign of climatic changes, but it is important to keep in mind that other factors, such as pests, availability of nutrients and competition, also may be important for a particular plant or animal population used for past climate reconstruction.

5.6 FEEDBACK MECHANISMS

The climate system consists of many processes and variables, and changes in one of them may affect other parts of the system. In other words, changes in one place may influence the system elsewhere. In such a complex system everything may affect everything, but to different degrees. Some of these mutual interactions can be described as feedback processes. A feedback mechanism may be a circular process where the state of one part $\mathcal{A}$ of the system affects another with a state $\mathcal{B}$. A change in $\mathcal{B}$ may subsequently influence the first part ($\mathcal{A}$). Such a feedback mechanism may be expressed as $\mathcal{A} \rightarrow \mathcal{B} \rightarrow \mathcal{A} \rightarrow \mathcal{B} \ldots$, thus describing a circular dependency (here referred to as type I feedback process).

Another definition of a feedback mechanism is a process that changes the sensitivity of the response to a change in the forcing F (here referred to as type II feedback process). For instance, if the change in the state of a process, denoted $\Delta\mathcal{A}$, depends on both the state of the object itself $\mathcal{A}$ and the forcing, then the feedback process may be illustrated by $\Delta\mathcal{A} \leftarrow f(\mathcal{A}, F)$, f being an unknown function.

The interdependency results in a regulation which can lead to an amplification or a smaller response, depending on the nature of the processes involved. Such amplifications may play a crucial role for solar–terrestrial relationships. The strength of the response may be described by a parameter λ also referred to as the *climate sensitivity*. Processes with a weak response tend to involve a negative feedback mechanism while a positive feedback causes amplification or oscillations. A number of different types of feedback mechanisms are outlined below.

5.6.1 Stefan–Boltzmann feedback

One example of a feedback mechanism (type II) involves a body's temperature and its black body radiation. The rate of energy loss due to black body radiation increases with the temperature, and the balance between long-wave radiation and solar irradiance is influenced by the emission temperature T_e. In other words, the temperature does not increase proportionally with an increase in the irradiance. If all other factors are neglected, then a simplified expression for the outgoing long-wave radiation can be expressed as a function of the emission temperature T_e: $F^{\uparrow}(\infty) = \sigma T_e^4$. The climate sensitivity is expressed in terms of one parameter $\lambda = (dF^{\uparrow}(\infty)/dT_e)^{-1}$.

Hartmann (1994) gives an expression for the climate sensitivity for the Stefan–Boltzmann feedback:

$$\lambda|_{\mathrm{BB}} = \left(\frac{\partial(\sigma T_e^4)}{\partial T_e}\right)^{-1} = (4\sigma T_e^3)^{-1} = 0.26\,\mathrm{K\,m^2/W} \qquad (5.13)$$

Assuming a radiative balance $S(1-A)/4 = F^{\uparrow}(\infty)$, $A = 0.3$, and $\Delta S = 5.7\,\mathrm{W/m^2}$, net forcing is equivalent to $1\,\mathrm{W/m^2}$. A change in S of $22\,\mathrm{W/m^2}$ (1.6%) will in this simple model produce a global mean temperature change of $\Delta T_e = 1^{\circ}\mathrm{C}$. Estimates of past variations in the total solar irradiance suggest an amplitude between 0.2% and 0.6% (Section 8.7). Hence, such a negative feedback process cannot explain the global mean temperature variations (of the order 1°C) seen in the palaeoclimatic data. The effect of the Stefan–Boltzmann feedback is, according to the simple model outlined above, to suppress the climatic response to variations in the TSI, and hence the direct influence of the solar activity.

5.6.2 Water vapour feedback

The most important feedback process in our climate system involves water vapour. When more energy is added to the climate system, part of the increase is used to enhance the evaporation of water, and the associated phase change keeps the temperature constant. However, the feedback mechanism also involves a change in the atmospheric water vapour concentration which subsequently affects the strength of the greenhouse effect. The water vapour concentration is sensitive to the surface temperature which again is affected by the greenhouse effect and the contribution from the atmospheric water vapour (type I).

In order to outline this process we need to consider the Clausius–Clapeyron equation, describing how the saturation vapour pressure e_s (a measure of the air's water content) varies with temperature:

$$\frac{de_s}{dT} = \frac{L}{T(\alpha_v - \alpha_l)}$$

The parameters α_v and α_l represent the specific volume of gas phase of fluid. The humidity, the water vapour pressure and temperature are related according to:

$$\frac{dq^*}{q^*} = \frac{de_s}{e_s} = \left(\frac{L}{R_v T}\right)\frac{dT}{T} \approx 20\frac{dT}{T}$$

A 1% (3 K) change in the temperature produces a 20% change in q^*. The relative humidity is assumed to be constant: RH$=$ $q/q^* \approx$ const. The terrestrial outgoing long-wave radiation (OLR) increases linearly with T_s because of the water vapour assuming radiative-convective equilibrium profiles with constant relative humidity (Hartmann, 1994)

$$\lambda|_{\text{FRH}} = \left(\left.\frac{dF^{\uparrow}(\infty)}{dT_s}\right|_{\text{FRH}}\right)^{-1} \approx 0.5\,\text{K m}^2/\text{W}$$

The presence of water vapour gives twice the climate sensitivity compared to the Stefan–Boltzmann feedback mechanism. This means that an increase in the surface temperature is reinforced by the increase in the atmospheric water vapour concentrations and consequently the enhanced greenhouse effect. This mechanism may therefore act to amplify variations in the TSI associated with the solar cycle assuming everything else is constant. The real climate system is of course much more complex than this, and more complicated interactions may play a role in the solar–terrestrial link.

5.6.3 Ice- and snow-albedo feedback

Ice and snow can only exist in cold regions as temperatures above freezing point favour melting and sublimation (direct transition from solid to gas phase such as from snow to vapour). Ice also reflects a large fraction of the sunlight and is therefore associated with a high albedo. The albedo of open sea and sea-ice differ by approximately 0.50: $A_{\mathrm{oce}} \approx 0.10$, whereas $A_{\mathrm{ice}} \approx 0.60$ (Hartmann, 1994). The more light reflected back to space, the less absorbed by the climate system, and the less available for warming the Earth. Cooler climates favour more ice and snow, which again increases the albedo. Thus the mechanism may be illustrated by the following sequence:

$$\text{more ice} \rightarrow \text{higher albedo} \rightarrow \text{cooler} \rightarrow \text{more ice} \ldots$$

The converse is also true:

$$\text{less ice} \rightarrow \text{lower albedo} \rightarrow \text{warmer} \rightarrow \text{less ice} \ldots$$

In other words, the effect of a change in the ice extent is reinforcing or has a positive feedback (type I). If this mechanism is unstable and no other factors play a role, one of two (albeit unrealistic) states are possible for the steady state: an ice-free world or complete ice cover.

There are annual variations in the global albedo because of seasonal changes in the snow-cover, but the effect of the snow albedo also depends on the vegetation (less important in forested areas). Old (dirty) snow reflects less light than fresh snow. Furthermore, the solar elevation is important at higher latitudes because there is little sunlight to reflect in the polar nights when the Sun is below the horizon.

Glacial periods imply more ice and higher albedo than at present, and more of the solar energy is reflected back to space without involving the climate system. The high albedo favours further cooling and in turn more ice, and glaciation may represent a strong positive feedback mechanism. However, the feedback mechanism is thought to be stable at present, but there is hypothetically a critical point beyond which the glaciation may take off and produce an ice-covered Earth (Hoffman and Schrag, 2000).

The solar energy absorbed by the climate system can be expressed (Hartmann, 1994) as: $Q_{\mathrm{ABS}}(x, T) = F^{\uparrow}(x, T) = \Delta F_{ao}(x, T)$ (here $x = \sin(\Phi)$ and s is a weighting function, defined by dividing the annual-mean insolation at each latitude by the global-average insolation[a]), and if an energy balance model is assumed, then:

$$Q_{\mathrm{ABS}}(x, T) = S[1 - A(x, T_s)]/4 \times s(x)$$

Both ice-free conditions and complete ice-cover are possible. The polar ice extent is sensitive to changes in S, and a change of 1.6% in the total solar irradiance may be sufficient to produce a complete glaciation according to this simple model. The critical point is reached when the ice extent reaches around 45° of latitude, and there have been speculations about whether the Earth has been close to this point during the last glacial maximum.

[a] The function $s(x)$ can be estimated approximately from $s(x) \approx 1.0000 - 0.2385 (3x^2 - 1)$ (Hartmann, 1994, p. 237).

A significant part of the meridional albedo profile is related to the mean cloud cover, and since the high latitude regions are more cloudy, the presence of ice has a smaller effect on the albedo near the poles than the simple models imply. Thus, the climate sensitivity associated with the ice feedback is reduced because of clouds. Another important aspect of the ice cover is its insulating property and its influence on the coupling between the oceans and the atmosphere as well as the interaction between the land surface and the air. The presence of ice inhibits heat and water fluxes, but also modifies mass and momentum exchange.

Bertrand and van Ypersele (1999) have suggested that the snow- and ice-albedo mechanism may be responsible for a link between solar activity and climate variability on inter-decadal (10–100 year) timescales. The empirical evidence of a solar signal in the sea-ice extent (Figure 5.17) supports the notion of the sea-ice reinforcing variations in the TSI in association with the solar cycle (Figure 4.13).

5.6.4 Cloud feedback

While clouds normally are white and reflect a large fraction of the incident sunlight, they may also warm the surface due to trapping outgoing long-wave radiation. The net effect of clouds tends to depend on the cloud characteristics: high clouds tend to favour warming whereas low clouds tend to cool the surface.

Clouds are themselves affected by the temperature and the circulation patterns, as the water vapour saturation pressure is strongly influenced by the temperature, and the cloud drop population is determined by the available water. Winds affect the transport of moisture but are also related to convergence and divergence (convection). The influence of the clouds on the climate is highly uncertain, but it is believed that it depends on the cloud type and whether they are high or low clouds (Section 7.10.4).

There have been speculations as to whether a reduction in the low clouds can explain the recent warming trend through the influence of galactic cosmic rays, and a reduced albedo (Section 7.10.5.1), but the more prominent night-time warming is difficult to explain with this mechanism. High clouds are believed to have a warming effect on the surface, but there is no well-known relationship between the TSI or solar activity and the high clouds.

5.6.5 Biochemical feedback

Biological activity is often dependent on a certain temperature range, and changes in the temperature may affect the various species. High activity can remove carbon from the atmosphere and hence reduce the (natural) greenhouse effect, but the production of methane and other chemical compounds is also influenced by living organisms.

One view, proposed by James Lovelock, is that the Earth, the biosphere and the entire climate system behaves like one living organism, a hypothesis known as *Gaia*. The idea is that the Earth-system constantly tries to produce optimal conditions for living mechanisms and is usually considered to be more of a hypothetical case rather than an actual mechanism.

A simplified model of the Earth-system "Daisy-world" consists of black and white daisies. The area covered by white daisies can according to a simple model be described by the following equation: $dA_w/dt = A_w(\beta x - \chi)$. Black daisies: $dA_b/dt = A_b(\beta x - \chi)$. The parameter $\beta = f(T)$, and the temperature is influenced by radiative equilibrium which is affected by the planetary albedo: $T = \eta(\alpha_p - \alpha_i) + T_e^4$. The white daisies reflect more light than the black daisies, but the white flowers can survive in a warmer climate than the black daisies. The albedo is related to the area covered by the flowers: $\alpha_p = A_g\alpha_g + A_w\alpha_w + A_b\alpha_b$.

The Daisy-world describes a stable system where the temperature and the flower populations are regulated though negative feedback. Hence, this mechanism may dampen the response of variations in the TSI if everything else is constant. However, there may be other bio-related mechanisms involving the interaction between forests and the hydrological cycle and cryosphere, as well as carbon dioxide feedbacks. The latter may have a similar effect to the water vapour feedback mechanism, i.e. a positive feedback.

5.7 THE OTHER PLANETS IN OUR SOLAR SYSTEM

5.7.1 The signature of solar variability from other planets

We can possibly learn more about our own climate by looking to the atmosphere on other planets. The same physical laws apply for the other planets and their climates, but the physical and chemical processes take place under different conditions. Although the solar forcing is weaker for the planets further away from the Sun, the variation in the insolation may be assumed to be coherent at the various planets. The most important differences for the different planets are that (a) they have different topography, surface roughness, and geographical features such as mountain ranges and oceans; (b) the atmospheres of the different planets have different chemical composition, and therefore different chemical and physical processes may be important; (c) the magnetic field differs among the planets; (d) the effect of gravity may not be the same because of different planetary mass.

Variations in Venus's brightness temperature should in principle be coherent with Earth's temperatures if they both are driven by changes in the solar energy output. A strong coherent signal may furthermore be indicative of common feedback mechanisms on Earth and Venus responsible for amplifying the effect. But it is also possible that a feedback mechanism only takes place on one planet. Hence, one question is: does the surface brightness temperature of other planets vary, and is it coherent with the temperature fluctuations on Earth? For such studies, one needs long, high-quality time-series to reduce the risk of coincidence.[23]

The origin of the planetary atmospheres is an interesting question for the understanding of the solar–terrestrial relationship. Interstellar gas, called the *primordial nebula*, is thought to be the material from which the Sun and the solar system is made. A long time ago, this gas rotated and contracted so that the Sun and the planets became the product of the condensing gas.

The general atmospheric compositions are different for the inner and the outer planets. The inner planets' atmospheres mainly consist of CO_2, N_2, H_2O and O_2 with traces of CO and O_3, whereas the outer giants are veiled in atmospheres of H_2 and He, with traces of CH_4 and NH_4. Not all the nebula is believed to have condensed, and the atmospheres of the outer planets is believed to have been formed by accretion of the surrounding nebula. The atmosphere of the inner planets, on the other hand, is thought to be due to volcanism and meteorite impacts. More details about the atmospheres of the other planets are given by Encrenaz (1997).

5.7.1.1 The presence of water

The distance to the Sun and the mass of the planets may be important for the differences in the ratio between deuterium and ordinary hydrogen (D/H). Venus has 120 times and Mars 6 times the ratio seen on Earth. High levels of deuterium are taken as an indication of out-gassed H_2O. One hypothesis is that the water has escaped from the planets closest to the Sun. The presence of water is of vital importance for shaping the climatic conditions and is the main ingredient in the oceans, clouds, ice sheets, hydrological cycle and life forms. The reader is referred to Encrenaz (1997) for a further discussion on water in the solar system.

5.7.1.2 Mercury

Of all the planets in our solar system, Mercury is the only one that does not have an atmosphere. This may be because the planet's mass is too small to keep the atmosphere in place. Mercury is also the planet closest to the Sun, and is subject to very intense solar winds that may blow away its atmosphere. The solar wind may blast away the atmosphere from the planets if there are no mechanisms to keep them in

[23] It is important that the observations are not made through Earth's atmosphere in order to eliminate "contamination" from our own climate.

place. One important mechanism may be a planetary[24] magnetic field that shields the planets from the solar wind. This implies that the planets with atmospheres must have interiors that can sustain dynamo processes. Sufficient planetary mass may furthermore be required to keep the atmosphere in place through gravitational forces. The Moon has a small mass and no magnetic field, and the Moon has virtually no atmosphere. Mars has a thin atmosphere, a slightly smaller mass than the Earth, but a very weak planetary magnetic field. The lightest atoms (H) escape first. The escape velocity depends on the planetary mass (gravitational field). Morton (1999) gives a further discussion on Mercury.

5.7.1.3 Venus

Venus has a surface pressure of about 90 atm and a dense atmosphere mainly consisting of CO_2. The planet has probably been subject to a runaway greenhouse effect resulting in a surface temperature $T_s > 740\,\mathrm{K}$. Strong winds have been observed near the cloud tops, but the surface winds are nevertheless weak. The diameter of Venus is 95% of that of the Earth and it has a mass which is 80% of the Earth's mass.

A day on Venus corresponds to 243 Earth days, and it is speculated that the slow rotation is the reason why Venus has no magnetic field. Venus is similar to Earth in other respects: similar densities and chemical composition, both have few craters which indicate that they are relatively young planets, and their masses and radii are similar.

5.7.1.4 Mars

Mars's axis of spin has a 25° inclination to the axis of its orbit. 25% of the atmospheric mass is shifted seasonally between the two hemispheres, which again implies a substantial transfer of heat and momentum between the two poles. Violent winds and dust storms are commonplace on Mars. The atmospheric dust is opaque and heats the atmosphere as well as acting as condensation nuclei for H_2O and CO_2. Mars has a thin atmosphere: the surface pressure is about 7 hPa. The composition of the atmosphere is 95% CO_2, 3% N, and 2% Ar, and the greenhouse gas warming is approximately 5 K.

$$S_{\text{Mars}} = 1370\,\mathrm{W\,m^{-2}} \times \frac{(1.50)^2}{(2.27)^2} = 598\,\mathrm{W\,m^{-2}}.$$

The mean temperature on Mars is 218 K, but its emission temperature is 216 K (the mean distance to the Sun is 2.27×10^{11} m). The large eccentricity ($\epsilon = 0.093$) causes large seasonal temperature fluctuations: $\Delta T = 30\,\mathrm{K}$ between aphelion and perihelion. The winter pole temperature is 140 K, whereas summer day temperature is around 300 K.

[24] Venus has no magnetic field, however.

Ice caps are seen on the poles of Mars, and these are believed to be mainly dry ice in alternating layers. The northern summer cap sublimes completely, whereas the southern polar cap never completely disappears. Further details about Mars's atmosphere can be found in Encrenaz (1997) or Houghton (1991).

6

Solar activity and the stratosphere

6.1 SYNOPSIS

The stratosphere may play a key role in the connection between solar activity and the terrestrial climate. The stratosphere is a high-altitude region where trace gases such as ozone absorb shortwave radiation including solar UV. Because of the absorption the stratosphere contains a region where the temperature increases with height. This region is also hydrostatically stable, acting like a "lid" on top of the troposphere below (Figure 5.9). The meridional temperature profile affects the circulation through the thermal wind balance. The stratosphere furthermore affects the planetary wave propagation as the waves' refractive index is dependent on the flow structure and the vertical density gradients.

The solar–terrestrial links based on the action of solar UV and ozone are primarily based on physical considerations. Observations of the stratosphere were sparse before the International Geophysical Year (IGY) in 1957–1958, and space-borne UV measurements did not start until the late 1970s. Although the observations are of high quality for the study of short-term variations, there are no high-quality records available for the study of long-term variations in solar UV.

In addition to radiative, thermodynamical, and chemical processes, the stratosphere also exhibits interesting dynamical features. However, as the chemical processes are the most likely candidates for participating in a mechanism linking solar activity to the terrestrial climate, we will focus on solar UV and chemical composition and then the thermal and dynamical response to the changes in the chemical processes.

Our understanding of the interaction between solar UV and ozone is based on both theory and measurements. The mechanism where stratospheric ozone plays a central role in the solar–terrestrial link is mainly inferred from model studies, although supported by recent observations. The theories describing the relationship between solar activity, (i) stratospheric ozone, (ii) the Quasi-Biennial Oscillation, and (iii) the Arctic Oscillation will be discussed below. Although these theories may seem to be the most promising explanations for a solar–terrestrial connection it is necessary to emphasise the need for long records of high-quality observations of the stratosphere for validation of these theories. The observational record is short and unlikely to provide an adequate representation of high-latitude processes when several factors affect the stratospheric ozone concentrations (Labitzke *et al.*, 2002). Satellite measurements, such as the MSU data indicate a general cooling trend in the stratosphere (Mears *et al.*, 2003) which may partly be a result of ozone depletion (near the poles and in early spring) and partly due to heat re-distribution as a result of an enhanced greenhouse effect. In general, the solar signal gets stronger with altitude from the lower stratosphere up to about 40 km which corresponds to about 3 hPa (Labitzke *et al.*, 2002). In order to get more accurate estimates of these trends, it is important to account for both the solar cycle signal as well as volcanoes. Conversely, to get a good estimate of the solar response, it is also necessary to take into account the trends and volcanoes (Labitzke *et al.*, 2002).

6.2 SOLAR ACTIVITY AND UV EMISSION

The solar cycle is accompanied by considerable changes in solar UV emission. Short-wave solar radiative energy with wavelength less than 300 nm varies with the 11-year Schwabe cycle with amplitudes that are several orders of magnitude greater than for visible light (Lean and Rind 1998). However, the fraction of the total energy represented by this part of the spectrum is only about 1% (15.5 $\mathrm{W\,m^{-2}}$ of 1366 ± 3 $\mathrm{W\,m^{-2}}$). Almost half (48%) of the solar electromagnetic energy lies within the 400–800-nm wavelength band, i.e. in the visible part of the spectrum.

The fractional change ($\Delta I/I$) in the radiation during a typical solar cycle with wavelength less than 100 nm is 10–20%, although the corresponding spectral irradiance change ($\Delta I = I_{\max} - I_{\min}$) is small (less than 0.1 $\mathrm{W\,m^{-2}}$). The change in the spectral irradiance associated with the solar cycle peaks at $\approx$400 nm (according to a spectral analysis by J. Lean). The light with wavelength less than 400 nm contributes with less than 9% of the total solar irradiance energy but 32% of the variations over a solar cycle.

The strong fractional variation in the short wavelength irradiation has implications for the chemical processes in the stratosphere where the UV light is absorbed. In particular, intensified UV radiation affects the ozone photo-chemistry and the local heating, hence modifying the ozone concentration. The ozone abundances regulate local heating rates and the irradiance flux at lower altitudes. Such changes may affect the climate indirectly by altering the radiative properties of the upper atmosphere and hence temperature and density gradients as well as upper layer wind structure. The short wavelength energy (UV) is absorbed above the troposphere, mainly by O_2, N_2, O, and O_3, and very little of this energy reaches the surface.

Solar variations are prominent in the thermosphere with amplitudes of 50–60 K at 120-km altitude and exceeding 100 K by 400-km height. Labitzke *et al.* (2002)

found correlations of the order 0.5–0.6 between zonally mean and annual average temperature and the 10.7-cm solar radio flux in the vertical region between the 200-hPa and 20-hPa levels and for the period 1968–1999 (NCEP/NCAR re-analysis). The strongest signal was observed in upper zonal bands between 45°S–45°N and weak signals in the vicinity of the southern subpolar jet. One weakness of zonal mean analysis is that zonal bands representing the low latitudes tend to involve a larger number of independent observations whereas those from the polar regions tend to consist of only a few (one?) independent realization (Benestad, 2005b). Taking zonal means therefore means that low-latitude bands provide more resistant measures (less noisy) than the high-latitude ones (more noisy), and care should therefore be taken when comparing zonal means over a range of latitudes. It has been proposed that these changes may affect the lower stratosphere through nonlinear coupling. This mechanism is still not well-documented and more work is needed to clarify its importance.

Most of the solar XUV radiation ($\lambda \approx 1\,\text{nm} \ldots 110\,\text{nm}$, although there is as yet no clear definition of the XUV spectral region) from the solar chromosphere and corona is absorbed in the thermospheric–ionospheric region at altitudes from 90 km to 600 km. Most of the absorbed energy is converted into heat, and the thermospheric temperature is lowest at the turbopause near 90 km and highest at the exosphere at about 1000 km. At the present, there are no XUV records that cover a complete solar cycle.

6.3 THE ROLE OF STRATOSPHERIC OZONE

Several studies have investigated the changes in solar radiation at the Earth's surface associated with stratospheric ozone variations. The observed total column ozone is between 1% and 2% higher at solar maximum than at solar minimum (Harrison and Shine, 1999). Attempts to reproduce the global column ozone during the Maunder minimum have resulted in values about 3% lower than present conditions (Wuebbles *et al.*, 1998). The increase since then is thought to have a negative radiative forcing ($-0.13\,\text{W}\,\text{m}^{-2}$) because the estimated changes are most prominent in the upper stratosphere. However, this effect may be offset by an increase in the total solar irradiance over the same period ($+0.5\,\text{W}\,\text{m}^{-2}$).

6.3.1 Chemical reactions

Ozone (O_3) is an unstable form of oxygen and has the important property that it absorbs UV radiation. There is ozone in the stratosphere as well as in the troposphere. The latter is due primarily to pollution and is harmful to human health. The stratospheric ozone, on the other hand, is natural and essential for life on earth as it blocks out harmful UV light which otherwise would destroy organic molecules in living tissue (at the surface). The stratospheric ozone also absorbs long-wave radiation and makes a small contribution to the natural greenhouse effect. But it

Table 6.1. Typical values for the reaction rates needed for estimating ozone concentrations (equation (6.2)) (Houghton, 1991).

J_2	$10^{-12} \dots 10^{-9}\,\mathrm{s}^{-1}$	increasing with altitude
k_2	$10^{-33}\,\mathrm{cm}^6\,\mathrm{s}^{-1}$	
k_3	$10^{-15}\,\mathrm{cm}^6\,\mathrm{s}^{-1}$	
J_3	$5 \times 10^{-3} \dots 10^{-2}\,\mathrm{s}^{-1}$	increasing with altitude

is important to distinguish the stratospheric ozone layer from the enhanced greenhouse effect due to increased atmospheric CO_2 concentrations.

The absorption of UV light is closely related to the ozone concentration. A model of the relationship between UV light and ozone concentration can be described by the simple Chapman (1930) theory.

$$\begin{aligned} O_2 + h\nu &\Rightarrow O + O && (J_2) \\ 2O + M &\Rightarrow O_2 + M && (k_1) \\ O + O_2 &\Rightarrow O_3 + M && (k_2) \\ O + O_3 &\Rightarrow 2O_2 && (k_3) \\ O_3 + h\nu &\Rightarrow O_2 + O && (J_3) \end{aligned} \tag{6.1}$$

The reaction rates (Table 6.1) for the various stages are denoted by J_2, k_1, k_2, k_3, and J_3 respectively, and M represents a third body that is required for the energy and momentum conservation (catalyst). For an equilibrium concentration between O_2 (n_2) and O_3 (n_3), the following relation must be satisfied:

$$n_3 = n_2 \sqrt{\frac{J_2 k_2 n_M}{J_3 k_3}} \tag{6.2}$$

The production of ozone requires UV light. Therefore, an intensification of the UV radiation may lead to enhanced ozone production. The stratospheric ozone concentration at different levels may play different roles for the climate. The heating in the upper stratosphere is mainly due to UV radiation being absorbed by ozone. The increase in the solar UV at sunspot maximum boosts the photochemical production of ozone, and since both the UV flux and the ozone concentration are increased during solar maximum, the upper stratosphere experiences local heating. The local heating is responsible for a meridional temperature gradient. The polar regions receive no sunlight during the polar night (winter). So any variation in solar UV can have no direct effect on the polar stratosphere during winter. Haigh (2003) proposed that the stratospheric ozone concentrations may be influenced by the level of solar activity, however, it is not well established how the variations are affected by changes in solar activity. The problem is partly lack of reliable data. One proposed mechanism is an increase in the emission of thermal infrared (TIR) radiation into the troposphere, due to the absorption of shortwave radiation by ozone and heating of the stratosphere. Estimates of the magnitude of the response caused by a modulation of ozone due to solar activity was made by Haigh (1994) and Myhre *et al.* (1998), who proposed that ozone increases produced a lower solar radiative forcing during solar maximum than during solar minimum ($-0.1\,\mathrm{W\,m^{-2}}$ and $-0.02\,\mathrm{W\,m^{-2}}$ respectively). Wuebbles *et al.* (1998) inferred a radiative forcing of $-0.13\,\mathrm{W\,m^{-2}}$ due to ozone increases since the Maunder Minimum. Matthes *et al.* (2003) state that in addition to great uncertainties in the observational data, important forcings (inputs) used for model simulations are poorly known.

6.4 THE "OZONE HOLE"

In recent years, the stratospheric ozone concentrations over the poles have diminished due to the destructive effect of chlorofluorocarbons (CFCs). The CFCs are man-made and act as catalysts for ozone destruction, depleting the ozone concentrations when the temperature is low and solid surfaces are present. The conditions for ozone destruction are favourable near the poles during spring when the first sunlight hits the frigid air after the long polar nights and when stratospheric clouds are present so that the chemical reaction (ozone destruction) can take place on the surface of the cloud particles. Volcanoes inject particles into the stratosphere and may also accelerate the ozone depletion. The area of depleted stratospheric ozone is popularly referred to as the "ozone hole".

The "ozone hole" was discovered by the British Antarctic Survey from Halley Bay (Antarctica) in 1985. It is a classic example of how important observations may go undetected due to automatic data control. American satellites had measured the ozone concentrations before the British, but since the measured concentrations were way below the expected values they were discarded before being analysed and checked.

The creation of an "ozone hole" has similar implications for the climate as with solar-induced ozone changes. The depletion of the stratospheric ozone results in a cooling of the polar lower stratosphere and thus an increase in the meridional temperature gradient. Shindell *et al.* (2001a) propose that the effect of the ozone depletion in the northern hemisphere on the planetary wave propagation is qualitatively similar to those produced by an enhanced greenhouse effect, and that the AO is strengthened. However, the northern hemisphere ozone depletion occurs primarily between late February and April, and the effect is thus seasonally dependent.

6.4.1 The theory of a link between solar activity and stratospheric ozone

In addition to the direct negative radiative forcing due to an increase in stratospheric ozone (the so-called "TIR-mechanism"), there may also be a dynamical response to changes (Haigh, 2003).

6.4.1.1 Dynamic response to chemical reactions

Haigh (1994) demonstrated that using a two-dimensional radiative-chemical-transport model that changes in the stratospheric ozone, caused by variations in the solar UV, may result in a highly nonlinear response in the local radiative balance across the tropopause (outside the tropics). UV light with a wavelength less than 242 nm ($1\,\mathrm{nm} = 10^{-9}\,\mathrm{m}$) forces photo-dissociation of the oxygen molecule (equation (6.1)). The troposphere at high latitudes receives less solar UV radiation in the northern winter during intense sunspot activity because of the solar-induced changes in the stratospheric ozone. Labitzke *et al.* (2002) used two GCMs to study the response in stratospheric ozone to changes in solar activity and hence solar UV. They examined vertical profiles and proposed that there are two types of impacts: a direct solar heating influence and changes in photochemical reactions. Models may not always give a true representation of the real world, and the ones used by Labitzke *et al.* (2002) had some systematic biases as they underestimated the peak response in the upper stratosphere by a factor of two which also was simulated about 5 km too low, otherwise the models produced realistic vertical ozone-response profiles in the middle stratosphere. In addition to the usual solar cycle, a 27-day (one solar rotation) signal could be found in the stratospheric photolysis and heating rates. Labitzke *et al.* (2002) also proposed that the solar response in temperature over the subtropics for different seasons (with positive values for solar maxima and negative for solar minima) was consistent with changes in the Hadley cell intensity and an intensified downward motion in the upper troposphere subtropics during solar maxima. Hence, the solar cycle may influence the large-scale meridional atmospheric (diabatic) circulation in the low latitudes.

6.4.1.2 Ozone variations and circulation patterns

Haigh (1999) conducted a series of model studies with solar-induced changes in stratospheric ozone and found that the Hadley cell is weakened and broadened in January months during solar maximum, and that this broadening is accompanied by a poleward displacement of the sub-tropical jet streams. These shifts in the circulation pattern result in bands of warming and cooling. Thus, variation in the stratospheric ozone may be one mechanism which may amplify the variation in the total solar irradiance. The notion of a latitudinal broadening of the Hadley cell was supported by independent analysis of Crooks and Gray (2005) based on ECMWF re-analysis data. One sign of an extended Hadley cell is that the northern hemisphere sub-tropical jet (centered at 30°N–40°N) becomes slightly weaker during periods of high solar activity, whereas mid-latitude (40°N–70°N) winds are strengthened.

Shindell *et al.* (1999) have suggested that changes in the UV radiation may influence the stratospheric chemistry and hence lead to changes in the upper stratospheric ozone concentrations (1–2% for the solar cycle). Changes in the ozone affect UV absorption and induce temperature changes as well as changes to the atmospheric circulation. As the ozone absorbs UV light it traps energy, which is converted to heat. Differences in the absorption rates at various locations result in horizontal and vertical temperature gradients, which in turn affect the mean circulation according to the thermal-wind equation. A manifestation of these responses is seen when analysing the 30 hPa geopotential heights which show a response between the equator and 30°N.

Enhanced UV associated with solar maximum can boost the stratospheric ozone production, depending on the time of year. Matthes *et al.*, (2003) estimated the maximum increase would by 3% at the 5-hP level from mid- to high latitudes. The solar variation in UV and ozone concentrations lead to a modulation of the shortwave heating in the stratosphere. The response to solar variations may be difficult to analyze and interpret due to its non-linear nature, however, Matthes *et al.* (2003) conducted a study based on different atmospheric models which pointed to a modulation of the shortwave heating rate that is a direct result of solar UV and stratospheric ozone changes. A direct response includes temperature increases as a result of an absorption of solar UV, whereas an indirect response is the change in temperature as a result of a change in circulation. In these numerical experiments, 11-year variations in the visible part of the spectrum (Chappuis band: 400–800 nm) were ignored. Their study was based on an idealised experiment and the model may have a bias that could affect the results, as the strength and structure of the heating range modulation appeared to depend on both the background ozone field and the model. Matthes *et al.* (2003) observed important model differences in the annual mean ozone at high latitudes, which could be caused by differences in the radiation schemes used in the models or different background ozone climatologies. The response in temperature and zonal mean zonal wind varied from model to model, however, contrasting zonal means at high latitudes can easily give the false impression of greater scatter than at lower latitudes (Benestad, 2005b). At low latitudes, however, the response was more consistent among the different models, with higher temperatures during solar maxima throughout the tropical, sub-tropical, and mid-latitude upper troposphere and stratosphere. There are also some suggestions that the evolution of the polar night jet oscillation (PJO) is affected by solar forcing in early winter (Kuroda and Kodera, 2002). The model studies of Matthes *et al.* (2003) indicated a stronger polar night jet during solar maxima by 4–6 m s^{-1}, but this may be smaller than observations and the model does not reproduce the downward propagation seen in the observations and confines it to too high latitudes without the observed equatorward tilt. Misrepresentation of this jet may create a cold bias in the high-latitude lower stratosphere. In the southern winter, the models indicated a pronounced wave-number-one (a wave for which the wave length equals the constant latitude circle length) response over Antarctica, but one model also yielded a wave-number-two (the wave length being half of the constant latitude circle length) response more consistent with the observations, albeit out of phase by 90 degrees. The largest response is expected in the northern summer stratosphere, but none of the models in Matthes *et al.*'s study was able to produce realistic magnitude of the response.

6.4.1.3 Ozone and planetary wave propagation

One consequence of photodissociation is a modification of the latitudinal temperature gradients, also leading to a modulation of the planetary wave activity and the stratospheric thermal wind structure. Planetary wave propagation is influenced by wind shear, and wave refraction is favoured by areas of increasing wind. The wave activity is most intense during the winter, and the planetary waves, according to Shindell *et al.* (2001a), propagate up from the surface. They are most pronounced in the northern hemisphere as they are excited by the interaction between topographic features and the atmosphere. The wave refraction diverts the waves towards the equator, which leads to an acceleration of the zonal wind. This wind anomaly extends further downward from the atmosphere to the surface over time as successive waves are affected by these induced winds. The equatorward refraction of the planetary waves produces a poleward transfer of angular momentum in the opposite sense to the wave energy propagation. The westerlies are enhanced as a consequence and the flow is diverted towards the pole by the surface friction, leading a cell of rising air in the polar region and a descent between 40°N and 50°N. The shift in atmospheric circulation leads to enhanced advection of warm maritime air from the oceans to the downstream continents. In other words, the solar forcing affects the surface winds and pressure according to this hypothesis. Shindell *et al.* (2001b) tried to reconstruct the global mean temperature difference between the present and the Maunder minimum according to the solar UV–ozone mechanism and estimated this to be ≈0.3–0.4°C. Regional temperature changes were found to be quite large. Foukal *et al.* (2004), however, has since claimed that there is a low correlation between 20th-century global temperature changes and a reconstruction of UV flux, and hence questioned whether solar UV effects can make a significant contribution to temperature variations at the surface.

The changes in the temperature and flow patterns, which are a result of changes in the UV absorption, affect the refractive indices of the planetary waves, so that these no longer can propagate into the stratosphere but are deflected equatorward. These waves are associated with both energy and angular momentum transport, and thus affect the energy distribution in the atmosphere. In contrast to the stratosphere, there is little change in the tropospheric meridional temperature gradient as a result of the ozone action. However, the troposphere is affected by the altered propagation of the planetary waves. The stratospheric winds control the planetary wave propagation, and tropospheric waves are less able to propagate into the middle atmosphere at northern mid-latitudes, and become trapped at lower latitudes. Thus, solar activity influences the troposphere indirectly through a redistribution of energy. The change in the circulation is also associated with changes in sea level pressure, and Shindell *et al.* (2001a) proposed that the sub-tropical Pacific and the Azores high-pressure systems may be strengthened as a result of strengthened winds. One consequence of the wave deflection may be that the tropics become warmer. One criticism of the dynamical response mechanism of stratospheric ozone may be that observed effects in the summer hemisphere are unlikely due to planetary wave activity (Haigh, 2003).

Thuillier (2000) suggested that solar UV radiation affects temperature, photochemistry and the dynamics of the stratosphere. UV radiation produces atomic oxygen by photo-dissociation of molecular oxygen, a product necessary for ozone production. An increase in the ozone concentration is equivalent to inserting a heat source in the stratosphere and the stratospheric ozone concentrations vary with the solar cycle. The stratospheric ozone affects the biosphere, and solar modulation in the UV flux at Earth's surface may affect organisms such as plankton and the carbon dioxide cycle.

Atmospheric tides with a semi-diurnal period and an amplitude of 1.5 hPa can be seen in the sea level pressure measurements. In the late 1960s R. S. Lindzen showed that the daily variation in the solar absorption by ozone between 20 and 80 km height, as well as absorption by water vapour in the lower atmosphere, can explain the atmospheric tides (Lindzen, 1990). The absorption results in a heating and expansion of the air and hence influences the sea level pressure.

6.5 THE THEORY OF A LINK BETWEEN THE QBO AND SOLAR ACTIVITY

6.5.1 Introduction

The Quasi-Biennial Oscillation (QBO) in the equatorial stratospheric zonal wind is believed to be driven by wave forcing, originating in the troposphere, fed by gravity waves from Earth's surface. In other words, the QBO is a manifestation of an interaction between the stratosphere and the troposphere. The QBO is a reversal of the winds that takes place about every other year, with a preferred periodicity of $\approx$29 months.

6.5.2 Sunspots and the QBO

The studies of Labitzke (1987) and Labitzke and van Loon (1988) can be considered as pioneering work on the link between the sunspots and the QBO, and research on this topic is still continuing today. Balachandran *et al.* (1999) assumed that solar UV light with wavelengths shorter than 300 nm changes by $\pm5\%$ over a solar cycle and used a GCM to investigate the effects of the changes in the solar irradiance on the lower stratosphere. They reported a clear increase in the sub-tropical geopotential height from solar minimum to maximum. A dipole pattern was apparent when the data were partitioned according to the QBO phase. They proposed that solar activity influences the stratosphere directly by altering the vertical temperature and zonal wind profiles. The stratospheric response furthermore affects the planetary wave propagation and perturbs the Eliassen–Palm flux divergence and hence the lower stratospheric and tropospheric circulation structures.

Salby and Callagan (2000) proposed that the QBO is related to the stratospheric temperature over the north pole (T_{NP}). They examined the equatorial wind at 45 hPa height and the polar temperature T_{NP} at 30 hPa and found a solar cycle modulation in the QBO. By dividing (stratifying) the temperature measurements according to whether the winds were westerly or easterly, they found for each subgroup a correlation with the sunspots, but not for the combined group. The correlation was in phase during westerly QBO phase and out of phase when the winds were from the east. The QBO period was reported to vary with the sunspots, with a shorter period during solar maxima and a longer period during high sunspot activity.

The physical explanation for the connection between sunspots and the QBO proposed by Salby and Callagan (2000) is that a convergence of polar motion, driven by wave absorption, causes a compressional warming in the (diabatically) descending air. The polar vortex is more disturbed at solar minima as a result of this and hence T_{NP} is affected. The QBO absorbs wave energy and acts as a regulating mechanism for the polar vortex. The meridional temperature profile is affected by changes in T_{NP}, and this profile is important for the thermal wind balance governing large-scale motion. The meridional temperature profile in the tropics may on the other hand be altered by changes in the photochemical processes and the stratospheric ozone concentrations. The hypothesis of Salby and Callagan (2000) is supported by evidence presented by Soukharev and Hood (2001), who applied cross-spectral analysis to study the coherence between the solar variation and terrestrial response. A more recent paper by Salby and Callaghan (2004) places more emphasis on the effect of planetary waves on the so-called residual circulation and a coupling between the polar and equatorial stratosphere. The residual circulation affects wintertime temperature through a downwelling and adiabatic heating near the poles. Salby and Callaghan (2004) also refer to the region where planetary waves become non-linear and produce mixing as the "critical region", or the "surf zone'. The effect of general mixing is to dampen waves and is a process that re-distributes energy as well as affecting the entropy (towards more disorder). The "critical line" is where the transition between linear and non-linear wave behaviour occurs, and the critical region is found poleward of the critical line in the winter hemisphere. It is proposed that the QBO can displace the critical line and thereby affect the residual circulation, but it also implies that ozone and the poleward transport of chemical compounds are strongly coherent with the anomalous forcing of the residual circulation.

The polar winter vortex is, according to Salby and Callaghan (2004), shaped by planetary waves and their interaction with the equatorial wind. In order to explain how solar activity affects the circulation, it is important to consider the QBO and the equatorial winds. Salby and Callaghan (2004) proposed that the QBO phase propagation, manifested as a downward migration of westerlies and easterlies, ceases near solar minimum. When this propagation stops, the duration of the phase aloft is prolonged and the QBO frequency is reduced. Near solar maximum, on the other hand, the stratospheric equatorial winds tend to change direction during the winter season. Hence, the solar activity affects the residual circulation through its effect on equatorial winds, the QBO, the critical region, and the dissipation of the planetary waves. They produced a statistical analysis based on running correlations in 3-year windows, however, such kinds of tests may not always yield objective and representative results (Katz, 1988). Their assessment of the significance was done through a Monte Carlo simulation, but it is not clear whether they chose an appropriate stochastic model (see Section 8.4.7). Crooks and Gray (2005) applied a regression analysis to the ERA-40 reanalysis from the European Centre for Medium-range Weather Forecasts (ECMWF) and noted that there is a possibility of a non-linear interaction between the solar variability and the QBO. Their regression results also suggested that volcanoes and ENSO could affect the QBO.

The QBO also alters the meridional gradient in the zonal wind in the lower stratosphere whereas UV variations alter the vertical gradient of the zonal wind. The refractive index of planetary waves is affected by gradients in both the zonal and vertical directions. Lean and Rind (1998) observed that changes in the refractive properties associated with the zonal winds, caused by solar activity, may result in induced warm and cool regions in the stratosphere, which subsequently alter the vertical stability of the stratosphere and troposphere. Results from model studies suggest that there is a link between this stability and generation of long planetary waves. The zonal band between 50°N–50°S of high correlations between 10.7-cm flux and 30-hPa heights found by Labitzke *et al.* (2002) is consistent with a 1–2 K warming in the lower stratosphere between solar maxima and minima (they analyzed the NCEP/NCAR re-analysis). Crooks and Gray (2005) applied a multiple regression analysis to the ECMWF re-analysis (ERA-40) and also found 1 K higher stratospheric temperatures during solar maxima compared with solar minima. They also found a zonal wind response consistent with these temperature anomalies and the thermal wind balance. The warming during solar maxima may have further implications for the QBO because of changes in the meridional temperature gradients that will affect the thermal wind equations.

6.6 THE THEORY OF A LINK BETWEEN THE AO AND SOLAR ACTIVITY

6.6.1 Introduction

Thompson and Wallace (1998) found zonally symmetric oscillations where the geopotential height anomalies at the poles are anti-correlated with sub-polar anomalies. They coined the term the *Arctic Oscillation* (AO), but the term *annular mode* is also used. It is hypothesised that the AO involves a coupling between the stratosphere and the troposphere. The AO is related to the NAO, and the NAO is often considered as a more regional feature of the AO. The AO is often defined by taking the first hemispheric empirical orthogonal function (EOF) of the SLP. An EOF analysis is similar to a principal component analysis, and is mathematically identical to the eigenvectors of the data in terms of their variance. Doubts have been expressed regarding the existence of the AO, based on the arguments that the patterns may be a mathematical artefact of the EOF analysis (Ambaum *et al.*, 2001).

6.6.2 A connection between solar activity and the AO

The northern lights have been explained in terms of solar activity and its effect on the Earth's magnetosphere, and there is little doubt that variations in solar activity have a profound influence on physical processes taking place high up in the atmosphere. At ground level, on the other hand, any direct influence from the sunspots is less evident. It is plausible that solar activity may affect processes in the upper atmosphere such as the chemical composition through the variations in the UV radiation. Shindell *et al.* (1999) proposed that solar-induced variations in the stratospheric ozone affect the propagation of planetary waves and thereby the distribution of heat. This mechanism may possibly be connected with the AO, for instance where the zonal wind anomalies are associated with the sub-polar maxima. Shindell *et al.* (2001a) used a climate model to study the effect of the solar cycle on climate and found that the solar cycle modulates the AO.

6.7 CRITICISM OF SOLAR–STRATOSPHERE HYPOTHESES

There is little doubt about a solar signature in the high-altitude atmosphere. The important question is: how does this information propagate down to the surface? The stratosphere mechanism assumes a stratosphere–troposphere coupling. Baldwin and Dunkerton (2001) have suggested that there is downward propagation of atmospheric disturbances from the stratosphere to the troposphere, and Shindell *et al.* (1999) suggest that the stratosphere may affect the tropospheric circulation by altering the refractive index for the planetary waves.

The fact that the solar UV radiation modulates the stratospheric ozone concentration means that this UV mechanism is only active on Earth's day-side. However, by influencing the planetary wave dynamics, the UV induced changes may also influence the night-time surface temperatures. But, to account for the recent long-term global surface warming, these hypotheses must be able to explain the diminishing diurnal cycle (IPCC, 2001) and the stronger night-time warming trends. The solar UV–ozone mechanism can perhaps explain some of the local variations observed over the recent decades and the strengthening of the NAO/AO. The hypotheses concerning the QBO are qualitatively similar to the hypothesis proposed by Shindell *et al.* (2001a), even though they may differ in detail. A further aspect is the fact that the stratosphere has cooled globally in recent decades which does not fit the concept of increased solar activity because stronger solar activity results in high solar UV emissions and stronger absorption in the lower stratosphere (Crooks and Gray, 2005). Man-made chemicals such as CFCs may have affected the temperatures to some degree (e.g., through depletion of the ozone layer particularly in the polar regions). The solar response in stratospheric temperature is most prominent over the tropics at an altitude of ∼40 km, with positive anomalies during solar maxima (Crooks and Gray, 2005).

The solar UV–ozone hypotheses may suffer from the lack of high-quality, long records of data for a thorough validation of the results. Extensive observation of the stratosphere started after the International Geophysical Year (IGY: 1957–1958). However, the existing decades of data and the observed response to the Pinatubo eruption support the model studies.

Table 6.2. Overview of the most recent major volcanic eruptions.

1815 April	Tambora	Indonesia 8°S
1883 August	Krakatoa	6°N
1902 May	Mt Pelee	Martinique
1902 October	Santa Maria	Guatemala
1903 February–March	Colima	Southern Mexico
1912 June	Katmai	Alaska
1963 March–May	Agung	8°S
1980 May	St Helens	Oregon 46°N
1982 March-April	El Chichón	Mexico 17.3°N
1991 June	Pinatubo	Philippines 15.1°N

6.8 VOLCANOES

Langley (1904) started bolometric measurements of solar irradiance in 1904 and found a 10% reduction in the solar radiation from the end of March 1903 to the end of the same year. This reduction also coincided with a drop in the mid-latitude northern hemisphere temperature. The subsequent warming was associated with the increased transparency of the atmosphere, and it is plausible that the drop also was a result of more dust in the atmosphere. The drop in temperature coincided with the eruption of Colima (Table 6.2). These results were among the first empirical evidence suggesting a direct relationship between the surface temperature and solar radiation.

Abbot and Fowle (1908) investigated the relationship between volcanoes and surface air temperature from 47 stations around the world and found that solar irradiance is diminished by the masses of volcanic dust injected into the atmosphere during an eruption. The drop in temperature is caused by small dust particles injected into the atmosphere which reflect and scatter the light from the Sun. There were clear minima during 1884–1885, 1890–1891, and 1903 in the pyroheliometric curve. Abbot and Fowle allegedly reproduced much of the temperature evolution over the 1880–1909 period by combining the volcanic effect with an inverted sunspot curve (Helland-Hansen and Nansen, 1920, p. 159).

Arctowski (1915) claimed that volcanic dust in general does not affect the temperature on Earth, except for exceptional cases such as the Krakatoa eruption. A study of the variations in the total solar irradiance ("solar constant") and the temperature from Arequipa in Peru (1905–1906) led Arctowski to the conclusion that variations in the solar irradiance brought about variations in the temperature (according to Helland-Hansen and Nansen (1920, p. 160): $0.015\,\mathrm{cal\,cm^{-2}\,min^{-1}} \rightarrow 1°\mathrm{F}$; the equivalent in SI units[1] is: $10\,\mathrm{J\,m^{-2}\,s^{-1}} \rightarrow 0.56°\mathrm{C}$). He did not consider the possibility of variations in the atmospheric transparency contaminating the analysis. A recent paper by Orlove *et al.* (2000) suggests that ENSO affects the

[1] 1 calorie = 4.18 J. The conversion between Fahrenheit (T_F) and Celsius temperature (T_C) scales is: $T_C = 5/9 \times (T_F - 32°\mathrm{F})$.

visibility of stars, which may have implications for ground-based (TSI measurements before the 1970s) measurements of total solar irradiance.

6.7.1.1 How volcanoes affect the climate

Carbon and sulphur injected into the atmosphere by volcanoes affect the atmospheric transparency and hence the radiative energy budget. The closer to the equator and the more powerful the eruptions, the stronger the effect on the atmosphere. Gas and dust have both a warming effect in the troposphere because of a weak greenhouse effect caused by some aerosols and a cooling effect because of increased albedo. Small particles scatter light but do not interact much with the long-wave radiation, thus increasing the albedo but not causing an enhanced greenhouse effect. Particles with radius $r < 1\,\mu$m may be responsible for a 3–4 W/m^2 reduction in net irradiation at the surface. Larger particles, on the other hand, are more important for the greenhouse effect. Small particles stay in the atmosphere for some time while aerosols with 1 μm radius fall at an approximate rate of 2 km/month. The effect of an eruption tends to be important for the subsequent months but almost negligible after 1–2 years.

The eruption of Pinatubo in June 1991 produced a short-term global cooling, in accordance with the model predictions. This incident has been used to test the climate models, and the evaluation came out favourably for the models. A climate model study of the effects of volcanism was carried out by Shindell *et al.* (2004), who prescribed the models with volcanic aerosols and TSI reconstructions derived from solar activity proxies. The response in their study to volcanic eruptions was a global cooling of the surface temperature and warming of the lower stratosphere as a result of increased absorption and reflection of incoming shortwave radiation. The annual mean cooling for a Pinatubo-type eruption was -0.35°C due to a radiative forcing of $-0.47\,\mathrm{W\,m^{-2}}$. Despite the general cooling, Shindell *et al.* (2004) found a winter warming over extra-tropical continents in the northern hemisphere after eruptions because the volcanic forcing strengthens the positive phase of the AO. They proposed that the warming of the stratosphere is strongest in the low latitudes and therefore the meridional temperature gradient near the tropospause is reduced as a consequence of an eruption. The change in temperature gradient affects the thermal wind by increasing the westerly winds in the same region. The effect of an enhanced westerly stratospheric flow affects the planetary wave propagation in such a way that upward propagating planetary waves are refracted toward the equator. The planetary waves play an important role in the re-distribution of angular momentum, and an equatorward refraction of planetary waves mediates a stronger poleward angular momentum flux that drives stronger westerly surface winds in the mid- and high latitudes and hence results in an strengthened AO phase. Shindell *et al.* (2004) argue that a long-term dynamical response to a volcanic eruption involves both an initially strengthened AO forced by stratospheric heating and then a delayed weakening of the AO forced by a surface cooling in the tropics.

A strong positive AO phase results in enhanced westerly advection of mild maritime air over the continents in winter as well as a cooling over continental to coastal areas in the east. The summer response to an eruption is a surface cooling, which is sometimes illustrated by the anecdotal summer of 1816 after Tambora often named "the year without a summer" in both Europe and North America.

The long-term effect of AO due to volcanic eruptions is opposite to that of a decrease in solar irradiance, according to Shindell *et al.* (2003, 2004), although both types of events lead to a global cooling of the surface. They argue that the initial AO response to volcanic eruption is an enhancement (i.e., more positive) of the AO, followed by a weakening (negative anomaly) after a three-year lag. The initial enhancement is due to a dynamical planetary wave feedback,[2] whereas the lagged converse effect is explained in terms of a reduction in the latitudinal temperature differences (weaker gradient) in the upper troposphere. The injected aerosols absorb solar energy and heat the stratosphere but cool the underlying troposphere with strongest effect at the lower latitudes. A reduction in the latitudinal temperature gradients implies a reduction in the strength of the westerly winds and thus a decrease in the AO (Shindell *et al.*, 2003). The different response is explained by the fact that the stratospheric and surface effects reinforce each other in case of a reduction of the TSI whereas they oppose each other after volcanic events. A model simulation of the pre-industrial climate variations by Shindell *et al.* (2004) suggested that the 17th-century solar forcing was $-0.19\,\mathrm{W\,m^{-2}}$ as opposed $-0.30\,\mathrm{W\,m^{-2}}$ globally due to volcanic eruptions. However, the much stronger response on more local scales and changes in the frequency of extremes was postulated to primarily be driven by solar variability. Crooks and Gray (2005) found a positive temperature response at all latitudes in the stratosphere due to volcanic forcing and a negative one in the troposphere. This pattern contrasted to a negative stratospheric response at high latitudes due to solar forcing. However, their analysis did not investigate lagged response. They found no evidence for a broadening of the Hadley cell due to volcanic activity, hence suggesting differences in the dynamic adjustments associated with solar and volcanic activity.

[2] See Section 6.4.1.

7

Solar magnetism and Earth's climate

7.1 SYNOPSIS

The hypotheses proposing that solar magnetism affects Earth's climate are often based on a physical model, but long, reliable data records are not always available for testing the hypotheses. Satellite data for cloud observations start in the early 1980s. The quality of these data furthermore may not be sufficiently good for testing some of these hypotheses. The time span of space-borne measurements of the total solar irradiation (TSI) is also short compared to the climatic and sunspot cycle timescales, and even the TSI measurements suffer from errors. The problem can be illustrated by the findings of Toma and White (2000) who found an apparent non-stationary relationship between the TSI and the solar activity proxies between cycles 22 and 23. Thus, one of the obstacles for these hypotheses is the lack of long records with high-quality direct measurements. There are nevertheless some long data records which may bear the imprint of a solar connection.

One way to start looking for a connection between solar activity and the clouds is to search for sunspot signals in precipitation records which tend to be longer (and with perhaps higher-quality measurements) than the cloud observations themselves. A 27-day variability in the rainfall may be a solar signal, but it is important not to confuse this timescale with the lunar orbit of 27.3 days.

There are also some long records of proxy data that can be used for assessing the conceptual models. Lockwood *et al.* (1999) used the aa-index (Mayaud, 1972) as a proxy for the long-term change in solar magnetism, but this record may be susceptible to contamination from internally generated geomagnetic variations. In addition to actual observations, some of these hypotheses may be tested in a laboratory.

7.2 NORTHERN LIGHTS AND THE SOLAR CYCLE

7.2.1 Introduction

The northern lights (the aurora borealis) have been observed by humans for much longer than the sunspots, although people in the high latitudes have been more frequently witness to this fascinating phenomenon than people at lower latitudes. The aurora has entered folklore in Scotland, Finland, Canada, North America, Greenland, Sweden and Norway (Brekke and Egeland, 1994). There are also the southern lights in the southern hemisphere which can be seen over New Zealand during active periods.

The northern lights are related to the Sun's activity but are also influenced by Earth's magnetic field. In the past, there have been various hypotheses proposed about the northern lights influencing the weather, but this notion is not taken seriously by the wider community today. Nevertheless, the northern lights activity may be taken as an indicator of the solar activity level. The controversial hypothesis today is whether magnetic fields from the Sun can affect the cloud formation on Earth indirectly by modulating the stream of cosmic galactic rays entering Earth's atmosphere.

7.3 EARTH'S MAGNETIC AND ELECTRIC FIELDS

7.3.1 Geomagnetic storms

Variations in the geomagnetic field, such as those due to electromagnetic disturbances during sunspot maximum, may generate currents in Earth's surface. Some examples include electric currents induced in telegraph lines and powergrids, and enhanced corrosion of pipelines. Geomagnetic storms occur when clouds of ions, protons and electrons (the cloud may be neutral although the particles themselves are charged) blown out from the Sun enter Earth's magnetic field. The electric charge of the individual particles will induce an electric current in the presence of the geomagnetic field in accordance with classical electromagnetics (Maxwell's equations). The force which the magnetic field exerts on these particles together with the collisions with molecules and atoms in Earth's atmosphere prevent most of them from reaching Earth's surface, and in doing so, part of the energy associated with these particles will be converted to the compression of the geomagnetic field. Chapman and Ferraro suggested in 1931 that the energy drives an electric ring-current that flows around the Earth (Kuiper, 1953, p. 442). Despite this, the correlation between the daily measurements of the geomagnetic activity is not correlated with the daily sunspot activity. However, there may still be a correlation between a certain size of sunspots, the types of flare, and the morphology (physical character and the evolution) of geomagnetic storms. Furthermore, at sunspot minimum, there may be geomagnetic storms despite the fact that almost no flares are seen.

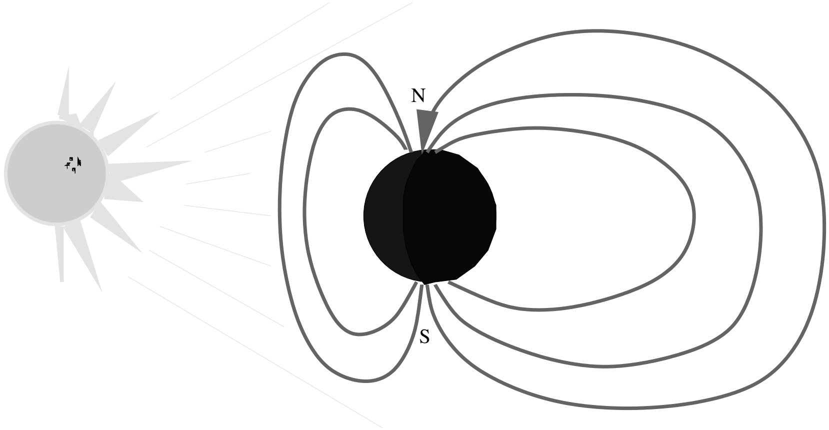


Figure 7.1. A schematic diagram showing how the magnetic field on the day-side is compressed by the solar wind whereas the night-side field is dragged out into a tail shape.

7.3.2 The geomagnetic field and solar wind

The Earth's geomagnetic field usually has a closed dipole structure, which means that the field lines are shaped like loops which emerge from the magnetic south pole and meet at the magnetic north pole where they penetrate the Earth's surface. The regions where the geomagnetic field has strongest influence on moving charges is known as the magnetosphere and includes the near space out to about ten Earth radii. The solar wind compresses the magnetic field on the day-side of the Earth whereas the field on the night-side is stretched out. A simplified schematic of the situation is shown in Figure 7.1.

One may regard the solar wind, the magnetosphere and the ionosphere as a single system where the energy and momentum are transferred from the solar wind to the Earth system. The coupling between the solar wind and the ionosphere is mediated by the geomagnetic field, where oppositely directed magnetic field lines "reconnect" (join). The reconnection takes place on the day-side of the Earth transforming the closed magnetic structure to an "open" field with one "end" connected to the Earth and the other attached to the solar wind. These reconnected open field lines move towards Earth's night-side where they stretch, due to the solar wind drag, thereby enhancing the so-called "tail lobes". Another reconnection takes place at the night-side, where magnetic lines anchored in the Earth join up forming a new closed-loop structure. The magnetic field carried by the solar wind also forms new connections. The closed terrestrial field must return (by convection in the magnetosphere interior and the ionosphere) to the day-side where new reconnection takes place. The convection towards the day-side is limited by collisions between ions and neutral atoms and molecules. The convective process extracts kinetic energy from the solar wind, and geomagnetic storms may occur when this convection is strengthened by prominent southward interplanetary magnetic field structures.

7.3.3 The magnetic field of other planets

Geomagnetic storms are related to the northern and southern lights phenomenon. Some of the other planets in our solar system also have magnetic fields, but Venus has no magnetic field (very slow rotation: 1 Venus day = 243 Earth days). Mars has a weak but extensive regional magnetic field (Mars Global Surveyor).

7.3.4 The Van Allen belts

The Van Allen belts are radiation zones around the Earth located at a distance between 650 and 650,000 km above the Earth. These belts were first discovered by J. A. Van Allen and their existence was established by Explorer 1 during the International Geophysical Year of 1957–1958. The belt consists of charged particles, mainly protons and electrons, that are trapped by the geomagnetic field lines extending from the south pole to the north pole. The inner belt is associated with intense high-energy protons with energies of the order 10–50 MeV, but is also smaller than the outer belt, located within 6500 km above Earth's surface. This belt is believed to be a by-product of the cosmic radiation. There are also particles with moderate energies (1–100 keV), some of which are related to the polar aurora. The particles' origin is believed to be from periodic solar flares carried away from the Sun by solar winds. A part of the belt dips into the upper region of the atmosphere over the southern Atlantic Ocean to form the *Southern Atlantic Anomaly* at about 250 km above the Atlantic Ocean off the coast of Brazil.

7.4 CHARGING MECHANISMS

7.4.1 Lightning

Thunderstorms trigger lightning all the time somewhere in the atmosphere and these lightning bolts involve electric currents between the ground, clouds, and ionosphere (80 to 300 km). The mean potential difference between the ionosphere and the ground is found to be around 400 kV. A global positive current of 4×10^{-12} A m^{-2} (for the whole Earth 2×10^3 A) flows downward through the atmosphere from the ionosphere to the ground. Both the ground and the ionosphere are good conductors, and excess charge is quickly redistributed evenly in the horizontal direction. This current would discharge the Earth–ionosphere potential difference in less than 10 minutes unless a constant recharging mechanism maintained the difference. It is believed that thunderstorms and lightning are responsible for the recharging of the ionosphere.

7.4.2 The atmospheric electric field

How the charges are separated in clouds to produce potential differences (positive at the cloud top, negative at the base) and lightning is still not known for sure. One hypothesis involves glaciation of cloud drops, which may result in a charge separation of light negative electrons and heavy positive ions. The mechanism is also known as the *polarisation mechanism* and explains the charge separation in terms of electric dipoles induced (polarised) in cloud particles by the mean (fair weather) electric field. When polarised ice pellets fall with the lower side being positively charged and top negatively charged, they bounce into smaller drops or ice particles and exchange charge through contact. Negative charge from the small particles is transferred to the large pellet but the positive charge is carried further aloft by the smaller droplets or ice crystals.

In addition, thermoelectric effects may play a role in the charge generation, as the concentration of dissociated ions in ice increases rapidly with temperature. The mobility of positive hydrogen ions is an order of magnitude greater than for the negative hydroxyl ion. Therefore, temperature gradients will induce different charge distributions and colder parts are more positively charged than warmer parts.

7.4.3 Cosmic rays

The galactic cosmic rays (GCR) are energetic particles (87% protons, 12% α-particles, and 1% heavier particles) that have their origin in other galaxies (remnants from supernovas). Cosmic rays were discovered in 1912 by V. F. Hess who used an electroscope to study electric discharge. An electroscope consists of two thin metal blades attached to a metal rod and a metal ball. The blades are situated in vacuum whereas the metal rod provides a connection to the ball outside the vacuum. When the metal ball, the rod and the blades are charged, equal charges in the blades repel each other, forcing the thin blades apart. Hess wanted to know why a well-insulated electroscope was discharged despite being in vacuum. One popular theory was that radioactive emission from the ground was the cause, but experiments with electroscopes carried in balloons 16,000 feet above the ground revealed that the discharge was faster aloft. The invention of the Geiger counter in 1928 enabled further studies on cosmic rays, and it was discovered that 60% of the cosmic radiation could penetrate more than 25 cm into a lead plate, suggesting that the cosmic rays were high-energy particles.

7.4.4 Interaction between cosmic rays and the air

The classical theory on atmospheric charge separation describes the ionisation through interaction between cosmic rays and the atmosphere or aerosols. There are two "types" of galactic cosmic rays: the *primary particles* from the galaxy and the *secondary particles* that are created through nuclear reactions caused by the collisions between the atmospheric molecules and the primary particles. It is the secondary particles that are responsible for most of the ion production in the atmosphere. The ion production rate increases with latitude due to the reduced shielding effect of the geomagnetic field, and is highest at around 10–20-km heights[1] due to

[1] The rate $\approx$1.5 ion pairs $\mathrm{cm^{-3}\,s^{-1}}$ at sea level to $\approx$300 ion pairs $\mathrm{cm^{-3}\,s^{-1}}$ at 13 km.

absorption by air molecules. The rate of ion production diminishes above 20 km because only primary particles are present at these levels and these are few in number. However, the solar ultraviolet and X-rays produce ions in the thin atmosphere above 90 km, giving rise to the ionosphere.

A different ion production mechanism involves radioactive decay of unstable elements in the ground and the air (rate ≈8 ion pairs $\text{cm}^{-3}\,\text{s}^{-1}$ over land). In this case, ions are formed close to the surface and the rate is not meridionally dependent.

7.5 AIRGLOW, SPRITES AND ELVES

There are some curious phenomena in the upper atmosphere that are associated with light emission. One is "airglow" which has been defined as "light other than the polar aurora, emitted from the upper atmosphere", but others have used the term to describe all the non-auroral, nocturnal, optical radiation. According to Houghton (1991), airglow is the radiation from photochemical reactions near the mesopause (≈90 km asl), such as $H + O_3 \rightarrow OH^* + O_2$. The OH molecule is "vibrationally" excited and radiates in the near-infrared, and this emission makes a significant contribution to the energy budget near the mesopause. There are also *sprites* and *elves* (Williams, 2001). C. T. R. Wilson predicted in 1920 the existence of these brief flashes of light high above large thunderstorms (30–100 km asl). The first observational evidence for the sprites and elves was presented by Boeck and Vaughan in 1990. The sprite phenomena are recognised as electrical discharge phenomena where a positive cloud-to-ground lightning strike has the same effect as a sudden deposition of negative charge into the lower part of the cloud. A corresponding positive image charge follows at an equal distance beneath Earth's surface, giving rise to a vertical dipole with a charge dipole equal to the transferred charge and the height z_{cb} of the cloud base. This dipole has an electric field which declines as z^{-3}, and when it exceeds a threshold value, determined by the air's dielectric strength, there will be a breakdown and a discharge. The dielectric strength of the air diminishes with lower density, and since the pressure decays exponentially with height, there will be a point where the air "breaks down" in terms of electrical resistance. While sprites tend to have a vertically elongated structure, elves are horizontally extensive with a doughnut shape. The vertical lightning return-stroke gives rise to an electric field that is shaped like a doughnut and is azimuthally symmetric about the lightning channel axis. One may speculate as to whether these phenomena may play a role in the climate system.[2] Williams (2001) states: "Current research on climate change gives emphasis to volatile, extreme weather events and their response to temperature change. Sprites and elves, the products of extreme lightning flashes, are clearly extreme events." Since these phenomena have their origin in the upper atmosphere and involve ionisation and electric fields, they may possibly provide a mechanism through which solar activity may influence the climate, although this is speculation

[2] One central question is how?

so far. There is also the possibility that these phenomena may be used as an indicator of a climatic response.

7.6 A HISTORICAL NOTE ON THE AURORA: THEORY AND OBSERVATIONS

7.6.1 Early scientific documentation

In the 18th century, the view on the aurora changed from being superstitious and religious to taking a more scientific aspect. Work started documenting the northern lights observations. The French physicist J. J. D. de Mairan (1678–1771) wrote in 1773 a book on the northern lights and proposed that the northern lights had been absent between 1621 and 1686. He also believed that the northern lights were the result of processes in the solar atmosphere, as opposed to a then-popular view that they were caused by various sulphuric gases of terrestrial origin. The Norwegian priest D. Schøth (1700–?) was also aware of the low northern lights activity between 1650 and 1710.

7.6.2 The aurora and geomagnetic disturbances

E. Halley (1656–1742) played a central role in unveiling the enigmatic aurora. He rejected a then-popular sulphur-gas hypothesis and proposed that the northern lights were caused by Earth's magnetism. Halley also brought an empirical aspect into the northern lights debate and discovered that the geomagnetic field had turned 15° westward in London between 1580 and 1683.

A. Celsius's brother-in-law, O. P. Hiorter (1696–1750), studied the variations in the geomagnetic field and the northern lights in 1741–1742, and documented an association between geomagnetic disturbances and northern lights activity. The Danish physicist H. C. Ørsted (1777–1851) made the important discovery in 1820 that a compass needle is affected by a nearby current.

In 1781 a Norwegian mathematics teacher D. C. Fester (1732–1811) reported that the highest frequency of northern lights occur during the autumn, but there is also greater frequency in spring than summer and winter.

The link between sunspots and the northern lights was probably first established when C. Hansteen (1784–1873) found an 11-year recurrence in the geomagnetic disturbances and H. S. Schwabe (1789–1875) discovered the 11-year cyclicity in sunspot activity. Hansteen also proposed in 1825 a link between the geomagnetism and the northern lights. S. Tromholt (1851–1896) showed that the northern lights are more frequent during high sunspot activity. He furthermore reported an association between disturbances on the telegraph lines and the northern lights, as these were most frequent in autumn and spring. H. Fritz (1830–1893) wrote a book entitled *Das Polarlicht* in 1881, where he documented the relationship between the sunspots and the northern lights in the period 1784–1871, and similar work was done by E. Loomis (1811–1889) for the period 1775–1873.

7.6.3 The discovery of day-side and night-side auroras

In contrast to earlier findings, Tromholt showed that the northern lights observations collected by S. P. Kleinschmidt (1814–1886) over Greenland between 1865 and 1880 suggested more frequent northern lights during low sunspot activity. The northern lights over Greenland were also seen during day-time, and Tromholt documented what may seem as two different types of northern lights: those on the day-side and those on the night-side, with the former anti-correlated with the sunspot number and the latter correlated with the sunspots. Loomis introduced the concept "the northern lights zone", and showed that the night-time northern lights were most prominent within a certain distance from the magnetic pole. Bigelow published in 1894 the discovery of the 27-day variations in the geomagnetic field.

7.6.4 Charged particles and the aurora

A. F. W. Paulsen (1833–1907) proposed in 1896 that the northern lights were produced by charged particles similar to the cathode rays. He also thought that the northern lights may influence cloud formation. K. Birkeland (1867–1917) proposed in 1896 that charged particles from sunspots are responsible for the northern lights. He suggested that these charged particles were guided towards the poles by the geomagnetic field and when they entered a sufficiently dense atmosphere they were decelerated in such a way that they emitted light. This theory was supported by a convincing laboratory demonstration with a ball inside an evacuated chamber. The ball contained an electromagnet and was covered with fluorescent paint and subjected to a high voltage. Two light-emitting rings could be observed around the magnetic poles of this "Earth" model.

7.6.5 The northern lights and the weather

By the turn of the 20th century, the northern lights were thought to influence the weather and were used as an argument for establishing an aurora observatory in 1899 at Halddetoppen in northern Norway. Since the northern lights are associated with activity on the Sun, the hypothesis of the aurora affecting the weather implies that the climate is also related to sunspots. The notion of the aurora influencing the weather, however, does not receive much support today.

7.6.6 The interplanetary magnetic field and the solar cycle

Earth's geomagnetic field interacts with the interplanetary magnetic field (IMF), which is a magnetic field of solar origin carried with the solar wind. The strength of these interplanetary magnetic fields is influenced by the solar cycle, as the magnetic fields associated with the sunspots and the solar activity are substantially stronger than the Sun's general magnetic field (Parker, 1997).

7.7 THE AURORA AND SOLAR ACTIVITY

7.7.1 The theory of day-side auroras

Auroras are light emission from excited particles caused by collisions between high-energy particles from space and Earth's atmosphere and the particles responsible for auroras are mostly electrons and protons. The high-energy particles are associated with the solar wind, and auroras on the day-side of the Earth are caused by the retarding solar wind particles. Moreover, protons in the corpuscular streams with velocities in the range 250–2000 km/s originating from the Sun do not have sufficient energy to explain the greatest auroral frequencies at heights of 100 km. It is proposed that these particles are subject to an additional acceleration ("after-acceleration mechanism") when approaching the ring-current around the Earth. The kinetic energy of many particles is postulated to be converted to electric energy, which in turn acts to accelerate a few particles. A result of the stream is a radial electric field between outer and inner parts of the ring-current that can trigger a discharge along the field lines and hence connect the ring-current with the polar latitudes of the Earth. It has been postulated that the associated acceleration of protons is sufficient for a penetration down to 100 km (Kuiper, 1953, p. 451).

7.7.2 The theory of night-side auroras

The night-side aurora, as opposed to the day-side aurora, may be caused by the acceleration of particles otherwise trapped in Earth's magnetospheric tail, which is dragged away from the Sun by the solar wind.

7.7.3 The aurora and the geomagnetic field

The aurora depends on the state of the geomagnetic field. The geomagnetic field lines are "closed" at low latitudes, which means that the lines "flow" out from the southern hemisphere and into the surface at the northern hemisphere (Figure 7.1). There are "open" structures near the poles, where only one "end" is attached to Earth's surface and the other end is somewhere out in space.[3] If the IMF has an orientation which is opposite to the geomagnetic field, then this will "close" the magnetic field lines further by connecting these lines with the IMF. An IMF with similar orientation to the geomagnetic field will result in an increasingly open magnetic field structure. The geomagnetic field plays an important role for the aurora, as it determines where the charged particles can penetrate far into Earth's atmosphere, such as near the polar regions where the magnetic fields are open. The magnetic field also restricts how the ionospheric currents flow. Lockwood (2002) proposed that the open solar magnetic flux may affect Earth's climate and cited studies based on ^{14}C and ^{10}B isotope records as some evidence supporting this

[3] This seems to contradict the Maxwell equation.

notion. He acknowledged the possibility of circular reasoning associated with the deposition of the isotopic ratios being affected by climate itself, but argued that the deposition processes were very different and it is unlikely that the different isotopes were affected similarly. A more likely explanation was a common factor, such as GCR. Changes in the open solar magnetic flux can be associated with both changes in the TSI and GCR, but Lockwood argued that it is difficult to discriminate one of them. There is no good explanation for why the open magnetic flux is related to TSI. It is possible that the open flux consists of a constant fraction of the total photospheric flux. In any case, a response in both TSI and GCR may have profound implications.

The open solar flux is defined as the magnetic field lines that connect the Sun and the Earth (both threaded by the same field line), and is often derived from surface magnetographs, taking the line-of-sight component and projecting this component onto a hypothetical surface referred to as the "coronal source surface". It emerges mainly from the active regions at the photosphere. A number of assumptions are also made: no currents in the corona between the photosphere and the source surface; and the surface is assumed to be spherical. Another way of estimating the open flux is by using the observation that the radial component of the heliospheric field is independent of heliographic latitude. According to Lockwood, current sheets exist in the heliosphere but no volume currents because of low plasma β (β is a measure of the ratio of the thermal to magnetic pressure) of the expanding solar wind. Thus, the radial field seen near Earth can be used to compute the total open flux threading a heliocentric sphere. The total flux, according to Lockwood, is the integral $\int \mathbf{B} \cdot d\mathbf{a}$, however according to Maxwell's equation ($\nabla \cdot \mathbf{B} = 0$) and Green's theorem, the integral strictly ought to be zero because magnetic fields exist as loops with no beginning or end. Lockwood distinguishes between outward and inward flux, and although not stated explicitly, it is probably assumed that only the outward component is considered. In that case, the integral gives a measure of how dense the magnetic fields fluxes are and thus an indication of the magnetic activity. It should also be noted that whereas many measurements of climatic variables are more direct, the estimates of open flux is subject to a number of assumptions, and may be biased by factors such as the magnetic fields of other planets in addition to instrument biases (biases have been detected in the MSU measurements of Earth's tropospheric temperatures). Lockwood observed that the open flux estimates were similar for two perihelion passes with the Ulysses spacecraft, despite very different sunspot activity between the two. According to his figure 2, there is no clear 11-year signal in the open solar flux inferred from IMF ($F_s = 2\pi R_l |B_{rl}|$, where $R_l = 1$ AU and B_{rl} is the near-Earth component of the IMF).

There have been some indications of a negative trend in the Forbush's GCR data between 1936 and 1958 which appears to be consistent with a coinciding reduction in the ^{10}Be isotope abundances (Lockwood, 2002). Similar conclusions can be drawn from high-altitude ionisation chamber measurements carried out by Neher between 1933 and 1965. Thus, Lockwood suggested that solar and enhanced greenhouse effect caused global temperature increases at different times, with solar activity changes explaining most of the warming in the interval 1907–1947 and anthropogenic contribution being dominant after 1967.

7.7.4 Aurora activity

7.7.4.1 The seasonality of the northern lights

The aurora is most active during autumn and spring, and the light emission has its origin 90 km to 150 km above Earth's surface, with higher activity in the evening and then lower after midnight. The night side northern (southern) lights tend to be most prominent in an oval approximately 23° from the magnetic pole at the night-side and about 12° at the day-side. However, when sunspot activity is high, the day-side and night-side aurora regions tend to be at the same angle away from the magnetic poles. The northern and southern lights tend to be coherent, and the southern lights are almost mirror images of the northern lights.

7.7.4.2 Auroral activity and the solar cycle

The night side northern (southern) lights are more frequent when there is high solar activity. Geomagnetic disturbances and auroral activity a few years before the sunspot minimum can be used to forecast the level of the subsequent sunspot maximum. The explanation for this predictive skill is that the next sunspot cycle starts with sunspots at high latitudes before the previous cycle has ended (see the "butterfly" diagram in Figure 4.3). The geomagnetic field is also more sensitive to sunspots further away from the solar equator (Brekke and Egeland, 1994, p. 123).

High auroral activity tends to recur at 27-day intervals, which corresponds to one solar rotation seen from Earth. There are also 27-day sequences in small geomagnetic disturbances (Kuiper, 1953, p. 445). While small storms show recurrence, large storms do not. Sometimes, sequences of the 27-day geomagnetic disturbance can last for up to ten solar rotations. These disturbances are caused by corpuscles (corpuscular streams), although sunspots' extensive magnetic fields may also interact with the geomagnetic field (Kuiper, 1953, p. 338). The corpuscles are associated with the sunspots (see Section 4.9.4). In order to conserve angular momentum, the solar wind is also expected to rotate, but with diminishing angular velocity away from the Sun.

7.7.4.3 The colours of the aurora

The aurora is usually associated with a greenish and reddish light (557.7, 630, and 636.4 nm) from the excitation of oxygen, but blue and violet colours can also be seen when nitrogen radicals are present (391.4, 427.8). Hydrogen gives blue and green light (487.1 and 656.3 nm). The aurora observations have been used to determine the chemical composition of the upper part of the atmosphere (90–150 km).

7.8 HISTORICAL CLIMATE INFORMATION FROM AURORA OBSERVATIONS

7.8.1 Aurora and the Maunder minimum

Proxy data indicate that temperatures were low during 1250–1850 ("Little Ice Age", Hartmann, 1994, p. 225), but this term is loosely defined, as Bard *et al.* (2000) refer to 1500–1750 as the "Little Ice Age", Lean and Rind (1998) 1450–1850, and Eddy (1976) 1645–1715. Other records, such as cosmogenic isotope ratios (^{14}C and ^{10}Be) suggest a high intensity of cosmic rays reaching Earth at similar times. One explanation may be that the intensification is related to changes in the shielding magnetic fields, such as those associated with the solar wind and sunspots. However, there is also a possibility that these proxies are contaminated by internal changes in the Earth's magnetic field *not* related to the solar activity. There are theories about the geomagnetic field being driven by a dynamo action, and the Earth's magnetic field has even been used as a model for explaining the behaviour of solar magnetism (Parker, 1997).

Mendoza (1997) has pointed out that even though there were no sunspots during the Maunder minimum, this does not necessarily mean that the Sun was in a non-cyclic state then. There have been some suggestions that the solar cycle persisted throughout the Maunder minimum although with drastically reduced amplitude (Rüdiger, 2000). Bard *et al.* (2000) tried to obtain an estimate for the total solar irradiance (TSI) from ^{14}C and ^{10}Be that suggested reduced TSI levels for the period between 1450 and 1850. They also found "solar minima" in 1900 and 1810 (Dalton minimum). In this respect, one important question is: Can variations in solar activity account for the entire cold period, 1250–1850?

7.8.2 The "Little Ice Age" and aurora activity

Eddy (1976) proposed that there was probably little solar activity during the "Little Ice Age" of 1645–1715 since there are no sunspot records from this period, and it has been hypothesised that the Sun was quiet during this interval. If this is true, then one would also expect a reduction in the number of incidents of northern lights for the same period. The Norwegian poet P. Dass (1647–1707), who lived in northern Norway, did not mention the northern lights and Brekke and Egeland (1994) suggest that the northern lights activity was low between 1640 and 1710 (Maunder minimum: 1645–1715) and during 1460–1550 (the Spörer period). Table 7.1 gives an overview of historical and literary references made to the northern lights. Some of these references can be found in poems, and assuming that the poet was inspired by own-life experience one can infer an approximate period for when the northern lights were present. However, such historical data are of course very uncertain and sketchy, and by no means evidence for high or low northern lights activity.

7.8.3 The magnitude and extent of the "Little Ice Age"

It is still not certain as to whether there was a global cooling during the "Little Ice Age" or whether this was a regional anomaly. Shindell *et al.* (2001b) and the results from the ADVICE project suggest that the NAO was in a negative phase during the "Little Ice Age". The study conducted by Shindell *et al.* (2001b) suggests that the global mean temperature may have been 0.3–0.4°C colder than today, as opposed to a 1–2°C regional cooling over the northern hemisphere continents. This estimate for the global cooling is considerably less than estimates by Jones and Briffa (2000).

7.9 THE MAUNDER MINIMUM AND THE QUIET-SUN THEORY

There are several accounts of the northern lights around 1662, strong northern lights in 1702, and auroras in 1706–1707 and a prominent aurora seen over large parts of Europe on March 6 1716. Moreover, was the Sun really "quiet" during the Maunder minimum?

Svensmark (2000) has argued by pointing to ^{10}Be records that there may have been cyclic magnetic behaviour during the Maunder minimum, and that lowest magnetic activity only occurs towards the end of this period: 1690–1715. The ^{10}Be records imply that the Sun cannot have been completely quiet during the "Little Ice Age" (Beer *et al.*, 1998). One may speculate as to whether the Sun may have been active for brief periods during the "Little Ice Age", or whether the Sun may have been active despite the absence of the sunspots. Another explanation may be that there *were* sunspots during this period, and that they were either not recorded or that the observational data were lost. (It would not be the first time in history that observational data had been lost; nor would it be the last!) Or, did the keen sunspot observers die; or were they forced to stop their observations due to famine or "political" or religious reasons? However, proxy-based solar activity reconstructions by Solanki *et al.* (2004) support the notion of low solar activity during the Little Ice Age, but also in periods before 1600. Shindell *et al.* (2004) prescribed solar irradiance reconstructions (Crowly, 2000) in a model simulation in order to reconstruct past temperature variations. They inferred that 75% of the global mean temperature change from the Maunder Minimum can be explained by a solar forcing and the remaining 25% was due to volcanic forcing.

One would expect to see an 11-year variation in the aurora frequency. A similar periodicity can be seen between 1706 and 1716, but there are only 4 years between 1702, when strong northern lights were seen over Bergen, and 1706. This account appears to be inconsistent with the notion of a quiet contemporary Sun and the hypothesis stating that the cold conditions were a consequence of the low solar activity. Was the solar cycle more irregular during the Maunder minimum, were the observations correct, or did the northern lights appear independently of the sunspot maximum?

Table 7.1. Northern lights recorded through history (*source:* Brekke and Egeland, 1994).

Year		Type	Remark
999		Early accounts	Sightings
1104		Early accounts	Sightings
≈1250		Written	Kongespeilet
1525	M. Luther	Historical	Religious interpretation
1563		Historical	Calais, France
1570		Illustration	Böhmen
1571		Historical	Domztice, Mid-Europe (far south)
1582		Historical	Bergen, Norway.
1584–1585	T. Brahe	Recordings	16–18 auroras/year Uranienborg, Denmark
1591	T. Brahe	Recordings	15 auroras observed Uranienborg, Denmark
1588	G. Olaus	Written	Sweden: religious interpretation of NL
1591, October 5		Print	NL over Nürnberg
1604		Drawing	Hungary. Probably inspired by NL
1629		Historical	Swedish king Gustavus Adolfus's campaign in Poland
1663		Illustration	Hungary
1664	C. Reitherus	Written	*Historico Geographica de Orbe Septentrional*
1664–1666	F. Negri	Written	Visit to Norway by Italian priest
1702, January 1		Historical	Prominent over Bergen, Norway
1706	T. Torfæus	Written	Book: *Grønlandia antiqva*
1707	O. Rømer	Records	Seen from Copenhagen
1716		Illustration	From Germany
1716, March 6			Prominent over Europe
1716, March 17		Illustrations	Danzig
1726, October 19		Illustration	Seen from Brevillepoint
1732, September 24	Celsius	Hypothesis	NL fire-bursting mountain at the north pole.
1741	J. H. Heitman	Written	Monogram on physical aspects
1767		Drawing	NL drawing
1782–1846	E. Tegnér	Poem	*Frithiof kommer til Kong Ring*
1782–1846	E. Tegnér	Poem	*Frithiof og Ingeborg*
1782–1846	E. Tegnér	Poem	*Frithiof tar arv efter sin Fader*
1796–1859	S. O. Wolff	Poem	*Nordhavet*
1779–1850	Oehlenschläger	Poem	*Thors Fiskerie*
1805–1875	H. C. Andersen	Fairy tale	*The Snowqueen*
1808–1845	H. Wergeland	Poem	*Stormen*
1838, August 23	R. C. Smith	Written	Book: *Travels in Norway*: NL seen from Oslo

1838–1839		Print	French expedition to Alta, Norway
1839, September 15		Historical	London, UK
1866	J. Moe	Poem	*Min Mening om Nordlys*
1872	J. Lie	Poetry	*Tremasteren Fremtiden eller Liv nordpaa*
1872–1874	C. Weyprecht	Written	Franz Josef Land expedition
1881	H. Fritz	Written	Book: *Das Polarlicht*
1881	J. A. Friis	Poetry	Book: *Lajla*
1882		Written	Book: *Three in Norway*
1885	S. Tromholt	Written	Book: *Under Nordlysets Straaler*
1891–1892	J. Lie	Poetry	*Jo i Sjøholmene* (Trold)
1892	G. Munthe	Painting	*Beilere* or *Nordlysets døtre*
1895	A. Garborg	Poem	*Det Vaknar* (Haugtussa)
1897	F. Nansen	Written	Book: *Fram over polhavet*
1904	K. Hamsun	Poetry	*Det vilde Kor*

7.10 MAGNETIC FIELDS, COSMIC RAYS AND CLOUD COVER

7.10.1 The cosmic ray and sulphate hypothesis

The cosmic ray flux measured near Earth's surface is high when the sunspot number is low and vice versa. There has been speculation about whether variations in the cosmic ray flux on Earth may affect cloud formation. Two different mechanisms have been proposed explaining how the cosmic rays can affect cloud formation: (i) Dickinson (1975) and (ii) Tinsley (1996). According to Dickinson's hypothesis, the cosmic-ray-induced ionisation near the tropopause may influence the formation of sulphate aerosol cloud condensation nuclei.

Wagner *et al.* (2001) examined the empirical data in the light of Dickinson's and Tinsley's hypotheses. They argue that the ionosphere–Earth current responds differently to the variations in the geomagnetic field than in solar activity.

7.10.2 The electro-freezing hypothesis

Tinsley *et al.* (1989) have proposed an atmospheric electrical mechanism as an explanation for observed correlations between the sunspot number and various indices of cyclone intensity. This mechanism proposes that cosmic rays ionise the air in the troposphere and the variations in the solar wind modulate the cosmic ray flux arriving in Earth's atmosphere.

The cosmic ray ionisation produces charged atmospheric aerosols which increases their effectiveness as ice nuclei. Thus, the induced changes in the ionisation favour the freezing of supercooled cloud drops ('electro-freezing'). This freezing releases latent heat that may modify the development of mid-latitude depressions. In 1996 Tinsley suggested that the galactic cosmic rays modulate the ionosphere–Earth current, and through this current modulate the ice nucleation rates (Tinsley, 1996).

7.10.3 The galactic cosmic rays and the climate

The hypothesis put forward by Tinsley (1996) implies no strong correlation between variations in the geomagnetic dipole and the climate. The Dickinson mechanism implies that the climate varies with the galactic cosmic ray flux, and it does not matter whether it is due to solar activity or internal variations in the geomagnetic field. Wagner *et al.* (2001) used a combination of low-pass-filtered ^{10}Be and ^{36}Cl as a proxy for the galactic cosmic ray flux and δ^{18}O and CH$_4$ as proxies describing the climate in the North Atlantic region. They found no significant correlation between these two sets of proxies. There is a marked peak in the radionuclide in the Laschamp period (36–41.5 kyr BP), but this event cannot be found in δ^{18}O nor CH$_4$. The Laschamp event was associated with high cosmic ray flux, but with little impact on the climate, and implies that the Dickinson mechanism may not be important. However, these findings are not inconsistent with the hypothesis proposed by Tinsley. GCR produces an increase in production of cloud condensation nuclei in the lower troposphere, but a decrease in the upper troposphere. Yu (2002) argued that the the altitude dependence is a result of a complex process depending on the ionisation rate itself, precursor gas concentrations, and ambient conditions.

7.10.4 Low and high clouds

The effect of clouds on the surface temperature can be illustrated using the heuristic example given below. The models described here are over-simplified, but nevertheless serve to illustrate the main principle of the role of low and high clouds. The difference between these two types is important for the models of solar–terrestrial links, including the hypothesis proposed by Svensmark and Friis-Christensen (1997).

A simple model based on energy balance considerations and a given atmospheric *lapse rate* (rate of change of temperature with height $\Gamma = dT/dz$) is derived below. The equations are based on radiative energy balance from the schematic diagrams for low and high clouds shown in Figure 7.2.

$$\frac{S(1 - A_{\rm ct})}{4} = \sigma T_{\rm ct}^4 \approx \sigma T_s^4 \tag{7.1}$$

The cloud top of low clouds may be assumed to have similar temperature to the ground, $T_{\rm ct} \approx T_s + \Gamma z, z \approx 0$. $T_{\rm ct}$ is the temperature of the cloud top and T_s is the surface temperature. Using this approximation and re-arranging equation (7.1), the surface temperature can be expressed in terms of the TSI and albedo:

$$T_s \approx \left(\frac{S(1 - A_{\rm ct})}{4\sigma}\right)^{\frac{1}{4}} \tag{7.2}$$

Thus, the effect of low clouds on the surface temperature can be inferred from the difference between the emission temperatures derived from a cloud-free ocean surface and an ocean surface covered by low clouds (equation (7.2)).

$$\Delta T_s(\text{low}) \approx \left(\frac{S}{4\sigma}\right)^{\frac{1}{4}}[(1 - A_{\text{ct}})^{1/4} - (1 - A_{\text{oce}})^{1/4}]$$

The albedo for cloud tops is assumed to be $A_{\text{ct}} = 0.5$, whereas the ocean surface reflects less radiation, $A_{\text{oce}} = 0.1$. The solar constant is taken as $1370\,\text{W}\,\text{m}^{-2}$, and the Stefan–Boltzmann constant is $\sigma = 5.67 \times 10^{-8}\,\text{W}\,\text{m}^{-2}\,\text{K}^{-4}$. Hence, the effect of the low clouds compared to cloud-free ocean surface is to reduce the surface temperature: $\Delta T_s(\text{low}) \approx -37\,\text{K}$. This estimate is clearly unrealistically high,[a] but nevertheless points to the importance of the clouds.

For high clouds, the cloud top temperature is substantially lower than the surface temperature.

$$T_{ct} \approx T_s + \Gamma z$$

The long-wave radiation emitted to space by the clouds (σT_{ct}^4) is equal to the amount directed back to Earth's surface. It is assumed that the cloud does not absorb solar radiation, and the only source of energy for the cloud is therefore long-wave radiation from Earth's surface (σT_s^4):

$$2\sigma T_{\text{ct}}^4 = \sigma T_s^4 \tag{7.3}$$

A radiative equilibrium between the insolation and Earth's black body radiation can be expressed as:

$$\sigma T_{\text{ct}}^4 = \frac{S(1 - A_{\text{ct}})}{4}$$

By using equation (7.3), the cloud top temperature can be substituted by the surface temperature:

$$\sigma T_s^4 = \frac{S(1 - A_{\text{ct}})}{2}$$

$$\Delta T_s(\text{high}) \approx \sqrt[4]{\left(\frac{S}{4\sigma}\right)}\,[\sqrt[4]{2(1 - A_{\text{ct}})} - \sqrt[4]{(1 - A_{\text{oce}})}]$$

The estimated effect of high clouds on the surface temperature is:

$$\Delta T_s(\text{high}) \approx 7\,\text{K}$$

This estimate is of course a crude simplification, but serves as a simple demonstration of why the high clouds are believed to have a warming effect on the surface.

[a] The radiative forcing is just one of several factors affecting the surface temperature.

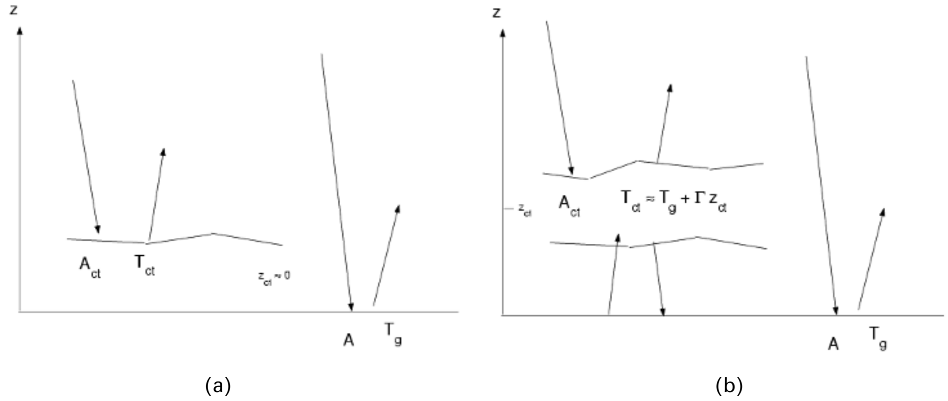


Figure 7.2. Schematic diagrams illustrating the different effects that low and high clouds have on the surface temperature.

The effect of radiative forcing associated with variations in the cloud cover depends on the cloud height. Low clouds tend to cool the surface because more solar energy is reflected back to space than upwelling long-wave radiation is absorbed and re-emitted back down to Earth. High clouds, on the other hand, increase the surface temperature (T_s) because the enhanced greenhouse effect is stronger than the albedo effect. Another factor playing a role is the number of cloud drops, as many small cloud drops have higher effective surface area than few large drops, given the same amount of water. Cloud thickness is also a factor affecting the radiation balance, with thick clouds being more important than thin clouds.

7.10.5 The hypothesis on cosmic rays and cloud droplet formation

7.10.5.1 Svensmark's hypothesis

The number of sunspots approaches zero at each minimum, but the geomagnetic field undergoes 11-year cycles superimposed on a slow monotonic trend. This trend, says Svensmark, is related to an intensification of the solar wind from the beginning of the century. The length of the sunspot cycle varies with the level of solar activity, and a short cycle is associated with high activity. The solar wind carries with it the solar magnetic field, which is dominated by small-scale solar features such as the sunspots. One hypothesis is that sunspot maxima are associated with more intense solar magnetic fields, some of which are carried away by the solar wind. This magnetic field is known as the interplanetary magnetic field (IMF).

It is generally acknowledged that galactic cosmic rays may produce rare and unstable (cosmogenic) isotopes, and according to physical theory these high-energy particles are believed to be deflected away from Earth by the solar magnetic field. Therefore, if this assumption is correct, the records of ^{10}Be and ^{14}C (Section 2.4.1.3) can be taken as proxy data indicating the past changes in the solar activity.

Cloud drops tend not to form spontaneously in clean air, but usually require particles, known as a *cloud condensation nuclei* (CCN), for drops to be created. The initial cloud drops grow by condensation after having reached a threshold size. The threshold where the growth takes off is described by the Köhler curve which describes the relationship between the saturation (excess humidity) and the cloud drop size. The rate of increase slows down in terms of their radius as they become bigger, and the time it would take for the drops to grow to 20 μm through pure condensation is 1.5–2.5 hours (Rogers and Yau, 1989, table 7.2). Thus, the initiation of rain requires something else to speed up the growth. There are two main theories of rain initiation: (i) *warm initiation* where the drops grow by collision and coalescence after reaching a critical size (Rogers and Yau, 1989) and (ii) *cold initiation* through freezing processes. Other mechanisms are also possible and may involve so-called giant nuclei (Benestad, 1994), and heterogeneous and homogeneous mixing (entrainment). The formation of cloud drops and initiation of rain are often associated with charge separation and lightning. Hence, a possible role of cosmic rays may involve the ionisation of the air. Such ionisation has been observed in the cloud chambers, where trails from these particles are seen in supersaturated air.

Svensmark (1998) suggested that galactic cosmic rays (GCR) affect Earth's climate by enhancing cloud formation. His analysis suggests that there is a correlation between the cloud cover and solar 10.7-cm radio waves, as well as the GCR. The former relationship is far from perfect, but the GCR is highly correlated[4] with a composite record for the cloud cover over the southern hemisphere (in parts extra-tropical) oceans over the period 1980–1995. The 10.7-cm radio flux is often used as a proxy for solar activity instead of sunspots and closely follows solar irradiance, soft X-rays and solar UV radiation. The GCR flux is measured by ionisation chambers at two sites:[5] Cheltenham/Fredericksburg and Yakutsk; and the neutron monitor in Climax, Colorado. The cosmic ray flux incident on Earth's surface is thought to be modulated by the IMF, which according to the theory shields the Earth against charged particles. There is a clear correlation between the solar cycle length and the variations in the cosmic ray flux count, which Svensmark (2000) takes as strong evidence for solar activity influencing Earth's climate. He showed that the coldest period in the 1000-year reconstructed temperature record by Jones *et al.* (1998b) coincided with extremely high ^{10}Be concentrations, and interpreted this to be indicative of a high rate of galactic cosmic rays and low magnetic activity. Svensmark proposed that fewer cosmic ray counts on Earth correspond to higher temperature. The proposed mechanism is that the enhanced cloud formation due to the GCR reflects more short-wave radiation back to space and thus results in a cooler climate. Svensmark presents a "back-of-the-envelope" calculation and estimates that the GCR mechanism may introduce cloud forcing of the order 0.5 W/m^2 in terms of

[4] The data has been low-pass-filtered by taking the 12-month running mean.

[5] As well as several other sites.

11-year-average increase over the period 1975–1989.[6] The Svensmark hypothesis lacks a detailed description of microphysical parameters.

Svensmark (2000) argues that the galactic cosmic ray counts follows Earth's northern hemisphere temperature more closely than other commonly used solar activity proxies. He compares the hemisphere temperature with (a) unfiltered solar cycle lengths, (b) 11-year running mean cosmic ray flux from ion chambers (measuring primarily the muon flux) and from the Climax neutron monitor, (c) 11-year running mean sunspot counts, and (d) reconstruction of the total solar irradiance (Lean *et al.*, 1995). Only the cosmic ray counts from the ion chambers do not show diverging tendencies with the hemisphere mean temperature after 1980. According to Svensmark, the Climax neutron counts, the solar cycle lengths, the low-pass-filtered sunspot numbers, and the low-pass-filtered solar irradiance do not show a convincing agreement with the temperature curve.

A study on the geomagnetic field by Lockwood *et al.* (1999) has suggested a systematic strengthening by a factor of around 100% in the last hundred years. They used measurements of the geomagnetic field from two sites near the opposite poles and derived an index, called the aa-index (Mayaud, 1972), by taking the difference between the magnetic fields. Each of the two magnetic records tends to show annual variations, which suggests that the magnetic field itself or the instrumentation is affected by changes in the solar irradiance or position with respect to the Earth, but the annual signal is almost absent in the aa-index, the difference between the two records. The geomagnetic field is influenced by the solar activity and the solar magnetic field. The magnetic field lines cannot cross each other ($\nabla \cdot \vec{B} = 0$), but the geomagnetic field combines with the magnetic field from the Sun. Geomagnetic storms are more prominent when solar activity is high and often recur with a periodicity of 27 days, the time the Sun takes to complete one cycle seen by an observer on Earth.

Wilson (1998) examined the geomagnetic aa-index derived from "K-values" from two observatories almost on opposite sides of the Earth and proposed that there is a good match between this index and the cycle-mean temperature.[7] The aa-index also correlates well with the sunspot numbers when cycle-mean averages are used, but the annual mean values do not have as high a correlation since the aa-index may have more than two peaks during a Hale cycle, whereas the sunspots always peak only twice. The aa-index minima often follow one year after the sunspot minima and peak values are delayed by around two years after sunspot maxima. Both the sunspot counts and the aa-index have a non-zero positive trend (aa: 0.067/year, R: 0.28/year), and Wilson applied a linear regression between the temperature and both the aa-index and Wolf sunspot number (R):[8] $\mathrm{T(aa,R)} = 8.56 + 0.0016\,\mathrm{R} + 0.018\,\mathrm{aa}$. The result suggested that solar cycle and geomagnetic forcing (also influenced by solar magnetism) can account for about

[6] The period 1975–1989 is a 15-year long time interval, but this estimate is presumably based on observations over 1969–1994.

[7] See Section 2.4.2 for the description of this index.

[8] Wilson gave the estimate of these coefficients to six digits!

8% of the variance in the temperature ($r = 0.275$), and the residual contained a small trend. Wilson found a strong anti-correlation between the length of the Hale cycle[9] and the Hale cycle averages of the annual mean temperature (trend not removed, $r = -0.89$). It is plausible that GCR may affect the ice-particle formation through their effect on the electric field (Carslaw *et al.*, 2002). Hence, a decrease in cosmic ray flux could reduce the production of ice-particles and hence result in a decrease in rainfall. This would, according to Carslaw *et al.*, also produce a change in cloudiness opposite to that observed, although they had not taken into account effects from latent heat release. It also remains to be established whether the effect of GCR can lead to detectable changes in the cloud properties, given these mechanisms.

Foukal *et al.* (2004) noted that the correlation between the IMF (aa-index) and irradiance is only significant on the monthly scale and not on the interannual timescale. They therefore concluded that relationships on longer timescales can only be speculative at the present stage.

Observational records of sunshine hours obtained by Pallé and Butler (2001), however, indicate a gradual decline between 1880 and 2000. According to Svensmark's hypothesis, a warming due to a reduction in the low cloud cover is expected to be associated with an increasing trend in the sunshine hours (sunshine factor). It is of course possible that a decrease in the low cloud cover may have been accompanied by an increase in the high cloud cover, but in that case it has not been resolved why there was a trend in the high cloud cover. The mean synoptic 1951–1990 cloud cover record at Armagh shows a trend towards more overcast conditions and Pallé and Butler (2001) found a strong negative correlation between the 1983–1994 cloud cover (%) and the sunshine factor. They also argued that the cloud cover over Ireland tends to be correlated with the North Atlantic cloud cover.

Pallé and Butler (2001) propose that the reduced sunshine factor and increased cloudiness may follow from rising sea surface temperature. Such a link would, however, involve a negative feedback, as increased cloud cover (assuming low clouds), according to Svensmark's hypothesis, favours a cooling of Earth's surface. It is therefore not clear as to whether the findings of Pallé and Butler (2001) really support the notion of a solar–terrestrial link or not. Although Solanki *et al.* (2004) found unusually high solar activity in the last 60 years, they stressed that solar variability was unlikely to be the prime cause of the strong global warming during the last 30 years. There seem to be some inconsistencies regarding the hypothesis of Svensmark.

There have been suggestions of brightness variations on planets like Neptune that have been associated with variations in solar activity. Svensmark has argued that these variations may well be due to modulations in cloudiness, in a similar way to those proposed on Earth.

[9] Trend not removed.

7.10.5.2 Svensmark's hypothesis and observed regional warming

Since the solar radiation per unit surface area is greater in the tropics than in the high latitudes, one may (to a first order approximation) expect a cloud–albedo mechanism to give the greatest response in the tropics, given a uniform modulation of the clouds with respect to latitude. The observations suggest greatest warming at the higher latitudes in the northern hemisphere (Hansen *et al.*, 1998a), which does not seem consistent with the tropics becoming warmer. But there may well be a latitudinal dependency of the cloud modulation, as the cosmic ray flux is expected to be more intense near the poles (see Section 7.12.0.1). There may well be other feedback mechanisms involved that may amplify the climatic response at higher latitudes (e.g. snow or sea-ice).

7.10.5.3 Cosmic rays and ice ages

Shaviv hypothesised that there could be a link between the galactic rays from the galaxy's spiral arms and climate. According to his ideas, our Solar System will be exposed to high levels of cosmic rays when crossing the spiral arms of the Milky Way. This hypothesis is controversial (Shaviv, 2004; Rahmstorf *et al.*, 2004) and makes a number of assumptions: equipartition between cosmic ray energy density and magnetic field, that cosmic rays exhibit simple diffusion with an assumed value of diffusion coefficient of $D = 10^{28}\,\mathrm{cm}^2\,\mathrm{s}^{-1}$. The hypothesis is explored with an ideal disc model with a simplistic ad hoc arm-description. According to this model, our solar system has different angular velocity to the arm of the galaxy. Shaviv suggested that there was a ~28% reduction in the cosmic ray compared with present. The empirical data used to test the hypothesis involved iron meteorites that have been exposed to cosmic ray bombardment (isotope ration $^{41}\mathrm{K}/^{40}\mathrm{K}$). The analysis assumes that the level of cosmic rays has been constant, but a gradual change could distort the age calibration.

7.10.6 Criticism of Svensmark's hypothesis

The solar–terrestrial link proposed by Svensmark has been controversial, and one criticism of Svensmark (1998) is that he did not account for all possible systematic errors in his paper due, for instance, to instrument drift or the instruments being affected by the geomagnetic fields at first, but at a later stage when new data indicated that the curves started to diverge after 1994, claimed there were problems with the ISCCP data (Marsh and Svensmark, 2002). Kristjánsson and Kristiansen (2000) pointed out that there were considerable uncertainties associated with the data on which Svensmark based his hypothesis. In the papers published by Svensmark there is an odd-looking missing data gap, where the cloud measurements from one satellite are high on either side but are also apparently different to other satellites' measurements in 1989. Furthermore, there is a substantial discontinuity between the measurements from the different satellites in 1992. So, the high cloudiness measurements since 1993 seem conspicuous.

The records are only a short series covering approximately two solar cycles, which makes it difficult to evaluate the relationship between the two curves. The time-series studied were derived from satellite observations and were only 7 years long. The value of a 0.5 W/m^2 net radiative forcing associated with 1% in the total cloud over is based on estimations with a high degree of uncertainty. The cloud data contain large errors due to difficulties in cloud-type detection (high clouds block the view of low clouds), instrument calibrations, and changes of satellites.

Laut (2003) argued that the article of Svensmark (1998) was misleading due to inappropriate combinations of two incongruous data sets (ISCCP and DMSP). Part of this argument rested on findings of Kristjásson and Kristiansen (2000) and part on the argument that the two records represent different physical quantities. Another criticism is the fact that variations in the cloud cover lag the GCR by more than half a year and that this delayed response cannot be explained physically. Furthermore, Laut argued that observations of low clouds from satellites are obscured by higher clouds and thus errors in cloud data could affect the analysis. A final remark on Svensmark's paper is that the method and results were not readily available so that Laut had to scan the original figures and the electronically digitise the curves ("remake") in order to replicate the results. The lack of transparencies should be avoided at all costs, as one criterion of the scientific process is the replication of past results and testing others methods and results. It is also important to comment on why there are gaps in the data or other unexpected features. Svensmark responded to Laut's paper with an article on his homepage (*http://www.dsri.dk/˜hsv/*) and objected to what he interpreted as an accusation of scientific dishonesty by Laut. He refused to accept the criticism and argued that the statement of "misleading character" and "incorrect handling of physical data" could only be traced to the inclusion of the DMSP data in one of his figures. Svensmark clarified in his response to Laut that the gap in the DMSP data was due to failure of one microwave channel on one of the satellites, however, this information should have been provided in the original paper to avoid speculations about the missing data. Part of Svensmark's defence was that essential data was unavailable at the time of publication, however, this should not excuse strong statements and too firm conclusions. Svensmark furthermore argued that the data description for the DMSP data was incorrectly represented as measurements of total cloud cover, and hence he was not being dishonest. Regarding Laut's statement about diverging GCR and cloud data after 1994, Svensmark claimed to have found calibration problems after 1994. He also argued that Kernthaler *et al.* (1999) did not detect the GCR–cloud relationship because they used the "flawed" ISCCP-C2 data. Hence, on several accounts he questioned the empirical data when they did not support his own ideas, however, it also seems that he neglected uncertainties and errors when they appeared to be in favour of his hypothesis. Svensmark did not address the caveat about the half-year lag between the GCR and cloud data on his home page.

Farrar (2000) argued that the high correlation was coincidental and was due to one dominant fluctuation caused by an El Niño. He found no correlation pattern which supports the claim made by Svensmark and Friis-Christensen (1997) that the clouds near the magnetic poles are most strongly affected.

Wagner *et al.* (2001) suggested that the seemingly good correlation in the 16-year long record analysed by Svensmark (a composite of four shorter series where none is longer than 7 years) may be a coincidence. They examined a longer record of synoptic cloud observations from Switzerland. This synoptic record exhibited a high correlation between the Climax cosmic ray record and cloud observations in the 1980–1995 period that Svensmark had analysed, but low correlation before 1980. The fact that it correlated with extensive areas in the North Atlantic region suggests that their cloud record may represent a larger area than just Switzerland. Haigh (2003) noted that the GCR and low clouds curves diverged after 1994, and hence weaken the case for a real correlation between the two.

Harrison and Shine (1999) reviewed some of the work on solar activity, cloud cover and climate. The cloud data used by Svensmark and Friis-Christensen (1997) was obtained from geostationary satellites and only covered the oceans between 60°S and 60°N. It has been argued that the signal detected by Svensmark and Friis-Christensen (1997) may actually be due to shifts in the circulation system rather than modulation of global cloudiness.

Kristjánsson and Kristiansen (2000) could not find support for Svensmark's hypothesis on the global total cloud cover over the mid-latitude oceans. On the other hand, there is a possibility that low marine clouds may vary with the cosmic ray flux. These clouds are associated with few cloud condensation nuclei and a large cooling effect, but Kristjánsson and Kristiansen (2000) argue that there is no known physical mechanism explaining how they are affected by cosmic rays. When a 43-year record of synoptic cloud observations of low cloud cover was compared with the cosmic ray record, they found a weak negative correlation. Their objections are supported by Bertrand and van Ypersele (1999) who argued that the detailed picture is missing on how atmospheric microphysical processes can connect solar activity and cosmic rays with Earth's planetary cloud cover.

Svensmark (2000) proposed that most of the recently observed global warming is due to a reduction in the low-level planetary cloud cover because of a reduction in the GCR flux and increasing solar activity. According to his hypothesis, the terrestrial temperatures are regulated by the planetary cloud cover, whose extent in turn is modulated by the GCR. The GCR flux is shielded by the interplanetary magnetic field dragged towards the Earth by the solar wind. It is postulated that an expansion in the cloud cover cools the surface by reflecting a larger fraction of the total solar irradiance back to space. Any mechanism involving the albedo implies strongest response in the day-time temperature. Observations, on the other hand, suggest a reduction in the diurnal temperature range where the night-time temperature has increased more than the day-time temperature (Houghton *et al.*, 2001). According to Svensmark's hypothesis, the warming is due to a reduction in Earth's albedo (reflected light), and therefore a long-term reduction in the low-level planetary cloud cover appears to be inconsistent with the observations (expect strongest day-time warming). The lack of long-term trend in GCR as well as in the aa-index, sunspot number, and unfiltered SCL since 1951, contradict Svensmark purported link between GCR and the recent global warming (Benestad, 2005a; Richardson *et al.*, 2002). The lack of trend in these proxies are furthermore inconsistent with the proposition by Willson (1997) that there is a trend in the TSI level during solar minima. Any hypothetical disagreement between the TSI level and these proxies would render past palaeoclimatic solar reconstructions invalid. Kristjánsson *et al.* (2002) pointed out that the correlations between GCR and cloud cover were only strong for long timescales such as 11 years and that the correlation between one-year, high-pass filtered data was negative. If GCR are to influence cloud formation through enhancing CCNs, then one would expect an almost instant response, due to clouds short lifetimes (minutes–hours), as well as positive correlation. Lockwood (2002) observed that correlations were greatest (0.65) when the cloud cover anomalies lagged the GCR data from Moscow neutron counter by 4 months, and Kristjánsson *et al.* (2002) found a one-month lag for TSI for IR data (−0.6) and a 4 month lag for the daytime low clouds (−0.4).

Usoskin *et al.* (2004) proposed that the GCR are correlated with mid-latitude low clouds as opposed to low latitudes. However, after de-trending the 1985–2000 period, they claimed to identify the solar 11-year cycle in the low cloud cover and that a strong trend in the tropical cloudiness masks the relationship between GCR and the low clouds. It is worth while keeping in mind that the tropics tend to be characterised by deep convection (high, thick clouds), some low stratus clouds also do exist off the west coast of South Africa and Peru. They found highest correlations in regions with generally small amounts of low clouds, and no correlations in regions characterised by low clouds such as off the Californian and Peruvian coasts (Kristjánsson *et al.*, 2004). Furthermore, their definition of "significant correlation" was suspicious as it was taken as "significance level >68%", and "highly significant correlation" was taken to be ">90%". What is more is that it is difficult to see how they obtained high correlations for the zonal means when the geographical distribution of correlation did not indicate high values, and their analysis did not point to any correlation in regions with extensive cloud cover. They also argue that because a regression constant between two standardised variables is close to unity, the variations of one is directly ascribed to variations in the other, even if the original physical units are not the same. However, this is not evident and their argument is not very convincing. Their conclusion is probably not supported by the evidence in the light of other factors possibly playing a role (e.g., TSI) and because of the lagged cloud response. Sun and Bradley (2002) argue that there is no solid relationship between GCR and low cloudiness. They examined an improved and more recent extended version of the satellite cloud data (ISCCP, D2 as opposed to C2) in addition to three surface-based cloud data going back to 1953 when GCR Climax measurements begin. They did not find any meaningful relationship between GCR and cloud cover over tropical and extra-tropical land areas since 1950 and no 11-year solar cycle. Regarding long-term trends, there was no clear overall tendency. They found an increase in the cloudiness over the contiguous U.S.A., a decrease over China, and no systematic change over the former USSR. Sun and Bradley dismiss the notion of GCR causing a global warming because this would require the low cloudiness undergoing a systematic decrease world wide. Ship observations, however, suggest a steady upward trend in the low cloud cover by 3.6% between 1952 and 1995 over the global oceans. According to Lockwood (2002), simulations with the Hadley Centre's

HadCM3 climate model, with a prescribed best-estimate of historic volcanic, solar, sulphate, and greenhouse gas forcings, suggest a decrease in low-altitude cloud cover after the mid-1970s. The lack of trend in the cloudiness reported by Sun and Bradley (2002), however, appears to be inconsistent with the HadCM3 results.

Sudden dips in the cosmic ray flux following solar flares are known as *Forbush* decreases, with a minima after a few hours and a return to normal levels within days. The Forbush events may originate from disturbances in the IMF and affect the geomagnetic indices such as the aa-index. Pallé and Butler (2001) found no evidence for increases in the sunshine factor following the Forbush decrease, using 50 years of data.

7.10.6.1 The weakness of the aa-index as a solar magnetism proxy

Ever since J. C. F. Gauss (1777–1855) started instrumental measurements of the geomagnetic field in 1820, the strength of Earth's magnetism has decreased systematically (Brekke and Egeland, 1994, pp. 92–93). Gauss had derived mathematical expressions for the magnetic field strengths for a given location on Earth's surface given a limited number of measurements. Hansteen's objection to Gauss's theory was that it did not correctly predict the observed regular daily magnetic variations, which include irregular short-term fluctuations as well as long-term variations. The diurnal magnetic variations were later associated with motion in the upper atmosphere across the magnetic field lines following the Sun. Modern theory on Earth's geodynamo suggests that Earth's magnetic field varies with time, and the field undergoes reversals with an average reversal interval of 10^5 years, but there is also a gradual variation in this periodicity (Buffet, 2000). In other words, the geomagnetic field changes internally, due to convection in Earth's core. Simple disk dynamo models of Earth's magnetic field suggest a chaotic evolution with reversal times of approximately 320,000, 120,000, and 50,000 years, but variability with higher frequency may also be present. Rapid magnetic drifts of the order 0.18°/year have also been reported by Walter Elsasser, and 12 evolving magnetic features have been identified in the geomagnetic field surface maps. Such features can be interpreted as an indication of 12 "small-scale" convective cells.

Over the period 1964–1996, Lookwood *et al.* (1999) reported a ~40% increase in the radial component of the interplanetary magnetic field (IMF). Richardson *et al.* (2002) disputed this trend, arguing that the apparent trend was due to lower than average fields during the 20th solar cycle (1964–1979). They argued that since 1976, the average IMF has in fact decreased slightly. They also argue that the IMF strength has been relatively constant since around 1954. Furthermore, the increase in the aa-index primarily took part before 1950. The conclusions of Richardson *et al.* (2002) are supported by independent measurements of cosmic rays and 10.7-cm flux. A similar but independent analysis by Benestad (2005a) also documented the lack of trend in the cosmic ray record from the Climax data, the 10.7-cm flux, as well as the aa-index since 1954. It is a theoretical possibility that the secular increase in the aa-index between 1900 and 1950 may have been due to a change in the solar wind structure, with CMEs occurring with increasing frequency. CMEs are responsible for the vast majority of large geomagnetic storms. However, Richardson *et al.* (2002) concluded that the increase in aa-index over the first half of the 20th century could not simply be due to a change in solar wind structure, but that changes in the IMF and/or solar wind speed could have been the cause.

A criticism of the hypothesis proposed by Lockwood *et al.* (1999) is that changes in the geomagnetic field implies that long-term geomagnetic measurements will be affected by internal changes of the geomagnetic field as well as changes to the solar magnetic field. Furthermore, currents in the ionosphere and the ground may produce magnetic fields which further contaminate the measurements. Even with measurements from many more locations, it may be difficult, if not impossible, to identify how much and which factor has contributed to the measured field intensity. Part of the problem may also be due to the lack of long data records from the entire globe and problems associated with getting homogeneous time-series. (Were there changes in the instrumentation? If so, how were the data merged?) Therefore, the proposition that the recent global warming trend is a result of an intensification of the solar magnetic field is at best based on circumstantial evidence. Long-term changes in the aa-index are not necessarily only due to changes in the solar activity, but may also be associated with variations in the geomagnetic field or other factors. Both the positions of the magnetic poles and the geomagnetic strength are known to vary over time (Brekke and Egeland, 1994, pp. 92–93).

Lockwood *et al.* (1999) found a long-term trend in the annual mean aa-index for the period 1964–1996, and a positive trend can also be inferred for the same time interval from the aa-index shown in Figure 7.3. However, there is no systematic change in GCR counts over the same interval, which suggests that the aa-index does not describe the strength of the solar magnetic field. It is therefore questionable whether solar magnetism may be derived from this index. The trend estimates in Figure 7.3 were calculated for the period 1951–2002, and the slope of these trends are not statistically significant from zero. The implication of zero trends in these records is sufficient to disprove the hypothesis proposed by Svensmark and co-workers.

7.11 THE MAUNDER MINIMUM

The Sun is believed to undergo variations on timescales of 100–1000 years. Edward Maunder observed that there was little solar activity in terms of sunspots between 1645 and 1715. This period is called the Maunder minimum. Later, Jack Eddy inferred from carbon-14 records that there may have been ten such periods of minimum solar activity during the last 7000 years. Thus, there may be modulations of the solar cycle with a periodicity of approximately 700 years. Furthermore, the Sun is believed to have been unusually active during eight periods within the last 7000 years, and the most recent peak is thought to have occurred during the 12th and 13th centuries. Eddy proposed that there is a correlation between the solar activity level and the temperatures in the northern temperate zone. The temperature difference between maximum and minimum has been estimated to be about 2–4°C, with cooler conditions coinciding with minimum solar activity. At the present, the Sun is

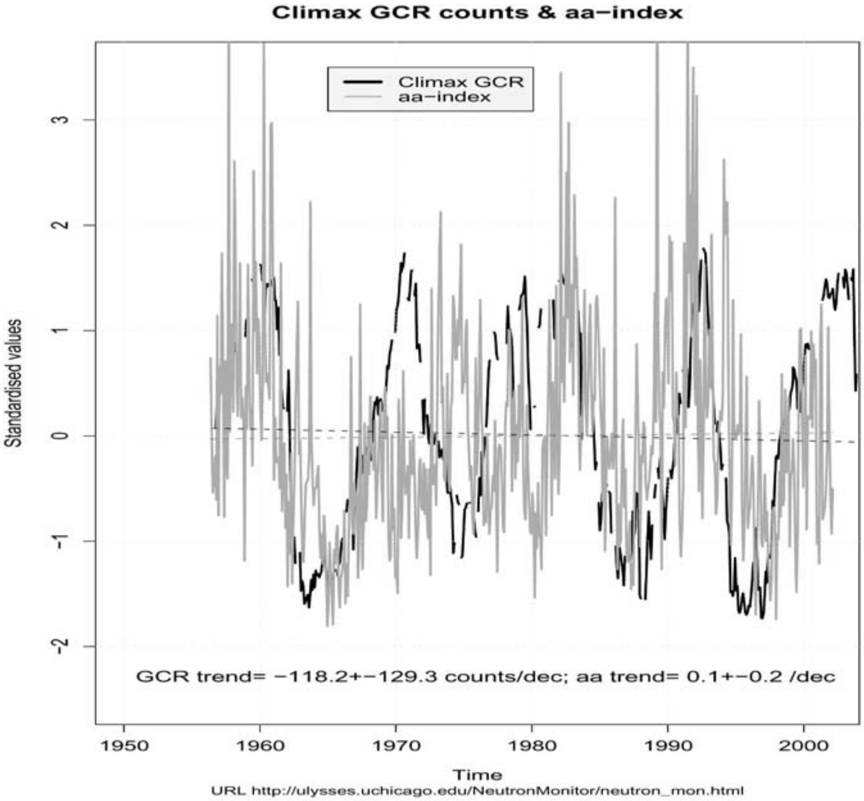


Figure 7.3. Time-series of standardised 30-day mean GCR measurements from Climax and the monthly mean aa-index for the period 1951–2002. The dashed lines show the best linear-fit, none of which is statistically significant. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and http://ulysses.uchicago.edu/NeutronMonitor/neutron_mon.html.

somewhere in the middle between minimum and maximum, and the activity level may be recovering slowly from the most recent Maunder minimum, when northern Europe may have been about 1°C colder than present.

Rüdiger (2000) discussed the nonlinear interplay between dynamo-induced magnetic fields and differential rotation in stellar convective zones. The Earth and Sun maintain two different types of dynamo where the solar dynamo exhibits a quasi-periodicity of about 11 years whereas the terrestrial magnetic field is "permanent". However, the geomagnetic field is strictly not permanent, but has much longer timescales than the Sun. The few sunspots observed during the Maunder minimum were reported to have an asymmetric latitudinal distribution lasting approximately 30 years. This persistent asymmetric sunspot tendency is a unique feature of the grand minima of the Maunder minimum, and may perhaps be associated with a parity change in the driving internal magnetic fields.

There have been studies of other stars than the Sun, and a paper by Baliunas and Jastrow (1990) suggests that there are similar stellar cycles on other stars than the Sun. In this study, the Ca-emission from 13 different stars with similar mass and estimated age as the Sun were compared and categorised according to their stellar cycle. Some of these stars exhibited a low level of starspots and were associated with low luminosity, whereas the more active stars were brighter. Of course, the study of starspots is much more difficult than that of sunspots, as they cannot be seen directly. Rather, their presence and prominence are inferred from the stars' brightness modulation (Byrne, 1992). It is important to distinguish the cause of brightness modulation from other factors, such as planets (i.e. blocking the line of view through their passage).

Starspots are known to be common in three star types: the so-called dMe flare stars (or BY Dra spotted), the RS CVn binary stars and the FK Com stars. The FK Com stars are single GIII stars with unknown mass and radius. The dMe flare stars include the KVe–MVe category, have an approximately similar mass to the Sun, and a radius of 0.2–0.9 times the Sun's radius, and can be either binary or single stars. The RS CVn binary stars have a combined mass similar to the Sun, but larger radius by a factor of 2–4. Starspots are more pronounced in rapidly rotating stars (dMe: 0.25–10-day period, RS CVn: 1–50 days, and FK Com: 1 day), but the neutral metal spectral lines are wider for slowly rotating dM stars than faster-spinning dMe stars.[a]

[a] Probably because chromospheric reversals in the line cores in the fast rotators will resemble the flat-bottom profiles usually interpreted as starspots.

7.11.1 The weakness of comparisons with "Sun-like" stars

Inferences made about long-term solar variability from comparisons with Sun-like stars and cosmogenic isotope records are only based on circumstantial evidence, and are associated with a high degree of uncertainty. First, it not clear whether all the so-called "Sun-like" stars really are similar to our Sun, because of different appearances in their stellar activity. The conclusions drawn from Baliunas and Jastrow (1990) are based on a small statistical sample, and the extrapolation from the small sample of stars to our Sun is at best tentative. Although Baliunas supports Eddy's theory, the probability that similar results could be obtained by random choices of 13-star samples may not qualify these results as statistically significant and our knowledge of starspot distribution must be treated with care (e.g. Byrne, 1992). Hence, as long as the origin of the solar cycle is not completely understood, the data is scarce, and it is difficult to validate the hypothesis through scientific tests, the similarity between the Sun and a handful of stars cannot be taken as solid scientific evidence for long-term changes in solar activity and TSI.

7.12 THE INFLUENCE OF CORPUSCULAR CLOUDS

Ludmány and Baranyi (2000) argue that the energy transferred to the Earth system by corpuscular clouds cannot be neglected and can in some instances be as important as the variations in the solar irradiation. The mean total energy flux associated with the TSI is around 10^7 times the energy associated with the corpuscular clouds. But the variations in the latter are proportionally much larger than the variations in TSI. The plasma in the corpuscular clouds is thought to interact with the terrestrial system through magnetic field reconnection, and hence the IMF may play a role.

7.12.0.1 Charged particles and the geomagnetic field

The sources of galactic cosmic rays are at great distances from Earth, and the particle flux may be assumed to be isotropic.[a] The influence of the Earth's geomagnetic field on these charged particles can be reduced to the calculation of allowed trajectories in a quasi-static dipole magnetic field. One starts the investigation with the basic electromagnetic theory: a charged particle that has the charge ze, where e is the charge of a proton (1.602×10^{-19} C) and the kinetic momentum $\vec{p} = \dfrac{m\vec{v}}{\sqrt{1-(|v|/c)^2}}$ is affected by static magnetic fields ($\vec{B}$):

$$\frac{d\vec{p}}{dt} = \frac{ze}{c}\vec{v} \times \vec{B} \tag{7.4}$$

The velocity can be decomposed into radial and tangential components: $\vec{v} = s\hat{s} + t\hat{t}$. Here $\hat{s}$ means the unit vector s with unit length: $|\hat{s}| = |\hat{t}| = 1$. In a uniform magnetic field, the particle path is a helix with constant speed $|\vec{v}| = \text{const.}$, but with a constant tangential acceleration towards the axis of the helix (centripetal acceleration):

$$\frac{d\vec{t}}{ds} = \frac{ze}{|\vec{p}|c}\hat{t} \times \vec{B} \tag{7.5}$$

The angle, α, between the momentum ($\vec{p}$) and the magnetic field ($\vec{B}$) is constant and the radius of the particle's orbit around the axis of the helix (parallel with $\vec{B}$) is $r = \dfrac{pc\sin(\alpha)}{zeB}$. The quantity $pc/(ze)$ is often referred to as the magnetic rigidity of the particle.

The geomagnetic field may be represented roughly by the dipole moment M, Earth's radius is a, and the geomagnetic latitude of the observer is λ. According to Williams (1960) a particle with rigidity greater than 60×10^9 V may arrive at any point on Earth and from any direction. The arrival at Earth's surface is more limited for particles with lower rigidities, and the critical condition that the particle must satisfy in order to reach Earth's surface is:

$$\frac{pc}{ze} = \Gamma_* \geq \frac{M}{a^2} \frac{\cos^4 \lambda}{(\sqrt{1 + \cos^3 \lambda} + 1)^2} \tag{7.6}$$

The cosmic ray flux depends on λ, and the threshold is lowest near the magnetic poles with $\lambda \approx 90^\circ$. This implies higher flux intensities near the poles than near the equator for cosmic galactic rays with magnetic rigidity energies similar to this critical threshold value Γ_*. Furthermore, if the galactic cosmic rays affect Earth's cloud cover, the mechanism is expected to be most evident over the polar regions.

[a] Independent of direction.

7.12.1 Solar activity and lightning

There has been speculation about whether geomagnetic fields can affect weather and climate in other ways than modulating cloud droplet formation via galactic rays. Magnetic fields exert a force on moving electric charges, or electric currents. An interesting question is whether there is a relationship between the geomagnetic field, lightning discharge on a global scale, and the solar cycle.

There is little work published on relationships between solar activity and lightning. Searches on the Internet give few hits. Schlegel *et al.* (2001) has found some indications that mid-latitude thunderstorm activity may be modulated by the solar cycle. One explanation may be the lack of long lightning records on a global scale. A link between solar activity and lightning is not impossible and far-fetched, even though such mechanisms are quite speculative at the present. Lightning discharges are related to the fair weather electric field between Earth's surface and the ionosphere (see Section 7.4.1 for details on this fair weather electric field). The ionosphere may be affected by both charged particles from the solar wind as well as magnetic fields of solar origin. Disruption of the ionospheric charge may affect the ionosphere-to-ground current through changes in the discharge frequency.

A physical mechanism linking upper atmospheric geomagnetic storm disturbances with tropospheric weather has been proposed by Sikka (1987) and it has been postulated that vertical mixing by turbulent eddy fluctuations results in the net transport upward of positive charges originating from lower levels accompanied by a downward flow of negative charges from higher levels.

7.12.2 The geomagnetic field and oceanic currents

Wollin has suggested that Earth is warmer when the geomagnetic field is weak, but proposed that the temperature is affected by the rate of change in magnetic field strength. It is speculated that magnetic field changes may affect electric charge transport in the oceans (oceanic electric currents) and hence the oceanic mass transport (ocean currents). Ocean currents influence heat transport, and may affect the geographic temperature distribution. There are ionised molecules in the oceans as well as in the ionosphere, and electric currents may also flow in salty oceans, where Na^+, Cl^-, H^+, and OH^- ions may act as conductors. Ocean currents transport these ions, and may as a result set up electric currents by means of the geomagnetic field and dynamo action.

7.12.3 The geomagnetic field and sea level pressure

The solar storm of March 13 1989 caused an increase in high-frequency fluctuations in surface atmospheric pressure, according to Selvam *et al.* (1997) who say: "Observational studies indicate that there is a close association between geomagnetic storms and meteorological parameters." Geomagnetic field lines closely follow the isobars of surface pressure (King, 1975).

As geographical differences in surface pressure represent net forces that drive the atmospheric circulation,[10] winds may be influenced by the geomagnetic field if the geomagnetic field influences the pressure. This hypothesis may explain the alleged dependence of westerly winds on the solar cycle, as the solar wind is enhanced and the IMF is stronger during sunspot maximum. Atmospheric circulation, including westerly winds, plays a role in the distribution of heat in the climate system, and one may speculate about whether variations in the circulation could be due to solar activity. Winds also drive ocean currents, and therefore involve atmosphere–ocean coupling. It remains to explain how the pressure field is influenced by the geomagnetic field. One mechanism may possibly involve an oceanic dynamo action, but this hypothesis is speculative, to say the least.

[10] Geostrophic flow.

8

A review of solar–terrestrial studies

8.1 SYNOPSIS

Any hypothesised link between solar activity and Earth's climate must be supported by empirical evidence. Observations are used both in the search for such relationships and for validation of physically based hypotheses. In either case, the science of solar and terrestrial relationships brings in concepts of data analysis and statistics. Data analysis can be thought of as consisting of three different stages: (i) the exploratory stage; (ii) descriptive stage; and (iii) the inferential stage. During data exploration, the researcher should be speculative and inventive in order to search for patterns or relationships in the data. At this stage, there is no need to apply critical tests in order to try to falsify the hypothesis (Karl Popper). The descriptive stage involves categorisation and putting the knowledge into order. This stage involves making hypotheses. The final part of the analytical approach involves the Popperian thoughts about "scientific method", whereby the researcher should be critical of her or his own findings and try to reject the hypotheses proposed in stage (ii). It is important to let the data speak for itself, and to set up tests which are not predisposed. Such tests are commonly made in statistics, where a *null-hypothesis* (which will be explained in more detail later) is tested and confidence levels are attributed to the descriptive statistics. Often, the null-hypothesis is that the results gained in stage (ii) could easily have been achieved by chance, and the goal is to show that the probability the results were due to pure chance is very small.

Statistical exploration and description can *never* prove that there is a physical link between two entities, although strong circumstantial evidence sometimes is found. There is never a guarantee that the data is not contaminated by external factors. In order to make robust claims about relationships or certain behaviour, it is necessary to formulate a hypothesis based on a physical explanation, and then apply the statistical analysis to test it. Although it is in principle never possible to prove a hypothesis through empirical studies, one may with certainty exclude those which fail the test.

It is important to reiterate the scientific criteria since the validity ("truth") of the conclusions drawn often hinges on the experimental set-up and the choice of analytical methods. The scientific criteria according to Feynman (1985) and K. Popper (Magee, 1973) are that the scientific tests must be objective and that in principle, any of a number of possible outcomes has an equal chance of being realised. As many of the current hypotheses on solar–terrestrial links are primarily based on empirical studies, it is important to first examine the analytical methodology on which they are based. Both the methodology and quality of the observations are of central importance in both solar and climate science, because if the analytical method (or experimental set-up) is not properly designed or the observations are contaminated, then the outcome may be pre-disposed. A systematic bias can be regarded as a pre-disposition of an experiment, whereby improper observational practice, unsuitable pre-processing, biased analysis or instrumentation errors affect the final outcome.

The distinction between pre-processing steps that are appropriate and those that are not is not always an easy one; it often depends on what one is looking for and what type of answers one is seeking. The purpose of pre-processing is often to eliminate factors that may affect the analytical results, such as inhomogeneities. There is occasionally the need for dealing with gaps of missing data. Can these be excluded without affecting the analysis or is it more appropriate to fill in the values? There are many ways to fill in data holes, and the type of method depends on what type of analysis is to be done. Often the annual cycle is very strongly present, and one may wish to remove this in order to emphasise weaker signals. The signal may be weak compared to the noise, and one may want to filter away all the irrelevant information (noise) so that the data series mainly contains information on the chosen object. It is important to stress that there must always be a good justification for pre-processing the data before the analysis. A high proportion of noise may imply a weak relationship or poor data quality.

It is also important that the analytical results are robust and not *too dependent* on the particular analytical method and pre-processing. If there is a strong and real relationship between two quantities, then this ought to be visible even in unprocessed data. The data processing should merely aim to bring out such a relationship more clearly from the background noise.

In summary, testing hypotheses requires that we know what we are looking for and there must be clear criteria for falsification (the "No" outcome) and verification ("Yes"). Once again, the tests should be completely objective and have no pre-disposition for either outcome, and the actual (true) observations should be the deciding factor. It is important to set up various hypotheses so that they can easily be tested and no outcome is favoured above the others. One key point is to end up with hypotheses consistent with the observations, the physics and the analytical results.

As mentioned above, a common practice in statistical analysis is to try to distinguish between two hypotheses. For instance, we may want to examine two quantities to see if they are related. There are two solutions: either they are related or they are not related. However, sometimes there is not a clear distinction between the two possibilities, but there may be a degree of truth in both. For instance, the two quantities may be influencing each other, but there may also be other factors that are more important. These other factors are "noise" that affect ("contaminate") the analysis.

A standard procedure of inferring statistical relationships involves a test statistic (such as the correlation coefficient) that is estimated for the data in question. The value of this test statistic is then tested against the *null-distribution* describing a range of values that can be expected if the *null-hypothesis* were true. There are so-called *parametric tests* which assume a certain condition and have analytical solutions, but the distinction can also be made by carrying out the same test on data that we know for sure are not related but nevertheless mimic the original data. The tests are made using computers that can construct a large number of surrogate data with similar properties as the original data, but which are nevertheless randomly arranged and by definition unrelated. This approach is called Monte Carlo simulation or re-sampling testing, and is based on a large number (10,000 or more) of Monte Carlo simulations. Technically speaking, the test statistics from these tests give the distribution of the null-hypothesis. All these activities are known as hypothesis testing, and are described in more detail by Wilks (1995) and von Storch and Zwiers (1999). The basic statistical background is essential for reviewing the history of the search for a solar–terrestrial connection.

8.2 A BRIEF HISTORICAL NOTE ON SOLAR–TERRESTRIAL LINKS

It is important to emphasise the importance of the quality of the data before searching for statistical relationships. A brief account of the historical solar observations is given in Section 2.3, and the climate data are discussed in Section 5.2. It is important to bear in mind what the data can reveal about the systems and what are the weaknesses. We know that the quality in general deteriorates as we go back in time. Some of the indices commonly used, such as the sunspot number, isotope records and the *aa*-index may furthermore be "contaminated" due to the fact that they may be influenced by more than just one factor. The early sunspot observations could be affected by atmospheric conditions, the isotope records by climate variations, and the magnetic indices by the geomagnetic field. Care must therefore be taken to avoid the risk of circular reasoning. Keeping the data quality in mind, we will now take a historical perspective on the study of solar–terrestrial links. There are two aspects to the variations in solar irradiance: (i) periodic variations in Earth's orbit and (ii) physical variations in the Sun itself. Only the latter will be considered here.

8.2.1 Historical account of sunspots and hypothesised links with Earth's climate

Sunspots have been intriguing scientists for centuries and there have been many books written on this subject. Table 8.1 gives a short summary of some of the

Table 8.1. Hypotheses about sunspots and their influence on Earth through history.

Year	Name	Observation/Hypothesis
1610	Harriot	Discovery of sunspots
1611	Fabricius	Publication on sunspots
1651	Galileo, Scheiner, and Riccioli	Sunspots affect temperature on Earth: $R_z \uparrow \rightarrow T \downarrow$
1776	Horrebour	Probably a sunspot periodicity that may be important for planets (unpublished)
1801	Herschel	Price of wheat influenced by sunspots
1826	Gruithuisen	Sunspot formation ceases → fine and settled weather Great sunspots variable weather: more scattered sunspots, lesser effect
1844	Gantrér	Years with many sunspots give cold conditions; sunspot periodicity ≈10 years
1852	R. Wolf	Sunspot periodicity of ≈11.1 years
1854	Fritsch	Lower temperature during sunspot increase and vice versa
1859	R. Wolf	Solar activity curve changed
1873	Koppen	Temperature peaks in the tropics 1.5–0.5 years before sunspot minimum. Minimum temperature in the tropics coincides with sunspot maximum
1875	Chambers	SLP in Bombay approximately correlated with sunspot number
1875	Blandford	Increase in insolation → increased evaporation → clouds and precipitation
1877	Hann	Difference in yearly extreme temperature varies directly with the sunspot number. Seasonal mean T anti-correlated with sunspot number
1878	Chambers	SLP low during sunspot maximum
	J. A. Broun	SLP over south Asia anti-correlated with sunspot number
1879	Blandford	Air mass exchanged between the tropics and extra-tropics during the solar cycle
1879	Hill	Yearly temperature variations in India greatest around sunspot minimum
	Archibald	SLP in St. Petersburg positively correlated with sunspot number
1880	Chambers	SLP variations follow months after sunspot number; links R_z with famines in India
1886	N. Lockyer	Spectroscopic investigation of the Sun → hottest at sunspot maximum
1888	Liznar	Diurnal temperature range lower during sunspot maximum
1891	Unterweger	Short sunspot periodicity of 26–30 days discovered
1894	Bigelow	Periodicity in variation in terrestrial gravitational forces of 26.68 days
1896	MacDowall	Daily temperature extremes more frequent around sunspot maximum
1897	Rizzo	Low summer T in Turin about 3 years after sunspot maximum, discernible minima in the geomagnetic field → 14-day periodicity
1900	Brückner	30-year variations in T, SLP and precipitation 34.8 ± 0.7 year periodicity from Sun (not sunspots)
	N. and W. Lockyer	Two solar spectral lines with opposite variations associated with the solar cycle Prominence cycle with timescale of 3.5–3.7 years
1901	W. Lockyer	Variations in sunspot cycle length (SCL)
1902	Richter	T in Europe compares with sunspot number, northern lights

		and the declination of the geomagnetic field
1903	Arrhenius	Time of biological activity in Sweden related to sunspots
	Nordmann	11-year T cycle in the tropics anti-correlated with sunspot number
	Angot	15 11-year T-cycle records anti-correlated, 2 correlated with sunspot number
1904	Langley	Solar radiation reduced by 10% from end of March 1903 to the end of the same year
1905	Schuster	Sunspot cycle of 33.375, 13.57, 11.125, 8.38, 5.625, 4.81, 3.78, and 2.69 years
1907	Abbot and Fowle	T-variations probably directly related to variations in insolation
1908	Bigelow	11-year cycle in the horizontal component of the geomagnetic field, temperature, humidity, and SLP over Europe 2.75-year cycle. Inversion of temperature related to prominences
	Hann	Winter and summer T high during sunspot minimum
	Newcomb	T maximum 0.33 years before sunspot maximum
1913	Abbot and Fowle	Volcanic eruptions block out light View that S lower during sunspot minimum probably wrong
	Humphreys	Small dust particles little greenhouse effect, but interferes with short-wave 11-year variation in violet and UV light with less at sunspot maximum. More O_3 and T increases
	Mielke	Strongest solar signal in the tropics and warmest years coincide with sunspot minimum
1908–1915	Arctowski	"pleions" and "anti-pleions": regions which vary out of phase and in-phase with the sunspot number 2.77-year cycle
1917	Kregnes	Magnetic storms and meteorological variations Variations in geomagnetic field at least as good predictors as the sunspots
	Meissner	Sunspot maximum gives low temperature and high rainfall over Berlin Different relationships during different seasons
1976	Eddy	Discovery of "the Maunder minimum"
1977	Kelly	Solar activity influence on SLP
1987	Reid	Solar activity influence on SST
1988	Labitzke and Van Loon	Solar activity influence on equatorial wind patterns
1989	Tinsley	Variations in SCL affect cyclogenesis
1991	Friis-Christensen and Lassen	Variations in SCL related to land surface temperatures

milestones in solar–terrestrial science.[1] One early review is given by Helland-Hansen and Nansen (1920), who devoted hundreds of pages to the topic.

Various scientists have been searching for a relationship between sunspots and the Earth's weather, starting with Galileo, Scheiner and Riccioli in 1651 (Helland-Hansen and Nansen, 1920, p. 147). The amount of effort that has gone

[1] A number of the references can be found in Helland-Hansen and Nansen (1920).

into investigations of possible associations of climatic variations with sunspots is illustrated by a quotation from Bray and Loughhead (1964, p. 238): "literally hundreds of such investigations [solar–terrestrial relationships] have been published in which this index [sunspot number] has been used. Moreover, it is the only index which reaches so far back in time." Most of the solar–terrestrial link propositions known today were put forward in the 19th century, when the sunspots records became sufficiently long. Wolf made a sunspot reconstruction extending back to 1611, but found no sign of sunspots between 1645 and 1715. One may speculate about whether the absence of sunspots from the records between 1645 and 1715 is due to lost records, to political or religious reasons, or to times of hardship and famine. Similar speculations have been made about observations of the northern lights according to Brekke and Egeland (1994), as there were few records of the northern lights between 1795 and 1825 over North America. Some suggestions are that this is attributable to there being few observers capable of writing or that famine prevailed in the early 19th century; but, of course, the features may have also been absent from the Sun.

It is certain that the hypothesised relationship between sunspots and weather on Earth, suggested by Galileo, Scheiner and Riccioli in 1651, could not have a solid empirical basis. By that time, the sunspot record was brief, the cyclical nature of the sunspots had not yet been discovered and the concept of statistical analysis must have been primitive. Therefore, the idea put forward by Galileo, Scheiner and Riccioli was most likely to have been based on pure speculation.

8.2.1.1 The solar cycle and surface temperatures

Table 8.1 suggests that speculations about a link between the sunspots and temperatures on the Earth dates more than two centuries back. In 1826, Gruithuisen thought that fine and settled weather followed the decline of sunspot activity. Gantrér found that the years coinciding with many sunspots were cold. He also proposed that there is a 10-year cycle in the sunspot recurrence.

Fritsch (1854) and Wolf (1859) related cold conditions with an increase in sunspot activity. In 1873, however, Köppen found that the temperature in the tropics peaked around a year before the minimum sunspot activity and that cold conditions coincided with sunspot maximum.

Köppen had in 1873 assembled temperature records from 403 stations representing 5 different climatic zones and 25 different regions over the whole world. He found that warm conditions had a tendency of taking place $\frac{1}{2}$ to $3\frac{1}{2}$ years before sunspot minimum. The timing of the warm episodes was more retarded further away from the equator. Cold anomalies, however, coincided with the timing of the sunspot maximum. These observations, therefore, were in good agreement with Gantrér, Fritsch and Wolf. Thus, an inverse relationship[2] between sunspots

[2] Here defined as anti-correlated, i.e. having a negative correlation coefficient: high sunspot number gives low temperature and vice versa.

and the temperature was observed for the period 1816–1859, but before and after this period, there was only a slight degree of correspondence. Furthermore, Köppen found in 1881 that the relationship over the 1859–1875 period disagreed with the results for the 1816–1859 period.

Wolf proposed that the solar activity had changed behaviour in 1859, but Blandford proposed in 1891 that the disagreement after 1860 was mainly due to lack of exact observations over the period 1875–1885. The temperature in Allahabad was allegedly inversely related to the sunspots with 3.7°C higher values at sunspot minimum than sunspot maximum.

Lockyer started the first spectroscopic solar studies and proposed that the Sun is probably hotter during the time of sunspot maximum, which seemed to contradict the earlier notion of the inverse relationship between sunspots and the temperature. Blandford (1875) explained this paradox by proposing that more solar energy leads to increased evaporation and more clouds. Thus, the concept of feedback mechanisms had already been brought into the study of the solar–terrestrial link in the 19th century. Increased cloud cover reflects more sunlight (increased albedo), but the increased evaporation of precipitated rainwater near the surface also reduces the surface temperature.

In 1877, Hann investigated the temperature of different seasons and related this to the sunspots. As a general rule, the temperature was anti-correlated with the sunspots, but the strength of correlation varied with the seasons.

S. A. Hill suggested in 1879 that the yearly variation in the temperature over northern India was greatest around the time of sunspot minimum and lowest during sunspot maximum. But this study was based on a short time interval and there were some large departures from this rule. Hann (1879) proposed that the yearly extreme temperature in Leipzig varied directly with the sunspot curve. At sunspot maximum, he observed the highest maximum temperature and lowest minimum temperature. Hann's observations were confirmed by Liznar (1880) who had carried out similar analysis on temperature records from eight other stations in Europe. The daily temperature in Vienna, Prague, Tuschaslau, Brünn and Trieste between 1857 and 1870 was lowest in 1859–1860 and 1870–1871 when sunspot activity peaked.

MacDowall observed in 1896 that "in sun spot maximum years a greater number as well of very hot as of very cold days occurs than in sun spot minimum years" (Helland-Hansen and Nansen, 1920, p. 155). After applying a 5-year smoothing filter he demonstrated a good correspondence between the August temperatures (from Geneva) and the inverted sunspot curve for the five cycles during 1830–1883. The curves went apparently in opposite directions after 1883.

Rizzo (1897) observed that high summer temperatures in Turin since 1752 tend to follow approximately 3 years after the sunspot maximum. His observations were reasonably consistent with the observation of warm events between $\frac{1}{2}$ and a $1\frac{1}{2}$ years before sunspot minimum made by Köppen, and colder conditions during the rising phase of the solar cycle (Gantrér, Fritsch and Wolf).

C. Nordmann (1903) investigated the annual mean temperature from 13 stations for the 31-year interval 1870–1900 and found an 11-year cycle. The temperature in the tropics had a negative correlation with the sunspot number. In 1903, A. Angot looked at 17 temperature records and found 15 which were anti-correlated, and 2 that had a positive correlation (Bombay and Barbados) with the sunspot number. Easton proposed in 1905 that the approach of cold winters in the last 300 years gave the best indication of the influence of prominent solar activity events on Earth's climate.

Hann suggested in 1908 that both winter and summer temperatures are high during times with high sunspot activity and low during sunspot minima. He had based his conclusions on 100 years of observations from Vienna.

In 1908 the data analysis became more advanced when Newcomb used a "special mathematical" technique to analyse various temperature series from widely different locations over the period 1871–1904. He found that warm periods appear 0.33 years before the sunspot minimum. These results are apparently inconsistent with Rizzo's observation of high summer temperatures following 3 years after sunspot maximum.

Abbot and Fowle (1908) estimated a mean value from 47 temperature records representing eight regions: North America (15), South America (1), middle and east Europe (8), North Africa (2), South Africa (2), North Asia (7), South Asia (6) and Australia (6). They were among the first to attempt to compute a "global" value, and managed to associate this mean curve with 11-year cycles anti-correlated with sunspots.

In several publications between 1908 and 1915, Arctowski discussed climatic and temperature variations and concluded that they varied in step with the variations in solar activity. There are pronounced 11-year cycles, but these do not vary coherently over the globe. The temperature varies inversely with the sunspots over most of the world, but there are some regions in which the temperature and sunspot number are positively correlated. Arctowski invented two new concepts: (i) pleions and (ii) anti-pleions. During the years of sunspot maximum, the pleions are regions of positive temperature anomalies, and can be regarded as small "islands" surrounded by cold conditions. The anti-pleions are those regions which are characterised by cold anomalies during sunspot minima, and these are surrounded by warmer than normal conditions elsewhere.

In the tropics (Arequipa, Peru), Arctowski found stations from which the temperature contained some irregularities with prominent timescales of 2.75 years. Such variability had also been reported by Bigelow and the Lockyer brothers. Arctowski thought these variations were unrelated to the 11-year cycle, but he did nevertheless think that they were caused by the shorter-timescale variations in solar activity. We will refer to these "2.75-year" variations when the El Niño Southern Oscillation is discussed in Section 9.2.

In 1913, Abbot and Fowle argued that the then-established view that the total solar irradiance ("solar constant") is greater at sunspot minimum than sunspot maximum was probably wrong. This conclusion was problematic for Arctowski, who had in 1907 put forward the hypothesis that the temperature variations were most probably directly related with variations in solar activity. In later papers, Arctowski examined the possibility of atmospheric circulation playing a more important role.

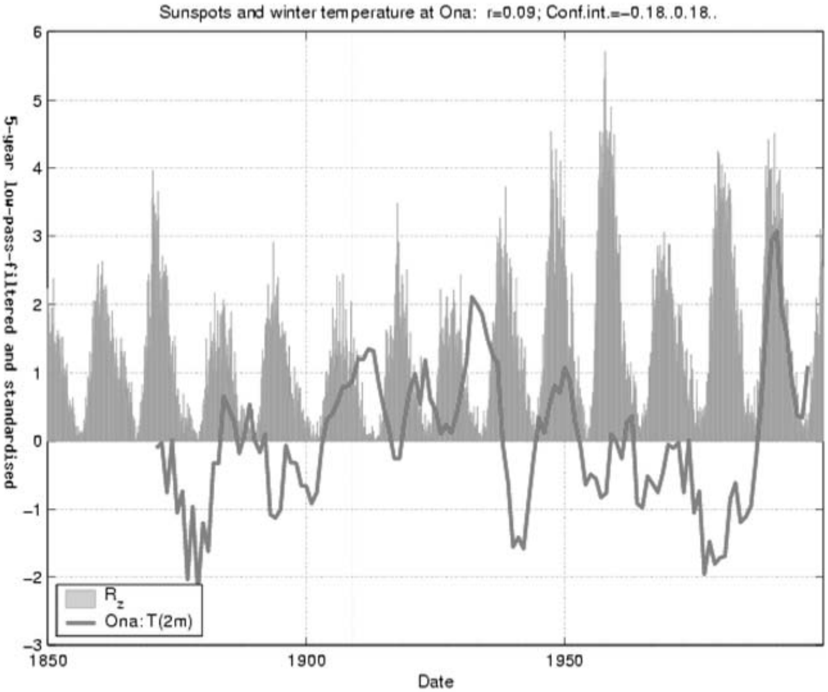


Figure 8.1. Correlation between the winter-time temperature at Ona and the sunspot number. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and the Norwegian Meteorological Institute.

In 1909 Helland-Hansen and Nansen published a study of the winter (1 November–30 April) temperature from the Norwegian lighthouse Ona over the period 1874–1907, in which they showed that the temperature changed in the same way as the sunspots so that high winter temperatures correspond to high sunspot numbers. Here we revisit this study of correlation between Ona winter temperature and the sunspot number, but with updated records. Figure 8.1 shows a time-series plot of standardised time-series representing the Ona winter-time temperature and R_z, and serves as an illustration of how statistical correlations can be explored. Although this example cannot be used to prove[3] or disprove the hypothesis on a connection between sunspots and Earth's climate, it demonstrates that the relationship that Helland-Hansen and Nansen described is either non-stationary or not robust. There are periods when the winter temperature at Ona seems to vary directly with the sunspot number, especially during the last solar cycle. Between 1900 and 1950, on the other hand, the temperature appears to be anti-correlated with the sunspot number. A similar switch between in-phase and out-of-phase variations was made as early as in 1920 by Helland-Hansen and Nansen (p. 151).

[3] If one searches among a vast number of independent time-series, the chances may favour one that is correlated with the sunspot record due to the problem of multiplicity (Wilks, 1995).

8.2.1.2 The solar cycle and sea level pressure (SLP)

C. Chambers noted in 1875 that the SLP measured in Bombay had similar periodicity to the sunspots, implying some sort of relationship (Helland-Hansen and Nansen, 1920, p. 166). His brother, F. Chambers (1878) proposed that a sunspot maximum in winter and summer subsequently results in low SLP. Broun (1878) suggested that the SLP over India and south Asia is negatively correlated with the sunspot number. The notion of a negative correlation between the sunspot number and SLP was supported by S. A. Hill (1879) and Blandford (1879, 1880). Blandford proposed that a larger portion of the tropical atmosphere "is transferred" to high latitudes during sunspot maximum, thus disturbing the atmosphere in the transition region between the tropics and the circumpolar zone. He also suggested that the most important factor lowering the SLP is an increased evaporation during solar maximum. However, E. D. Archibald (1879) found evidence suggesting that the SLP in St Petersburg is positively correlated with sunspots (Helland-Hansen and Nansen, 1920, p. 167). In 1880, F. Chambers put forward the notion that the SLP variations lag the sunspot number by several months. He also connected famines in India with these SLP variations (Helland-Hansen and Nansen, 1920, p. 169).

8.2.1.3 The solar cycle and geomagnetism

Some of the first accounts of a relationship between the solar cycle and Earth's magnetic field can be traced as far back as the late 19th century. In 1891, Unterweger thought he had found a short period of 26–30 days in the sunspots and in solar activity, but also a periodicity of 69.4 days. This cycle was partly related to the Sun's rotation (29.56 ± 0.5 days), but was thought not to be a direct product of solar rotation. F. Bigelow found in 1894 a periodicity in the terrestrial magnetic forces with a similar timescale (26.68 days). Discernible minima were measured on day 1–2, 5, 9, 15, 20, and 24 of each 27-day solar rotation. Conversely, maximum intensity was observed on the 3rd, 7th, 11th, 14th, 16th, 19th, 22nd, and 26th day. Although, it was scarcely mentioned by Bigelow, he had uncovered a 14-day cycle. However, he suggested that the Sun sends unequal quantities of energy during its rotation, depending on the meridian. When he looked at the temperature records, he found that in some places the temperature varied coherently with the magnetic force, but he also found cases where the temperature and the magnetic field were anti-correlated.

William Lockyer studied the magnetic epochs and the variations in the sunspot cycle length, and observed in 1901 that the time between minimum and maximum sunspots varies regularly with a cycle of 35 years. Similar periodicity in the sunspots and the horizontal magnetic intensity was found by Brückner in 1902, and in 1905 Schuster estimated the period to be 33.375 years.[4] Other periodicities were also found: 15.57, 11.125, 8.38, 5.625, 4.81, 3.78, and 2.69 years.

[4] The accuracy of this estimate is incredible, which raises the question of how errors and uncertainties were accounted for.

8.2.1.4 Prominences and temperature

Helland-Hansen and Nansen (1920, p. 141) presented empirical evidence pointing to a relationship between prominences and surface temperature on Earth. Past studies by Bigelow have suggested that the temperature (1873–1900) in the tropics increases with the number of prominences, but in places such as Japan, China, southeast Russia, central Europe, Iceland and east Greenland, the temperature was observed to drop when there were many prominences (Helland-Hansen and Nansen, 1920, p. 152). There were furthermore places where the temperature did not indicate any relationship with the prominences, such as the higher parts of India, middle Siberia, southwestern Russia, and the east coast of the USA. Temperature is not the only meteorological element that seemed to be affected by solar activity. The Lockyer brothers (1904–1908) suggested that the air pressure in some regions (e.g. the Indian Ocean) varies in phase with the prominences, but in other parts of the world (e.g. South America) it varies out of phase with them.

In 1908, Bigelow reported an 11-year cycle in the horizontal magnetic field intensity over Europe, but also similar periodicities in the air temperature, vapour pressure, and air pressure in different regions of the United States. The temperature and the vapour pressure were inversely related to the prominences and variations in the magnetic field. This 11-year cycle found in the USA was most pronounced along the Pacific coast and was weak east of the Rockies. This inverse relationship was explained as being caused by horizontal atmospheric circulation induced by solar activity. Bigelow also found shorter timescales of about 3 and 2.75 years which were pronounced everywhere. However, this was proportional to the magnetic variations and the prominences in the western USA but inversely related to those in the east. Thus, there was a phase displacement in these meteorological quantities towards the east.

According to Helland-Hansen and Nansen (1920, p. 154), Brückner published in 1900 a classical study on climatic variations since 1700, and remarked on a pronounced cycle of 36 years in the air temperature, pressure and rainfall. His investigations also extended to archaeological data such as records of ice conditions of rivers, wine harvest dates and the frequencies of strong winds, and determined the period of the cycle to be 34.8 ± 0.7 years.[5] These climatic variations, Brückner thought, were not related to sunspots, but rather oscillations in the energy coming from the Sun. This cycle furthermore appeared to be less pronounced in the tropics than in the higher latitudes.

8.2.1.5 Solar-phenomenological links

Flammarion considered the times of grape formation in France, the times of vintage, and the blossoming times for different plants, and related these to sunspots. In Sweden, S. Arrhenius (1903) also looked at phenological phenomena. These studies suggested that spring months in years of high sunspot activity were warmer than when there were few sunspots.

[5] This accuracy is questionable, considering the short instrumental record (1700–1900) and the inherent uncertainty associated with archaeological data.

Table 8.2. Number of references to studies of solar–terrestrial links from various periods, selected from their bibliographies. Only the references with a reference to solar–terrestrial links and variations in TSI (usually determined from the titles) have been counted.

Review	Before 1940	1940–1959	1960–1979	After 1980	Total
Lean and Rind (1998)	0	0	5	61	66
Bertrand and van Ypersele (1999)	0	0	6	47	53
Harrison and Shine (1999)	0	1	3	46	50
Helland-Hansen and Nansen (1920)	51				51
This book	26	2	5	101	134

8.3 RECENT STATISTICS ON SOLAR–TERRESTRIAL LINKS

8.3.1 A renaissance for solar–terrestrial links

By the middle of the 20th century, enthusiasm about the idea that solar activity controls our climate appear to have waned to give place to more sceptical dispositions. A review of studies on solar activity and Earth's climate in recent works suggests that most of the citations on solar–terrestrial studies allude to recent studies[6] (Table 8.2). There are surprisingly few citations for the period before 1960. This drop in citations may reflect the contemporary trend of little research on solar–terrestrial links. However, the book by Helland-Hansen and Nansen (1920) lists at least 51 citations to publications on solar–terrestrial work, all published before 1920, suggesting that much of the early work seems to have been "forgotten" by the later generations of scientists. The question is therefore whether there also were similar studies carried out between 1920 and 1980 that have been "lost" in various archives, or if there really was a drop in the number of published works on solar–terrestrial links (or the poor quality of some of the earlier papers?). There is a book of proceedings about solar–terrestrial relations published in 1965 (Orther and Maseland, 1965), but this discusses solar physics, the interplanetary medium, the ionosphere, auroras and magnetic storms rather than the climate near the Earth's surface. It is interesting to look at different explanations for why the interest in the solar–terrestrial links declined. The search for a solar–terrestrial link did often lead to over-stated claims and was turfed on dubious statistics (Haigh, 2001) which over time may have given the subject a bad name. In 1917 the polar front theory was introduced by V. Bjerknes and colleagues. Weather development was successfully described in terms of fronts separating cold and warm air, which led to dramatically improved weather prognoses. Thus, at the time sunspots seemed to play a lesser role.

Some accounts from the period 1920–1980 seem to support the idea of a change in attitude towards solar–terrestrial links. Godske (1956) describes the attempts by

[6] Selected according to the title of the publication.

H. C. Willett to use the sunspots to predict the temperature for the next 10–15 years as brave and foolish. Willett, however, predicted cooling from the 1950s to 1960–1965, which was not far from the truth. Monin (1972) was also sceptical towards the idea of solar activity's influence on Earth's climate. There were nevertheless some publications on solar–terrestrial links in this period. Clough (1943) wrote a paper on the relationship between the solar cycle length and climatic variations and Bjerknes (1959) speculated about whether decadal and centennial variability in the ocean may be due to solar influences.

It is now recognised that the weather is largely governed by "internal" atmospheric variability ("chaos"), but over long timescales the weather statistics may still be influenced by external factors. For instance, Bond *et al.* (2001) reported a centennial-scale cycle in the North Atlantic drift ice over the past 12,000 years and proposed that these southward excursions of the ice edge were associated with periods of reduced solar output. Thus, the success of the polar front theory does not exclude the possibility of other factors playing important roles on longer timescales. One example of how external forcing may affect the weather statistics is the seasonal cycle, causing colder and stormier conditions during winter away from the tropics. This acknowledgement brings us up to the recent work on solar–terrestrial links. But, in order to appreciate the various results, we will first give a brief review of some relevant statistical concepts.

8.4 RECENT WORK AND HYPOTHESES

A comprehensive report written by Harrison and Shine (1999) gives an up-to-date review of more recent progress in understanding how solar activity may affect Earth's climate. They divide this subject into several categories: the direct effects of changes in the solar irradiance, effects of solar-induced changes in the middle atmosphere, and cosmic rays and their connection to Earth's climate. There are also some recent papers, such as by Lean and Rind (1998) and Bertrand and van Ypersele (1999), that give a brief but comprehensive review of the most up-to-date developments in this field. Haigh has also published a number of recent papers concerning stratospheric response to solar variability.

8.4.1 Basic statistical concepts

Much of the knowledge about solar–terrestrial relationships rests on empirical observations and data analysis. Therefore, it is important to have a basic understanding of the statistical concepts employed in order to understand the significance of the hypothesised relationships. The following section gives a brief overview of the most basic concepts and definitions of a number of statistical quantities.

In many cases, the objective is to study the relationship between some climate element and some time-series. The most pronounced variations in Earth's climate are associated with the diurnal cycle (night and day) and the seasonal cycle. These periodic variations are driven by well-known variations in the insolation at a

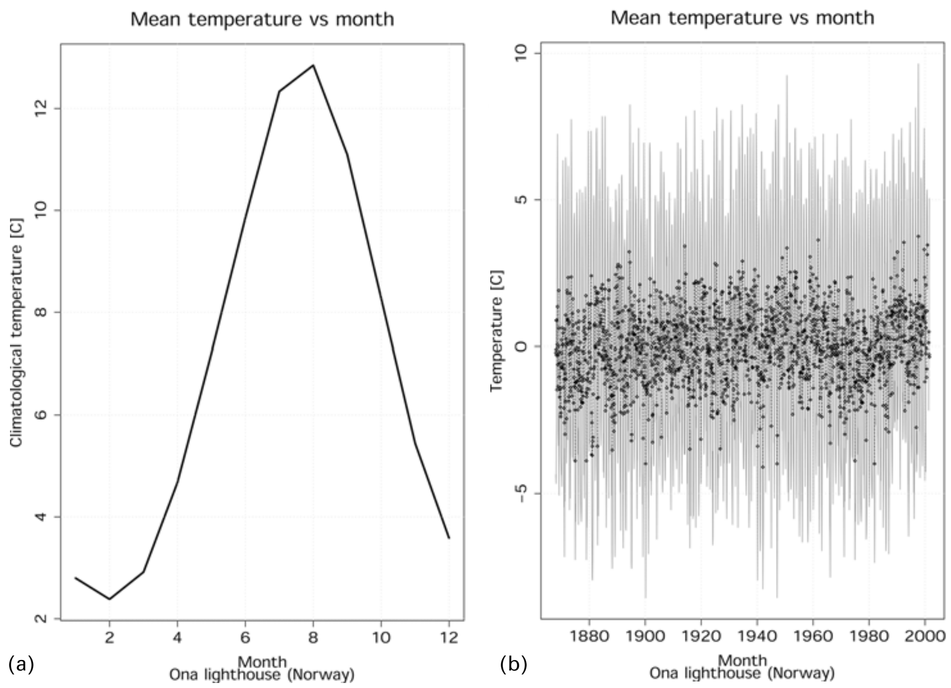


Figure 8.2. (a) The temperature climatology at Ona lighthouse in Norway. (b) The anomalies (black diamonds), are defined as the difference between the total temperature and the climatology. Also shown in grey are the temperature deviations from the all-record mean value (anomaly + climatology). The data were obtained from the Norwegian Meteorological Institute's climate database, and the time interval is 1868–2001.

given location, and tend to be stronger than other signals such as the influence of solar activity, and may therefore complicate the analysis (‘‘contaminate’’ the results). In such cases, it may be desirable to ‘‘filter’’ out the seasonal or diurnal cycle since we already know they are unrelated to the phenomena that we are studying. The seasonal or diurnal cycle can easily be removed from the record so that the signal of interest can be enhanced.

8.4.1.1 The seasonal cycle

Typical seasonal values are often referred to as climatology, and are estimated by taking the mean value of the respective calendar months. The climatology of the temperature at Ona lighthouse is shown in Figure 8.2(a) as an illustration, where the temperature is plotted against the 12 months January to December.

8.4.1.2 Anomalies

Anomalies are the deviation of a variable from its mean value, and have by definition zero mean value.

Here, we denote the anomalies by a prime:

$$x' = x - \bar{x} \tag{8.1}$$

It is also common in climate studies to define the anomaly with respect to the seasonally varying mean value (climatology), for example that anomalous values for the January month are all January observations minus the climatological value for January (mean value for all January months). The temperature anomalies at Ona lighthouse are shown as black diamonds in Figure 8.2(b). The anomalies are compared with the annual cycle (grey).

8.4.1.3 Power and variance

Sometimes it is of interest to know how strongly a quantity changes or fluctuates over time, for instance, if one would like to know the amplitude, i.e. the strength, of a signal. One measure for the signal strength is the power of the time-series, which is a common measure for magnitude of the variations. In statistics, the power is also referred to as the variance. The standard deviation is the square-root of the variance and gives another measure of the magnitude of the variations. Figure 8.3 shows the

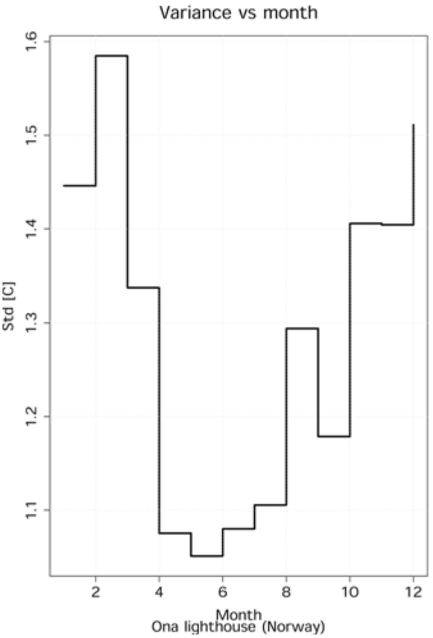


Figure 8.3. The standard deviation of monthly mean temperature as a function of calendar month. Strongest inter-annual variations are seen during winter (February) and weakest in late spring (May).

standard deviation of the temperature for each calendar month recorded at Ona lighthouse. This figure indicates that there is more pronounced inter-annual variations during winter than in summer.

The mathematical definition of the variance is:

$$\sigma^2 = \frac{1}{(N-1)} \sum_{n=1}^{N} (x_n - \bar{x})^2 \tag{8.2}$$

The power estimated in the time-domain (equation (8.2)) is the same as the power estimated with spectral methods (frequency-domain). A physical analogy of power is the mean energy flow, and for some wave signals such as light, these become identical. The standard deviation, σ, is defined as the square root of the variance. Sometimes, if the signal strength is not important, analysis is carried out on dimensionless quantities. One process of making a series dimensionless is referred to as standardisation.

8.4.1.4 Standardisation

A standardised time-series has by definition zero mean value and a standard deviation of one. Thus, by subtracting the mean from the data and then dividing by its standard deviation, a time-series is standardised: $\hat{x} = (x - \bar{x})/\sigma$. The standardised series is dimensionless, and is handy when comparing various time-series with different physical units or variance, as shown in Figure 8.4 (see colour plate section) which shows the sunspot number curve[7] (bars) as well as the standardised SCL (black), global mean SST (blue) and the global mean temperature (red).

In the meteorological and climatological communities, the word "normalisation" is often used when meaning "standardisation", but in the statistics community, normalisation also implies imposing a normal distribution onto the data (through a transformation).

8.4.2 Linear trends

Sometimes it is of interest to examine the long-term rate of change of a given quantity. The simplest type of change is a constant rate of change, i.e. where a variable changes at the same rate regardless of time. The term "linear trend" is often used to describe this constant rate of change, or the slope m of a straight line: $y = mx + c$. The slope m and constant c can be found through an ordinary regression analysis (see Section 4.8.2.1). The best-fit linear trends of the hemisphere mean temperature (dark grey) are shown as dashed lines in Figure 8.5: the slopes of

[7] The sunspot number is technically not standardised since its mean value is greater than zero.

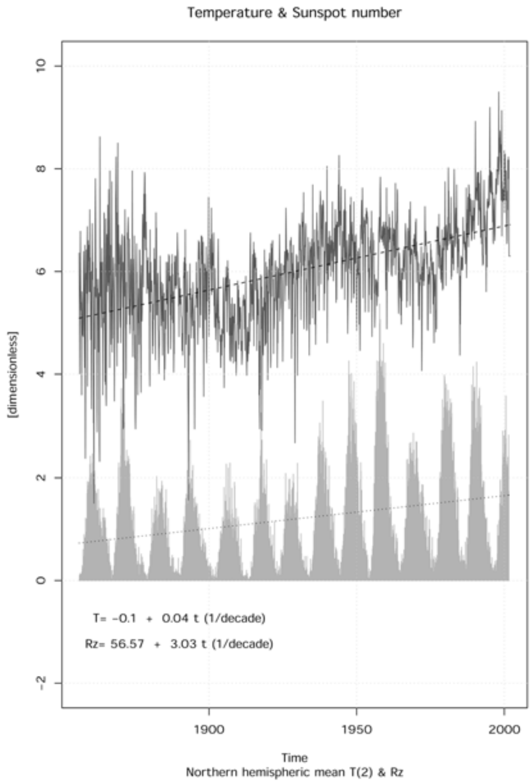


Figure 8.5. Estimated linear trend (black) in the northern hemisphere mean temperature anomalies from Jones (dark grey) suggests a mean warming of 0.04°C/decade warming over the period 1856–2001. A similar analysis carried out for the sunspot number (grey bars) gives a trend of 3.03 numbers/decade. The data shown have been standardised. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and CRU.

these trends correspond to 0.04°C/decade. A corresponding trend analysis for the sunspot number (R_z is shown in grey) yields a trend estimate (dotted line) of 3.03 spots/decade. It is important to keep in mind that these trend estimates only give an approximate description of the long-term changes over the time interval for which the trend analysis was applied. The northern hemisphere temperature shown in Figure 8.5, for example, has a weaker warming rate in the beginning of the record and faster changes towards the end. Regarding the sunspot number, the highest number 254 was observed in 1957, and the maximum sunspot number at solar maxima has decreased since then, and may perhaps continue to do so in the near future.

8.4.3 Correlation

Correlation analysis is a common method used to examine the similarities between two quantities.

There are various ways of making a correlation analysis, however the most common method is the *Pearson correlation*, $r(x_n, y_n)$ which is defined as:

$$r(x_n, y_n) = \frac{\sum_{n=1}^{N} x'_n y'_n}{\sqrt{\sum_{l=1}^{N} x'^2_l \sum_{m=1}^{N} y'^2_m}} \tag{8.3}$$

In equation (8.3) x' and y' are anomalies of x and y respectively and are defined according to equation (8.1).

The correlation estimate is invariant with respect to scaling or constant offset: $r(x, y) = r(a_1 x + c_1, a_2 y + c_2)$. The analysis shown in Figure 8.4 gives a (Pearson) correlation of $r(R_z, T[2m]) = -0.18$ between the *unfiltered* monthly mean sunspot record and *low-pass*-filtered global mean temperature. For the SCL and temperature the correlation is $r(\mathrm{SCL}, T[2m]) = -0.16$. The former result is considered as statistically significant at the 5% level, as there is a 95% probability of two similar but random numbers achieving a correlation of 0.05. The correlation between the sunspot number and the global mean temperature is usually reported to be positive and the TSI is positively correlated with R_z. The questions are then: what do the results from such an analysis mean, and how reliable are the confidence estimates? The answer to these questions can be explored through hypothesis testing, discussed in Section 8.4.3.1.

The results for SCL and temperature, on the other hand, are not statistically significant (lower correlation and far fewer data points). The analysis in Figure 8.4 disagrees with the conclusion of Friis-Christensen and Lassen (1991) who suggested there is a relationship between the SCL and the terrestrial temperature. We will return to the study of Friis-Christensen and Lassen (1991) in more detail in Section 8.4.8 and use it as a case study after a brief review of statistical significance, the effects of de-trending and filtering the data prior to correlation analysis, and Monte Carlo simulations.

8.4.3.1 Statistical significance of correlation

Analysis of an *infinite* series provides a definite answer of whether two quantities are related or not, as a non-zero correlation implies that the two series always will fluctuate in synchronisation to some extent. However, infinite series of observations available for analysis do not exist, and we have to make do with *finite* series. In the analysis of finite series, there is a possibility that there is a coincidental resemblance between two quantities examined. The shorter the series, the higher the risk of accidental resemblance. It is therefore important to assess whether the analytical results are likely to be due to chance or whether such a fluke is unlikely. This is known as *significance testing* or *statistical inference*.

It is important to stress that inference tests only give a "rule-of-thumb" measure of the distinction between the null-hypothesis and the alternative hypothesis.

The statistical significance of the Pearson correlation (r) of two series with a small number of independent data points ($N \sim 20$) can be estimated using the formula (Press *et al.*, 1989, p. 533):

$$t = r\sqrt{\frac{N-2}{1-r^2}} \tag{8.4}$$

Here the parameter t is a test statistic that is a measure of the statistical significance and follows a t-distribution which can be found in standard statistics textbooks. Equation (8.4) is not always an appropriate test of statistical significance. It is, for instance, not a good choice for long time-series.

When all else fails, a significance test may be carried out using brute force and letting the computer crunch through the numbers. Such approaches include the Monte Carlo integration approach. Before the Monte Carlo integrations can be carried out, the data must be properly pre-conditioned since the outcome of these tests may be biased unless proper care is taken. Time-series may for instance have to be de-trended prior to these integrations in order to avoid the test being pre-disposed. Correlation analysis may also give misleading results, especially if an unsuitable filter is applied to the series before the correlation analysis. These two types of biases will be discussed in Section 8.4.7.

8.4.4 Correlation and de-trending the data

One common mistake is to apply correlation and regression analysis to series which contain long-term trends if it is not known *a priori* that these are part of the signal. Such trends may bias the analysis by, for instance, inflating the correlation scores if they are caused by external factors, especially for short data series. It is often wise to de-trend the data (here the least-squares fit linear trend has been subtracted) in order to reduce such biases.

De-trending can be justified in terms of a simple mathematical model: let $y = mx + c$ [1] describe a linear relationship between two quantities. Both quantities may have a long-term trend with short-term variations superimposed on this: $y = y_{dt} + y_{tm}$, $x = x_{dt} + x_{tm}$. The term y_{tm} contains the mean of y as well as the best-fit linear trend, and y_{dt} is the de-trended part of y with a zero mean. Equation [1] implies that

$(y_{dt} + y_{tm}) = m(x_{dt} + x_{tm}) + c$, or $y_{dt} = mx_{dt} - (y_{tm} - x_{tm} - c)$ [2]. If the linear trends in x and y are related (not influenced by external factors), then the expression in parenthesis in equation [2] equals zero and we are left with $y_{dt} = mx_{dt}$ [3], which relates short-term excursions in x to short-term variations in y. If, on the other hand, the trends are influenced by external factors, then the non-de-trended analysis will be biased by these factors. It is less likely that short-term variations are correlated with other (external) factors by chance than that long-term trends (linear slopes) are, because the former have more degrees of freedom.[a]

[a] It is less likely that two random sophisticated patterns resemble each other than two random simple patterns.

A demonstration of the inflation of the correlation score by non-zero long-term trends can be made by letting a computer[8] generate two stochastic series and then add a linear trend to these (Figure 8.6(a,b)). The correlation is computed for these non-zero trend series as well as the de-trended series (Figure 8.7(a,b)). Hence, it is important to remove the long-term trend in the data (a process called "de-trending") before calculating the correlation if the trends may be caused by factors other than those explained by the hypothesis. Correlation analysis on short data series with substantial non-zero trends often produce misleading results if these trends are due to external factors. The chance of a random trend being positive or negative is 50%–50%.

8.4.5 Correlation and filtering the data

Low-pass-filtering may introduce spurious results if care is not taken. Low-pass-filtered signals tend to give too high correlation, and sometimes low-pass-filtering may produce spurious undulation, an effect known as the *Slutsky–Yule effect*. It is therefore important to examine the sensitivity of the analysis to the filter type and width. The effect of low-pass-filtering[9] on the correlation analysis is easy to demonstrate on a home computer by repeating the analysis many times but after applying different filters. Figure 8.8 (see colour plate section) shows the results of series of computations where the effect of the filter window width has been explored systematically. This demonstration shows clearly that the correlation estimate is inflated by low-pass-filtering.

The same analysis has been repeated but applying the low-pass filter only to the temperature (to remove the contamination from the annual cycle) but not to the sunspot number (blue). Both records have been de-trended in this example. The results explain the negative correlation between the sunspot numbers and the temperature in Figure 8.4, as the correlation estimates are scattered about zero. Hence,

[8] For instance using an analytical tool such as R, S-plus, Matlab, Octave, or IDL.

[9] Also referred to as smoothing.

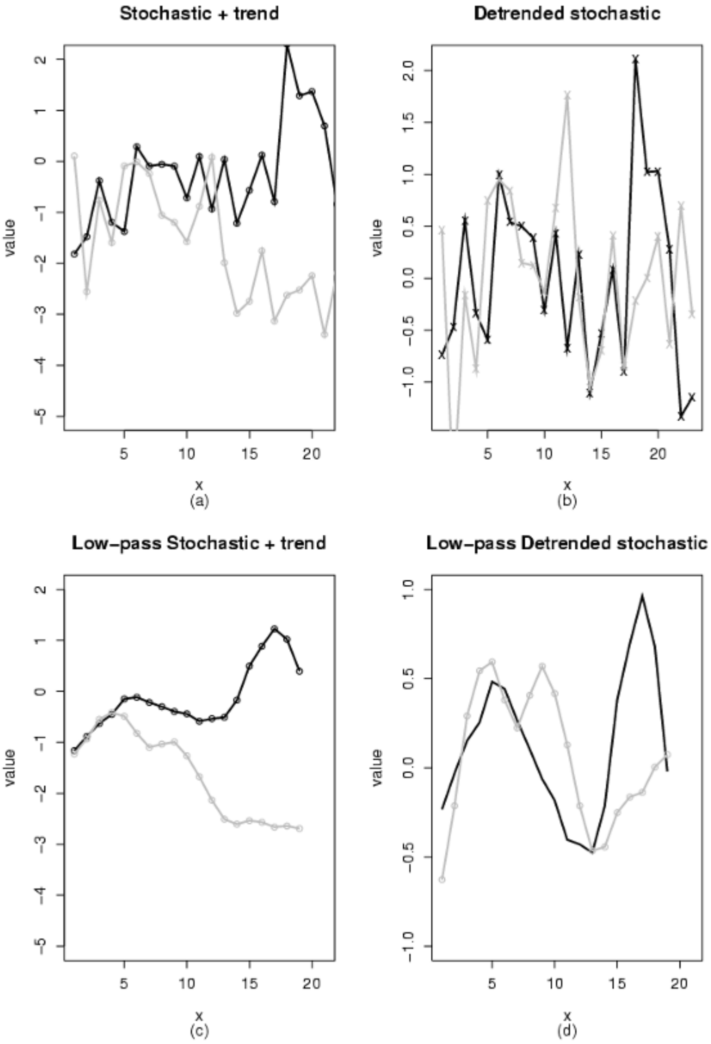


Figure 8.6. Two stochastic series where the numbers have been produced using a random number generator. The panels show the series with non-zero trend (a), de-trended series (b), low-pass-filtered series before the trends have been removed (c), and low-pass-filtered de-trended series (d).

this demonstration shows that if only the temperature is smoothed, it is possible to get *negative* correlation estimates, as seen in Figure 8.4. The demonstration can be repeated for stochastic (by definition unrelated) series as well as ‘‘synthetic’’ data with an imposed relationship as in Figure 8.9 (see colour plate section), which shows the results for pairs of unrelated stochastic series (light grey), and partially related series (red and blue). For a strong relationship, the correlation is close to unity for all

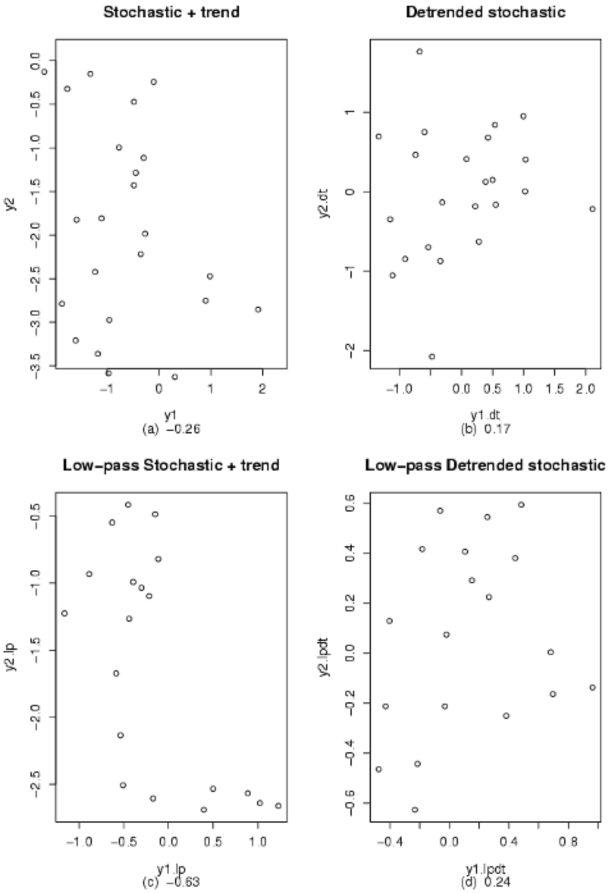


Figure 8.7. Correlation between two stochastic series. The panels show the scatter plots for the series with non-zero trend (a), de-trended series (b), low-pass-filtered series before the trends have been removed (c), and low-pass-filtered de-trended series (d). Any correlation is purely coincidental.

different filter widths and hence results are robust. The weakly related series, on the other hand, give fluctuating correlation estimates as the filter width is varied. Hence, if the signal-to-noise ratio is low, then the estimates fluctuate strongly as in Figure 8.8. The resemblance between synthetic correlation curves with weakly imposed relation and the correlation curve for the actual observations in Figures 8.8 and 8.9 supports the hypothesis that solar activity has a weak but significant influence on the terrestrial temperature. Reid (1987) suggested that there is a positive correlation between the 11-year running average of the global mean sea surface temperature and similarly 11-year low-pass-filtered sunspots. It is important to be aware of the effect of filtering on the outcome of a correlation analysis.

In order to examine whether long-term trends may affect the outcome of correlation analyses, the analysis in Figure 8.8 was repeated using non-de-trended series as well as de-trended series. For the de-trended series, two different types of filters have been employed: a binomial filter and a square moving average filter (also known as "box-car" or "MA" filter[10]). Figures 8.11 and 8.12 show some examples of the different filtering techniques for the sunspot record and the temperature. The different filter types give different details, but the low-frequency behaviour is similar. The correlation between the sunspot number and the northern hemisphere mean temperature is $\approx$0.1 for unfiltered series, but increases to $\approx$ 0.5 for 11-year low-pass-filtered series based on the MA filter. The binomial filter yields strongly variable correlation estimates for long timescales. For longer filter window widths, the non-de-trended estimates approach unity whereas the de-trended estimates are unstable (not robust). The unfiltered estimates are contaminated by the seasonal cycle, which "pulls down" the true estimate.

It is important to keep in mind that the low-pass filtering "throws away" much of the variation and the smoothed curve describes a small fraction of the fluctuations. Figure 8.10 shows how the fraction of variance in the filtered data to unfiltered data varies with the smoothing of two filter types. According to this analysis, the 10-year low-pass-filtered global mean temperature describes $\sim$20% of the total (unfiltered) variance, and a hypothetical perfect correlation between the 10-year low-pass-filtered global mean temperature and the sunspot numbers would suggest that the sunspot number may hypothetically account for about 20% of the variations. In other words, if the data have been subject to a filter, the correlation estimates should be presented together with the fraction of the variance in the filtered data to the variance in the unfiltered data. Figure 8.10 suggests that the 11-year cycle is weak in the temperature record, suggesting a relationship between the solar cycle and the temperature must be weak. The difference between the variance associated with the different filtered curves (Figures 8.11 and 8.12) is also consistent with the different correlation estimates in Figure 8.8.

There are many different ways of filtering the data. The effect of the different types of filtering may seem subtle if only judged from the appearance of the time-series (Figure 8.11). Figures 8.11 and 8.12 show two different window shapes: binomial and box-car. High-pass-filtering may be done by subtracting a low-pass filter from the unfiltered data. It is important to keep in mind that the box-car filter may not be appropriate for producing high-pass-filtered data because it may produce spurious ringing effects.[11]

Which correlation value is representative of the true correlation between solar activity indices and the global mean terrestrial temperature? The fact that monthly fluctuations in sunspot number affects the results to such a degree (e.g. the difference between the results in Figure 8.8 where $T(2m)$ are low-pass-filtered and R_z unfiltered) suggests that it is not the sunspots that affect the terrestrial climate, but some other mechanism with a similar 11-year cycle to the Schwabe cycle. The estimates

[10] Moving-average filter.

[11] High-frequency sinusoids.

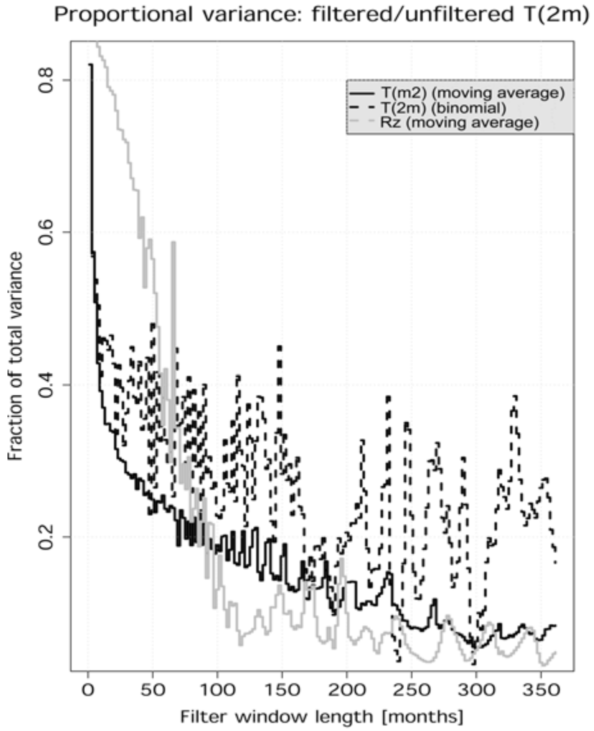


Figure 8.10. The proportion of variance in filtered temperature and sunspot number to unfiltered data. The two curves show the results derived using the binomial and square filters respectively. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and CRU.

become pseudo-stable in the vicinity of the 11-year window length, which supports this hypothesis. This observation is also in agreement with the hypothesis that the cycle-mean sunspot number is related to the temperature over the same interval. Both the number of sunspots in the solar cycle as well as the terrestrial climate may be affected by such a mechanism. One explanation may be the modulation of the TSI, the solar UV or the solar radio emission. White (2000) proposes that modulation of solar X-ray and UV light is by magnetic structures and thermodynamic conditions in the solar atmosphere.

8.4.6 Autocorrelation and lag-correlation

When applying a low-pass filter to a series we increase the autocorrelation by making adjacent data points more similar to each other. Thus in series with non-zero autocorrelation the successive data points are not entirely independent of each other. This means that each data point cannot be regarded as an independent observation and the degrees of freedom associated with the series is less than the number of observations. Successive observations are related to each other because the quantity

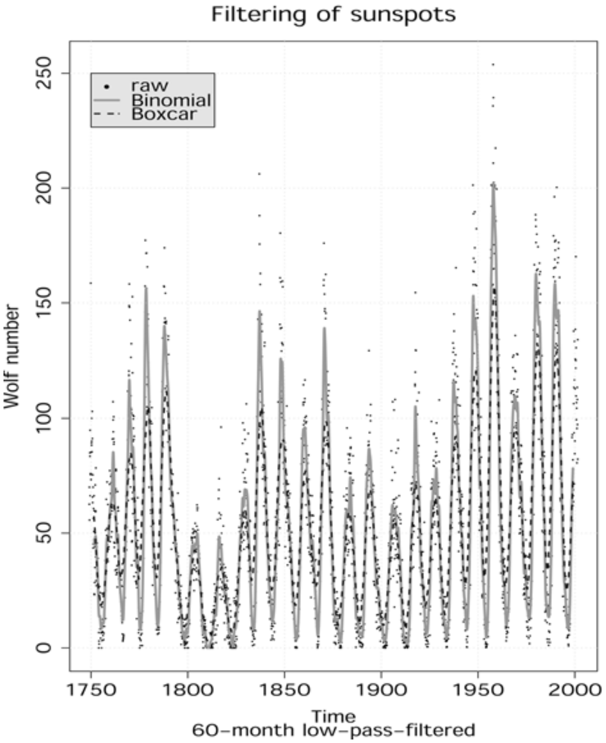


Figure 8.11. Example of low-pass-filtering of the sunspot number using two different kinds of filters. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

changes slowly compared to the frequency of observations. The successive relationship is often referred to as persistence, and the quantity called *autocorrelation* or *serial correlation* is a measure of how similar the successive measurements are. In other words, the autocorrelation gives an indication of how slowly the variable changes compared to the time interval between the observations.

The autocorrelation, $a = r(x_n, x_{n+l})$, is a dimensionless quantity defined as:

$$a = \frac{\sum_{n=1}^{N-1}\left(x_{n+1} - \frac{1}{N-1}\sum_{l=2}^{N} x_l\right)\left(x_n - \frac{1}{N-1}\sum_{l=1}^{N-1} x_l\right)}{\left(\sum_{n=2}^{N}\left(x_n - \frac{1}{N-1}\sum_{l=2}^{N} x_l\right)^2\right)\left(\sum_{n=1}^{N-1}\left(x_n - \frac{1}{N-1}\sum_{l=1}^{N-1} x_l\right)^2\right)} \quad (8.5)$$

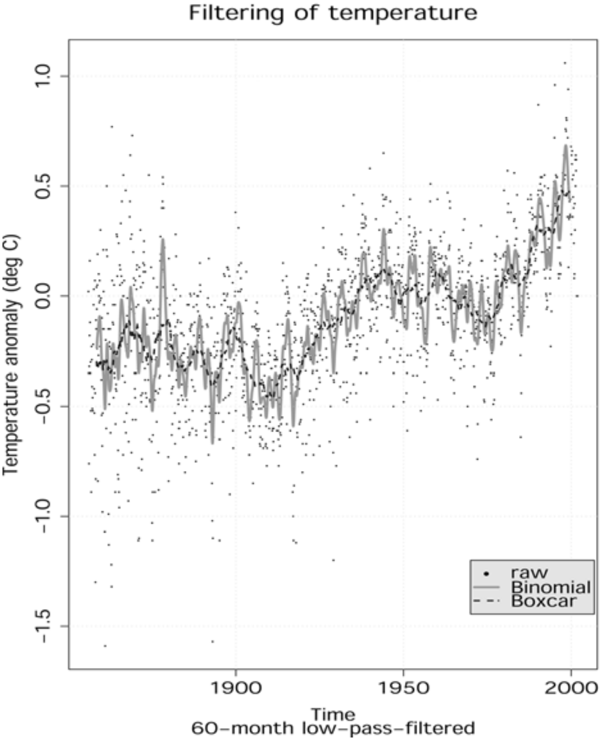


Figure 8.12. Example of low-pass-filtering of the northern hemisphere mean temperature using two different kinds of filters. Data from CRU.

An easy method for estimating the autocorrelation is to make two identical copies of a time-series, remove the first element(s) of the first series and a similar number at the end of the other, and then perform an ordinary correlation analysis. In other words, the autocorrelation is identical to a lagged-correlation of two identical time-series.

Figure 8.13 shows the autocorrelation functions for the northern hemisphere mean temperature (a) and the Wolf number (b). At lag zero, the autocorrelation is 1, whereas it diminishes at large lags. But the non-zero correlation at lags greater than zero implies a degree of *persistence*; there is "memory" in the system. We can see periodicities in both series shown in Figure 8.13: the annual cycle in (a) and the 11-year solar cycle in (b).

8.4.6.1 Lag-correlation

By lag-correlation, $r(x_n, y_{n+l})$, we usually mean a correlation analysis between two quantities, but where the entries of one of the variables have been time-shifted with a given number of positions (lag). One easy way of applying lagged-correlations is by removing the first (M) element(s) of the first series (i.e. x_1 for lag $= 1$) and the last

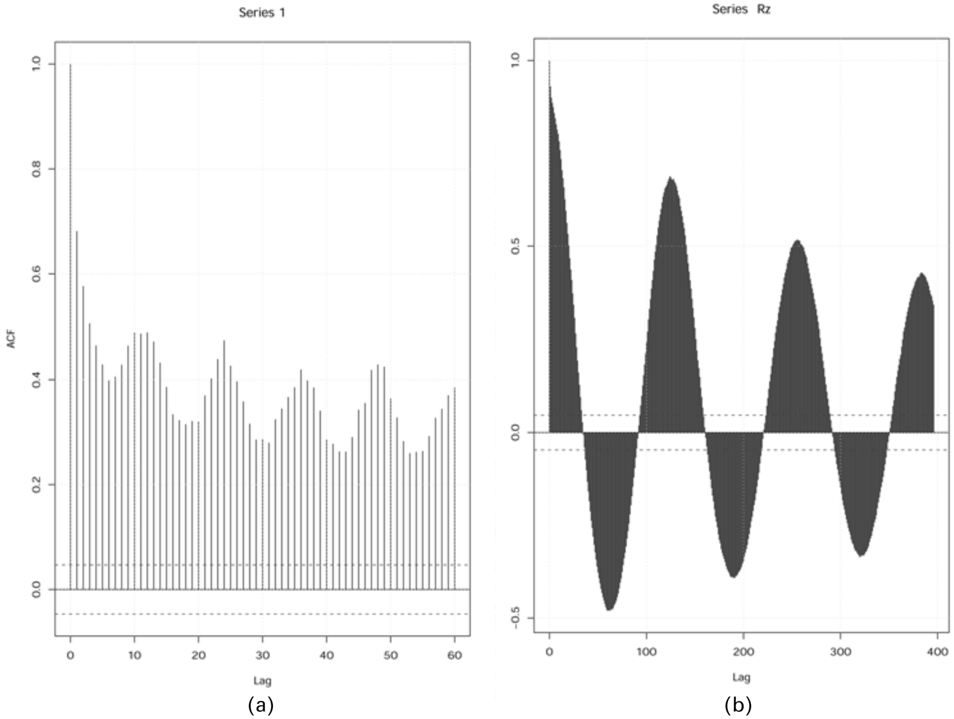


Figure 8.13. Autocorrelation of (a) the northern hemisphere mean temperature and (b) the sunspot number. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and CRU.

(ones) of the second series (i.e. y_N for lag $= 1$) and then performing a regular correlation analysis.

8.4.6.2 Sub-sampling

Analytical tests often assume *independent* observations of a given quantity. For series with non-zero autocorrelation, this criterion is not fulfilled. But, it is possible to obtain independent observations by sub-sampling the data by picking a subsection of the data. This is done by selecting every *n*th observation, where *n* is the distance between the data points corresponding to zero correlation.

For monthly mean climate data, one effective way to obtain a set of independent observations is to pick, for instance, every January month as there tend to be low inter-annual correlations.

8.4.7 Monte Carlo simulations

One usual way by which the results are judged is by trying to distinguish them from results obtained from an identical analysis but on two data sets that are known for sure to be unrelated. By carrying the analysis on many similar random data sets, it is possible to obtain a probability distribution for the analytical results, and from this distribution confidence levels may be estimated. This type of testing is analogous to the classical statistical example of demonstrating probabilities by throwing dice. But by conducting these "dice throwing" exercises, one studies cases where there is no correlation by definition (random numbers). Thus, any correlation is purely coincidental. The null-hypothesis is that there is no real relationship between the quantities tested and that the outcome are subject to statistical fluctuations (e.g. any correlation is coincidental), and the test aims to falsify this proposition. It is not possible to prove that the null-hypothesis is wrong, but it is feasible to show that it is unlikely that it is true. The computer in this case is an invaluable tool, as it can produce a large set of random numbers and keep track of them in a short time. Hence, the computer can generate numbers that follow the null-hypothesis: "The two series are not related". This type of exercise is often referred to as Monte Carlo integration.

One common type of Monte Carlo integration is to make a large number of surrogate time-series that mimic the original data in terms of autocorrelation. The autocorrelation is estimated for the respective quantities, and a large number of AR1 (red noise) series with similar autocorrelation are synthesised.[12] These surrogate data consist of stochastic processes, which are unrelated as they are made using random number generators.[13] Figure 8.14(a) shows a histogram of the correlation scores estimated from 10,000 such tests, and this histogram gives an estimate of the null-distribution.

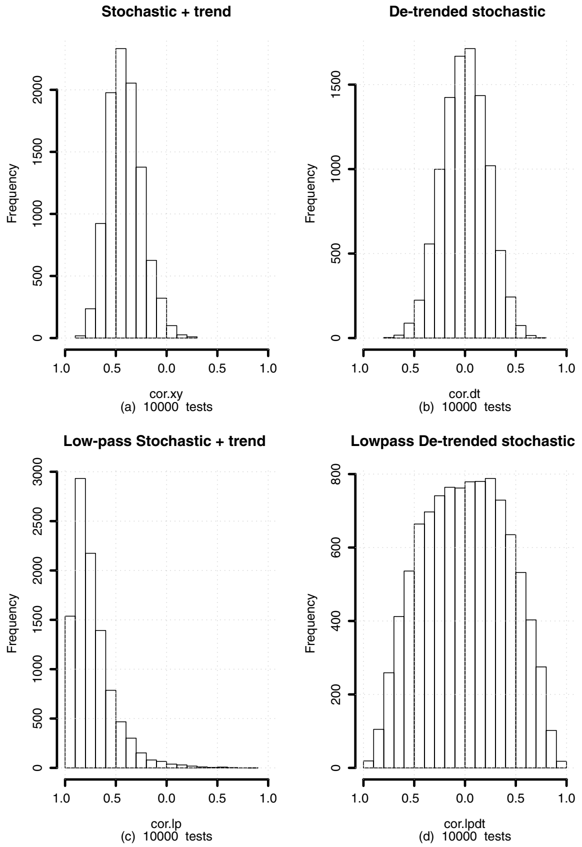
The choice of the analysis set-up can affect the outcome of a statistical test. It is possible to add a deterministic[14] trend to each of the two populations of stochastic series and show that the whole null-distribution is shifted if the trends are non-zero. Two such series are shown in Figure 8.6(a). De-trended versions are shown in Figure 8.6(b). Low-pass-filtering may also pre-dispose the outcome of the Monte Carlo tests, and smooth signals tend to increase the spread of the null-distribution[15] (Figure 8.14(d)). This effect may be demonstrated by low-pass-filtering the stochastic series. The effect of the smoothing is illustrated in Figure 8.6(c) for the non-zero trend cases and Figure 8.6(d) for the de-trended versions. Figure 8.7 shows the scatter plots and the correlation estimates of the corresponding series. For such short series, the presence of a non-zero trend can affect the correlation estimates considerably. The process illustrated in Figures 8.6 and 8.7 is repeated 10,000 times and a record is kept of the correlation scores. Hence, the null-distribution is estimated for the correlation of two uncorrelated series with a similar number of data points as the series analysed by Friis-Christensen

[12] An AR1 series x can be synthesised using the formula $x_n = ax_{n-1} + (1 - a)\eta$, where a is the autocorrelation and η represents random numbers from a random number generator.

[13] Assuming that the random number generator is capable of producing "truly" random numbers.

[14] The same trend is added to all of the 10,000 stochastic series.

[15] It has the same effect as increasing the autocorrelation.



(a) 10000 tests

(b) 10000 tests

(c) 10000 tests

(d) 10000 tests

Figure 8.14. The effect of pre-processing on the null-distribution of correlation between two uncorrelated series. The panels show the distribution of a large number of correlation studies between series with non-zero trend (a), de-trended series (b), low-pass-filtered series before the trends have been removed (c), and low-pass-filtered de-trended series (d).

and Lassen (1991) (Figure 8.14). It is clear that de-trending and low-pass-filtering (smoothing) both affect the results: low-pass-filtering increases the probability of coincidentally obtaining high correlations as the distribution gets wider with the filter and non-zero long-term trends shift the location of the distribution away from zero.

The case study shown in Figure 8.14 illustrates the need for careful consideration about how to set up the Monte Carlo simulations for data with non-zero trends, and there seems to be no clear-cut answer for the cases where the *a priori* knowledge about the processes is limited. The safest procedure is therefore to pre-process the data in a way that reduces the effect of the linear trend (i.e. by de-trending or by differencing transformation). It is essential to always give detailed documentation of the set-up of a Monte Carlo test.

8.4.7.1 Bootstrapping

The null-hypothesis may also be tested by synthesising the test data from the original observation, but in a randomly scrambled form. This type of test is called a *re-sampling* test or a *bootstrap test*, and is essentially similar to the Monte Carlo approach discussed above. It is, however, important that each time-series entry is an independent realisation (zero autocorrelation).

8.4.8 The solar cycle length and terrestrial temperature

It has commonly been assumed that the solar constant varies by 0.1% over a solar cycle, an estimate deduced from satellite measurements of the solar output between sunspot maximum and minimum. Friis-Christensen and Lassen (1991) argued that there is no *a priori* reason to believe that the long-term changes in the solar irradiance are adequately represented by the sunspot number and suggested that slow modulation of the solar activity level may be related to the global temperature. These Danish scientists presented an analysis of 11-year smoothed time-series which suggests that land temperature variations are caused by changes in the length of the sunspot cycle. Lockwood (2002) suggested that longer solar cycles allow the open magnetic fluxes to decay away to a greater extent whereas shorter cycles cause an accumulation in the open flux.

There have been numerous attempts to search for relationships between solar cycle length and climate, such as Clough (1943), Friis-Christensen and Lassen (1991) and Hoyt and Schatten (1993). Recent statistical analyses (Friis-Christensen and Lessen, 1991; Hoytt and Schatten, 1993; Lassen and Friis-Christensen, 1995; Thejll and Lassen, 1999) have come up with correlation coefficients in the range 0.79–0.95 between the surface temperature since 1850 and solar activity proxies. Pallé and Butler (2001) confirmed the high correlation for the period 1984–1991, but this correspondence ceased to exist after 1991. Friis-Christensen and Lassen (1991), received a great deal of attention for their article in *Science* about the correlation between the solar cycle length (SCL, also known as epochs) and northern hemisphere temperature. Although similar correlations have been reported earlier (Clough, 1943), none had such high correlations as Friis-Christensen and Lassen (1991). The Danish scientists had used a low-pass filter (with weights 1-2-2-2-1) to pre-process the epochs. Since the work by Friis-Christensen and Lassen (1991) and Hoyt and Schatten (1993) does not say whether the data were de-trended or not prior to the analysis, it is reasonable to assume that they were not. The inspection of the figures in Friis-Christensen and Lassen (1991) suggest that the data contained significant long-term trends.

Criticism of the work of Friis-Christensen and Lassen (1991) has been put forward on several points: (i) the data records were not de-trended before the analysis (see Figure 8.14), (ii) filtering of epochs is not justified (see Section 8.4.9), (iii) the determination of the epochs is associated with a great deal of uncertainty and (iv) estimates for future solar maxima were highly uncertain. The good correlations they found may have been artificially inflated by the use of an inappropriate filtering technique and a questionable extrapolation of observations leading to a distorted series of the solar cycle length. Another weak point of the hypothesis they proposed is that there is no *well-established* physical explanation for how the SCL influences Earth's climate. Some hypotheses have been proposed relating the SCL inversely to the solar convection intensity and thus to the variations in the solar output. However, records of space-borne TSI spanning several solar cycles are required in order to evaluate this hypothesis against the observations.

8.4.8.1 Criticism on non-zero trends

Part of the problem here is that we do not know *a priori* how much of the trend can be attributed to solar activity, and how much is for instance due to anthropogenic "greenhouse gases", changes to the landscape, or other factors. One way to reduce the effect of external factors such as an enhanced greenhouse warming is to *de-trend* the series, i.e. subtract a best-fit linear trend. But, before illustrating how this is done, we will recap the correlation analysis.

If Friis-Christensen and Lassen (1991) did not remove the linear trend but applied a low-pass filter to the data, then the null-distribution for their case would look like the one in Figure 8.14(c). By using low-pass-filtered values for only the SCL curve but not the cycle mean values for the temperature, only one of the curves is essentially low-pass-filtered. In their case, the appropriate null-distribution would be that of the low-pass-filtered non-de-trended series because an inspection of their temperature curve suggests a high autocorrelation in the temperature curve. A repeated analysis on the same data after de-trending and using no filtering is shown in Figure 8.4. In this case, the correlation between the SCL and the temperature does not qualify as statistically significant at the 5% level. The results of Friis-Christensen and Lassen (1991) do not therefore appear to be robust. This example shows that the analysis carried out by them cannot prove there is a relationship between the SCL and the terrestrial surface temperature, but this does not falsify the hypothesis proposing that the temperature is related to the SCL either.

8.4.8.2 Criticism on filtering

Another reason for the differences in the correlation results is that some of these studies used a filter to smooth the data prior to the analysis. Unless there are strong justifications for doing so on physical grounds, filtering should be avoided in correlation studies because it involves an addition of subjective information. Laut and Gundermann (2000) criticised the work by Friis-Christensen and Lassen (1991) and argued that their unacceptable mixing of filtered and unfiltered data gives a false impression of a good fit. They tried to reproduce the proposed relationship between the solar cycle length and global and northern hemisphere mean temperature using unfiltered SCL, but failed to find such a relationship. Lassen and Friis-Christensen (2000) responded to the criticism of Laut and Guntermann, arguing that Laut and Guntermann gave the incorrect impression that the conclusions of the 1991 paper relied almost entirely on the data after 1970 and they disagreed with the criticism of how to combine the two temperature series. Furthermore, Lassen and

Friis-Christensen argued that their criticism gave the wrong impression that their conclusions ruled out anthropogenic contribution to the global warming. They did acknowledge that new data suggested a lesser role for solar activity in terms of the Earth's surface temperature. However, Lassen and Friis-Christensen did not explain that their original "12221" filter had an effective filter width of roughly 55 years (5 epochs a ∼11 years) whereas the temperature series was smoothed by approximately 11 years. Since their filters were centred, this has some odd (and incorrect) implications: (1) that for any given point in time, temperatures should to some degree be influenced by the state of solar activity in the future, or (2) that somehow longer timescales on the Sun are warped into shorter timescales on Earth. Other peculiar points in their paper was the claim that it would be misleading to include newer data that did not fit the temperature well in the regression analysis, since their analysis did not involve a validation with independent data. However, it is not entirely clear how they carried out their analysis, and in a paper of 2003, Laut maintained their criticism of the work by Friis-Christensen and Lassen (1991) as well as later work published in Lassen and Friis-Christensen (2000). Part of their new criticism included the lack of transparency of methodology and difficulties replicating their results: Laut had to "remake" their results by scanning the figures and then electronically digitising the curves. He criticised their work for the unacceptable filtering method where the two last data points were unfiltered and the two before them only partially filtered. The unfiltered data fluctuated strongly about the filtered curves, and smoothed features were not visible in the original data. He also criticised them for merging two incongruous data sets (the Groveman/Landsberg/Borzenkova and the Jones curve) and in an unjustified ad hoc fashion lifting the Groveman/Landsberg/Borzenkova curve to achieve a good match. Laut criticised the results of the Lassen and Friis-Christensen (2000) paper further for using a linear regression with undisclosed zero weighting for the first 230 years of the solar record. When they repeated the regression with equal weights to all the data and omitting maximum–maximum epochs (which are considered to be less reliable than minimum–minimum), they obtained a poor fit. Furthermore, Laut argued that Lassen and Friis-Christensen (2000) and Thejl and Lassen (2000) had made some arithmetic errors in the updated data and that these errors had implications about trends in the filtered SCL series after 1950: with his corrected values, there was no trend. Thejll and Lassen (2002) wrote an erratum acknowledging errors in two of their figures. The fact that the curve representing the filtered SCL changed character over time with the inclusion of new data implies that the results are not robust.

Benestad (2005a) estimated new values for SCL using a combination of Fourier approximation, multiple regression, and simple calculus for determining the local maxima and minima. The results from this new analysis were fundamentally different to the impression from the main results of Friis-Christensen and Lassen (1991): the new results, which involved no smoothing, suggested that the Sun had been unsettled before 1900 and more stable since then. Thus, the smoothing applied by Friis-Christensen and Lassen (1991) seems to have contaminated their analysis. Furthermore, Benestad (2001) tried to use the solar cycle to predict the temperature where the SCL record had not been filtered but was de-trended prior to the model

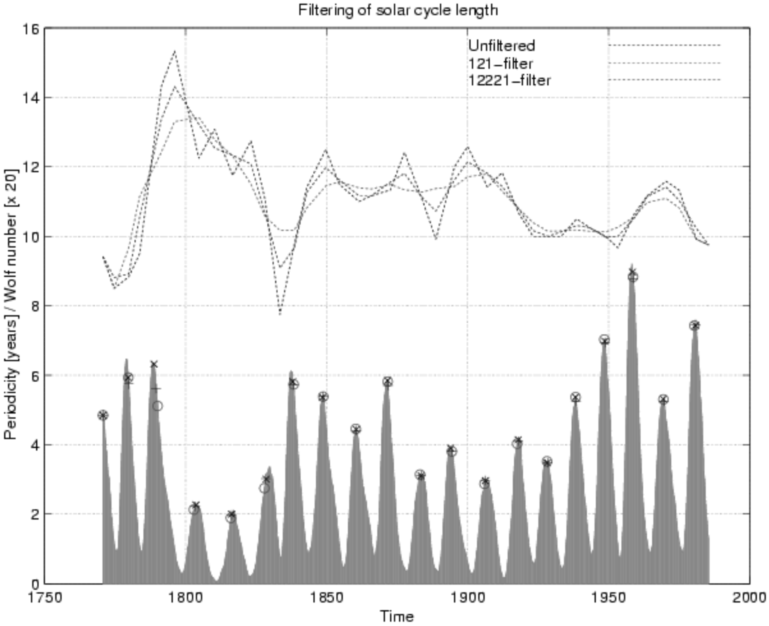


Figure 8.15. The sunspot record and estimated SCL using the max-max-min-min. Also shown are the filtered versions and the timing of the corresponding maxima. It is evident that the implication of the filtering of SCL is a shift in the sunspot maxima.

calibration (Figure 8.7). The filtering of the solar cycle cannot be justified on analytical (Figure 8.15) or physical grounds. A filter in this case shifts the apparent timing of the maxima and minima, as a reconstruction of the times of sunspot maxima and minima using a filtered curve will give a squeezed and stretched version of the original one. In other words, the timing of the maxima and minima can be reconstructed from the SCL curves, and using smoothed SCL records gives a shift in the timing of the maxima and minima (marked with '+' and '○' in Figure 8.15). The filtering will also introduce a nonlinear transform where cycles before/after long adjacent epochs have an equal influence on a given cycle to those before/after short adjacent epochs. Damon and Peristykh (2005) used a constant filter on the SCL record and their filtered curve did not exhibit changes that could account for the marked increase in the mean northern hemispheric temperature since 1950. They criticised Friis-Christensen and Lassen (1991) for failing to use the northern hemispheric paleoclimate record as a reference baseline.

Mursula and Ulich (1998) suggested that the estimates of the SCL are associated with significant uncertainties, and suggested an alternative to an approach based on the combination of intervals between successive minima and maxima respectively (min-min-max-max). However, they reported that the new SCL estimates did not alter the conclusion of Friis-Christensen and Lassen (1991), which may also suggest that the analytical test is pre-disposed (correlation analysis on low-pass-filtered data with non-zero trend) and does not "let the data speak for itself".

Table 8.3. Year, month and sunspot numbers corresponding to the maxima and minima shown in Figure 4.4. Also shown in the first column is the sunspot cycle number.

Cycle	Maxima			Minima		
n	Year	Month	R_z	Year	Month	R_z
				1755	5	0
1	1761	5	107	1766	6	3
2	1769	10	158	1775	2	0
3	1778	5	239	1784	5	6
4	1787	12	174	1798	5	0
5	1804	10	62	1809	10	0
6	1817	3	96	1822	1	0
7	1830	4	106	1833	6	1
8	1836	12	206	1843	2	4
9	1847	10	180	1855	9	0
10	1860	7	117	1867	1	0
11	1870	5	176	1878	8	0
12	1884	1	92	1889	11	0
13	1893	8	129	1901	4	0
14	1907	2	108	1912	2	0
15	1917	8	154	1923	8	0
16	1928	7	98	1933	8	0
17	1938	7	165	1944	4	0
18	1947	5	201	1954	1	0
19	1957	10	254	1964	7	3
20	1969	3	136	1976	7	2
21	1979	9	188	1986	6	1
22	1990	8	200	1996	10	1
23	2000	7	170			

8.4.9 Considerations on solar cycle lengths

The study of time intervals in statistical analysis is sometimes problematic. One question arises when considering the relationship between a time-series containing the annual cycle and another which does not. If the analysis does not start and stop at the same time of the season, then the seasonal cycle may affect the outcome of the analysis. In particular, the correlation study between the SCL and the northern hemisphere mean temperature may be affected by the seasonal cycle if the cycle mean temperature is used and the epochs start and end in different seasons. Table 8.3 lists the year and month of the sunspot maxima and minima respectively. For instance, cycle 16 starts in August 1933 during solar minimum and ends in April 1944. The corresponding cycle mean temperature would have a cold bias due to the over-representation of winter months. Conversely, cycle 18 that starts in April 1944, ends in January 1954 and is associated with an over-representation of summer months. One solution may be to use a fixed window width, i.e using a 10-year mean temperature centred at the mid-point between the maxima (Figure 8.22). An alternative is to remove the seasonal cycle prior to the analysis, hence using temperature anomalies.

Wilson (1998) has identified signs which may be interpreted as evidence for long-term solar cycle forcing (1844–1992) and secular variation in the Armagh observatory temperature record. The temperature record used in his study has a high correlation with the northern hemisphere temperature, and hence Wilson implied that a similar solar influence on the northern hemisphere could be inferred from the high correlations reported for the one site in Ireland. Solar and secular signals were identified, and after the removal of these, the residual consisted of white noise. The cyclic averages of the annual mean temperature were highly anti-correlated with the length of the Hale cycle (even-odd numbered cycle), with a correlation coefficient of $r = -0.89$ at a significance level of less than 2%. The Armagh observatory temperature record is considered to be of high quality.

Wilson assumed the null-hypothesis, i.e. that the cycle-mean temperatures are randomly distributed in time, and then subjected the data to runs- and t-statistic hypothesis tests. In both cases, the distribution of these quantities failed the tests, suggesting that the cycle-mean temperatures are not random, but are showing a slow warming trend of 0.04°C/decade. Some critics will argue, however, that it is easy to demonstrate with similar tests on finite sequences of random walk processes that these are not random either, so in this case the implications of Wilson's results are not clear cut.

Wilson's results may be criticised because the linear regression was performed on non-de-trended time-series. We do not know how much of the trend can be attributed to solar forcing, how much to enhanced greenhouse effects, and how much to other factors. The high inverse correlation may have been inflated by the trends and the small effective lengths of the time-series.

There are various ways to carry out such analyses as Wilson's, for instance by using different ways of estimating the solar cycle lengths and different approaches to estimate the temperatures. It is common to low-pass-filter the temperature and sunspot record or take the cycle mean.

It is important to be aware of the definition of "cycle-mean temperatures", as if the cycle starts in the autumn and ends in the spring, then the mean value will in most cases be biased towards lower values in the northern hemisphere because the epoch will contain one more winter than summer season. Using annual mean temperatures to estimate the cycle-mean temperature, as Wilson did, avoids this problem, but these estimates may not account for all of the cycle or are slightly influenced by the adjacent cycles if the solar cycles do not start and end during the same season (which they do not). The definition of sunspot cycle length is also problematic (Mursula and Ulich, 1998) as their estimation may have errors of up to 6 months. There are various ways of estimating the SCL, such as estimating the time intervals between successive minima or maxima. The two types of estimates may be combined (min-min-max-max method). The cycle median sunspot number may be used instead of the min or max and Mursula and Ulich (1998) argue that this quantity gives a better estimate of the SCL. Alternatively, one can use a spectral

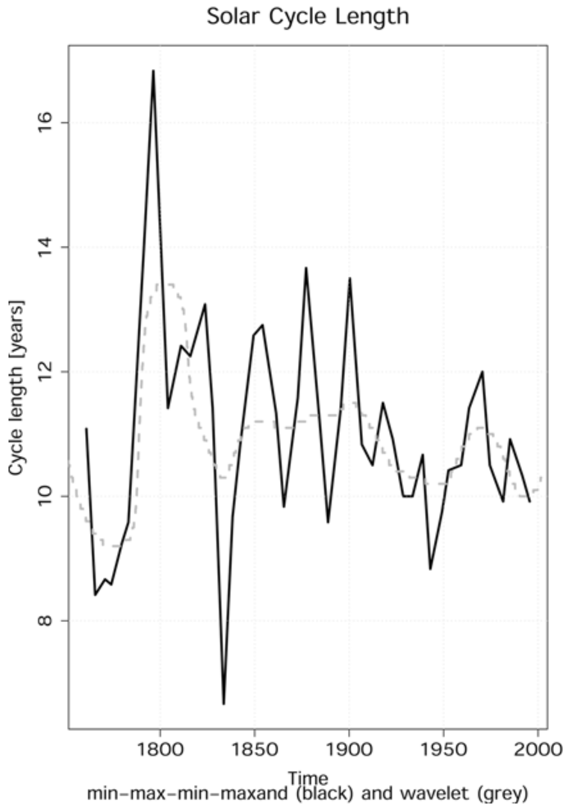


Figure 8.16. Different reconstructions of SCL: based on max-max-min-min (black) and wavelet analysis (grey). Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

analysis called *wavelet* analysis to estimate the solar cycle lengths. A comparison between SCL estimates based on the max-max-min-min and the wavelet methods is shown in Figure 8.16. It is interesting to note that Figures 8.15 and 8.16 indicate substantially more dramatic variations in the SCL before 1850 than during most of the recent century when the (unfiltered) SCL values have been relatively stable. The relatively strong variations in the SCL can also been discerned in the wavelet analysis (Figure 4.9, see colour plate section) and in figure 1 (thin lines) from Damon and Peristykh (2005). In this context, it is important to ask whether there could be an association between (unfiltered) SCL and terrestrial temperatures, and if this is the case, whether the SCL record suggests suggests more pronounced temperature variations prior to 1850. The SCL record has extreme values in ~1800 (SCL > 16) and 1830–1840 (SCL < 8).

8.4.10 Correlation studies and pitfalls

An example illustrating the dangers associated with correlation studies is given in Figure 8.17 (see colour plate section) which shows the correlation map between the

SCL and temperatures from a climate model. The temperatures in the climate model have realistic timescales, but are not related to the SCL because variations in solar activity are not taken into account by the model. There are some regions where the correlation is high, and 9% of the area in the upper panel is associated with correlations outside the 95% confidence interval (i.e. correlations that qualify as 5% significant). After the trends have been removed, the area of significant correlation is reduced to 6% (lower panel). However, out of 100 independent tests of stochastic series, one expects to find 5 that qualify as being significant at the 5% level of confidence. This is known as *the problem of multiplicity*.

In Figure 8.18 (see colour plate section) the same correlation is shown for temperature reconstructions (Benestad, 2000a) from actual observations. Although there is a considerable drop in the correlation values after de-trending the series, there are some small regions where the local de-trended correlation appears to be significant. If there was a strong and real correlation, we would expect to see large areas with high correlations after the data have been de-trended. Hence, these results do not supports Friis-Christensen and Lassen (1991) hypothesis. This case study, however, illustrates how the presence of a (prescribed in the model integration) trend may affect the outcome of statistical analysis if care is not taken.

Lean and Rind (1998) argue that the influence of climate forcing factors other than those associated with solar activity may result in variations with apparent solar–terrestrial correlations. Such factors may include volcanoes in the 19th century. A model simulation of climate using the GISS $8^\circ \times 10^\circ$ GCM forced with a solar irradiance reconstruction based on a combination of the average amplitude of the solar cycle and short-term variations associated with the 11-year cycle gave a 0.3–0.5°C global warming as a response to a prescribed 0.24% increase in the solar irradiance since the Maunder minimum. However, Lean and Rind (1998) concluded that less than one-third of the recorded increase in the global mean surface temperature since 1970 can be attributed to changes in the Sun.

The results from a study by Cubasch *et al.* (1997), which involved a climate model (GCM) forced with the solar irradiance construction by Hoyt and Schatten (1993), suggest that the solar-induced warming during the past 30 years was 0.16 ± 0.09 K compared to a warming of 0.54 ± 0.13 K due to the enhanced greenhouse effect. The observed temperature change since the pre-industrial era is estimated to be 0.6 ± 0.2 K (Houghton *et al.*, 2001).

8.4.10.1 The combination of natural factors and anthropogenic forcing

Recent model studies with comprehensive climate models suggest that there is not just one factor causing climate fluctuations. Some British scientists have explained the recent 100 years of global temperature change in terms of both solar and anthropogenic (greenhouse gas) forcing (Tett *et al.*, 1999; Gillet *et al.*, 2000). However, changes to Earth's surface such as deforestation and expanding agriculture may also alter the climate on a permanent basis (Hansen *et al.*, 1998a). Schlesinger and Ramankutty (1992) found strong circumstantial evidence for inter-cycle variations in solar irradiance having contributed to observed temperature changes since 1856, but also found that the dominant contribution to the long-term change since the 19th century was from greenhouse gases. Tiny particles suspended in the air, known as aerosols, can affect the energy balance directly by absorbing or reflecting light. They can also affect the climate indirectly through their influence on the cloud drop and rain formation. There are various types of aerosols, some of which are of natural origin (dust, organic compounds, salt, ammonia NH_4, and sulphur) and some that are produced by humans (black carbon, or soot, and sulphate SO_4). Volcanoes inject large amounts of dust particles into the upper atmosphere, and are known to produce a cooler global climate (Hartmann, 1994). However, the effects from aerosols tend to be relatively short-lived[16] as long as there is no continuous production of them, and the cooling due to volcanoes tends to last for one to two years. Aerosols may have different effects depending on their location. High up in the polar atmosphere, they may facilitate the destruction of stratospheric ozone when the temperatures are sufficiently low and the ozone molecules are exposed to UV radiation.

8.5 VALIDATION OF PREDICTIONS

One common way of validating a prediction involves calibrating the models using only some of the data, and using the remaining (independent) data to compare with the model predictions. Correlation, explained variance (ANOVA), and root-mean-error statistics are often used to quantify the model "skill".

Figure 8.19(a) shows a prediction study where a regression model is calibrated for the period 1857–1957, and the temperatures since 1959 are used for evaluation. The calibration data were de-trended prior to the regression. The standard deviation of the predicted temperature is 0.013°C which is about 0.005% of the global mean temperature (288 K). It was stated in Section 5.4.1 that a proportional change of 0.1% in the TSI can explain about 0.025% of the temperature (288 K), which is about 0.07°C. Thus, the variations in the TSI associated with the solar cycle are more than strong enough to explain the predictions. The smaller magnitudes seen in the predictions may suggest that the variations are damped. The sunspots cannot predict the rapid recent long-term warming. In Figure 8.19(b) the analysis was repeated, but for 11-year mean values.

The empirical model based on the 11-year mean values (Figure 8.19(a)) reproduces most of the warming before 1950 as well as some of the most recent warming (linear slope in the independent data points). Regression analysis yields a low estimate for the probability of the fit, being a result of a coincident. However, this estimate assumes that the data points are independent, and does not take into account the autocorrelation (i.e. the *p*-value estimate is too low). Furthermore, there is a "jump" between the dependent and independent predictions that is not seen in the observations. It is furthermore important to recall that the 11-year low-pass-filtered temperature only accounts for ~20% of the temperature variations,

[16] A lifetime less than 10 years.

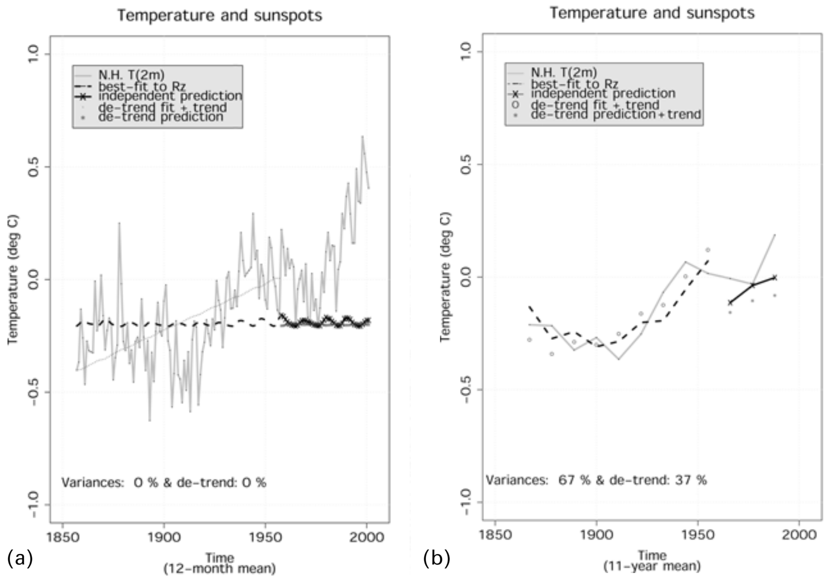


Figure 8.19. Using the sunspots to predict the northern hemisphere mean temperature with models based on annual mean temperature and sunspot number (a) and 11-year means (b). The best-fit is not statistically significant at the 5% level for the unfiltered data, but the *p*-value is 0.01 for the de-trended low-pass-filtered data. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and CRU.

although the smoothed curve describes most of the long-term trends and relate these results to the results in Figure 8.8. The robustness of the regression results can be tested by repeating the analysis with a slightly different calibration period (Figure 8.20(a)). Although the fit still appears to be good for the dependent data, the similarity between the curves for the independent period has disappeared because the above-mentioned "jump" is now seen in the predictions (the independent predictions no longer describe a trend). The regression was repeated further with 10-year-long (Figure 8.20(b)) and 12-year-long (not shown) filter widths, in addition to using an 11-year filter width. The justification for these additional tests is that SCL is not exactly 11-years long, but recent SCL values vary between 9 and 12 years (Figure 8.16). Furthermore, it is important to test whether spurious signals have been introduced by the low-pass-filtering, such as the Slutsky–Yule effect. The regression analyses with different filter widths produce curves for the independent validation period that are substantially different for the different tests. These results therefore do not appear to be robust.

A similar analysis to that in Figure 8.19 was repeated with the sunspot number replaced by the aa-index (Figure 8.21). The predictions based on the aa-index give a better description of the global mean temperature than the unfiltered R_z, and the standard deviation of the independent predictions is 0.07°C. In other words, the amplitude of these 11-year variations can, according to the estimates in Section 5.4.1, be accounted for by the changes in the TSI associated with the 11-

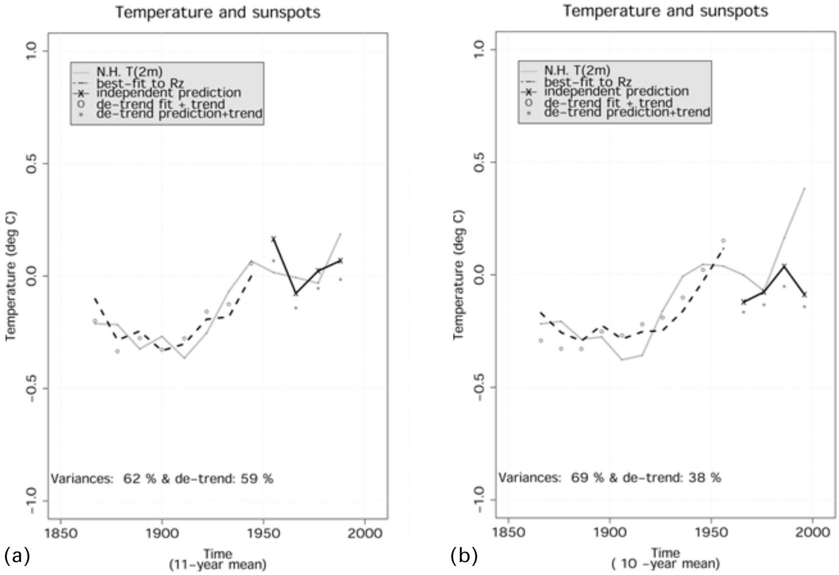


Figure 8.20. Similar to Figure 8.19, but using a slightly different subset of data for calibration (a) and using a 10-year-long low-pass filter instead of 11 years.

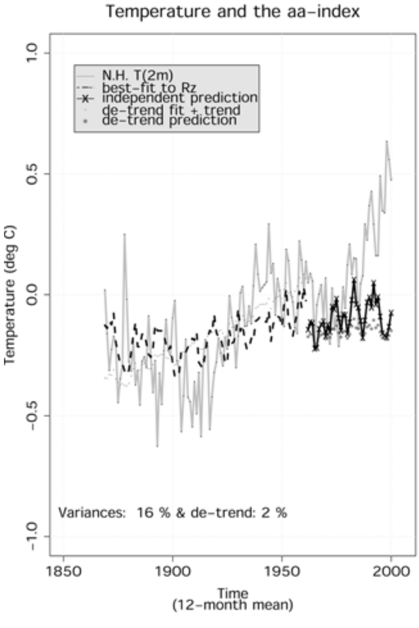


Figure 8.21. Using the aa-index to predict the northern hemisphere mean temperature with models based on annual mean temperature and sunspot number. For non-de-trended analysis, the *p*-value is 0, and the *p*-value is 17% for the de-trended series. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and CRU.

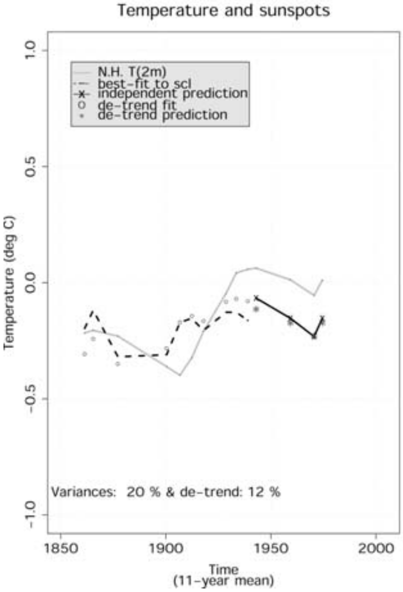


Figure 8.22. Similar to Figure 8.19, but for SCL-based prediction. For non-de-trended analysis, the *p*-value is 0.19, and the *p*-value is 0.32 for the de-trended series (see Section 8.4.9). Thus, the results do not suggest a high statistical significance. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and CRU.

year solar cycle. It is also possible that clouds are involved in the connection between the aa-index and the temperature (Section 7.10.5.1). The aa-index cannot reproduce the long-term warming trend.

The statistical connection between the sunspot data and temperature can be explored further by using these solar proxies for the reconstruction of the past temperature. Benestad (1999) reconstructed temperature and SST based on a multiple regression against the R_z and SCL. The reconstruction based on the SST indicated strong temperature variations in the 18th and 19th centuries compared to the most recent century, but the reconstruction based on the global mean land temperature suggested much weaker fluctuations. The predicted warm events in the mid-1700s and mid-1800s were not realistic as the magnitude of the recent temperature variations were much smaller than seen in the observations. Figure 8.22 shows the same analysis as in Figure 8.19 but repeated with SCL instead of the sunspot number. The results suggest low statistical significance for the SCL-based best-fit to the temperature curve. The independent predictions reproduce the main features of the corresponding part of the temperature curve, but this evaluation only includes four data points and the best-fit is not very similar to the temperature, suggesting that the apparent similarity may be coincidental.

The Sun's irradiance level is thought to have increased from the 1800s to a maximum in the 1930s to 1940s. Hoyt and Schatten (1993) used observations

Table 8.4. The timing of peak solar activity (TSI) during the 20th century according to various solar indices. The spread in timing of the maximum suggest there is considerable uncertainty regarding the relationship between these indices and the TSI variations.

Proxy	Year of 20th century maximum
Sunspot structure (U/W) (Hoyt and Schatten, 1993)	1934
Fraction of sunspots without umbra (Hoyt and Schatten, 1993)	1933
Solar cycle length (1-2-1 filter) (Hoyt and Schatten, 1993)	1937.5 ± 5.5
Solar cycle length (Hoyt and Schatten, 1993)	1940
Rate of solar cycle decay (Hoyt and Schatten, 1993)	1920–1931
Equatorial solar rotation (Hoyt and Schatten, 1993)	1924–1934
Wolf sunspot number R_z (Table 8.3)	1957

from Nimbus 7 and the 1979–1990 satellite derived temperature record from Spencer and Christy (1990) and estimated the climate sensitivity to variations in the solar irradiance to be 1.67°C/(% change in TSI). The proxy models described in their paper only predict changes in the solar energy output during the last century that are less than 0.2% of the solar constant. However, the estimates of the solar variations are highly uncertain (Table 8.4), and most estimates of the solar irradiance changes over the 20th century are in the region 0.14% to 0.38%. Nevertheless, Hoyt and Schatten (1993) concluded that these magnitudes are too low to directly account for the 0.5°C warming found in the observational record, as a change of 0.14% in the solar forcing can only account for a 0.23°C change. But, if the solar variations involve feedback mechanisms such as changes in the ice- and snow-cover, or variations in the plant population and hence the plant absorptivity, then the solar variations may be amplified. Hoyt and Schatten (1993) also speculated that if the surface wind velocities are reduced due to changes in the length of the day then there may be an increase in the absorption of the water surface, or the planetary albedo may change as the plants' orientation is affected by changes in the winds.

Kelly and Wigley (1992) investigated the relationship between the solar cycle periodicity and Earth's global mean surface temperature and compared the solar contribution to temperature change with forcing due to increased greenhouse gas forcing. They used an upwelling-diffusion energy-balance climate model to reconstruct the global mean temperature. Only a small portion of the long-term trends could be attributed to changes in the solar cycle length. In fact, their results suggest that the solar term can account for more of the variance than just the greenhouse gases on their own. But, if the forcing is due to both greenhouse gases and solar forcing, then the solar forcing term accounts for less of the variability than the greenhouse gases do. They argued that the climate sensitivity implied by solar forcing alone is implausibly high, and the overall credibility of the solar-only forcing experiments is low because it is unreasonable to neglect the well-established cause of the greenhouse gases forcing. Furthermore, the results showed considerable sensitivity to the filtering method used to calculate the solar epoch record, which

Table 8.5. Estimated change in the solar forcing (TSI) between 1750 and the present. The results are taken from the Intergovermental Panel of Climate Change (IPCC) Third Assessment Report (TAR). The estimates given in the earlier IPCC report (Second Assessment Report; SAR) are the same as those given in IPCC TAR, indicating that there has been no recent improvements in terms of reducing uncertainties associated with TSI changes

	Global mean radiative forcing (Wm^{-2})	Uncertainty (%)
IPCC 1995	+0.30	67
IPCC 2001	+0.30	67

implies large uncertainties in the cycle-length-based reconstructions. More recently, Solanki *et al.* (2004) proposed that at most 30% of the global warming since 1970 can be attributed solar activity. Three studies by Meehl *et al.* (2003, 2004) and Stott *et al.* (2003) respectively used GCMs to try to assess how strongly Earth's climate responds to changes in the solar activity. The US study by Meehl *et al.* used the American Parallel Climate Model (PCM) to make several numerical experiments with different combinations of solar activity, sulphate forcing, and greenhouse gases, whereas Stott *et al.* employed a British climate model from the UK MetOffice (HadCM3) and applied volcanic, solar, and greenhouse gas forcing. Both studies suggested that the early century warming (the rapid temperature increase in the northern hemisphere before 1940: Stott *et al.* 0.29 K per century) were likely due to solar forcing while the warming after 1970 was best explained in terms of increased greenhouse gas concentrations. Stott *et al.* estimated that the solar forcing over 1950–1999 was 16–36% of that of the greenhouse forcing, depending on which solar proxy was used. The US study suggested that the response from solar forcing and greenhouse gases was non-linear because the solar forcing is less homogeneous than an enhanced greenhouse effect. These results suggested that solar forcing in isolation did not reproduce much of the early century warming, but in combination with sulfate and the increased greenhouse gas concentrations, the model could account for the strong temperature increase between 1920 and 1940. In their study, the effects of volcanoes were neglected, however, volcanic eruption could have resulted in cooling in the late 19th and early 20th centuries (Stott *et al.*, 2003). Enhanced temperature gradients could have implications for monsoon systems and conversion zones. The British study, on the other hand, suggested that the combined response from solar and greenhouse forcing is linear, and that the total effect is approximately a linear sum of the two. However, they also suggested that the climate models may underestimate the true climatic response to the solar forcing. None of these studies accounted for solar induced changes in the stratospheric ozone (e.g., Haigh, 2003) or cosmic rays. One interesting aspect of the US study was the suggestion that changes in the total solar irradiance related to solar activity may affect the cloud cover. There are also large uncertainties in the other TSI reconstructions, and this uncertainty has not been reduced in the time of the two most recent IPCC reports (Table 8.5).

8.6 TOTAL SOLAR IRRADIANCE STUDIES

8.6.1 The "Little Ice Age" and TSI

Part of the 20th century warming may be a recovery from cold conditions after the "Little Ice Age". A study was conducted by Bertrand and van Ypersele (1999), based on a 2-dimensional (sector-averaged) global climate model to illuminate the effect of solar variability on Earth's surface temperature. The model suggested that changes in the reconstructed solar irradiance, based on three different empirical formulae (Willson and Hudson, 1988; Hoyt and Schatten, 1993; Lean *et al.*, 1995), can account for 0.045–0.22°C of the 0.53 ± 0.07°C warming recorded in the 20th century (7.5–33.6% of the temperature trend). The simulation of the Maunder minimum produced no more than 0.37°C cooling for 1645–1715. The 10–12 year oscillations found on Earth, on the other hand, appeared to be unrelated to the variations in solar forcing (see Figure 8.17 for a demonstration of how 11-year variations may seem to be correlated with the solar cycle). The longer 88-year Gleissberg cycle (Eddy, 1976) was a more prominent feature in the model results, and Bertrand and van Ypersele (1999) concluded that the variations in the solar irradiance are most important on centennial timescales. They suggested snow-cover and ice albedo feedback mechanisms as possible candidates for amplifying the solar signal (Section 5.6.3). Further evidence corroborates the notion that amplifying feedback mechanisms must have been involved if the "Little Ice Age" (LIA) was entirely due to changes in solar activity. Haigh (2003) inferred from a multiple-regression analysis between zonal mean temperatures on the one hand, and the solar variability, volcanic forcing, the QBO, the NAO, and ENSO on the other, that the response to solar variability varies geographically, and that other mechanisms than direct radiative heating must be involved. An independent piece of supporting evidence was given by Bond *et al.* (2001) who claimed to have found empirical evidence in ice drift tracers that the LIA was linked to changes in solar irradiance. They observed a LIA response in the proxies and a correlation of 0.44–0.56 between "stacked marine records" ^{14}C and ^{10}Be from Greenland ice cores.

Harrison and Shine (1999) discuss five different reconstructions of the solar irradiance based on various proxies for the solar activity: sunspots, faculae and background emission from the photosphere. It is noted that the various reconstructions diverge by $5\,\mathrm{W\,m^{-2}}$ on the estimation of the present-day irradiance. But, four of the five reconstructions nevertheless suggest that the solar output has been more or less constant throughout the latter part of the 20th century. The forcing implied by the four series is the range given by IPCC (1995) of $0.3 \pm 0.2\,\mathrm{W\,m^{-2}}$ since 1850.

8.6.1.1 Simple energy balance models

Many simple energy balance models (EBMs) do not describe the complex situation on Earth very well, as the complexity of the climate system and the feedback mechanisms have been ignored. Rind *et al.* (1999) have estimated a $\delta T_s \approx 0.45$°C in response to a 0.25% increase in the total solar irradiance, using a more sophisticated model. The simple EBMs can nevertheless be used to demonstrate the effect of the basic radiative mechanisms. The EBM shown in Figure 8.23(a) (see colour plate section) takes into account variations in the TSI associated with changes in solar activity (but no increases in the CO_2 concentrations), describes a warming from 1900 to 1950 and an almost constant temperature since 1950.

A simple energy balance model can be used to describe the temperature evolution, and Harrison and Shine (1999) expressed the model in terms of one equation, where C_s is the heat capacity of Earth's surface, t is time, λ is the climate sensitivity, and ΔF is the radiative forcing. The black body radiation has been approximated by a linear function in ΔT.

$$C_s \frac{d\Delta T}{dt} = \Delta F - \lambda \Delta T \tag{8.6}$$

Harrison and Shine (1999) assumed a value of $4 \times 10^8\,\mathrm{J\,K^{-1}\,m^{-2}}$ for C_s, roughly representative of a planet covered by a 100 m deep mixed layer ocean. The climate sensitivity was taken as $0.67\,\mathrm{K(W\,m^{-2})^{-1}}$, which is arrived at by assuming a temperature increase of 2.5°C for a doubling of CO_2 from pre-industrial concentrations. These parameters are rough guesses, and the model must demonstrate that the solutions are not too strongly dependent on these values. It is therefore important to apply a range of values for these parameters and compare the different results in order to make conclusions about how the global mean temperature responds to changes in solar forcing. Figure 8.23(a) (see colour plate section) shows the result of the energy balance model described by equation (8.6) with different values for the heat capacity C_s. A lower value for the heat capacity gives a greater temperature increase since 1900, but also unrealistic 11-year variations with unrealistic amplitude. Harrison and Shine (1999) show that the solutions using this simple model, forced with the four different solar reconstructions and the parameter values given above, suggest that a solar induced warming since 1900 ranges between 0.2 and 0.3°C, similar to the increase seen in Figure 8.6. A fifth reconstruction based on the formula proposed by Reid (1997) suggests stronger warming, but the evolution of this solution is notably different from the four others. The albedo feedback can also be included in these EBM (see box below).

Harrison and Shine (1999) bring attention to the fact that there are different ways of deriving past total solar irradiance, and the different approaches must be thoroughly evaluated in order to shed light on the differences between the solutions. Usually, there are three steps: (i) defining the variability of solar irradiance in terms of a proxy, (ii) estimation of the magnitude of the irradiance variations associated with this variability, and (iii) using recent measurements for an absolute reference level.

The reconstruction of Reid (1997) is criticised by Harrison and Shine (1999) for using climatic information in its derivation, which may render this reconstruction inappropriate for climate studies, as doing so may introduce an aspect of circular logic into the solution (this reconstruction may be perfectly all right for other uses). The magnitude of the irradiance changes in the second step is sometimes inferred from a reference to the Maunder minimum, and Reid (1997) estimated the slope of the linear fit by assuming that the cold conditions in the mid-17th century (1°C colder than at present) were purely due to changes in solar activity. The estimate for the global mean cooling during the Maunder minimum is not well known and factors such as internal variability, volcanism, changes in the greenhouse gas concentrations and land surface changes are not adequately dealt with, thus implying a high degree of uncertainty in the calibration of such models relating global mean temperature to the insolation.

The surface of the Earth may not be in energetical equilibrium with the rest of the universe since terrestrial heat is accumulated or lost from the planetary surface and the albedo is a function of temperature. A simple EBM that takes into account the effects of Earth's heat capacity (c) and the albedo (A) feedback (Section 5.6.3) can be expressed as

$$[1 - A(T_s)]\pi r_e^2 S(t) = 4\pi r_e^2 \sigma T_s^4(t) + 4\pi r_e^2 c\rho d_*(t)\frac{dT_s(t)}{dt} \tag{8.7}$$

The albedo is also assumed to be a function of temperature. In equation (8.7), only the upper d_* metres of the surface are influenced by the fluctuations in the solar irradiance on the timescale of interest (decadal). This parameter is a function of the forcing timescale and the soil's thermal conductivity and heat capacity (c). Although d_* is assumed to be the mean value for the planet, it is bound to vary geographically, with larger values over the oceans (due to water being more transparent than soil, and mixing processes). Both the parameters $d_*(T_s)$ and $A(T_s)$ may vary with seasons due to seasonally varying soil moisture, snow- and ice-cover, and vegetation, while c is taken as a constant.[a]

One may explore equation (8.7) by letting $S(t)$ vary with the solar cycle using a formula by Wilson and Hudson (1988, equation (4.23)). A simple assumption for $d_*(T_s)$: $d_*(T_s)$ is small during winter when the ground is snow-covered and there is frost in the ground. A simple function is assumed for the albedo where $A(T_s) = 1/1 + \exp[-\lambda(T' - T_0)]$, λ is a tunable sensitivity factor. The variable T' is the temperature anomaly, and $A(T' = 0) = 0.3$ so that $T_0 = 0.85/\lambda$.

[a] Variations in the soil's heat capacity are taken into account by using an effective depth d_* which is a function of temperature.

In Figure 8.23(b) (see colour plate section), the aspects of an albedo feedback and heat capacity are brought into the EBM model through a simple idealistic representation of an albedo. The albedo is assumed to be a function of temperature as it is assumed to be affected mainly by the snow, ice-cover and the surface's heat capacity. The influence of the temperature on the albedo is shown in Figure 8.24(a) (see colour plate section). Figure 8.24(b) shows how the albedo varies as a consequence of the fluctuations in the temperature and the TSI. The different curves represent different solutions with different albedo sensitivities (Figure 8.24(b)), and these results show that the emission temperature is greatly affected by the albedo sensitivity. The difference between the constant-albedo-solution (black curve in Figure 8.23(b)) and a solution with an albedo-response to temperature (red curve) suggests that the albedo-feedback may amplify the climatic response. The solutions are not as sensitive to different values for heat capacity as the choice of albedo-representation. For some configurations, the solution is stable and the fluctuations are small, whereas for stronger sensitivity (blue and grey curves) the runaway situations may take place where the temperature drops into an extremely cold or jumps into a warm state.

Kristjánsson *et al.* (2002, 2004) proposed that changes are amplified through affecting low clouds. They observed that the low cloud cover from the ISCCP data were negatively correlated with TSI (−0.5 after the annual cycle in cloud cover has been removed), however, the statistical significance was not very high. They nevertheless found a stronger correlation between the low cloud cover and TSI than from GCR (0.3 when annual cycle has been subtracted). Hence, they argued that the low clouds are modulated by TSI rather than GCR, and the apparent correlation with the latter is a result of TSI and GCR both being affected by changes in solar activity. A similar difference in correlation strengths was noted by Lockwood (2002), however, he pointed out that the two estimates were not statistically different. Kristjánsson *et al.* suggested that an increase in TSI results in higher SST, which subsequently reduces the static stability in the lower atmosphere. They showed that the low cloud cover is correlated with the degree of static stability. Thus, high TSI increases the SST, which reduces the static stability and hence reduces the low stratus cloud cover and reduces the albedo. In summary, the TSI–low cloud hypothesis proposes a positive feedback mechanism.

8.6.1.2 Recap of TSI reconstructions

In summary, there is a wide range of estimates for the solar output during the Maunder minimum, as suggested by the entries in Table 8.6. A comparison between the solar irradiance composite from satellite observations by Fröhlich and Lean (1998a) and solar irradiance reconstructions by Hoyt and Schatten (1993) suggests that the upward trend between the solar maxima in the reconstruction is not present in the composite data (1978–1997). The difference may be due to a regression component associated with the solar cycle length, which questions the relationship between the total solar irradiance and the solar cycle length (Harrison and Shine, 1999, p. 11).

Harrison and Shine (1999) cite work by Clough (1943) in their review on hypothesised links between the solar cycle length and climate variations. It is difficult to gather empirical evidence for stronger solar intensity related to shorter

Table 8.6. Estimated reduction in the total solar irradiance during the Maunder minimum as compared to the present state (0.37% = 5.5 W m^{-2}, 1.23% = 15 W m^{-2}, 0.24% = 3.3 W m^{-2}, 0.2% = 2.7 W m^{-2}).

	Proxy	Range (%)	Mean
Baliunas and Jastrow (1990)	Calcium line	0.2–0.24	1.5 W m^{-2}
Hoyt and Schatten (1993)			
Lean *et al.* (1995)			0.24%
Mendoza (1997)	Radii and rotation	0.37–1.23	
Nesme-Ribes and Manganey (1992)			0.5%
Solanki and Fligge (1998)			4 W m^{-2}
Soon *et al.* (1994)	Calcium line	0.2–0.7	0.37%

solar cycles, as direct space-borne measurements of the total solar irradiance do not go sufficiently far back in time. Correlation studies between these quantities are also likely to be associated with large uncertainties due to the difficulty of determining the solar cycle length (Mursula and Ulich, 1998) and the correct level of irradiance (Fröhlich and Lean, 1998a). Kelly and Wigley (1992), on the other hand, found some support for a link between the SCL and climate variability as the inclusion of a solar SCL term improved the fit between the global mean surface temperature and a simulation using an upwelling-diffusion energy balance model which also models the enhanced greenhouse gas forcing.

In a recent paper, Scafetta and West (2005) argued that changes in TSI have contributed at minimum ~10–30% of the global surface warming between 1980–2002. In their analysis they chose to use the TSI estimates by Willson and Mordvinov (2003)[17] where the TSI estimates for cycles 22–23 (1980–1991) are higher than cycles 21–22 (1991–2002) by 0.45±0.10 W m^{-2}. Their conclusion therefore hinges on the choice of TSI reconstruction. An obvious weakness of their analysis is that no solar proxies such as sunspot number (Figures 2.3, 4.13; Table 8.3), 10.7-cm flux (Figures 4.11), GCR (Figure 7.3; Benestad, 2005a), nor the aa-index (Figure 2.5: lower during cycle 22–23 minima than during 21–22 minima) do indicate an increase in solar activity, implying that a systematic and long-term increase in TSI must be unrelated to ordinary solar activity. Furthermore, the stratosphere has been cooling rather than warming, most of which in the lower stratosphere can be associated with stratospheric ozone depletion rather than enhanced solar UV associated with increased solar activity. The analysis done by Scafetta and West also suffers shortcomings as they in essence compare the magnitudes of two short band-pass filtered series within similar bandwidths. The fact that the filtered curves representing global temperature and solar activity did not exhibit one-to-one relationships may suggest that they are not directly related and that the estimation of climate sensitivity

[17] See Section 4.8.

based on the ratio of their amplitudes is not representative. Similar wiggles are given by the set up of the analysis, when similar band-pass filtering is employed for two series, and thus the question of statistical significance is crucial. Scafetta and West also argue that the climate response is stronger at lower frequency as the ocean damps the faster changes. It is well-known that the ocean variability often resemble red noise processes, but nevertheless, the most pronounced climate variations are the annual variations[18], ENSO related variability, and volcanoes. These have short timescales compared with 11 or 22 years, and it remains a challenge to explain how the response associated with these events are relatively pronounced whereas those with 11-year solar activity are heavily damped if the climate sensitivity is as high as their estimates ($0.11\pm0.02 - 0.17\pm0.06\,\mathrm{K}/(\mathrm{W\,m^{-2}})$ depending on 11-year or 22-year timescales). The oceanic response tends to be in the upper mixed layer (e.g., ENSO[19]), and does usually not involve the deep sea.

8.7 COMPARISONS WITH STELLAR STUDIES

8.7.1 The life cycle of a star

As the stars age, they go through several life-stages: young stars start as burning balls primarily of hydrogen that is converted to helium. Young stars are believed to rotate faster and more erratically than our Sun at the present, but slow down with age. As the hydrogen in their core is used up, the stars start to burn helium and heavier elements, resulting in a different energy production. At this stage, the stars start to bulge and grow larger. Lower temperatures also give a red colour. When all the material is burned up, the stars may end up as white dwarves, if their mass is not sufficiently large to produce a supernova (a magnificent explosion). The lifetime of a star is estimated to be of the order of 10 billion years.

The Sun, which is about 4 billion years old, is 30–40% brighter than it was just after formation, due to more recent burning of helium. It is also expected that the Sun will slowly become brighter and grow in size, and in a couple of billion years it will engulf the innermost planets (Mercury and Venus). However, this evolution is so slow that it will not be noticeable on Earth for a long time. It is nevertheless postulated that Earth's climate has been exposed to a slow, long-term increase in brightness since Earth's formation, but geological records suggest that the climate was not significantly colder in the beginning of Earth's existence, despite the weaker solar irradiation. This hypothesis is of course difficult to verify, and there is room for errors in the interpretation of the geological records, in assumptions associated with these, and in the stellar models.

[18] On regional scales.

[19] See Section 9.2.

8.7.2 Inferring the solar evolution from stellar studies

Ionised Ca-emissions can be used as a proxy for magnetic activity on the Sun as well as for stars. The Ca-emissions, it has been argued, closely track the facular brightness variations that control the long-term solar radiative output. Some solar reconstructions belong to the "astrophysical" class, deriving a value for the Maunder minimum irradiance using emissions in the core of the Fraunhofer calcium line at 393.4 nm. An empirical relationship between emissions in this line and the solar output from recent measurements is then extrapolated to a state with no sunspots, and this method gives a reduction of $1.5\,\mathrm{W\,m^{-2}}$ for the Maunder minimum.

Thirteen "sun-like" stars[20] have been monitored since 1966 (for 23 years), and some of them do not exhibit a stellar cycle. The Ca-emissions from the ones without a stellar cycle are low compared to the stars with stellar cycles. Baliunas and Jastrow (1990) have found evidence that may suggest there is low energy production in non-cycling stars (four stars were classified as "magnetically flat" whereas nine show a range of magnetic activity similar to that on the Sun). It has been proposed that the Sun with its contemporary 11-year cycle is more similar to the brightest of these stars with more stellar activity. Baliunas and Jastrow (1990) suggested that the non-cyclic stars are in a similar condition to the Sun during the Maunder minimum, and argued that low solar activity is accompanied by a reduction in the solar output. Lean and Rind (1998) plotted the facular brightening (Ca-emission) against the total irradiance and found a best-fit slope. From this slope they made an extrapolation to a Maunder minimum state and estimated that the total solar irradiance was 0.24% ($-3.3\,\mathrm{W\,m^{-2}}$) less during the Maunder minimum than now. Other studies have suggested a corresponding change in the range between 0.2% and 0.6%, and results from studies of other stars suggest that even larger variations in the luminosity are possible. There is, however, a spread in the calcium emissions from the non-cycling stars suggesting a range of irradiance values from $0.6\,\mathrm{W\,m^{-2}}$ to $5\,\mathrm{W\,m^{-2}}$ below present solar output. Soon *et al.* (1994) estimate the range to be $2.7\,\mathrm{W\,m^{-2}}$ to $9.5\,\mathrm{W\,m^{-2}}$ (mean of $5.5\,\mathrm{W\,m^{-2}}$).

The Sun is considered to be among the most active of the Sun-like stars and it is not clear whether the batch of non-cyclic "Sun-like" stars really are representative of the Sun during the Maunder minimum. Foukal *et al.* (2004) questioned the validity of the stellar-based irradiance reconstructions as they claimed that the bimodal distribution of stellar magnetic activity proposed by Baliunas and Jastrow (1990) is not found when more homogeneous star samples are studied in more detail. Furthermore, they argued that wide-band photometry of 18 solar analogs in a study by G. Henry did not show evidence of luminosity variations greater than 0.05%.

The criterion for being "Sun-like" is that the star has similar age and mass as the Sun, but not similar magnetic behaviour. A trivial point must be made here as it is important to ascertain that it is *not* because of the low intensity of the light (Ca-

[20] These stellar data were supplemented by sporadic observations from 61 other stars, starting from 1978.

emission) or some other reason that the cyclic variations are not seen, but that this feature really is absent in those non-cyclic stars. Also, the observations must not be affected by changes in the atmospheric transparency, but this can often be checked by using some "neutral" stars as baselines. It is therefore important to note that the stellar records in Baliunas and Jastrow's study were short (23 years) and a periodicity of about 9 years implies 2–3 stellar cycles. There have also been some suggestions that the Maunder minimum was not caused by solar changes but rather by volcanic dust.

8.7.3 The "snowball Earth" effect

Hypotheses have been proposed about a "snowball Earth" effect, most recently by two Harvard scientists, P. Hoffman and D. Schrag, suggesting that the Earth froze over completely about 600 million years ago (just before the Neoproterozoic era when recognisable animal life appeared), when the Sun was assumed to be 6% fainter than now (Hoffman and Schrag, 2000). There were four similar occasions between 750 and 580 million years ago, according to the Harvard scientists, when the Earth was completely frozen. However, the Earth is believed to have had a more habitable climate both before and after these events. The "snowball Earth" hypothesis was motivated by the surprising discoveries of glacial debris near the sea level in the tropics. Other geological clues include rocks with high iron content, which are believed to form only under conditions with little oxygen, and carbon isotopic ratios in the rocks revealing the signature of prolonged periods of low biological activity during these events. Calcium- and magnesium-carbonate minerals are also found just above the glacial debris, and have been interpreted as evidence for an extremely hot climate following these extreme cold events. The orientation of tiny grains in rocks sensitive to magnetic fields suggests that the rocks were formed in locations where the magnetic field was approximately horizontal. If the geomagnetic field resembled the present one in any way, one would expect that rocks formed near the poles would bear the signature of an almost vertically aligned magnetic field. The "snowball Earth" effect hypothesis is very controversial, and other researchers argue that other empirical evidence suggests that the oceans did not freeze over and that the conditions were quite normal (Adler, 2001).

The "snowball Earth" effect is based on a positive albedo feedback mechanism involving snow and ice, and the hypothesis is that this effect may be subject to a runaway situation where a cooling results in an extension of the ice- and snow-cover (see Figure 8.23(b)). When the area of the planet covered by ice and snow reaches a critical point, then the runaway takes over as an increasingly large proportion of the incoming solar energy is reflected to space. Atmospheric CO_2 produced by volcanoes may counterbalance the albedo effect, but CO_2 is normally deposited on rocks, for instance when silicate rocks are eroded and exposed. The carbon is removed from the atmosphere and from sea-water when it combines with calcium and magnesium to produce carbonate sediments. The "snowball Earth" condition can only last as long as the CO_2 concentrations do not accumulate in the atmosphere. The end of these glacial periods has been explained by such an accumulation of CO_2 in the atmosphere due to the insulating effects of the ice-sheets on the continents. The liquid water required for the erosion of rocks was frozen in the ice, shutting down the CO_2-sink. One interesting question is how such cold events may get initiated, even if they still may be regarded as speculative at this stage. There are no obvious solar features which explain such episodes, although the gradual brightening of the Sun may explain why there have not been more recent cases of the "snowball Earth" effect. It is proposed that the continents during the time of these global glacial periods were clustered near the equator, and remained ice-free sufficiently long for the glaciation to reach its critical runaway point.

8.8 IS THERE A RELATION BETWEEN SUNSPOTS AND RAINFALL?

Szocs and Kosa-Kiss (2001) propose that the solar particles are important for Earth's climate and that the flux intensity reaches a maximum when the sunspots are crossing the central solar meridian. If this is correct, then there ought to be a 27-day signal in the climate system. Szocs and Kosa-Kiss (2001) also argue that cyclones on Earth are affected by the solar particles flux, and hence there may be a signal in the rainfall records. Tinsley *et al.* (1989) proposed one mechanism that may account for a link between sunspots and cyclones. One may search for a solar signal in different data series through wavelet and other spectral analyses. Figures 8.25 (see colour plate section) and 8.26 show results from spectral analysis on the daily Oslo rainfall. The lack of a solar signal in a single series from only one location cannot be used as evidence against a solar–terrestrial link, but on the other hand, can a prominent solar cycle in a few locations give an indication of such a link? The most prominent signal in these figures is the annual cycle (365.25 days). There is, however, also a spectral peak at around 7000–8000 days, corresponding to 19–22 years. There is also a minor spectral peak at 1000–1100 days (around 3 years). Thus, the analysis may suggest that there is some rainfall variability that *may* be related to the Hale cycle, although a solar link would be rather tentative without a physical explanation. However, there is little sign of a spectral peak at 27 days (the period of one solar rotation). On the longer timescales, drift-ice tracers indicative of cool icy waters in the North Atlantic have been compared with high-resolution $\delta^{18}O$ measurements from stalagmites in Oman. The comparison has suggested reduced rainfall and diminished monsoon activity in periods with low solar activity (6300, 7400, 8300, 9000, and 9500 years ago; Bond *et al.*, 2001).

8.9 REQUIREMENTS OF SOLAR–TERRESTRIAL HYPOTHESES

Any study of a connection between solar variability and Earth's climate must involve historical observations, not only for developing empirical models but also for testing physically-based hypotheses. It is therefore important to have long records of high-quality observations of both the Sun and Earth's climate and to

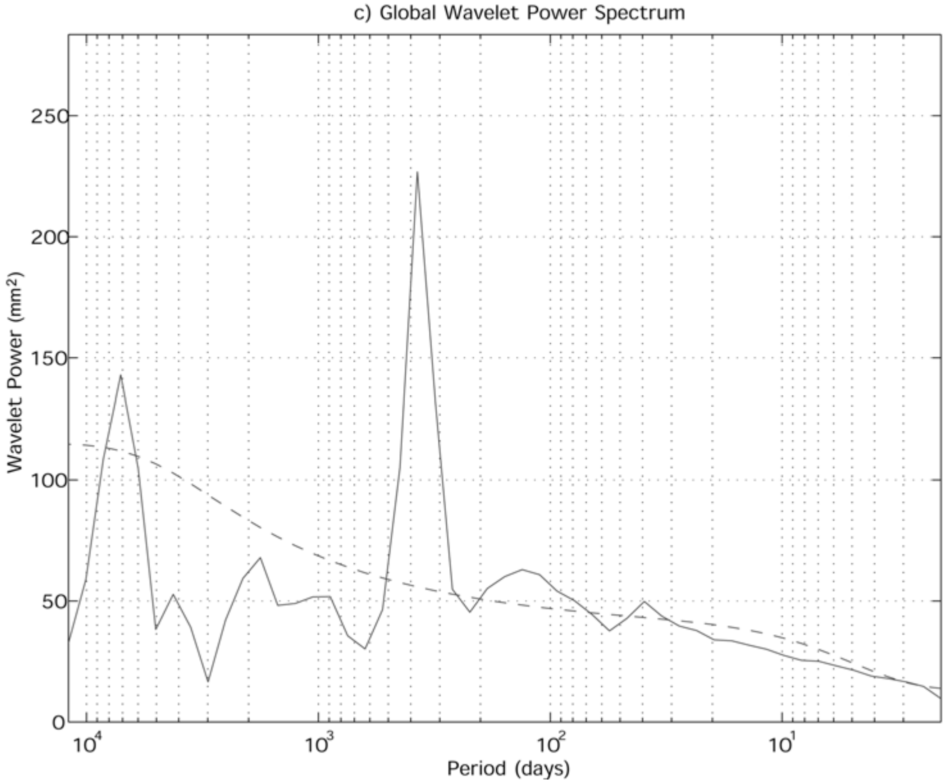


Figure 8.26. Spectral analysis of daily precipitation in Oslo, exhibiting power peaks at 1 year and 7000 days (19 years). Data from the Norwegian Meteorological Institute.

be able to formulate a physical model explaining the causality. It is unfortunate that there are few long-term observations that really are suitable for such studies and leading climate scientists have called attention to the state of the climate observational network. The observational programmes have been intended for use in weather prediction and short-term meteorological studies, and not for monitoring small, gradual, long-term climatic changes. Many observation sites have been subject to instrument replacements, displacements and encroachment of human development. There are vast regions with no observations going further back than 50 years, and some of these voids are filled in by statistical and dynamical analysis. The most complete long-term data sets that exist are surface measurements of temperature, sea level pressure, and sea surface temperature. Similarly, there are few high-quality, long-term solar observations, and one of the few direct solar quantities that has been recorded for more than 100 years is the sunspot activity. There have been measurements of the geomagnetic field for more than 100 years, and these records also hold some information about solar activity. Most efforts to find links between solar activity and the Earth's climate are therefore likely to involve the sunspots or the geomagnetic field.

9

Solar activity and regional climate variations

9.1 SYNOPSIS

Earth's climate encompasses natural variations on regional scales, and it is the sum of these regional climates that constitutes the global mean. There is a number of regional climate variations which each have characteristic patterns and traits that distinguish them from the rest of the climate system. These patterns and associated temporal variations are often referred to as *modes* and are of course affected by the rest of the climate system as well as exerting an influence on the rest of the climate themselves. Hence, it is not always easy to isolate these regional patterns. The modes of natural variability include El Niño Southern Oscillation (ENSO), the North Atlantic Oscillation (NAO), the Southeast Asian monsoon, the Quasi Biennial Oscillation (QBO), the Arctic Oscillation (AO) and the Pacific Decennial Oscillation (PDO). The relationship between the sunspots and the QBO is discussed in Section 6.5.2 and the AO is discussed in Section 6.6.

Numerical models of Earth's climate produce similar modes of natural variations to those found in the real climate system when the external forcing is constant, suggesting that the changes may be internal to the climate system (Houghton *et al.*, 2001). The coupling between various climate components associated with different timescales, such as the oceans, atmosphere and ice-sheets, can produce such variations. Climatic variations associated with both ENSO and the NAO are thought to arise from unstable air–sea coupling, nonlinearities or delayed action in the oceans. They are influenced by geographical features such as the shape and size of the ocean basin, as well as the latitude. The annual cycle has a profound effect on ENSO and the monsoon. If the solar irradiation undergoes changes, one would expect such changes also to affect Earth's climate. One important question is whether the variations in solar activity may affect the modes of natural variability.

9.2 EL NIÑO SOUTHERN OSCILLATION AND SOLAR ACTIVITY

Earth's climate is subject to natural variations which may affect the global temperature. One such phenomenon is the El Niño Southern Oscillation (ENSO), which is seen as inter-annually varying pressure, SST, and wind anomalies that are associated with timescales between 2 and 8 years. An El Niño event is the warm phase of the ENSO in terms of the SST in the eastern tropical Pacific. Thus, El Niños imply that the surface waters in the eastern tropical Pacific are warmer than normal. El Niños tend to start in the northern spring, and ENSO is therefore "tied" to the seasonal cycle.

It has been proposed that ENSO may be related to external forcing, such as solar activity. This hypothesis will be examined here, but in order to do so, a good knowledge of the nature of ENSO is required. ENSO has been associated with various traits and the dominant features include an oscillating pressure dipole located over Tahiti in the central equatorial Pacific and Darwin in northern Australia, SST anomalies in the eastern Pacific, and the reversal of the trade winds over the central and western Pacific. ENSO may also be associated with inter-annual variations in the south Asian monsoon, the position of the intertropical convergence zone (ITCZ), the hurricane frequency in the tropical Atlantic, the pressure system over north America, the Pacific North America pattern (PNA), and droughts over parts of Africa and Australia. The driving force behind ENSO is still not entirely understood, but various hypotheses have been proposed.

Easterly surface winds (blowing from the east) along the equator push the surface water westward, raising the sea level in the west where warm surface water is piled up. These easterly winds explain why the sea surface (the "warm pool") in the west is warmer than in the east. The prevailing easterly wind stress over the equator is thought to be responsible for the downward sloping thermocline[1] from the east to the west.

9.2.0.1 The tropical ocean–atmosphere system

Easterly trade winds near the equator force surface currents and are responsible for an *Ekman drift* which results from the balance between the surface friction and the Coriolis force[2] (Figure 9.1). The Ekman drift is often associated with divergence in the equatorial ocean mixed layers (Gill, 1982) and it is thought that the upwelling only takes place in the upper layer of the ocean. In order to conserve mass, deep water must flow into the mixed layer from below. The vertical advection of sub-surface water is often referred to as Ekman pumping and produces a narrow band of upwelling along the equator (Gill, 1982; Philander, 1989); this upwelling varies with ENSO as the trade winds change. The

[1] The thermocline is the depth in the upper ocean where temperature changes most rapidly in the vertical direction. It is often taken as the interface between the surface waters and the deeper ocean layers.

[2] The Coriolis forcing term is zero and changes direction on the equator.

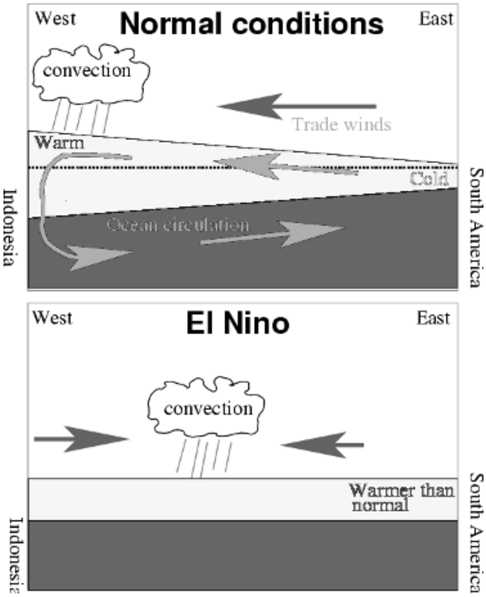


Figure 9.1. A schematic illustration showing the essential differences between the "normal conditions" and El Niño conditions. Under normal conditions, the stronger east-to-west surface winds drag the equatorial surface water westward, causing a pile-up of warm surface water in the west. The warmer western surface is accompanied by atmospheric convection and a west-to-east pressure gradient at depths below the sea level in the east. The pressure gradient drives the west-to-east equatorial undercurrent (EUC). During El Niños, the trade winds relax, the eastern tropical Pacific becomes anomalously warm, the surface water is no longer piled up in the west, and the atmospheric convection is shifted to the central tropical Pacific.

inter-annual variations in the upwelling are accompanied by inter-annual SST anomalies (Figure 9.2). A combination of upwelling and a shallow thermocline may be responsible for the equatorial cold tongue in the eastern Pacific.[3] The ocean surface will be cooled if the vertical advection extends down through the thermocline when the thermocline is shallow. The warm pool in the western Pacific is associated with a deep thermocline.

9.2.0.2 Delayed action oscillator hypothesis and oceanic waves

McCreary (1983) proposed a conceptual model of ENSO where Rossby waves generate low-frequency oscillations which may be associated with ENSO. He conducted a coupled ocean–atmosphere model study, based on a model of the Pacific Ocean surface layer and an atmospheric model. The atmospheric model is formulated as two wind patches, which represent the *equatorial* zonal wind stress and

[3] The thermocline may be shallow because of upwelling.

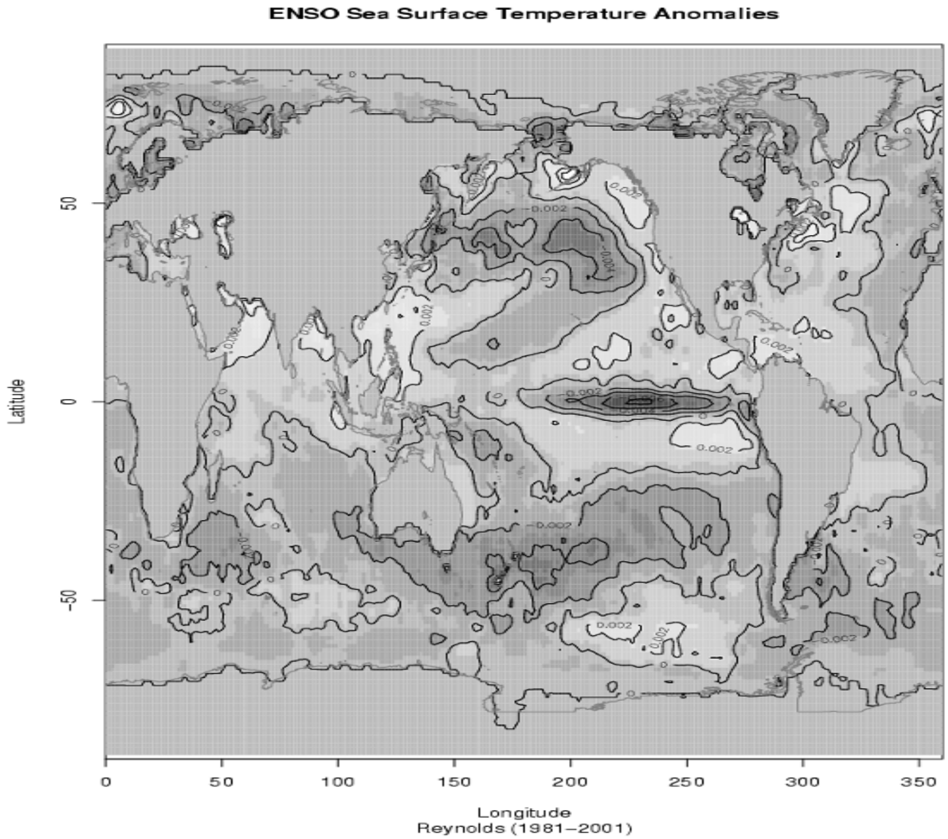


Figure 9.2. Geographical distribution of SST anomalies associated with ENSO. Data from Climate Diagnostic Center (CDC), USA.

the *extra-equatorial* zonal wind stress respectively (Figure 9.3). The surface winds in the model respond to changes in the SSTs. For instance, equatorial easterly wind anomalies appear over the central ocean when the east ocean is cool and the thermocline in the west is deeper than in the east (J. Bjerknes's Walker circulation). When the east ocean is warm, on the other hand, the atmospheric model simulates extra-equatorial easterlies (enhanced Hadley circulation).

The oceanic response to the changes in the wind stress in the central *equatorial* Pacific is westward-propagating equatorial Rossby waves. Easterly wind anomalies produce downwelling Rossby waves. When the *extra-equatorial* easterlies switch on, however, westward-propagating extra-equatorial downwelling Rossby waves are excited. These extra-tropical downwelling Rossby waves are accompanied by equatorial upwelling Rossby waves. Both upwelling and downwelling waves influence the thermocline depth.

The Rossby waves may reflect as Kelvin waves when they reach the western boundary. The upwelling or downwelling properties are assumed to be conserved

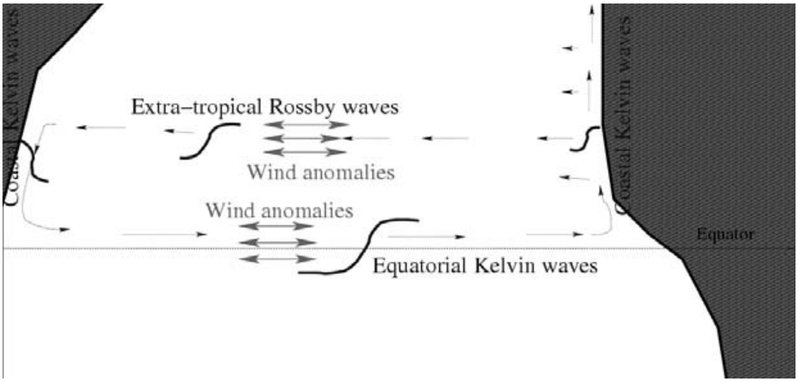


Figure 9.3. A schematic illustration of the delayed oscillator.

during a reflection. The Kelvin waves may deepen or shoal the thermocline as they propagate eastward, depending on whether they are downwelling or upwelling. A deepening of the thermocline is often associated with a warming at the sea surface in the eastern Pacific.

The Kelvin waves may become coastally trapped Kelvin waves when they arrive at the eastern boundary. However, some coastal waves can radiate wave energy in the form of westward propagating Rossby waves as they move poleward. These Rossby waves eventually reach the western boundary, and subsequently reflect as equatorial Kelvin waves.

In short, the delayed oscillator hypothesis describes a delayed negative feedback mechanism, in which warm SSTs in the east Pacific eventually are reversed by the appearance of extra-equatorial easterlies and the generation of upwelling Kelvin waves. McCreary and Anderson (1984) and McCreary (1983) showed that timescales of a few years can be accounted for by these wave processes. More recently, it has been pointed out that the propagation speeds of the lowest baroclinic mode Kelvin waves (2.5 m/s) and Rossby waves (–0.8 m/s) imply too short an ENSO period if only the first baroclinic modes[4] are important; however, the second baroclinic modes may also play a role and can account for the timescale of 3–8 years. Neelin (1991) and Kirtman (1997) offer possible explanations for these discrepancies and describe the "slow SST–fast wave" and "fast wave–slow SST" modes. The former hypothesis assumes that the thermodynamics adjust slowly to the changes in the thermocline variations and the latter implies that the atmosphere reacts slowly to the SSTs. Kirtman (1997) also argued that the slower extra-equatorial Rossby waves may be important for the delayed oscillator mechanism.

[4] Oceanic waves can be decomposed into vertical modes, also known as baroclinic modes. These are different solutions of the wave equations that may be present independently of each other, represent wave solution with different phase speeds, and describe different depth-profiles in terms of the wave flow. A discussion of these modes can be found in Gill (1982) and Philander and Pacanowski (1981) pp. 160–173.

Schopf and Suarez (1988) proposed a similar "delayed action oscillator" hypothesis, in which the easterly wind anomalies prevailing in the extra-equatorial regions in the central Pacific force westward-propagating Rossby waves. They used a $2\frac{1}{2}$-layer ocean model coupled to a 2-layer atmospheric model, and found irregular oscillations that involved uncoupled Rossby waves and Kelvin waves travelling between the western boundary and the active forcing region. The irregularities were believed to be caused by nonlinearities in the atmosphere.

More than one Rossby wave can give irregular oscillating solutions similar to ENSO. The Rossby waves and the Kelvin waves are thought to be excited by changes in the wind stress. The causes of fluctuations in the winds are not well known, but it is believed that the winds in the tropics are strongly affected by the SSTs. One way of looking at the coupling process is in terms of the heat flux from the ocean to the atmosphere which increases the energy budget of the atmosphere. Thus, heat loss in the ocean may lead to an increase in latent and atmospheric potential energy, which in turn can be converted to kinetic energy. Ultimately, heat fluxes from the ocean to the atmosphere can give rise to winds. If the heat fluxes depend strongly on the SSTs, then the winds are also likely to be linked to the SSTs.

9.2.0.3 The coupled slow mode mechanism

A warm SST anomaly in the equatorial Pacific is usually associated with westerly zonal wind stress anomalies located to the west of the SST anomaly (Figure 9.4). Westerly winds may cause zonal advection of SSTs, and may therefore be responsible for increasing the SSTs eastward of the warm anomaly. The anomalous winds can also reduce the divergent flow at the equator, and hence reduce the upwelling. Lower upwelling rates are usually associated with warmer SSTs in the east Pacific. Conversely, the surface water to the east of the SST anomaly may cool due to diverging Ekman transport (positive zonal wind anomaly, stronger divergence, increased upwelling, cooling).

If the upwelling dominates the SST changes, then the coupled mode moves slowly westward. However, if the advection has a stronger influence on the SSTs, the SST anomaly may propagate eastward. In other words, the coupled modes are sensitive to the type of heat equation that describes the thermal processes (SST in the surface layer) in the ocean.

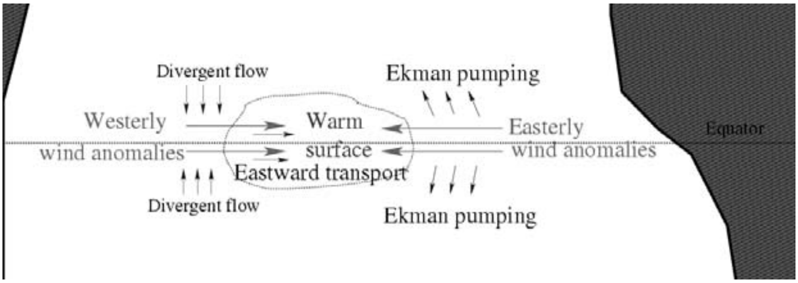


Figure 9.4. A schematic illustration of the coupled mode.

The physical mixed layer processes present may be crucial for the coupled Kelvin modes. The mixed layer temperature has a strong relation to the thickness of the mixed layer (Kraus and Turner, 1967). The depth of the base of the mixed layer is in turn affected by the turbulent mixing (wind stress) and the buoyant stability (insolation). Pacanowski and Philander (1981) proposed that the mixing also depends on the Richardson number of the flow, i.e. the current shear and hydrostatic stability.

The slow mode hypothesis, described by Hirst (1986, 1988), Anderson and McCreary (1985), and Wang and Weisberg (1994) assumes that the SSTs are dominated by the divergence and the convergence of the surface wind stress. Hirst (1986) observed that the character of the unstable modes depends on the relative position between the atmospheric heating and the wave crest. Kelvin waves in his model were unstable and first meridional mode ($m = 1$) Rossby waves were damped if the atmospheric heating was centred near the wave crest, but the opposite was true when the atmospheric heating source was displaced westward by a quarter to a half wavelength of the wave crest. The behaviour of the modes depended on whether advection or upwelling contributed most to the SST changes because these terms controlled the location of the atmospheric heating relative to the wave crest. The growing modes can be found by computing the eigenvalues and eigenvectors of the coupled ocean–atmosphere system (Hirst, 1988).

Anderson and McCreary (1985) and Wang and Weisberg (1994) showed that the ocean model can sustain growing oscillations when the ocean basin is sufficiently wide, and when the atmospheric wind pattern over the ocean contains both a divergent pattern over one region and a convergent field over another region of the ocean. In their study, the coupled mode travelled slowly eastwards as it grew.

The eastern edge of the warm pool and the westerly winds progressed eastward at a speed of 0.6 m/s for a few months during late 1991, and Kessler and McPhaden (1995) suggested that this may have been caused by a coupled mode mechanism.

9.2.0.4 Local ocean–atmosphere feedback mechanisms

The heat fluxes between the oceans and the atmosphere involve evaporation, radiation, and thermal conduction. It is difficult to measure or estimate the heat fluxes, and they are often parameterised in a coupled model. The heat fluxes are thought to depend strongly on the surface winds as well as the respective temperatures of the atmosphere and the oceans.

Regions of warm SST are associated with warm and moist air that will ascend if the atmosphere is unstable, and hence can give rise to convergence near the sea level. If the growth of the convection intensity and spatial extent is unstable, the convergence zone may become large in a short time and thus affect the large-scale circulation (i.e. Madden–Julian oscillation (MJ0), tropical cyclones, ITCZ). It is possible that convection may play a role in ENSO phenomenon. For instance, past studies suggest that the position of the ITCZ convective centre is correlated with the ENSO cycle (Philander, 1989, p. 33) and that it can alter the winds in the western tropical Pacific. The convective centre of the ITCZ may perhaps be part of a feedback process.

Local processes involving cloud formation, evaporation rates, precipitation and topography, as well as local wind fields, may also be central to the ENSO evolution. One of these is addressed in the Ramanathan hypothesis (Ramanathan and Collins, 1991), in which cirrus clouds affect the local radiation balance and act as a thermostat.

A nonlinear (chaotic) atmospheric response can have more than one stable solution, and the convergence zone may act as a trigger that decides in which state the atmosphere must be (McCreary and Anderson, 1991). The nonlinearity may reside in the coupling (lower boundary conditions of the atmosphere) as well as in the internal atmospheric dynamics, and may be responsible for instabilities in certain situations.

9.2.0.5 Alternative explanations for ENSO

Barnett (1983) suggested that ENSO could be a result of the interaction between the Indian monsoon and the Pacific trade wind field. He proposed that the monsoon and the trade winds expand and contract in anti-phase. Thus, the trade winds are weak when the monsoon expands into the western Pacific.

The variations in the heat content in the upper layer of the tropical oceans are slow compared with the characteristic timescale of the atmosphere. There are suggestions that a subsurface memory is important for ENSO. Zebiak and Cane (1987) described an ENSO model, in which the inter-annual variability is linked to the subsurface memory of the system. They demonstrated that the model was sensitive to the way in which the ocean and the atmosphere were coupled.

9.2.0.6 Stochastic ENSO models

Lau (1985) proposed that intra-seasonal (intra-seasonal timescale of 40–50 days) variability associated with the Madden–Julian Oscillation (MJO) may play a central role in triggering ENSO events. Westerly inflow to the west of enhanced convective regions over the western Pacific may affect the SSTs in the eastern Pacific through remote forcing. This hypothesis may therefore involve oceanic intra-seasonal Kelvin waves that link the forcing in the west with the ocean response in the east.

Wyrtki (1985) proposed that warm water is accumulated in the western Pacific until the warm water pool becomes unstable to high-frequency atmospheric forcing. This "pile-up theory" could also be an important mechanism for ENSO.

Penland and Sardeshmukh (1995) proposed that ENSO could be explained in terms of linear dynamics and white noise. They derived a linear model of the tropical SSTs using an inverse modelling technique (POPs). Most of the El Niño events, except for the warmest ones, could be explained by the linear model and they suggested that the linear system need not be unstable to explain the growth of SST anomalies. The warm and cold events were explained by a constructive interference of several damped linear modes of SSTs. The implication of this study is that stochastic forcing is an essential part of ENSO. However, they found that the random forcing could not be white in both time and space, but rather the forcing must have a certain spatial structure.

Moore and Kleeman (1997), Blanke *et al.* (1997) and Eckert and Latif (1997) conducted various experiments with different (hybrid) coupled models where they introduced high frequency stochastic noise. They all argued that the noise affected the ENSO events and reduced the predictability of ENSO. In the model of Moore and Kleeman (1997) the inter-annual variability was sensitive to "stochastic optimals" that produced wind patterns which resembled westerly wind bursts. Their results implied that the intra-seasonal[5] Kelvin waves may play an important part in triggering the ENSO events. ENSO affects these intra-seasonal Kelvin waves (Benestad, 1997), and it is not impossible that they play a role in a coupling between physical processes associated with intra-seasonal and inter-annual timescales.

9.2.0.7 ENSO and the Madden–Julian oscillation

Kessler *et al.* (1995) analysed a 10-year-long time-series of SST and 20°C isotherm depth,[6] and found that variations in the intra-seasonal energy in the ocean coincided with the so-called Madden–Julian oscillation (MJO). The amplitude of the intra-seasonal waves was, however, modulated by a low-frequency inter-annual signal. During El Niño onset years, the convection extended further east, and this was thought to give more fetch to the westerlies, which in turn gave rise to unusually intense downwelling intra-seasonal Kelvin waves. Thus, intra-seasonal variability seemed to be modulated by ENSO. Kelvin waves were present during all phases of ENSO, but they were stronger during the warm phase.

Kessler and McPhaden (1995) found various propagating features in the 20°C depth field at different timescales. Low-pass-filtered observations of 20°C isotherm depth showed a slow propagation with a speed of 0.1 m/s (a large-scale phenomenon). However, the same observations with higher time resolution suggested that these large-scale features were composed of many small scale phenomena. The small-scale features included equatorial Kelvin waves from the western boundary with phase speeds = 2.4 m/s.

The MJO usually appears over the central Indian Ocean as an organised region of deep convection and is accompanied by westerly wind bursts. The first features of the westerly wind bursts can often be described as a westward moving easterly wave in the surface winds over the tropical western Pacific (Riehl, 1954; Kindle and Phoebus, 1995). As the convective activity intensifies, it moves eastward with a speed of 3–6 m/s. When the MJO events propagate past the edge of the west Pacific warm pool, the convection and surface winds become very weak.

Kessler *et al.* (1995) suggested that the changes in the amplitude of intra-seasonal variability may be a mechanism by which the Pacific is affected by low-frequency signals from outside the Pacific basin (i.e. the Asian monsoon and the

[5] Timescale of 30–90 days.

[6] The 20°C isotherm is often taken as proxy for the thermocline in the tropics. The isotherms are "surfaces" (depths or heights) of constant temperature.

Indian Ocean). However, McPhaden *et al.* (1986) suggested that the wind-burst events may even have been a symptom rather than a mechanism triggering El Niño because the variations in the SSTs in the Indian Ocean seem to lag behind those in the Pacific.

During El Niño phases, the location of the wind bursts tends to be shifted closer to the equator and eastward (Kindle and Phoebus, 1995). There is only a vague suggestion of the individual intra-seasonal wind events propagating eastward into the central Pacific. However, successive events often extend further west than the preceding event (Kessler *et al.*, 1995). Kessler and McPhaden (1995) proposed that the reason why the Madden–Julian oscillation may penetrate further east during the warm ENSO phase is that the associated westerlies develop longer fetch as the warm pool extends further east.

9.2.1 How solar activity may physically affect ENSO

9.2.1.1 Physical considerations

When considering possible roles that solar activity may have in terms of natural variability such as ENSO, the physical aspects should be established before considering empirical evidence. Since the ENSO phenomenon still is elusive regarding driving mechanisms, it is important to consider various explanations for how the tropical inter-annual variations may be generated.

Oceanic equatorial intra-seasonal Kelvin waves play a central role in several of the ENSO models (see Section 5.5.2.4), such as the delayed oscillator mechanism and the stochastic forcing hypotheses. The coupled mode mechanism may also involve intra-seasonal Kelvin waves through a feedback mechanism proposed by Kessler *et al.* (1995). However, the relationship between the intra-seasonal Kelvin waves and the inter-annual timescales is not yet fully understood, and there may be room for other factors, such as external forcings. It has been demonstrated that inter-annual variations in the coupled atmospheric–oceanic systems can adequately be explained in terms of the above-mentioned ENSO model, without the need of external forcing. So, the question is whether changes in the solar irradiance *may* bias the system or if sunspots and faculae, with approximately similar timescales as intra-seasonal winds and Kelvin waves, *may* influence ENSO. In this respect, which of the ENSO models really play a role in nature? Can, for instance, the solar cycle affect the triggering of El Niño events in the stochastic ENSO models?

Any hypothesis relating solar activity to ENSO is still speculative and there is as yet no widely accepted theory on such a link. Most ENSO models do not take solar activity into account. Therefore, the mechanisms outlined below must be regarded as highly hypothetical.

In the 19th and early 20th centuries, before ENSO was a well-known concept, there was a belief that 4-year variations in the tropics were a consequence of solar activity. For instance, Arctowski (Helland-Hansen and Nansen, 1920, p. 159) proposed in 1915 that the tropical 2.75-year temperature variations were forced by shorter variations in solar activity. It is now generally acknowledged that this feature is primarily due to internal variability through air–sea interaction, since ENSO models which do not take solar activity into account reproduce the regional climate variations with realistic timescales and magnitude. So, the question is not so much whether these fluctuations are caused by solar activity, but whether changes in the Sun are influencing the course of ENSO.

One of the more obvious explanations as to how solar activity may influence ENSO involves the accompanying changes to the TSI. If ENSO is due to unstable air–sea coupling, then a small perturbation in solar energy may be sufficient to affect the system. If ENSO is a "chaotic" system, then fluctuations in the energy may produce large changes later on. However, increased solar activity may not necessarily be associated with a systematic shift in the ENSO system due to nonlinear actions. Therefore, a lack of correlation cannot be taken as evidence of there being no link between sunspots and ENSO.

The trade wind system that is responsible for the east–west asymmetry in terms of SST and precipitation may be regarded as a link between the Hadley and Walker cells and ENSO. As the Hadley and Walker cells provide a foundation for the interannual variations associated with ENSO, any changes to these systems may theoretically also affect the character of ENSO. Haigh (1996) has proposed that an increase in the UV radiation accompanying intensified solar activity may result in stratospheric heating (see Chapter 6). The stratospheric winds are strengthened as a consequence and the tropospheric sub-tropical jets displaced poleward. The latitudinal boundaries of the Hadley cell determine the position of these jets, a poleward shifting of which implies a similar shift in the descending limbs of the Hadley cell. Furthermore, such displacements may also affect the positions of the mid-latitude storm tracks. Similar solar-induced effects on the stratosphere have been suggested by Shindell *et al.* (1999) and Salby and Callagan (2000).

Another way that solar activity may affect ENSO is through a modification of the clouds or insolation. Ramanathan and Collins (1991) have put forward a controversial hypothesis stating that the clouds act as a thermostat that may play an important role for ENSO. Svensmark (1998) and Marsh and Svensmark (2000) argue that the low cloud cover may be regulated by solar activity through interplanetary magnetic fields accompanied by the solar wind and their shielding effects on galactic cosmic rays (see Section 7.10.5.1). It is therefore plausible that there may be a link between solar activity and ENSO facilitated by clouds if Svensmark's hypothesis is true. Farrar (2000) reported a high correlation between the cloud cover over the central tropical Pacific and the NINO3 ENSO index.

The question whether solar activity can trigger or terminate ENSO events is a difficult one. It is not impossible that changes in solar activity may lead to changes in the surface winds, and these may produce Kelvin waves that may kick off an event. Comparing the northern spring solar activity (i.e. sunspot number, flares) when the El Niños are usually initiated, to subsequent winter ENSO indices (the mature phase of ENSO) may illuminate this question.

White *et al.* (2000) reported that the decadal fluctuations in the global mean upper ocean temperature varied in phase with the decadal variations in the Sun's irradiance ($\pm 0.5\,\mathrm{W\,m^{-2}}$) over the past 100 years. The amplitude in the sea temperature is about 2–3 times greater than theoretical estimates based on a transient balance between TSI and Stefan–Boltzmann radiation. Higher decadal temperatures are found to accompany the reduction in the trade wind intensity across the tropics (hence decreasing the global average latent heat flux) and a considerable part of the upper ocean temperature variations is not related to the solar cycle but is driven internally by the nonlinear atmosphere–ocean systems such as ENSO. White *et al.* (2000) conducted a model study based on a delayed action oscillator with a theoretical solar-induced Stefan–Boltzmann upper ocean temperature response superimposed, and found a weak decadal signal coherent with the solar cycle.

Mann *et al.* (2005) conducted a study with an intermediate coupled ocean–atmosphere model for the tropical Pacific forced with volcanic and solar forcing reconstructed for the past 1000 years. They found that the model favoured El Niño states after volcanic events, but also that acombination of solar and volcanic forcing may explain the some of the past evolution of ENSO. They concluded that the ENSO response was likely a consequence of a dynamic adjustment to radiative forcing.

Landscheit (2000) argued that ENSO is subject to strong solar forcing, and makes an association between the ENSO phase and solar rise-to-maximum and fall-to-minimum. He also linked ENSO to the 22-year Hale cycle and the magnetic reversals in the sunspot activity. Perturbations in the rate of change in the solar orbital angular momentum (about the solar system's centre of mass) recur, according to Landscheit, at quasi-periodic intervals and may somehow initiate phase reversals in climatic phenomena. He predicts the next El Niño to appear around year 2002.9 ± 0.4 and the La Niña[7] phase will be predominant in the current 22-year cycle. The ideas of Landscheit have been viewed with scepticism.

9.2.1.2 Empirical relationships between ENSO and solar activity

After the physical considerations have been outlined, it is important to look at the empirical evidence supporting or excluding the hypotheses. In conjunction with the statistical analysis, it is also important to keep in mind the data quality and possible errors associated with the ENSO indices and the solar activity. Therefore, the results must be robust and not sensitive to the data set or method. There are various independent ENSO indices associated with the SST (NINO3 or NINO3.4) and SLP (SOI: standardised pressure difference Tahiti–Darwin) respectively, and these should point to the same conclusions if the results are robust.

Figure 9.5 shows a comparison between three ENSO index time-series and the sunspot number. The figure does not show any clear association between the sunspot number and ENSO. Figure 9.6 shows a scatter plot between the ENSO indices and the sunspot number. The correlation between spring-time R_z and winter ENSO index is also shown, and none of these analyses suggest that there is a connection between ENSO and solar activity. Moreover, these analyses indicate that the hypothesis

[7] A La Niña is the opposite to an El Niño, i.e. cool eastern tropical Pacific surface water and strong trade winds.

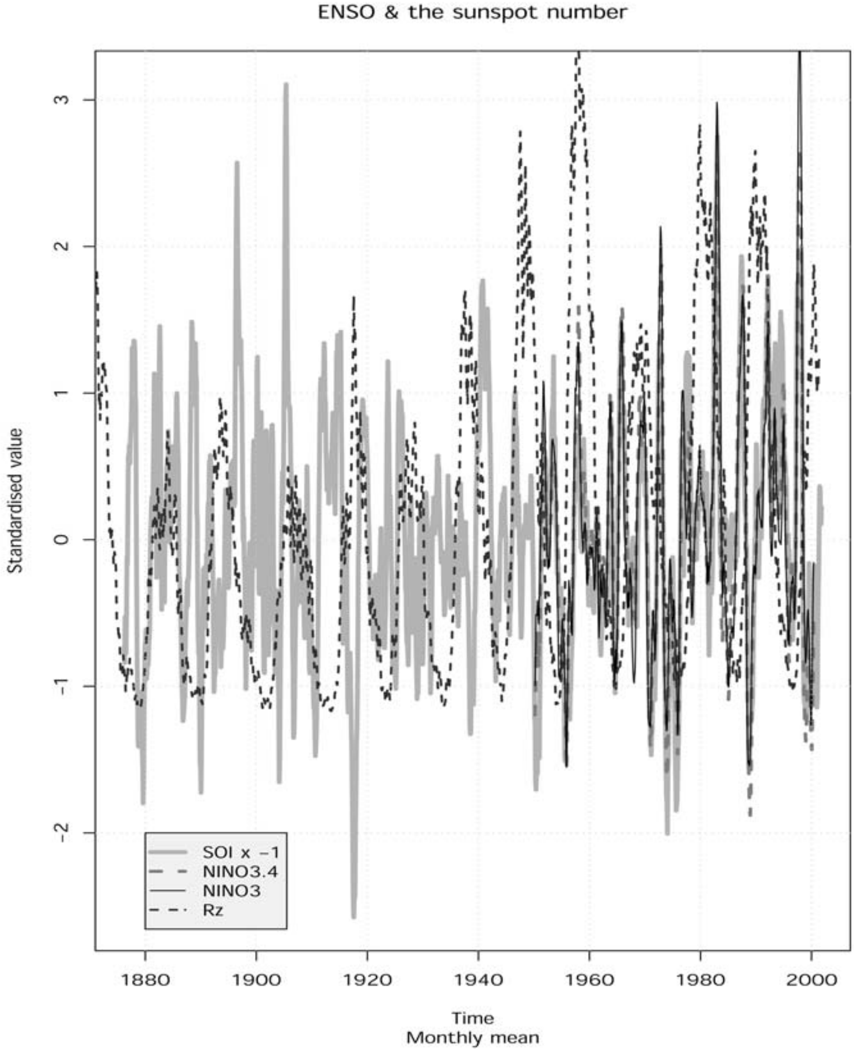


Figure 9.5. Time-series of an ENSO index (NINO3: spatial mean SST anomaly over 150°W–90°W 5°N–5°S) and the sunspot number. There is no apparent correlation between the El Niños and the sunspot number. Also shown are NINO3.4 (SST 170°W–120°W 5°N–5°S) and SOI. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and Bureau of Meteorology, Australia.

states that the variations in TSI associated with the solar cycle do not produce a clear direct response in ENSO. However, if the evolution of ENSO is chaotic, then a small perturbation in the climate system may lead to larger changes later on, and such a link may not yield a high correlation between the sunspots and the ENSO indices.

Hoyt and Schatten (1993) have argued that changes in the solar irradiance will produce coherent variations in the temperature between the two hemispheres, but

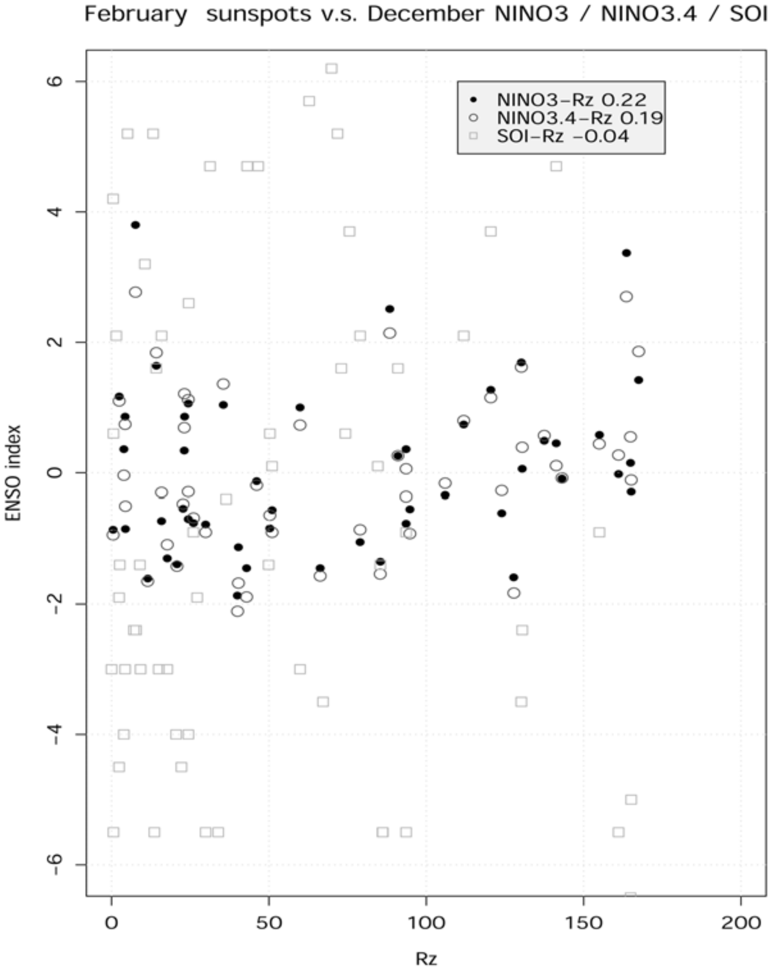


Figure 9.6. Scatter plot between February R_z and following December NINO3 for the period 1950–2000 and SOI for the period 1876–2001. The correlation between the monthly mean sunspot number for January–December was 0.11, 0.19, 0.17, 0.15, 0.11, 0.13, 0.10, 0.09, 0.19, 0.17, 0.06, and 0.12 respectively. As the El Niño events start during northern spring, the high correlations in September and October are coincidental since the solar activity during these months takes place after the events are set off. None of the correlation values qualifies as statistically significant at the 5% level. Also shown are NINO3.4 and SOI. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and Bureau of Meteorology, Australia.

that their amplitude will differ because the northern hemisphere contains more land area whereas a larger portion of the southern hemisphere is covered by oceans. They applied a correlation analysis on de-trended series of northern and southern hemisphere temperatures and found a correlation of 0.55 at a 99.9% significance level (0.80 for non-de-trended series). Thus the coherent variations between the two hemispheres was used to support the hypothesis of the presence of an external forcing source. ENSO involves a large region of the tropics in both hemispheres and its variations can be seen in the global mean temperature record. Hence, temperature variations associated with ENSO will produce inter-hemispheric correlations. It is easy to demonstrate from climate model simulations without solar forcing that climatic variations in the two hemispheres may be correlated without a solar forcing (Benestad, 2000b). Hoyt and Schatten (1993) compared the annual mean northern hemisphere temperature from Hansen and Lebedeff (1988) with modelled solar irradiance and argued that approximately 71% of the temperature variations during the past 100 years can be accounted for by secular changes in the solar irradiance.

9.3 THE ROLE OF SOLAR ACTIVITY IN THE SOUTH ASIAN MONSOON SYSTEM

9.3.1 The monsoon

The south Asian monsoon is the wet season which starts around the beginning of June in southern India. The main mechanism behind the monsoon seems to be understood and this involves the land–sea temperature contrast. One may, however, again ask the question whether sunspots *may* after all play a role, such as biasing the rain statistics. The monsoon has been observed and studied for a long time and the all-Indian summer rainfall is still hard to predict. The lack of success at predicting monsoon characteristics may be due to an incomplete knowledge of important details of the monsoon system or due to its chaotic nature. The south Asian monsoon is associated with increased rainfall and southwesterly winds, and is essentially a large-scale sea breeze situation driven by the land–sea temperature contrast. A simplified description of the monsoon is that air is heated more over land than over sea, and the warming over land triggers convection, cloud formation, pressure gradients and a large-scale sea breeze. The total monsoon rainfall varies from year to year, and the wet season usually consists of rainy spells interspersed with sunny days.

Past studies have suggested that the monsoon rainfall may be influenced by factors other than the sea surface and land temperatures. A relationship between ENSO and the amount of precipitation during the monsoon has been suggested by Barnett (1983), but this link does not yet have a solid physical explanation. Kumar *et al.* (1999) found a strong anti-correlation between the El Niño episodes and the all-Indian rainfall before 1988, but more recently this relationship seems to have broken down. There have also been studies indicating that there is a link between the snow-cover over the Himalayas and the monsoon. Thus, if this link is real and the snow-cover is influenced by solar activity, then there may be an indirect association between the solar cycle and the monsoon. Again, the physical link between the solar activity and the snow-cover must be established as well as between the snow-cover and the monsoon.

Wang *et al.* (2005) inferred a connection between solar forcing and the south Asian Monsoon from oxygen isotope ratios (δ^{18}O) from stalagmites from the

Dongge Cave in southern China. They argued that δ^{18}O-ratios were low for heavy Monsoon rains and high for dry conditions. A correlation between the oxygen isotope ratio and δ^{14}C of 0.30 was interpreted as evidence for some of the variability in the Asian Monsoon being associated with solar changes. If the carbon-14 ratio is affected by climate itself in any way, this could give a false impression of a link with solar variability (a kind of circular reasoning), however, Wang *et al.* (2005) did not rule out such possibilities in their paper.

9.3.2 Relation to sunspots

Sir Norman Lockyer, one of the founders of the journal *Nature*, attempted in the late 19th century to use sunspot information to predict the monsoon rains, but with little success. Since then, the mainstream view in the field of meteorology and climatology is that the monsoon is a phenomenon which is internally driven, and that external factors such as solar activity are of little importance. One may speculate about whether the small change in the solar irradiance due to the Schwabe cycle may affect the land–sea temperature contrasts or the snow-cover over the Himalayas. If the sunspots affect ENSO, then there may be an indirect way that the sunspots also affect the monsoon, provided there is a connection between the two weather systems. More speculative hypotheses may involve cosmic rays and cloud cover or magnetic field disturbances.

9.4 THE NORTH ATLANTIC OSCILLATION AND SUNSPOTS

It has been known since the 18th century that when it is cold in east Greenland the climate over Denmark is usually mild. The Danish missionary H. E. Saabyr in 1770–1778 noted that severe winters in Denmark concurred with mild conditions over Greenland. There is also a near out-of-phase relationship between the sea level pressure variation over Iceland and the Azores (Figure 9.7). There is a semi-permanent low-pressure system sitting over Iceland, whereas the Azores are usually blessed with high-pressure conditions. The north–south difference in the sea level pressure drives so-called geostrophic winds, which are mainly from the west and bring mild maritime air over northern Europe (Hurrel, 1995). When the pressure levels over Iceland and the Azores fluctuate, the Icelandic low tends to deepen while the Azores high is strengthened (or vice versa), increasing (decreasing) the meridional sea level pressure gradient, which subsequently enhances the westerly winds over the North Atlantic Ocean and the Norwegian Sea. This dipole pressure system (Figure 9.7) tends to oscillate and is often referred to as the North Atlantic Oscillation (NAO) and is most pronounced during winter.

The Norwegian meteorologist J. Bjerknes proposed that a coupling between the ocean and the atmosphere is responsible for the year-to-year variations in the SLP pattern over the North Atlantic. It is still unclear whether the ocean temperature anomalies have significant influence on the atmospheric circulation sufficient to explain the NAO phenomenon. It is possible that NAO is internally atmospheric,

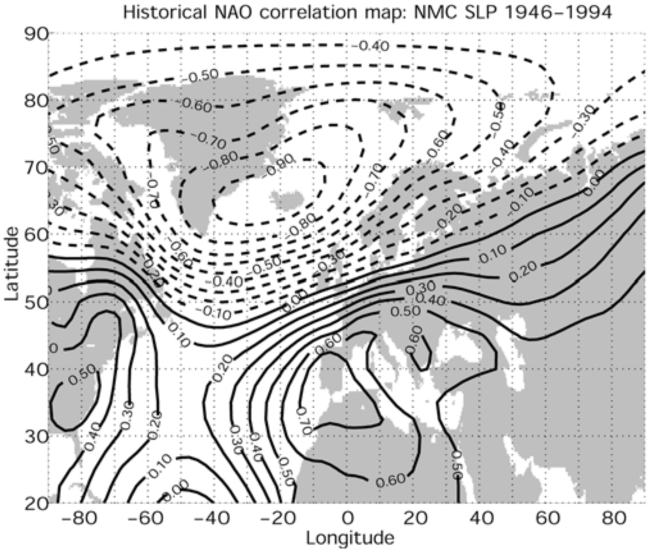


Figure 9.7. The correlation between local SLP and the NAO index (standardised pressure difference between Iceland and Lisbon) exhibits the spatial NAO structure consisting of a north–south dipole. Data from NMC (now NCEP) and CRU.

and the fluctuations are chaotic due to nonlinear dynamical interactions. It is also possible that the changes in the energy and moisture supply from the oceans due to changing sea surface temperatures may influence the NAO state or its course.

The NAO has a considerable effect on climate variability over northern Europe, and especially during the winter. A positive NAO phase, usually identified in terms of a positive NAO index (NAOI), is associated with lower than normal sea level pressure (SLP) over Iceland and higher SLP over the Azores, stronger westerlies over the Norwegian Sea, enhanced precipitation and mild conditions over northern Europe. The centres of action (local maximum and minimum SLP) are not geographically fixed, but may "wander" around. A positive NAO is also accompanied by dry and cold conditions over Iberia, and the NAO affects the position of the North Atlantic storm tracks.

9.4.1 How may the NAO be affected by solar activity?

It is important to note that the NAO can be reproduced in coupled ocean–atmosphere models with no solar activity, and therefore the solar activity is not required in order to explain the NAO. The question is therefore: can the solar activity perturb the NAO so that it alters its state or course and what would be the physical explanations for such a link? As the NAO is still an elusive phenomenon, an improved physical understanding is required in order to propose such mechanisms. So, any relation between the solar activity and the NAO must be purely speculative at this stage. If the sea-ice influences the NAO, one such mechanism could be through a

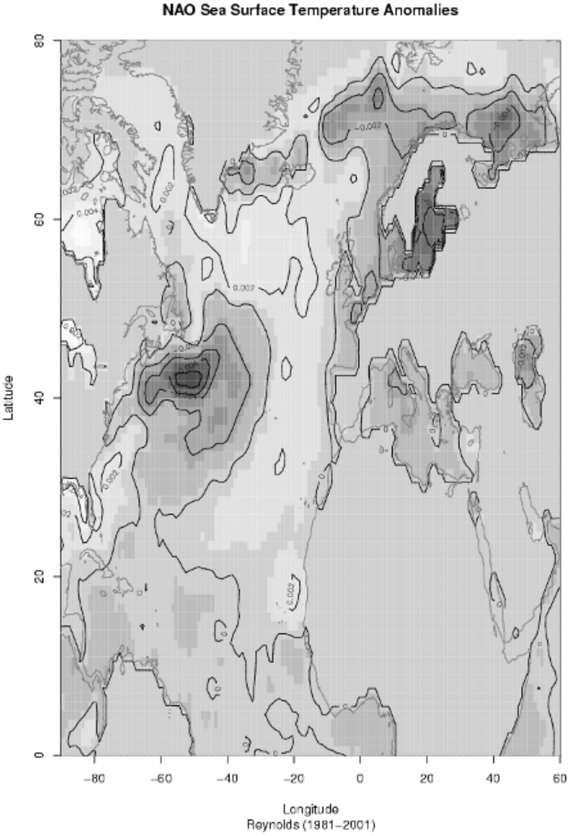


Figure 9.8. Anomalous SSTs related to the winter-time (December–April) NAO over the period 1981–2001. The pattern was derived from a regression analysis on 3 of the SST PCs, and the fit is associated with an $R^2 = 0.15$, and a p-value of 0.001. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and Reynolds and Smith (1994).

connection between the solar activity and the sea-ice. A weak solar signal can be seen in the sea-ice data (Figure 5.15 in Section 5.5.3).

Another mechanism where sunspots affect the NAO may involve sea surface temperatures. Reid (1987) reported a statistical relationship between the sunspots and the sea surface temperatures, and Figure 9.8 demonstrates that there is indeed a connection between the sea surface temperature and the NAO. It is not impossible that clouds may act as a medium through which the solar activity may affect the NAO. Svensmark and Friis-Christensen (1997) have proposed that the cloud cover is modulated by the solar activity via galactic cosmic rays (see Section 7.10.5.1). There have also been suggestions that the NAO is related to the Arctic oscillation, which may be driven or influenced by stratospheric processes, and there is little doubt that the upper atmosphere is influenced by the solar activity. The Arctic oscillation is discussed in Section 6.6, and the next section will focus on empirical evidence that can be used to support the hypotheses about a solar NAO connection.

Ogi *et al.* (2003) suggested that the spatial structure of the sea level pressure variations in the northern hemisphere was influenced by the level of solar activity. They argued that during solar maximum, the pattern had the more hemispheric character associated with the Arctic Oscillation (AO), but the signal was confined to the Atlantic region during solar minima just like the NAO pattern. They also suggested the wintertime NAO has a significant effect on climate conditions during spring and summer when the solar activity was high, but not during low solar activity. They explained that the mechanism by which this link could exist was through a systematic influence on snow and ice conditions in the Barents Sea region. Although supported by empirical data, their analysis involved short data series, and it remains to be seen whether the purported relationship holds for the future. Dima *et al.* (2005) applied multivariate techniques (Empirical Orthogonal Function, Principal Oscillation Pattern, and Singular Spectrum Analysis) to SST and SLP data, and their interpretation of their results was that the AO mode responded to solar forcing but the NAO was due to air–sea coupling. The solar signal was found to be weaker than internal variability. They proposed that solar response in the stratosphere affected the lower troposphere through large-scale annular stirring from the stratosphere above. One potential shortcoming of the analysis of Dima *et al.*, however, was that they applied a 9–14-year, band-pass filter to the data prior to most of the multivariate techniques and did not test the sensitivity of the results to the choice of preprocessing.

9.4.1.1 Empirical relationship between the NAOI and sunspots

Figure 9.9(a) shows a comparison between the NAOI and the sunspot number since 1821. The curves have been 11-year low-pass-filtered for clarity as it is difficult to tell whether the curves resemble each other from just the unfiltered series. Panel (b) of Figure 9.9 shows a spectrogram of the NAOI, which suggests that there is no particular timescale that characterises the NAO.[8] The lack of a common characteristic timescale in the NAO and the solar activity suggests that the NAO is not influenced strongly by the solar activity. Figure 9.10 presents scatter plots of unfiltered December-to-February mean values of the NAOI and the sunspot number respectively. Also shown in both figures are the annual mean NAOI. There is no resemblance between the NAOI and sunspot curves in Figure 9.9 and the correlation analysis suggests that these series are uncorrelated. Hence, the empirical evidence suggests there is no linear relationship between the sunspots and the NAO.

9.5 THE GULF STREAM AND SUNSPOTS

The Gulf Stream is known to be responsible for a considerable part of oceanic heat transport and hence influences sea surface temperatures over substantial parts of the North Atlantic Ocean. A comparison between the meridional excursions of the Gulf

[8] The NAO has sometimes been dubbed the "Noisy Atlantic Oscillation".

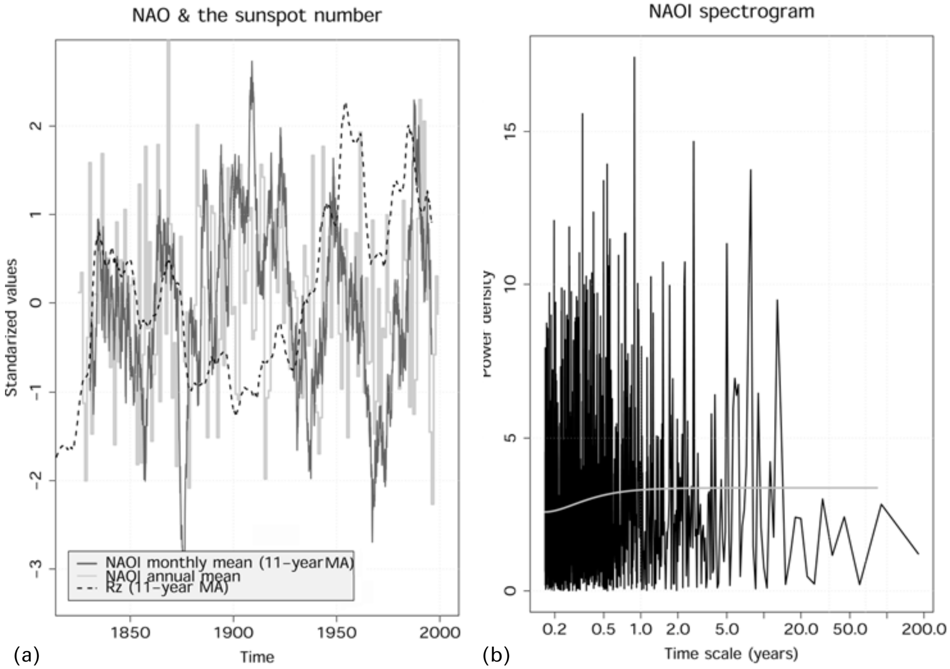


Figure 9.9. (a) A comparison between standardised 11-year low-pass-filtered NAOI (dark grey) and sunspot number (black). Also shown is the annual mean NAOI (light grey "step" type). (b) A spectrogram showing the power density associated with the various timescales. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and the CRU.

Stream just off the American continent and the sunspot number (Figure 9.11) may be used for studying links between solar activity and oceanic circulation. One intriguing observation is the tendency for a northerly displacement of the Gulf Stream after sunspot maxima and before sunspot minima. The big question is whether the pattern is coincidental. If both series are characterised by fluctuations with similar timescales, then a comparison between a small number of cycles may yield an apparent connection between the series. Figure 9.12 gives the power spectrum of the Gulf Stream record. The power spectrum exhibits a spectral peak in the vicinity of 8–11 years which is similar to the solar cycle. This intriguing comparison is no evidence of a relationship between the two series as it stands. In order to establish whether there is a link between solar activity and the Gulf Stream displacement, a physical explanation must be sought: how can the sunspots affect the Gulf Stream position? (Why do they not influence the NAO in the same way?) What sets the preferential timescale of the Gulf Stream latitude? These are important questions which must be answered before a link between the two records can be established.

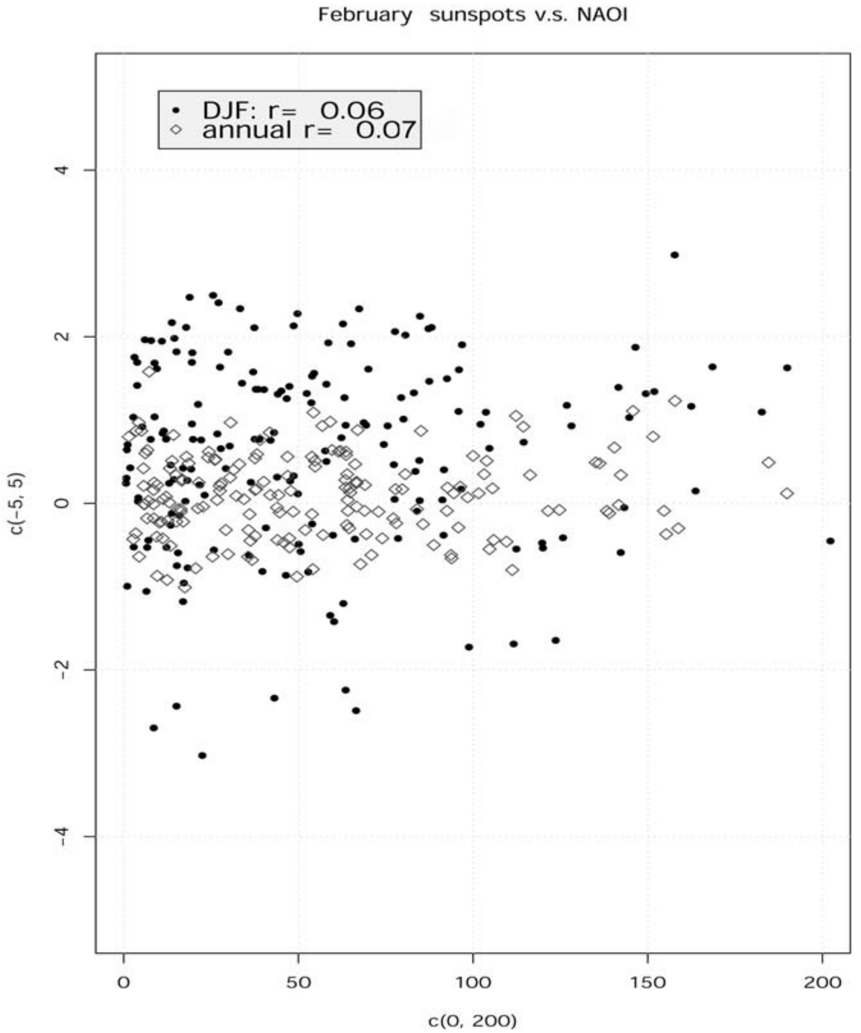


Figure 9.10. Scatter plot showing the correlation between the winter (DJF) NAOI and corresponding sunspot number and the same analysis for the annual mean values. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and CRU.

9.6 PACIFIC DECADAL OSCILLATION AND SOLAR ACTIVITY

Recently, attention in climate research has been on decadal variations in the North Pacific and a possible discovery of a new mode of natural variability (Zhang *et al.*, 1997; Mantua *et al.*, 1997). Variations with timescales of the order of 10 years may potentially be related to the 11-year solar cycle. Figure 9.13 shows a comparison between an index of the *Pacific Decadal Oscillation* (PDO) and the sunspot record.

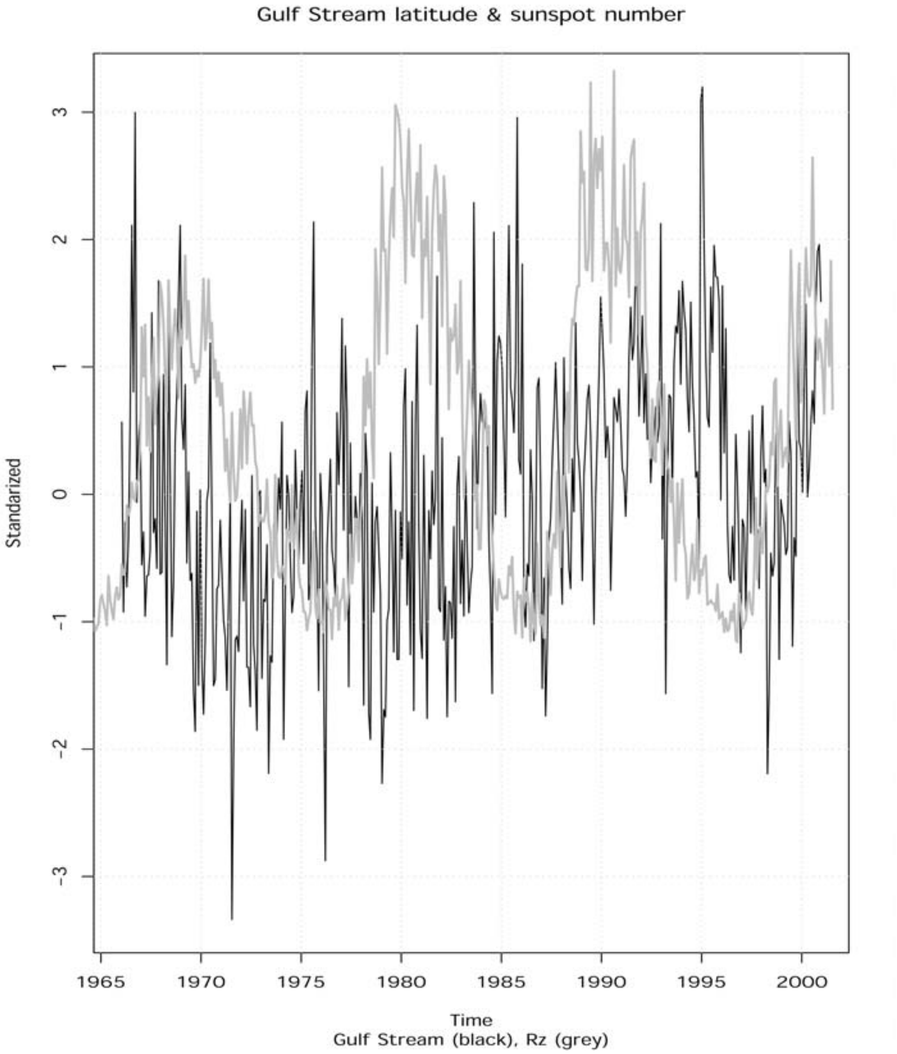


Figure 9.11. A comparison between the monthly mean sunspot number and the latitudinal excursions of the Gulf Stream near the east coast of the USA. The Pearson correlation between two series is −0.04 (p-value = 0.376). Sunspot data from ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and the Gulf Stream position from Taylor (1995).

The PDO index has been derived by taking the (standardised) values of the leading PC of monthly SST anomalies in the North Pacific Ocean north of 20°N. A correlation between the PDO index and the sunspot record of −0.11 with a p-value of zero suggests that there may be a weak link between the PDO and the solar cycle. However, before such a relationship can be established, a physical model must be in place that can explain in detail how the solar activity can affect the PDO. It was

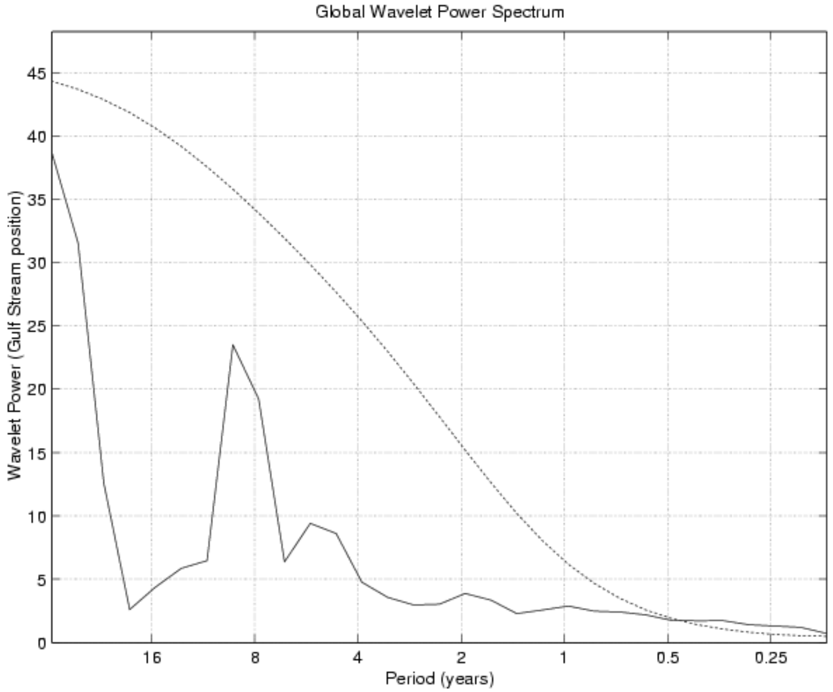


Figure 9.12. Spectral analysis of latitudinal excursions of the Gulf Stream indicating a preferred timescale of 8–11 years. Data from: Taylor (1995).

shown in Figure 8.17, in Section 8.4.10, that two data records may have a high correlation even if they are unrelated.

9.7 LAND–SEA CONTRASTS

Shindell *et al.* (2003) conducted a number of GCM experiments to investigate the temperature response to pre-industrial forcings (solar and volcanic). They found that solar forcing (through stratospheric effects, Section 6.4.1) may explain a large statistically significant response that is of same sign as proxy data and exceeds internal (unforced) variability. Their model results produced stronger cooling over the continents during the little ice age (LIA), consistent with indications based on proxy data. However, the model anomalies were more pronounced than the paloeclimatic data. The response to volcanic forcing, on the other hand, gave no consistent pattern of temperature change over longer intervals, although eruptions can be seen as brief spikes of cooling in the proxy and model data. The proxy data may implicitly be smoothed as a limited number of eigenvectors were used. It is also important to note that continental proxies tend to be more sensitive to the summer season and not give

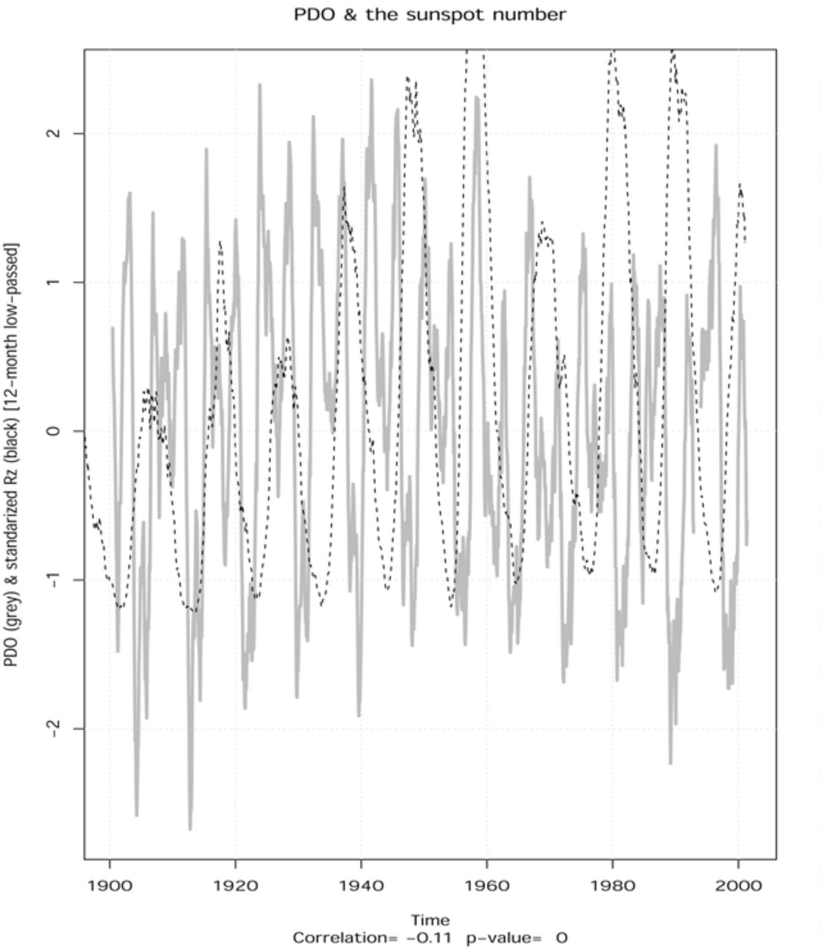


Figure 9.13. A comparison between the sunspot record and the PDO index. Data from ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and ftp://ftp.atmos.washington.edu/mantua/pnw_impacts/INDICES/PDO.latest.

an adequate representation of, for example, the winter warming associated with volcanic eruptions.

9.8 OTHER EXTERNAL CLIMATE FORCINGS

9.8.1 The anthropogenic perturbation to the natural balance

Historical temperature measurements indicate that the global mean near-surface temperature has warmed by about 0.6° ± 0.2 since 1860 (Houghton *et al.*, 2001).

Meanwhile there is a cooling trend in the stratosphere. The warmest year on record[9] on a global level was 1998 and the global mean sea level is estimated to rise by 0.2–0.3 mm/year. Reports on glaciers and ice-sheets suggest that many of them are getting thinner and are receding, although some glaciers have grown (in Norway). At the same time, the atmospheric CO_2 concentrations have increased as a result of an increase in anthropogenic energy consumption and the burning of fossil fuels. Recent emissions of CO_2 enhance the carbon-12/carbon-14 ratio, as fossilised organisms have low carbon-14 content.

The accumulation of CO_2 and other greenhouse gases in the atmosphere leads to a vertical redistribution of solar energy and hence a warming of Earth's surface and a cooling of the upper atmosphere. Greenhouse gases such as CO_2, CH_4, CFCs and water vapour absorb up-going infrared radiation emitted from Earth. The molecules of these atmospheric trace gases re-radiate on average half of the absorbed energy back to ground. There is little doubt that the greenhouse effect is real, as an energy balance between incoming solar and emitted long-wave radiation points to an emission temperature (256 K, see Section 5.4.3) approximately 30°C lower than the estimated global mean temperature (288 K). A local manifestation of the greenhouse effect can be experienced directly in the high-latitude winter nights: Cloud-free nights tend to be much colder than when it is cloudy. The clouds trap the heat radiated by the Earth's surface and then re-radiate some of this heat back to the surface. Nevertheless, the question as to whether the addition of man-made gases may have a discernible influence on the climate system has been controversial in some quarters.

9.8.2 Orbital parameters

The geological climate archive suggests that Earth has experienced ice ages and warm periods. Some valleys are remnants of ice carving into the mountain, and moraines are witness to the retreat of glaciers. The different climatic epochs were first explained by Milankovitch (1941) in terms of changes in solar energy due to variations in Earth's orbital parameters.

9.8.2.1 The Milankovitch cycles

Earth's eccentricity (ϵ) varies periodically with a periodicity of 97,000 years, and the tilt of its axis (obliquity, ε) undergoes similar cycles but with a timescale of 40,000 years. The spinning Earth also precesses like a spinning top, and the precession of the perihelion (ω) completes every 21,000 years. The reason for the precession is the torque due to the interaction between the Earth, the Sun and the Moon, and the precession may be regarded as the Earth's wobble (Kleppner and Kolenkow, 1978, p. 300). The current values of the orbital parameters are listed in Table 9.1.

Milankovitch proposed that the changes in the orbital parameters led to slight

[9] Before 2005.

Table 9.1. Orbital parameters: from Peixoto and Oort (1992), Houghton (1991) and Kleppner and Kolenkow (1978). The range reflects the different values for the periods given by the various sources.

Parameter	Period (kyr)	Max./Min. value	Present value
Eccentricity	97–100	0–0.05	$\epsilon = 0.0167$
Obliquity	40–41	22–24.5°	$\varepsilon = 23.27°$
Precession	21–26	−23.5–23.5°	$\omega = 0$ for Polaris reference frame[a]

[a] The axis is pointing towards the North star.

variations in the summer-time insolation in the polar regions. The variations in solar energy were thought to be sufficient to cause growth and retreat of the ice-sheets, and thereby affect the climate system through the ice feedback mechanism (see Section 5.6.3).

Jenkins (2000) has carried out model simulations with different angles of obliquity (54°–70° which are much higher than today) which suggests that high obliquity angles favour a warmer climate.

9.8.3 The Moon

Even the Moon influences Earth's climate through its gravitational pull (Kleppner and Kolenkow, 1978, p. 350). The most pronounced effect is that of the tidal waves in the ocean. The Moon may have important subtle consequences for the climate as the tides may be important for mixing processes (Egbert and Ray, 2000; Wunch, 2000), which again affect the ocean circulation and the SSTs. The oceans may subsequently affect the atmosphere. The tides have periodicities of 12 hours + 25 minutes, 27.3 days (Fleagle and Businger, 1980, p. 19), but also timescales of 18 and 34 years (Lean and Rind, 1998).

9.8.4 Meteorite impacts

Meteorites originate from the asteroid belt. It has been speculated that comet collisions may have been the origin of the inner planet atmospheres (Encrenaz, 1997). Meteorite impacts on Earth may be immensely destructive, causing widespread forest fires and explosions. The impacts also whirl up dust into the atmosphere and may affect the atmospheric transparency. It has been proposed that a massive meteorite impact may have killed off the dinosaurs during the end-Cretaceous mass extinction at about 65 million years ago (Vines, 1999).

10

Synthesis

There are clear solar signatures in the atmosphere above the stratopause. For instance, there is strong empirical evidence pointing to solar activity influencing the northern lights, and the solar link is explained by undisputed theoretical models. High solar activity is accompanied by more intense solar winds which result in disturbances in the geomagnetic field. There is also little doubt that the cosmic ray flux incident on the Earth's surface is modulated by solar activity.

Although recent solar–terrestrial studies tend to make few references to related studies earlier than 1950, there was active research on connections between the solar cycle and climate variations in the late 19th and early 20th centuries. Recent empirical studies point to many of the same findings as the studies carried out in the 19th century. So, have we become any wiser about solar activity and the Earth's climate since the 19th century? We now think we know much more about the Sun itself and have a greater understanding about how the sunspots arise and what they "do". Furthermore, we also have more data and observations to analyse, new and better analytical methods, and more computing resources. Thus, the three "pillars" (observations, methods and theory) of scientific knowledge introduced in Chapter 1 have been reinforced since the 19th century, even though the conclusions about solar–terrestrial links may be the same.

One major advantage that the modern scientist has over the 19th-century scientist is the computer. Which can be used to test analytical methods. This book gives an example of how this may be done by exploring the importance of filtering in correlation analysis. Computers may also be used to examine observational data and control its quality. In Section 5.2.1.2 an example is given on how station records can be examined for the effect of urban encroachment. We have since the 19th century learned more about various shortcomings of the observational records, and now to a greater extent acknowledge the fact that the observations may contain errors. The theories themselves can be tested using computers as long as there is a mathematical formulation of the system. Thus, the scientific knowledge of modern times is much more robust than in the 19th century, since all of the three "pillars of knowledge" may be reinforced through computer-based studies. Finally, the computer provides an access to the Internet, where an abundance of information, data, and numerical tools are freely available (see Table 11.5 in the Appendix). The modern computer is the scientists' "Swiss Army knife".

We now know that statistical analysis, and correlation studies in particular, should be viewed with caution, as there have been several examples of tests suggesting a link between two quantities which in fact are unrelated. Recent analyses nevertheless seem to confirm a link between solar activity and the climate, albeit a weak one for timescales less than 100 years. The TSI varies in phase with the sunspot cycle and this signal has a weak influence on the climate. There seem to be similar weak 11-year solar signals in several independent observations: the land surface temperature record, sea surface temperatures, the MSU data, and the sea-ice extent. Had there been a strong link, it would probably have been established long ago.

There has been some debate about indirect solar influences on climate, possibly mediated through interplanetary magnetic fields, cosmic rays and cloud formation. However, such a mechanism cannot yet explain the recent warming trend which is strongest at night-time. This caveat by itself does not prove the cosmic rays hypothesis wrong, but the fact that this is still strongly disputed in the scientific communities suggests that a link between cosmic rays and the climate is still not established. The solar influence on stratospheric ozone may, on the other hand, be one mechanism through which variations in solar activity may affect the climate. Another mechanism may be variations in the sea-ice extent, as such variations may affect the planetary albedo and the coupling between the oceans and the atmosphere. However, it is still not clear how important these mechanisms are for the real climate system.

Whether the long-term variations in the Earth's climate are related to changes in solar activity is much more uncertain than the solar effects on 11-year timescales. The effect of the solar activity is small compared to other factors, such as internal variations, which makes it hard to identify solar signals in the data. The terrestrial temperature record suggests that there was an interval of global warming between 1900 and 1930, which the climate models cannot explain in terms of the enhanced greenhouse forcing alone. It has been proposed that this may be due to an increase in solar activity. The sunspot number at the solar maxima increased up to 1957, then dropped slightly and has levelled off since. Other proxies for solar activity suggest peak activities at slightly different times, suggesting that these indices only have a "loose" connection to the underlying solar physics. It is still not certain whether the observational records are of the sufficiently high quality required for studies on long-term relationships between solar activity and variations in our climate. The present analytical methods and empirical data cannot establish whether there really is a link between the solar epochs (SCL) and the climate. So far, the studies claiming such a connection do not stand up to the tests of objectivity.

It is interesting to note the lack of trends in solar proxies such as GCR, 10.7-cm solar flux, sunspot number, the aa-index, and SCL during the most recent 50 years (e.g., Figures 4.10, 4.11, 7.3; Richardson *et al.*, 2002; Kristjánsson *et al.*, 2004; Benestad, 2005a). There is furthermore no clear trend in the IMF-based open solar magnetic flux since 1950 (figures 2 and 7 of Lockwood, 2002). However, the 11-year running average of open flux derived from both the aa-index and the IMF peaked in 1987, and has decreased since then (Lockwood did not explain how he dealt with the end points of the series). The lack of trend in GCR constitutes strong evidence that the most recent global warming is not due to changes in the GCR, and hence falsifies the claim by Svensmark (1998, 2000). An increase in solar activity is also generally associated with enhanced solar UV and a warming of the stratosphere (Haigh, 1999; Crooks and Gray, 2005), at odds with the observed stratospheric cooling. Willson and Mordvinov (2003) argue that there has been an non-negligible increase in TSI during the solar minimum of solar cycles 22 and 23, however, other studies disagree with their results. If there really has been a trend in TSI over the last couple of decades, this must have implications for all solar TSI reconstructions where the TSI is based on the solar indices. Hence, if Willson and Mordvinov (2003) are correct, this may invalidate a large number of studies of historic TSI construction and its likely effects on climate. Another challenge is also to explain how the TSI would change while (virtually) all other solar indices do not exhibit corresponding behaviours.

Palaeo-proxies suggest that there have been pronounced long-term variations in the prehistoric past. A part of these variations may be explained in terms of changes in Earth's orbit, tilt and wobble, but there are also a number of events which cannot be explained by the orbital parameter theory. Volcanoes, meteor impacts, geological changes, changes in the composition of the atmosphere, and internal variations may account for some of the fluctuations, as well as variations in solar activity. Since the prehistoric data is both scarce and associated with a high degree of uncertainty, it is hard to distinguish between the various factors that may have played a role. It is very plausible that solar activity has been one important factor.

It is important to acknowledge the fact that a climatic response to several forcing factors will probably *not* be a linear sum of the isolated effect of each individual forcing, but that the net effect will more probably be a result of complicated nonlinear interactions between different effects (Meehl et al., 2003). Different forcings will probably involve the same climatic feedback mechanisms. Therefore, it is wrong to assume that solar activity is unimportant for terrestrial temperature trends, even if an enhanced greenhouse effect can explain a similar warming as observed in the past century. It is equally wrong to argue that the anthropogenic greenhouse effect is insignificant if variations in solar activity may produce a similar temperature change to that seen in the observational records. Hence, without detailed and accurate knowledge about the complex climate system and solar activity, the anthropogenic warming theory cannot be taken as an antithesis to solar–terrestrial links.

Anybody with access to the Internet may get information on solar–terrestrial relationships through the various search engines.[1] It is probable that most readers have access to a personal computer, and using computers in learning should not be restricted to professional scientists. One "criterion of science" is that results must be

[1] For example, *Google* or *Netscape*.

repeatable, and I hope that some readers will attempt to repeat some of the demonstrations in this book in order to become convinced.

In summary, a weak but apparently robust solar cycle signal can be found in various independent observations using various statistical tests. Furthermore, there are physically-based explanations which suggest how solar activity may influence our climate. Thus, all the three criteria for ‘‘scientific knowledge’’ proposed in Chapter 1 are in place for solar–terrestrial links. Finally, it is important to re-iterate the fact that a solar–terrestrial link does not diminish the importance of other factors, such as volcanism, meteorite impacts, aerosols, landscape changes, and changes in the atmospheric greenhouse gases. Neither is it certain that the response to the various factors would be linear; most scientific tests (e.g. correlation or regression analysis) tend to assume a linear relationship.

11

Appendix

Table 11.1. Symbols and abbreviations.

S	TSI	Total Solar Irradiance (“solar constant”)
L_m	SCL	Solar Cycle Length
D_{spot}		Sunspot decay rate
R_z		Wolf/Zurich sunspot number
$R_{\max}$		The maxima Wolf/Zurich sunspot numbers
$R_{\min}$		The minima Wolf/Zurich sunspot numbers
T_F		Sunspot cycle fall-time
T_R		Sunspot cycle rise-time
Niño 1+2		0–10°S, 90–80°W
Niño 3		5°N–5°S, 150–90°W
Niño 3.4		5°N–5°S, 170–120°W
Niño 4		5°N–5°S, 160°E–150°W
nm	10^{-9} m	Nanometre

Table 11.2. Physical constants.

Gravitational constant	$G = 6.67259 \times 10^{-11}\,\mathrm{m^3\,kg^{-1}\,s^{-2}}$
Unified atomic mass constant	$m_u = 1.6605402 \times 10^{-27}\,\mathrm{kg}$
Mass of proton	$m_p = 1.6726231 \times 10^{-27}\,\mathrm{kg} = 938.27\,\mathrm{MeV}$
Mass of electron	$m_e = 9.1093897 \times 10^{-31}\,\mathrm{kg} = 0.51100\,\mathrm{MeV}$
Mass of neutron	$m_n = 1.6749286 \times 10^{-27}\,\mathrm{kg} = 939.57\,\mathrm{MeV}$
Planck constant	$h = 6.6260755 \times 10^{-34}\,\mathrm{J\,s}$
	$\hbar = h/(2\pi) = 1.05457266 \times 10^{-34}\,\mathrm{J\,s}$
Molecular gas constant	$R = 8.314510\,\mathrm{J\,K^{-1}\,mol^{-1}}$
Boltzmann constant	$k = 1.380658 \times 10^{-23}\,\mathrm{J\,K^{-1}}$
Stefan–Boltzmann constant	$\sigma = 5.67051 \times 10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$
Avogadro constant	$N_A = 6.022 \times 10^{23}\,\mathrm{mol^{-1}}$
Heat capacity of water	$c = 4.12\,\mathrm{KJ/(kg\,K)}$
Mean radius of Earth	$r_E = 6.37 \times 10^{6}\,\mathrm{m}$
Mass of Earth	$m_E = 5.98 \times 10^{24}\,\mathrm{kg}$
Mass of the Sun	$m_S = 1.99 \times 10^{30}\,\mathrm{kg}$
Radius of the Sun	$R\odot = 7 \times 10^{8}\,\mathrm{m}$
Mean radius of Earth's orbit	$a = 1.49 \times 10^{11}\,\mathrm{m}$

Table 11.3. Solar system statistics.

	Mean distance	Radius	Mass		
Planet	(×1000 km)	(km)	(kg)	Eccentricity	Period
Mercury	57,910	2,439	3.30×10^{23}	0.206	88 days
Venus	108,200	6,052	4.87×10^{24}	0.007	225 days
Earth	149,600	6,378	$5,98 \times 10^{24}$	0.017	365.25 days
Mars	227,940	3,397	6.42×10^{23}	0.093	2 years
Jupiter	778,330	71,492	1.90×10^{27}	0.048	12 years
Saturn	1,426,940	60,268	5.69×10^{26}	0.055	29 years
Uranus	2,870,990	25,559	8.69×10^{25}	0.051	84 years
Neptune	4,497,070	24,764	1.02×10^{26}	0.007	165 years
Pluto	5,913,520	1,160	1.31×10^{22}	0.252	248 years
Moon	384	1,738	7.35×10^{22}		27.3 days

Table 11.4. Data acronyms and sources.

ACRIM	Active Cavity Radiometer Irradiance Monitor (*Spacelab 1, ATLAS 1,* SMM)
ANOVA	ANalysis Of VAriance
AO	Arctic Oscillation
ASCCP	International Satellite Cloud Climatology Project
ATSSR	Along Track Scanning . . . ?
AVHRR	Advanced Very-High REsolution Radiometer (*NOAA-7*)
CCN	Cloud Condensation Nuclei
COADS	Comprehensive Ocean Atmosphere Data Set
CRU	Climate Research Unit (University of East Anglia, UK)
DMSP	Defence Meteorological Satellite Programme
ECMWF	European Centre for Medium Weather Forecasting
EBM	Energy Balance Model
ENSO	El Niño Southern Oscillation
ERB	Earth Radiation Budget (*Nimbus-7*)
ERBS	Earth Radiation Budget Satellite
EUC	Equatorial Under Current
ESA	European Space Agency
GISS	Goddard Institute for Space Studies (NASA), New York, USA
IGY	International Geophysical Year
IMF	Interplanetary Magnetic Field
IPCC	Intergovernmental Panel on Climate Change
ISCCP	International Satellite Cloud Climatology Project
GCM	General Circulation Model
GCM	Global Climate Model
GCR	Galactic Cosmic Rays
GNU	"GNU it's Not Unix" – Free Software Foundation
LIA	Little Ice Age
LTE	Local Thermodynamic Equilibrium
MJO	Madden–Julian oscillation
MSU	Microwave Sounding Unit
NAO	North Atlantic Oscillation
NCAR	National Center for Atmospheric Research
NCEP	National Center for Environmental Prediction (USA)
NOAA	National Ocean Atmospheric Administration
OLR	Outgoing Long-wave Radiation
PDO	Pacific Decadal Oscillation
PV	Potential Vorticity
QBO	Quasi-Biennial Oscillation
RH	Relative humidity
RMS	Root Mean Square
RMSE	Root Mean Square Error
R_z	Wolf sunspot number
SCL	Solar Cycle Length
SDAC	Solar Data Analysis Center
SLA	Sea Level height Anomaly
SLP	Sea Level Pressure
SMM	Solar Maximum Mission
SMMR	Scanning Multichannel Microwave Radiometer (*Nimbus 7*, 1978–1987)
SOHO	SOlar and Heliospheric Observatory
SOI	Southern Oscillation Index
SSM/I	Special Sensor Microwave Imager (DMSP, 1987–)
SST	Sea Surface Temperature
TAR	Third Assessment Report (IPCC, 2001)
TSI	Total Solar Irradiance
UARS	Upper Atmosphere Research Satellite
UV	Ultraviolet radiation
WMO	World Meteorological Organisation
WCRP	World Climate Research Experiment

Table 11.5. Internet sites with relevant information and data. There is an abundance of sunspot-related literature on the Internet that can be found through Internet search engines.

Climate blog	http//:www.RealClimate.org
IPCC	http//:www.grida.no/climate/ipcc_tar/
GISS temperature	http://www.giss.nasa.gov/data/update/gistemp/
MSU	http://vortex.nsstc.uah.edu/data/
NASA	http://www.nasa.gov/
NASA Climatology Interdisciplinary Data Collection	http://daac.gsfc.nasa.gov/CAMPAIGN_DOCS/FTP_SITE/
NGCD–NOAA, USA	ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA
R language	http://cran.r-project.org/
SDAC	http://umbra.nascom.nasa.gov/sdac.html
Statistics	http://www.gfi.uib.no/ñilsg/kurs/klimaanalyse.html
SOHO	http://soho.nascom.nasa.gov/
Solar–terrestrial indices	http://nssdc.gsfc.nasa.gov/space/model/solar/solar_index.html
Svensmark	http://dsri.dk/˜hsv/
University of East Anglia (CRU): climate observations	http://cru.uea.ac.uk/cru/cru.htm
University of Tromsø (northern lights)	http://geo.phys.uit.no/

Bibliography

Abbot, C. G. and Fowle, F. E. (1908) Radiation and Terrestrial Temperature. *Annals of the Astrophysical Observatory of the Smithsonian Institute*.

Abbot, C. G. and Fowle, F. E. (1913) Volcanoes and climate. *Ann. of the Astrophys. Observ. Smithson. Inst.*

Abetti, G. (1957) *The Sun*. Faber & Faber, London.

Adler, R. (2001) Melting away. *New Scientist*, 15 December, 15.

Alfvén, H. (1942) *Ark. f. Mat. Astr. och Fys.*, **29B**(2).

Ambaum, M. H. P., Hoskins, B. J. and Stephenson, D. B. (2001) Arctic Oscillation or North Atlantic Oscillation? *Journal of Climate*, **14**, 3495–3507.

Anderson, D. L. T. and McCreary, J. P. (1985) Slowly propagating disturbance in a coupled ocean–atmosphere model. *Journal of the Atmospheric Sciences*, **42**, 615–629.

Angot, A. (1903) On the simultaneous variations of sun spots and of terrestrial atmospheric temperatures. *Monthly Weather Review*.

Archibald, E. D. (1879) Barometric pressure and sunspots. *Nature*.

Arctowski, H. (1910) Variations in the distribution of atmospheric pressure in North America. *Bulletin of the American Geographical Society*, **XLLL**.

Arctowski, H. (1915) Volcanic dust veils and climatic variations. *Ann. New York Acad. Sci.*

Arrhenius, S. (1896) On the influence of carbonic acid in the air upon the temperature of the ground. *Philosophical Magazine and Journal of Science*, 236–276.

Arrhenius, S. (1903) *Lehrbuch der kosmischen Physik*.

Arvesen *et al.* (1969) *Applied Optics*, **8**(11), 2215–2232.

Balachandran, N. K., Rind, D., Lonergan, P. and Shindell, D. T. (1999) Effects of solar cycle variability on the lower stratosphere and troposphere. *Journal of Geophysical Research*, **104**(D22), 27321–27339.

Baldwin, M. P. and Dunkerton, T. J. (2001). Stratospheric harbingers of anomalous weather regimes. *Science*, **294**, 581–584.

Baliunas, S. and Jastrow, R. (1990) Evidence for long-term brightness changes of solar-type stars. *Nature*, **348**, 520–522.

Bard, E., Raisbeck, G., Yiou, F. and Jouzel, J. (2000) Solar irradiance during the last 1200 years based on cosmogenic nuclides. *Tellus*, **52B**, 985–992.

Barnett, T. P. (1983) Interaction of the monsoon and the Pacific tradewind system at interannual time scale. Part I. The equatorial zone. *Monthly Weather Review*, **111**, 756–773.

Beer, J. (2000) Polar ice as an archive for solar cycles and terrestrial climate. *The Solar Cycle and Terrestrial Climate* (pp. 671–676). European Space Agency, Noordwijk, The Netherlands.

Beer, J., Tobias, A. and Weiss, B. (1998) An active Sun throughout the Maunder minimum. *Solar Physics*, **181**, 237–249.

Benestad, R. E. (1994) *Study of Large Drops in Maritime and Continental Cumuli*. M.Phil. thesis, New Mexico Institute of Mining and Technology, Socorro, NM 87801, USA.

Benestad, R. E. (1997) *Intraseasonal Kelvin Waves in the Tropical Pacific*. Ph.D. thesis, Oxford, UK.

Benestad, R. E. (1999) Solar activity and global sea surface temperatures. *Astronomy & Geophysics*, **40**(June), 14–17.

Benestad, R. E. (2000a) *Analysis of Gridded Sea Level Pressure and 2-meter Temperature for 1873–1998 Based on UEA and NCEP Re-analysis II* (KLIMA 03/00). DNMI, PO Box 43 Blindern, 0313 Oslo, Norway.

Benestad, R. E. (2000b) On solar–terrestrial correlation studies: Pitfalls and real signals. In: A. Wilson (ed), *The Solar Cycle and Terrestrial Climate* (pp. 477–480). ESTEC, Noordwijk, The Netherlands and European Space Agency, Noordwijk, The Netherlands.

Benestad, R. E. (2001) A comparison between two empirical downscaling strategies. *Int. J. Climatology*, **21** (November), 1645–1668. DOI 10.1002/joc.703.

Benestad, R. E. (2005a) A review of the solar cycle length estimates. *Geophys. Res. Lett.*, **32**, L15714, doi:10.1029/2005GL023621.

Benestad, R. E. (2005b) On latitude profiles of zonal means. *Geophys. Res. Lett.*, **32**, L19713, doi:10.1029/2005GL023652.

Bertrand, C. and van Ypersele, J.-P. (1999) Potential role of solar variability as an agent for climate change. *Climatic Change*, **43**, 387–411.

Bigelow, F. (1894) Inversions of temperature in the 26.68 day solar magnetic period. *American Journal of Science*, **XLVIII**(3).

Bigelow, F. (1908) The relations between the meteorological elements of the United States and the solar radiation. *American Journal of Science*.

Bjerknes, J. (1959) Temperaturforandring i Golfströmmen i tidsrummet for klimaforbedringer i Norden. *YMES*.

Blandford, H. F. (1875) *Journal of the Asiatic Society of Bengal*.

Blandford, H. F. (1879) Report on the Meteorology of India in 1878.

Blandford, H. F. (1880) On the barometric see-saw between Russia and India in the sunspot cycle, *Nature*, **XXI**.

Blanke, B., Neelin, J. D. and Gutzler, D. (1997) Estimating the effect of stochastic wind stress forcing on ENSO irregularity. *Journal of Climate*, **10,** 1054–1063.

Bond, G., Kromer, B., Beer, J., Muscheler, R., Evans, M. M., Showers, W., Hoffmann, S., Lotti-Bond, R., Hajdas, I. and Bonani, G. (2001) Persistent solar influence on North Atlantic climate during the Holocene. *Science*, **294**, 2130–2135.

Bowler, S. (1999) Geochemical markers. *New Scientist*, 18 September, Inside science.

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Bray, R. J. and Loughhead, R. E. (1964) *Sunspots* (Vol. 7). Chapman & Hall, London.

Brekke, A. and Egeland, A. (1994) *Nordlyset* [*Northern Lights*]. Grøndahl & Dryers Forlag (Norwegian).

Broun, J. A. (1878) Sunspots, atmospheric pressure and the sun's heat. *Nature*.

Buffet, B. A. (2000) Earth's core and the geodynamo. *Science*, **288**, 2007–2012.

Byrne, P. B. (1992) *Sunspots: Theory and Observation* (Series C: Mathematical and Physical Sciences, Vol. 375, pp. 63–74. Kluwer, Dordrecht, The Netherlands.

Carslaw, K.S., Harrison, R. G. and Kirkby, J. (2002) Cosmic rays, clouds, and climate. *Science*, **298**, 1732–1737.

Chambers, C. (1875) *Meteorology of the Bombay Presidency*.

Chambers, F. (1878) Sun-spots and weather. *Nature*.

Chambers, F. (1880) Abnormal variations of barometric pressure in the Tropics, and their relation to sun-spots, rainfall and famines. *Nature*.

Chapman, S. (1930) *Mem. R. Met. Soc.*, **3**, 103.

Chavez, F. P., Strutton, P. G., Friederich, G. E., Feely, R. A., Feldman, G. C., Foley, D. G. and McPhaden, M. J. (1999) Biological and chemical response of the equatorial Pacific Ocean to the 1997–98 El Niño. *Science*, **286**, 2126–2131.

Clough, H. W. (1943) The long period variations in the length of the 11-year solar period and on current variations in terrestrial phenomena. *Bull. Amer. Meteor. Soc.*, **24**, 154–163.

Couzin, J. (1999) Landscape changes make regional climate run hot and cold. *Science*, **283,** 317–319.

Cowling, T. G. (1934) *Mon. Not. R. Astron. Soc.*, **94,** 39–48.

Crowley, T. J. (2002) Causes of climate change over the past 1000 years. *Science*, **289**, 270–277.

Crooks, S. A and Gray, L. J. (2005) Characterization of the 11-year solar signal using a multiple regression analysis of the ERA-40 dataset. *J. Climate*, **18**, 996–1014.

Cubasch, U., Voss, R., Hegerl, G. C., Waskevitz, J. and Crowley, T. J. (1997) Simulations of the influence of solar radiation variations on the global climate with an ocean–atmosphere general circulation model. *Climate Dynamics*, **13**, 757–767.

Damon, P. E. and Peristykh, A. N. (2005) Solar forcing of global temperature change since AD 1400. *Climatic Change*, **68**, 101–111.

de Toma, G. and White, O. R. (2000) From solar minimum to solar maximum: changes in total and spectral solar irradiance. *The Solar Cycle and Terrestrial Climate* (pp. 45–50). European Space Agency, Noordwijk, The Netherlands.

Dickinson, R. E. (1975) Solar variability and the lower atmosphere. *Bull. Amer. Meteor. Soc.*, **56,** 1240–1248.

Diego, F. (1999) Total solar eclipses: Magic, science and wonder. *Physics World*, July, 31–36.

Dima M., Lohmann, G. and Dima I. (2005) Solar-induced and internal climate variability at decadal time scales. *Int. J. Clim.*, **25**, 713–733.

Drew S. T., Schmidt G. A., Miller R. L. and Mann M. E. (2003) Volcanic and solar forcing of climate change during the preindustrial era. *J. Clim.*, **16**, 4094–4107.

Eckert, C. and Latif, M. (1997) Predictability of a stochastically forced coupled model of El Niño. *Journal of Climate*, **10**, 1488–1504.

Eddy, J. A. (1976) The Maunder minimum. *Science*, **192**, 1189–1202.

Egbert, G. D. and Ray, R. D. (2000) Significant dissipation of tidal energy in the deep ocean inferred from satellite altimeter data. *Nature*, **405**, 775–778.

Ellner, Stephen, P. (2001) Review of R, Version 1.1.1. *Bulletin of the Ecological Society of America*, **82**(2), 127–128.

Encrenaz, T. (1997) Planets: Something in the air. *Physics World*, **10**(10), 29–34.

Farrar, P. D. (2000) Are cosmic rays influencing oceanic cloud coverage – or is it only El Niño? *Climatic Change*, **47**, 7–15.

Feynman, R. P. (1985) *Surely You're Joking Mr. Feynman*. Norton, New York.

Fleagle, R. G. and Businger, J. A. (1980) *An Introduction to Atmospheric Physics* (2ndd edn, International Geophysics Series Vol. 25). Academic Press, Orlando, FL.

Foukal P., North, G. and Wigley, T. (2004) A stellar view on solar variations and climate. *Science*, **306**, 68–69.

French, A. P. and Taylor, E. F. (1989) *An Introduction to Quantum Physics* (MIT Introductory Physics Series). Chapman & Hall.

Friis-Christensen, E. and Lassen, K. (1991) Length of the solar cycle: An indicator of solar activity closely associated with climate. *Science*, **254**, 698–700.

Fritsch, K. (1854) *Über das Steigenund Fallen def Lufttemperatur binnen einer analogen elfjährigen Periode, in welcher die Sonnenflecke sich vermindern oder vermehren*. Wien.

Fröhlich, C. and Lean, J. (1998a) The sun's total irradiance: Cycles, trends and related climate change uncertainties since 1976. *Geophys. Res. Lett.*, **25**, 4377–4380.

Fröhlich, C. and Lean, J. (1998b) Total solar irradiance variations. *New Eyes to See Inside the Sun and Stars* (pp. 89–102). IAU.

Gaffen, D. J., Santer, B. D., Boyle, J. S., Christy, J. R., Graham, N. E. and Ross, R. J. (2000) Multidecadal changes in the vertical temperature structure of the tropical troposphere. *Science*, **287**, 1242–1245.

Gentleman, R. and Ihaka, R. (2000) Lexical scope and statistical computing. *Journal of Computational and Graphical Statistics*, **9**, 491–508.

Gibson, J. K., Kallberg, P., Uppala, S., Hernandez, A., Nomura, A. and Serrano, E. (1997) *ERA Description* (ERA Project Report Series). ECMWF.

Gill, A. E. (1982) *Atmosphere–Ocean Dynamics*. Academic Press, San Diego, CA.

Gillet, N. P., Allen, M. R. and Tett, S. F. B. (2000) Modeller and observed variability in atmospheric vertical temperature structure. *Climate Dynamics*, **16**, 49–61.

Gleick, J. (1987) *Chaos*. Cardinal.

Godske, C. L. (1956) *Hvordan blir været?* (1st edn). J. W. Cappelen, Oslo.

Götz, G., Mészáros, E. and Vali, G. (1991) *Atmospheric Particles and Nuclei*. Akadémiai Kiadó, Budapest, Hungary.

Granger, C. W. J. (1957) A statistical model for sunspot activity. *Astrophys. J.*, **126**, 152–158.

Haigh, J. D. (1994) The role of stratospheric ozone in modulating the solar radiative forcing of climate. *Nature*, **370**(August), 544–546.

Haigh, J. D. (1996) The impact of solar variability on climate. *Science*, **272**, 981–984.

Haigh, J. D. (1999) A GCM study of climate change in response to the 11-year solar cycle. *Quarterly Journal of the Royal Meteorological Society*, **125**, 871–892.

Haigh, J. D. (2001) Climate variability and the influence of the Sun. *Science*, **294**, 2109–2110.

Haigh, J. D. (2003) The effects of solar variability on the Earth's climate. *Phil. Trans. R. Soc. Lond. A*, **361**, 95–111.

Haigh, J. D. (2004) *Fundamentals of the Earth's Atmopshere and Climate* (Geophysical Monograph, 141). American Geophysical Union, doi:10.1029/141GM06.

Hale, G. E. and Ellerman, F. (1906) The five-foot spectroheliograph of solar observatory. *Astrophys. J.*, **23**, 54.

Hale, G. E. and Nicholson, S. B. (1938) *Magnetic Observations of Sunspots 1917–1924* (Part I). Carnegie Institute.

Hann, J. (1908) *Handbuch der Klimatologie* (Vol. I).

Hansen, J. E. and Lebedeff, S. (1988) Global surface air temperatures: Update through 1987. *Geophys. Res. Lett.*, **15**, 323.

Hansen, J. E., Sato, M., Lacis, A., Ruedy, R., Tegen, I. and Matthews, E. (1998a) Climate forcings in the Industrial era. *Proc. Natl Acad. Sci.*, **95**, 12753–12758.

Hansen, J. E., Sato, M., Ruedy, R., Lacis, A. and Glascoe, J. (1998b) Global climate data and models: A reconciliation. *Science*, **281**(14 August), 930–932.

Hansen, J. E., Ruedy, R., Sato, M., Imboff, M., Lawrence, W., Easteling, D., Peterson, T. and Karl, T. (2001) A closer look at United States global temperature change. *Journal of Geophysical Research*, **106**(D20), 23947–23963.

Harrison, R. G. and Shine, K. P. (1999) (February). *A Review of Recent Studies of the Influence of Solar Changes on the Earth's Climate* (Tech. rept. HCTN6). Hadley Centre for Climate Prediction & Research, UK Meteorological Office, Bracknell, RG12 2SY, UK.

Hartmann, D. L. (1994) *Global Physical Climatology*. Academic Press, San Diego, CA.

Helland-Hansen, B. and Nansen, F. (1920) *Temperature Variations in the North Atlantic Ocean and in the Atmosphere* (Smithsonian miscellaneous collections Vol. 70, No. 4). Smithsonian Institute.

Herschel, W. (1801) Observations tending to investigate the nature of the sun, in order to find the causes of symptoms of its variable emission of light and heat. *Philosophical Transactions*. London.

Hill, S. A. (1879) Über eine zehnjährige Perioden in der jährige Änderung der Temperatur und des Luftdruckes in Nord-Indian. *Zeitschr. der Österreich Gesellsch. für Meteorologie*.

Hirst, A. C. (1986) Unstable and damped equatorial modes in simple coupled ocean–atmosphere models. *Journal of the Atmospheric Sciences*, **43**, 606–630.

Hirst, A. C. (1988) Slow instabilities in tropical ocean basin–global atmosphere models. *Journal of the Atmospheric Sciences*, **45**, 830–852.

Hoffert, M. I., Callegari, A. J. and Hsieh, C.-T. (1980) *Journal of Geophysical Research*, **85**, 6667–6679.

Houghton, J. T. (1991) *The Physics of Atmospheres* (2nd edn) Cambridge University Press, Cambridge, UK.

Houghton, J. T., Ding, Y., Griggs, D. J., Noguer, M., van der Linden, P. J., Dai, X., Maskell, K. and Johnson, C. A. (2001) *Climate Change 2001: The Scientific Basis* (Contribution of Working Group I to the Third Assessment Report of IPCC). Intergovernmental Panel on Climate Change (available at www.ipcc.ch).

Howe, R., Christensen-Dalsgaard, J., Hill, F., Komm, R. W., Larsen, R. M., Schou, J., Thompson, M. J. and Toomre, J. (2000) Dynamic variations at the base of the solar convection zone. *Science*, **287**, 2456–2460.

Hoyt, D. V. and Schatten, K. H. (1993) A discussion of plausible solar irradiance variations 1700–1992. *Journal of Geophysical Research*, **18**, 18895–18906.

Humphreys, W. J. (1913) Volcanic dust and other factors in the production of climatic changes, and their possible relation to ice ages. *Bulletin of the Mount Weather Observatory*, **6**.

Hurrel, J. (1995) Decadal trends in the North Atlantic Ocsillation and relationships to regional temperature and precipitation. *Science*, **269**, 676–679.

IPCC (1995) *The Second Assessment Report* (Technical Summary). WMO & UNEP.

IPCC (2001) *IPCC WGI Third Assessment Report* (Summary for Policymakers). WMO.

Jenkins, G. S. (2000) Global climate model high-obliquity solutions to the ancient climate puzzles of the Faint-Young Sun Paradox and low Proterozoic Glaciation. *Journal of Geophysical Research*, **105**(D6), 7357–7370.

Jones, P. D., Raper, S. C. B., Bradley, R. S., Diaz, H. F., Kelly, P. M. and Wigley, T. M. L. (1998a) Northern Hemisphere surface air temperature variations, 1851–1984. *J. Clim. Appl. Met.*, **25**, 161–179.

Jones, P. D. and Briffa, K. R. (2000) Temperature trends from instrumental and proxy indicators for the last millenium. *The Solar Cycle and Terrestrial Climate* (pp. 179–180). European Space Agency, Noordwijk, The Netherlands.

Jones, P. D., Briffa, K. R., Barnett, T. P. and Tett, S. F. B. (1998b) High-resolution paleoclimatic records for the last millennium: Interpretation, intergration and comparison with general circulation model control-run temperatures. *The Holocene*, **8**, 455–471.

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., Iredell, M., Saha, S., White, G., Wollen, J., Zhu, Y., Chelliah, M., Ebisuzaki, W., Higgins, W., Janowiak, J., Mo, K. C., Ropelewski, C., Wang, J., Leetmaa, A., Reynolds, R., Jenne, R. and Joseph, D. (1996) The NCEP/NCAR 40-year reanalysis project. *Bull. Amer. Meteor. Soc.*, **77**(3), 437–471.

Kaplan, A., Cane, M. A., Kushnir, Y., Clement, A. C., Blumenthal, M. B. and Rajagopalan, B. (1998) Analyses of global sea surface temperature 1856–1991. *Journal of Geophysical Research*, **103**(C9), 18567–18589.

Katz, R.W. (1988) Use of cross correlations in the search for teleconnections. *Journal of Climatology*, **8**, 241–253.

Kelly, P. M. and Wigley, T. M. L. (1992) Solar cycle length, greenhouse forcing and global climate. *Nature*, **360**(November), 328–330.

Kernthaler, S., Tuomi, R. and Haigh, J. (1999) Some doubts concerning a link between cosmic ray fluxes and global cloudiness. *GRL*, **26**, 863–865.

Kessler, W. S. and McPhaden, M. J. (1995) Oceanic equatorial waves and the 1991–93 El Niño. *Journal of Climate*, **8**, 1757–1774.

Kessler, W. S., McPhaden, M. J. and Weickmann, K. M. (1995) Forcing of intraseasonal Kelvin waves in the equatorial Pacific. *Journal of Geophysical Research*, **100**, 10613–10631.

Kindle, J. C. and Phoebus, P. A. (1995) The ocean response to operational wind bursts during the 1991–1992 El Niño. *Journal of Geophysical Research*, **100**, 4893–4920.

King, J. W. (1975) *Aeronautics and Astronautics*, **13**(4), 10.

Kirtman, B. P. (1997) Oceanic Rossby wave dynamics and the ENSO period in a coupled model. *Journal of Climate*, **10**, 1690–1704.

Kleppner, D. and Kolenkow, R. J. (1978) *An Introduction to Mechanics. Engineering Mechanics*. McGraw-Hill.

Köppen, W. (1873) Über mehrjährige Perioden der Witterung. *Deutsche Rundschau für Geografie und Statistik*.

Kraus, E. B. and Turner, J. S. (1967) A one-dimensional model of the seasonal thermocline. *Tellus*, **XIX**, 98–105.

Kristjánsson, J. E. and Kristiansen, J. (2000) Is there a cosmic ray signal in recent variations in global cloudiness and cloud radiative forcing? *Journal of Geophysical Research*, **105**(D9), 11851–11863.

Kristjánsson, J. E., Kristiansen, J. and Kaas, E. (2004) Solar activity, cosmic rays, clouds and climate – An update. *Adv. Space Research*, **34**, 407–415.

Kristjánsson, J. E., Stable, A. and Kristiansen, J. (2002) A new look at possible connections between solar activity, clouds and climate. *Geophys. Res. Lett.*, **29**(23), 2107, doi:10.1029/2002GL015646.

Kuiper, G. P. (ed.) (1953) *The Sun* (4th edn). University of Chicago Press, Chicago.

Kumar, K. K., Rajagopalan, B. and Cane, M. A. (1999) On the weakening relationship between the Indian monsoon and ENSO. *Science*, **284**, 2156–2159.

Kuroda, Y. and Kodera, K. (2002) Effect of the solar cycle on the polar-night jet oscillation. *J. Met. Soc. Japan*, **80**, 973–984.

Labitzke, K. (1987) Sunspots, the QBO, and the stratospheric temperature in the north polar region. *Geophys. Res. Lett.*, **14**, 535–537.

Labitzke, K. and van Loon, H. (1988) Association between the 11-year solar cycle, the QBO, and the atmosphere, I. The troposphere and stratosphere on the northern hemisphere winter. *J. Atmos. Terr. Phys.*, **50**, 197–206.

Labitzke, K., Austin, J., Butchart, N., Knight, J., Takahashi, M., Nakamoto, M., Nagashima, T., Haigh, J. D. and Williams, V. (2002) The global signal of the 11-year solar cycle in the stratosphere: Observations and models. *J. Atm. Solar-Terrestrail Phys.*, **64**, 203–210.

Landscheit, T. (2000) Solar forcing of El Niño and La Niña. *The Solar Cycle and Terrestrial Climate* (pp. 135–140). European Space Agency, Noordwijk, The Netherlands.

Langley, S. P. (1904) On a possible variation of the solar radiation and its probable effect on terrestrial temperatures. *Astrophysical J.*

Lassen, K. and Friis-Christensen, E. (1995) Variability of the solar cycle length during the past five centuries and the apparent association with terrestrial climate. *J. Atmos. Terr. Phys.*, **57**, 835–845.

Lassen, K. and Friis-Christensen, E. (2000) Reply to the article: Solar cycle lengths and climate: A reference revisited, by P. Laut and J. Gundermann. *J. Geophys. Res. – Space*, **105**, 27493–27495.

Lau, K.-M. (1985) Subseasonal scale oscillation, bimodal climatic state and the El Niño/Southern Oscillation. *Coupled Ocean–Atmosphere Models* (Elsevier Oceanography Series Vol. 40, pp. 29–40). Elsevier.

Laut, P. (2003) Solar activity and terrestrial climate: An analysis of some purported correlations. *J. Atmos. Sol-Terr. Phys.*, **65**, 801–812.

Laut, P. and Gundermann, J. 2000. Solar cycle lengths and climate: A reference revisited. *Journal of Geophysical Research*, **105**(A12), 27489–27294.

Lean, J. (2000) Evolution of the Sun's spectral irradiance since the Maunder minimum. *Geophys. Res. Lett.*, **27**, 2425–2428.

Lean, J. (2005) Living with a variable Sun. *Physics Today*, June, 32–38.

Lean, J. and Rind, D. (1998) Climate forcing by changing solar radiation. *Journal of Climate*, **11**, 3069–3094.

Lean, J., Beer, J. and Breadley, R. (1995) Reconstruction of the solar irradiance since 1610: Implications for climate change. *Geophys. Res. Lett.*, **22**, 3195–3198.

Lindzen, R. S. (1990) *Dynamics in Atmospheric Physics*. Cambridge University Press, Cambridge, UK.

Liznar, J. (1880) Beziehung der täglichen und jährlichen Temperaturschwankung zur 11-jährigen Sonnenflecken periode. *Sitzungsberichte der Wien Akad.* **LXXXII (***Nature* **XXIII**, p. 133).

Lockwood, M. (2002) *Long-term Variations in the Open Solar Flux and Possible Links to Earth's Climate* (ESA-SP-508). ESA publications, Noordvijk, the Netherlands, pp. 507–522.

Lockwood, M., Stamper, R. and Wild, M. N. (1999) A doubling of the Sun's coronal magnetic field during the past 100 years. *Nature*, **399**(June), 437–439.

Lopez, R. (2001) Driven to extremes. *New Scientist*, 6 October, 38–42.

Lorenz, E. (1967) *The Nature and Theory of the General Circulation of the Atmosphere* (Publication 218). WMO, Geneva.

Ludmány, A. and Baranyi, T. (2000) Comparative study of the atmospheric effects driven by irradiance vs. corpuscular radiation. *The Solar Cycle and Terrestrial Climate*. European Space Agency, Noordwijk, The Netherlands.

MacDowall, A. B. (1896) Sonnenflecken und Sommertemperaturen. *Meteorologische Zeitschrift*.

Magee, B. (1973) *Popper* (13th edn, Fontana Modern Masters). Fontana Press, London.

Maltby, P. (1992) *Sunspots: Theory and Observation* (Series C: Mathematical and Physical Sciences Vol. 375, pp. 103–120). Kluwer, Dordrecht, The Netherlands.

Mann, M. E., Cane, M. A., Zebiak, S. E. and Clement, A. (2005) Volcanic and solar forcing of the tropical Pacific over the past 1000 years. *J. Climate*, **18**, 447–456.

Mantua, N. J., Hare, S. R., Zhang, Y., Wallace, J. M. and Francis, R. C. (1997) A Pacific interdecadal climate oscillation with impacts on salmon production. *Bull. Amer. Meteor. Soc.*, **78**, 1069–1079.

Marsh, N. D. and Svensmark, H. (2000) Low cloud properties influenced by cosmic rays. *Physical Review Letters*, **85**(23), 5004–5007.

Marsh, N. D. and Svensmark, H. (2002) GCR and ENSO trends in ISCCP-D2 low cloud properties. *J. Geophysical Research*, **108**(D6), 4195, doi:10.1029/2001JD001264.

Matthes, K., Kodera, K., Haigh, J. D., Shindell, D.T., Shibata, K., Langematz, U., Rozanov, E. and Kurooda, Y. (2003) GRIPS Solar Experiments Intercomparison Project: Initial results. *Papers in Meteorology and Geophysics*, **54**, 71–90.

Mattig, W. (1958) Zur Linienabsorption im inhomogenen Magnetfeld der Sonnenflecken. *Z. Astrophys.*, **44**, 280.

Mayaud, P.-N. (1972) The *aa* Indices: A 100-year series characterizing the magnetic activity. *Journal of Geophysical Research*, **77**(34), 6870–6874.

McCreary, J. P. (1983) A model of tropical ocean–atmosphere interaction. *Monthly Weather Review*, **111**, 370–389.

McCreary, J. P. and Anderson, D. L. T. (1984) A simple model of El Niño and the Southern Oscillation. *Monthly Weather Review*, **112**, 934–946.

McCreary, J. P. and Anderson, D. L. T. (1991) An overview of coupled ocean–atmosphere models of El Niño and the Southern Oscillation. *Journal of Geophysical Research*, **96**, 3125–3150.

McPhaden, M. J., Freitag, H. P., Hayes, S. P., Taft, B. A., Chen, Z. and Wyrtki, K. (1986) The response of the equatorial Pacific Ocean to a westerly wind bust in May 1986. *Journal of Geophysical Research*, **93**, 10589–10603.

Mears, C. A., Schabel, M. C. and Wentz, F. J. (2003) A reanalysis of the MSU channel 2 tropospheric temperature record. *Journal of Climate*, **16**, 3650–3664.

Meehl, G. A., Washington, W. M., Wigley, T. M. L., Arblaster, J. M. and Dai, A. (2003) Solar and greenhouse gas forcing and climate response in the twentieth century. *J. Clim.*, **16**, 426–444.

Meehl, G. A, Washington, W. M., Ammann, C. M., Arblaster, J. M., Wigley, T. M. L. and Tebaldi, C. (2004) Combinations of natural and anthropogenic forcings in twentieth-century climate. *J. Climate*, **17**, 3721–3727.

Meissner, O. (1917) Über die Beziehung der Temperatur zur Sonnenfleckenperiode. *Ann. der. Hydr. u. Marit. Meteor.*, **XLV**.

Mendoza, B. (1997) Estimations of Maunder minimum solar radiation and Ca ii H and K fluxes using rotation rates and diameters. *Astrophys J.*, **483**, 523–526.

Michard, R. (1953) Contribution à l'étude physique de la photosphère et des taches solaires. *Ann. Astrophys.*, **16**, 217.

Mielke, J. (1913) Die Temperaturschwankungen 1870–1910 in irem Verhältnis zu der 11 jährigen Sonnenfleckenperiode. *Achiv der Deutschen Seewarte*, **XXXVI**(3).

Milankovitch, M. (1941) *History of Radiation on the Earth and its Use for the Problem of Ice Ages* (special publication). Serbian Academy of Geography [in German].

Monin, A. S. (1972) *Weather Forecasting as a Problem in Physics* (English translation from Russian). MIT Press.

Moore, A. M. and Kleeman, R. (1997) Stochastic forcing of tropical interannual variability: A Paradigm for ENSO (Private communication).

Morton, O. (1999) Mystery of the missing atmosphere. *New Scientist*, 20 November, 35–38.

Mursula, K. and Ulich, T. (1998). A new method to determine the solar cycle length. *Geophys. Res. Lett.*, **25**, 1837–1840.

Myhre, G., Stordal, F., Rognernd, B. and Isaksen, I. (1998) Radiative forcing due to stratospheric ozone. *Proceedings of the 18th Quadrennial Ozone Symposium* (R. D. Bojkov and G. Visconti, eds), pp. 813–816. Parco Scientifico e Technologico d'Abruzzo, L'Aguila, Italy.

Neelin, J. D. (1991) The slow sea surface temperature mode and the fast wave limit: Analytical theory for tropical interannual oscillations and experiments with a hybrid coupled model. *Journal of the Atmospheric Sciences*, **48**, 584–606.

Nesme-Ribes, E. and Manganey, A. (1992) On a plausible physical mechanism linking the Maunder minimum to the little ice age. *Radiocarbon*, **34**, 263–270.

Newcomb, S. (1908) A search for fluctuations in the Sun's thermal radiation through their influence on terrestrial tamperature. *Transactions of the American Philosophical Society*.

Nishiizumi, K. and Caffee, M. W. (2001) Beryllium-10 from the Sun. *Science*, **294**, 352–354.

Nordmann, C. (1903). La période des taches solaries et les variations des températures moyennes annuelles de la terre. *Comptes Rendus* (reprinted in *Monthly Weather Review*, **XXXI**, 1903).

Ogi, M, Yamazaki, K. and Tachibana, Y. (2003) Solar cycle modulation of the seasonal linkage of the North Atlantic Oscillation (NAO). *Geophys. Res. Lett.*, **30**, CLM 8–1.

Orlove, B. S., Chiang, J. C. H. and Cane, M. A. (2000) Forecasting Andean rainfall and crop yield from the influence of El Niño on Pleiades visibility. *Nature*, **403**(January), 68–71.

Orther, J. and Maseland, H. (eds) (1965) Introduction to Solar Terrestrial Relations. *Proceedings of the Summer School in Space Physics Held in Alpbach, Austria, July 15–August 10, 1963* (organised by the European Space Research Organisation, ESRO). D. Reidel, Dordrecht, The Netherlands.

Pacanowski, R. C. and Philander, S. G. H. (1981) Parameterization of vertical mixing in numerical models of tropical oceans. *Journal of Physical Oceanography*, **11**, 1443–1451.

Pallé, E. and Butler, C. J. (2001) Sunshine records from Ireland: Cloud factors and possible links to solar activity and cosmic rays. *International Journal of Climatology*, **21**, 709–729.

Parker, E. 1997. Mysteries of the Sun. *Physics World*, October, 35–40.

Peixoto, J. P. and Oort, A. H. (1992) *Physics of Climate*. American Institute of Physics, New York.

Penland, C. and Sardeshmukh, P. D. (1995) The optimal growth of tropical sea surface temperature anomalies. *Journal of Climate*, **8**, 1999–2023.

Petrovay, K. (2000) What makes the Sun tick? The origin of the solar cycle. *The Solar Cycle and Terrestrial Climate* (pp. 3–14). European Space Agency, Noordwijk, The Netherlands.

Philander, S. G. (1989) *El Niño, La Niña, and the Southern Oscillation*. Academic Press, New York.

Philander, S. G. H. and Pacanowski, R. C. (1981) Response of equatorial oceans to periodic forcing. *Journal of Geophysical Research*, **86**, 1903–1916.

Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1989) *Numerical Recipes in Pascal*. Cambridge University Press, Cambridge, UK.

Pruppacher, H. R. and Klett, J. D. (1978) *Microphysics of Clouds and Precipitation*. D. Reidel, Dordrecht, The Netherlands.

Rahmstorf, S., Archer, D., Ebel, D. S., Eugster, O., Jouzel, J., Maraun, D., Neu, U., Schmidt, G. A., Severinghaus, J., Weaver, A. J. and Zachos, J. (2004) Reply. *Eos*, **85**, 511.

Ramanathan, V. and Collins, W. (1991) Thermodynamic regulation of ocean warming by cirrus clouds deduced from observations of the 1987 El Niño. *Nature*, **351**, 27–32.

Reid, G. C. (1987) Influence of solar variability on global sea surface temperatures. *Nature*, **329**, 142–143.

Reid, G. C. (1997) Solar forcing of global climate change since the mid 17th century. *Climatic Change*, **37**, 391–405.

Reynolds, R. W. and Smith, T. M. (1994) Improved global sea surface temperature analysis using optimum interpolation. *Journal of Climate*, **7**, 929–948.

Reynolds, R. W. and Smith, T. M. (1995) A high-resolution global sea surface temperature climatology. *Journal of Climate*, **8**, 1571–1583.

Richardson, I. G., Cliver, E. W. and Cane, H. V. (2002) Long-term trends in interplanetary magnetic field strength and solar wind structure during the twentieth century. *J. Geophys. Res.*, **107**, SSH12.

Richter, C. M. (1902) Sonnenflecken, Erdmagnettismus und Luftdruck. *Meteorologische Zeitschrift*.

Riehl, H. (1954) *Tropical Meteorology*. McGraw-Hill, New York.

Rind, D., Lean, J. and Healy, R. (1999) Simulated time-dependent climate response to solar radiative forcing since 1600. *Journal of Geophysical Research*, **104**, 1973–1990.

Rizzo, G. B. (1897) *Sulla Relazione per le Macchie Solari a la Temperatura dell'Aria*. Torino.

Rogers, R. R. and Yau, M. K. (1989) *A Short Course in Cloud Physics* (3rd edn). Pergamon Press, Oxford.

Rüdiger, G. (2000) The dynamo theory for the Maunder minimum. *The Solar Cycle and Terrestrial Climate*. European Space Agency, Noordwijk, The Netherlands.

Salby, M. and Callagan, P. (2000) Connection between the solar cycle and the QBO: The missing link. *Journal of Climate*, **13**, 328–338.

Salby, M. and Callaghan, P. (2004) Evidence of the solar cycle in the general circulation of the stratosphere. *J. Clim.*, **17**, 34–46.

Santer, B. D., Wigley, T. M. L., Gaffen, D. J., Bengtsson, L., Doutriaux, C., Boyle, J. S., Esch, M., Hnilo, J. J., Jones, P. D., Meehl, G. A., Roeckner, E., Taylor, K. E. and Wehner, M. F. (2000) Interpreting differential temperature trends at the surface and in the lower troposphere. *Science*, **287**, 1227–1232.

Scafetta, N. and West, B. J. (2005) Estimated solar contribution to the global surface warming using the ACRIM TSI satellite composite. *Geophys. Res. Lett.*, **32**, L18713.

Schlegel K., Diendorfer G., Thern S. and Schmidt M. (2001) Thunderstorms, lightning and solar activity in Middle Europe. *J. Atmos. Sol-Terr. Phys.*, **63**, 1705–1713.

Schlesinger, M. and Ramankutty, N. (1992) Implications for global warming of intercycle solar irradiance variations. *Nature*, **360**(November), 330–333.

Schmitz, B. (2000) Plankton cooled a greenhouse. *Nature*, **407**, 143–144.

Schopf, P. S. and Suarez, M. J. (1988) Vacillations in coupled ocean–atmosphere model. *Journal of the Atmospheric Sciences*, **45**, 549.

Schwarzschild, B. (2001) Isotopic analysis of pristine microshells resolves a troubling paradox of paleoclimatology. *Physics Today*, December, 16–18.

Selvam, A. M., Fadnavis, S., Athale, S. U. and Tinmaker, M. I. R. (1997) Enhancement in surface atmospheric pressure variability associated with a major geomagnetic storm, *Proc. IAAA, 8th Scientific Assembly with ICMA and STP Symposium, Uppsala, August 9–15*.

Shaviv, N. J. (2002) Cosmic ray diffusion from the Galactic spiral arms, iron meteorites, and a possible climatic connection? *Phys. Rev. Lett.*, **89**(5), 051102.

Shaviv, N. J. (2004) Comment. *Eos*, **85**, 510.

Shindell, D., Rind, D., Balachandran, N., Lean, J. and Lonergan, P. (1999) Solar cycle variability, ozone and climate. *Science*, **284**, 305–308.

Shindell, D. T., Schmidt, G. A., Miller, R. L. and Rind, D. (2001a) Northern hemisphere winter climate response to greenhouse gas, ozone, solar and volcanic forcing. *Journal of Geophysical Research*, **106**(D7), 7193–7210.

Shindell, D. T., Schmidt, G. A., Mann, M. E., Rind, D. and Waple, A. (2001b) Solar forcing of regional climate change during the Maunder minimum. *Science*, in press.

Shindell, D. T., Schmidt, G. A., Miller, R. L. and Mann, M. E. (2003) Volcanic and solar forcing of climate change during the Preindustrial Era. *Journal of Climate*, **16**, 4094–4107.

Shindell, D. T., Schmidt, G. A., Miller, R. L. and Mann, M. E. (2004) Volcanic and solar forcing of climate change during the predindustrial era. *J. Clim.*, **16**, 4094–4107.

Sikka, D. R. *et al.* (1987) *Adv. Atmos. Sci.*, **5**(2), 217.

Simpson, S. (1999) (October) Deserting the Sahara. *Scientific American*.

Slutz, R. J., Lubker, S. J., Hiscox, J. D., Woodruff, S. D., Jenne, R. L., Steurer, P. M. and Elms, J. D. (1985) *Comprehensive Ocean–Atmosphere Data Set; Release 1* (Tech. rept). Climate Research Program, Boulder, CO.

Solanki, S. K. and Fligge, M. (1998) Solar irradiance since 1874 revisited. *Geophys. Res. Lett.*, **25**(3), 341–344.

Solanki, S. K. and Fligge, M. (2000) Long-term changes in the solar irradiance. *The Solar Cycle and Terrestrial Climate* (pp. 57–60). European Space Agency, Noordwijk, The Netherlands.

Solanki S. K., Usoskkin, I. G., Kromer, B., Schussler, M. and Beer, J. (2004) Unusual activity of the Sun during recent decades compared to the previous 11,000 years. *Nature*, **431**, 1084–1087.

Soon, W. H., Baliunas, S. L. and Zhang, Q. (1994) A technique for estimating long-term variations of solar total irradiance: Preliminary estimates based on observations of the Sun and solar-type stats. In: Nesme-Ribes (ed.), *The Solar Engine and its Influence on Terrestrial Atmosphere and Climate* (NATO ASI Vol. 125). Springer-Verlag, Berlin.

Soukharev, B. E. and Hood, L. L. (2001) Possible solar modulation of the equatorial quasi-biennial oscillation: Additional statistical evidence. *Journal of Geophysical Research*, **106**(D14), 14855–14868.

Spencer, R. W. and Christy, J. R. (1990) Precise monitoring of global temperature trends from satellites. *Science*, **247**, 1558.

Stewart, J. Q. and Panofsky, H. A. A. (1938) The mathematical characteristics of sunspot variations. *Astrophys J.*, **88**, 385–407.

Stott, P. A., Allen, M. R. and Jones, G. S. (2002) Estimating signal amplitudes in optimal finger printing. Part II: Application to general circulation models. *Clim. Dyn.*, **21**, 493–500.

Stott, P. A., Jones, G. S. and Mitchell, J. F. B. (2003) Do models underestimate the solar contribution to recent climate change? *J. Clim.*, **16**, 4079–4093.

Strang, G. (1995) *Linear Algebra and its Application*. Harcourt Brace & Company, San Diego, CA.

Sun B. and Bradley R. S. (2002) Solar influences on cosmic rays and cloud formation: A reassessment. *J. Geophys. Res.*, **107**(D14), doi:10.1029/2001JD000560.

Svensmark, H. (1998) Influence of cosmic rays on Earth's climate. *Physical Review Letters*, **81**(22), 5027–5030.

Svensmark, H. (2000) Cosmic rays and Earth's climate. *Space Science Review*, **93**, 155–166.

Svensmark, H. and Friis-Christensen, E. (1997) Variations of cosmic rays flux and global cloud coverage. A missing link in solar–climate relationships. *J. Atmos. Sol. Terr. Phys.*, **59**, 1225.

Sýkora, J., Badalyan, O. G. and Obridko, V. N. (2000) "Coronal holes" (recorded from 1943) – a source of solar-induced terrestrial response? *The Solar Cycle and Terrestrial Climate* (pp. 95–100). European Space Agency, Noordwijk, The Netherlands.

Szocs, H. L. and Kosa-Kiss, A. (2001) Observational evidence for contributions to cyclogenesis due to short-term variations within the lower troposphere caused by sunspots. *Journal of Meteorology*, **26**(261), 241–249.

Taylor, A. H. (1995) North–south shifts of the Gulf Stream and their climatic connection with the abundance of zooplankton in the UK and its surrounding seas. *ICES Journal of Marine Science*, **52**, 711–721.

Tett, S. F. B., Stott, P. A., Allen, M. R., Ingram, W. J. and Mitchell, J. F. B. (1999) Causes of twentieth-century temperature change near the Earth's surface. *Nature*, **399**, 569–572.

Thejll, P. and Lassen, K. (1999) *Solar Forcing of the Northern Hemisphere Land Air TemptERATURE: New Data* (Scientific Report 99-9). DMI, Copenhagen.

Thejll, P. and Lassen, K. (2000) Solar forcing of the northern hemisphere land air temperature: New data. *J. Atm. Solar-Terrestrial Physics*, **62**, 1207–1213.

Thejll, P. and Lassen, K. (2002) Erratum to: Solar forcing of the northern hemisphere land air temperature: New data (in: *J. Atm. Solar-Terrestrial Physics*, **62**, 1207–1213). *J. Atmos. Solar-Terrestr. Physics*, **64**, 105.

Thomas, J. H. and Weiss, N. O. (1992) *Sunspots: Theory and Observation* (Series C: Mathematical and Physical Sciences Vol. 375, pp. 3–60) Kluwer, Dordrecht, The Netherlands.

Thompson, W. J. and Wallace, J. W. (1998) The Arctic Oscillation signature in the wintertime geopotential height and temperature fields. *Geophys. Res. Lett.*, **25**, 1297–1300.

Thuillier, G. (2000) Absolute UV radiation, its variability and consequences for the Earth's climate. *The Solar Cycle and Terrestrial Climate* (pp. 69–78). European Space Agency, Noordwijk, The Netherlands.

Tinsley, B. A. (1996) Solar wind of the global electric circuit and apparent effects on cloud microphysics, latent heat relase, and tropospheric dynamics. *J. Geomagn. Geoelectr.*, **48**, 165–275.

Tinsley, B. A., Brown, G. M. and Scherrer, P. H. (1989) Solar variability influences on weather and climate: possible connection through cosmic ray-flux and storm intensification. *Journal of Geophysical Research*, **94**, 14783–14792.

Tobias, S. and Weiss, N. (1999) Solar magnetic field poses problems. *Physics World*, December, 56.

Torrence, C. and Compo, G. P. (1998) A practical guide to wavelet analysis. *Bull. Amer. Meteor. Soc.*, **79**, 61–78.

Trenberth, K. and Stepaniak, D.-P. (2004) The flow of energy through the Earth's climate system. *Quart. J. R. Met. Soc.*, **130**, 2677–2701.

Unterweger, J. (1891) *Über die kleinen Perioden der Sonnenflecken und ihre Beziehung zu eingen periodischen Erscheinungen der Erde*. Vienna.

Usoskin, I. G., Marsh, N., Kovaltsov, G. A., Mursula, K. and Gladysheva, O. G. (2004): Latitudinal dependence of low cloud amount on cosmic ray induced ionization. *Geopyhs. Res. Lett.*, **31**(16), L16109, doi:10.1029/2004GL019507.

Usoskin, I. G., Solanki, S. K., Schussler, M., Mursula, K. and Alanko, K. (2003) A millenium scale sunspot number reconstuction: Evidence for an unusually active Sun since the 1940's. *Phys. Rev. Lett.*, **91**(21), 211101.

v. P. Gruithuisen, Fr. (1826) Naturwissenschaftlicher Reisebericht. *Kastner's Archiv für die gesammete Naturlehre*. Nürnberg.

van Geel, B. and Mook, W. G. (1989) High-resolution ^{14}C dating of organic deposits using natural atmospheric ^{14}C variations. *Radiocarbon*, **31**, 151–155.

Verschuren, D., Laird, C. R. and Cumming, B. F. (2000) Rainfall and drought in equatorial east Africa during the past 1,100 years. *Nature*, **403**, 410–414.

Vines, G. (1999) Mass extinctions. *New Scientist*, December, 1–4, Inside science.

von Storch, H. and Zwiers, F. W. (1999). *Statistical Analysis in Climate Research*. Cambridge University Press, Cambridge, UK.

Wagner, F., Bohncke, S. J. P., Dilcher, D. L., Küscher, W. M., van Geel, B. and Visscher, H. (1999) Century-scale shifts in early Holocene atmospheric CO_2 concentration. *Science*, **284**, 1971–1973.

Wagner, G. W., Livingstone, D. M., Masarik, J., Muscheler, R. and Beer, J. (2001) Some results relevant to the discussion of a possible link between cosmic rays and the Earth's climate. *Journal of Geophysical Research*, **106**, 3381–3387.

Waldmeier, M. (1955) *Ergebnisse und Probleme der Sonnenforschung*. Geest u. Portig, Leipzig.

Wang, C. and Weisberg, R. (1994) On the "slow mode" mechanism in ENSO-related coupled ocean–atmosphere models. *American Meteorological Society*, **7**, 1657–1667.

Wang, Y., Cheng, H., Edwards, R. L., He, Y., Kong, X., An, Z., Wu, J., Kelly, M. J., Dykoski, C. A. and Li, X. (2005) The Holocene Asian Monsoon: Links to solar changes and North Atlantic climate. *Science*, **308**, 854–857.

Webb, S. (1999) Measuring the universe. *Astronomy and Astrophysics*. Springer-Praxis, Chichester, UK.

White, O. R. (2000) Magnetic sources of solar variability. *The Solar Cycle and Terrestrial Climate* (pp. 27–37). European Space Agency, Noordwijk, The Netherlands.

White, W. B., Dettinger, M. D. and Cayan, D. R. (2000) Global average upper ocean temperature response to changing solar irradiance: Exciting the internal decadal mode. *The Solar Cycle and Terrestrial Climate* (pp. 125–129). European Space Agency, Noordwijk, The Netherlands.

Wilks, D. S. (1995) *Statistical Methods in the Atmospheric Sciences*. Academic Press, Orlando, FL.

Williams, E. R. (2001) Sprites, elves, and glow discharge tubes. *Physics Today*, November, 41–47.

Williams, R. W. (1960) *Fundamental Formulas of Physics* (Vol. 2). Dover Books, New York.

Williams, R. W. (1960) *Cosmic Rays and High-energy Physics* (Chap. 23, pp. 544–562).

Willson, R. C. (1997) Total solar irradiance trend during solar cycles 21 and 22. *Science*, **277**, 1963–1965.

Willson, R. C. and Hudson, H. S. (1988) The Sun's luminosity over a complete solar cycle 21 *Nature*, **332**, 810–812.

Willson, R. C. and Mordvinov, A. V. (2003) Secular total solar irradiance trend during solar cycles 21–23. *Geophysical Research Letters*, **30**(5), 1199, doi:10.1029/2002GL016038.

Wilson, R. M. (1998) Evidence for solar-cycle forcing and secular variation in the Armagh Observatory temperature record (1844–1992). *Journal of Geophysical Research*, **103**, 11159–11171.

Wittmann, A. D. and Bianda, M. (2000) Drift-time measurements of the solar diameter 1990–2000: New limits of constancy. *The Solar Cycle and Terrestrial Climate* (pp. 113–116). European Space Agency, Noordwijk, The Netherlands.

Wolfstein, L. 1998. Neutrino mass discovered. *Physics World*, **11**(7), 17.

Wuebbles, D. J., Wei, C. F. and Patten, K. O. (1998) Effects on stratospheric ozone and temperature during the Maunder minimum. *Geophys. Res. Lett.*, **25**, 523–526.

Wunch, C. (2000) Moon, tides and climate. *Nature*, **405**(June), 743–744.

Wyrtki, K. (1985) Water displacements in the Pacific and the genesis of El Niño cycles. *Journal of Geophysical Research*, **90**, 7129–7132.

Yu, F. (2002) Altitude variations of cosmic ray induced production of aerosols: Implications for global cloudiness and climate. *J. Geophys. Res.*, **107**, SIA 8.

Zebiak, S. E. and Cane, M. A. (1987) A model ENSO. *Monthly Weather Review*, **115**.

Zhang, Y., Wallace, J. M. and Battisti, D. S. (1997) ENSO-like interdecadal variability: 1900–93. *Journal of Climate*, **10**, 1004–1020.

Zwaan, C. (1992) *Sunspots: Theory and Observation* (Series C: Mathematical and Physical Sciences Vol. 375, pp. 75–100). Kluwer, Dordrecht, The Netherlands.

Exercises

CHAPTER 1

1.1 What are the 3 ‘pillars’ on which scientific knowledge is built?
1.2 What are the criteria for being scientific?

CHAPTER 2

2.1 (a) What instrument measures the total solar irradiance?
(b) How does it work?
2.2 What 3 fundamental types of information can be derived from electromagnetic radiation?
2.3 What are the four fundamental physical quantities?
2.4 (a) What is the *continuous spectrum*?
(b) What causes its width?
2.5 (a) What are *Fraunhofer lines*?
(b) What is *spectroscopy*?
(c) What does the *magnetograph* measure?
2.6 How can the wavelength of light provide information about motion?
2.7 What is *seeing correction*?
2.8 What can an east–west hemispheric asymmetry in observed number of sunspots indicate?
2.9 (a) What are *cosmogenic isotopes*?
(b) How are they produced?
2.10 How are isotope ratios measured?
2.11 (a) What is *carbon dating*?
(b) What instrument is involved?
2.12 What is *wiggle match*?

2.13 (a) What does a high $\delta^{18}C$ ratio indicate?
(b) What does a high $\delta^{13}C$ ratio indicate?
2.14 Which factors may influence $\delta^{10}Be$?
2.15 What is the $^{18}O/^{18}O$ ratio a proxy for?
2.16 What is the *aa-index*?

CHAPTER 3

3.1 What is the material which the Sun consists of?
3.2 (a) How is energy produced in the Sun?
(b) Where in the Sun?
3.3 Explain the energy transfer from the region where solar energy is produced to the Earth.
3.4 (a) How is the temperature of the photosphere determined?
(b) What is the estimated value?
3.5 (a) Describe *facular brightening*.
(b) When is the radiation greatest – when there are many or few sunspots?
3.6 What observed features at the photosphere suggests convective processes?
3.7 Describe the *chronosphere*.
3.8 What are *plages* and *spicules*?
3.9 (a) Describe the *corona*.
(b) How is the shape of the corona affected by solar activity?
3.10 (a) What is the *solar wind*?
(b) What does it do with the solar magnetic field?
3.11 What characterises the general solar rotation?
3.12 What phenomena is thought to be responsible for a net bipolar solar magnetic surface field?
3.13 Discuss the origin of the general solar magnetic field.
3.14 What is the essence of the *Alfvén theorem*?
3.15 (a) What is a *prominence*?
(b) What is a *flare*?
3.16 What is a *corona mass ejection (CME)*?

CHAPTER 4

4.1 What two characteristic do sunspots have that make them differ from the photosphere?
4.2 What is a *penumbra* and *umbra*?
4.3 What is the Schwabe cycle?
4.4 What is the relationship between pores and magnetic fields?
4.5 (a) Are the photospheric granules bigger or smaller in a sunspot?
(b) What is a moat cell?

4.6 Is the magnetic field associated with sunspots weaker or stronger than for the rest of the photosphere?

4.7 What is the *Wilson effect*?

4.8 What is the *Evershed effect*?

4.9 What is meant by *leading* and *following* spots?

4.10 What are bipolar groups?

4.11 (a) What is the spot zone?
(b) How does the latitude of the sunspots vary?

4.12 Where do the spots during solar minimum typically appear on the Sun?

4.13 (a) What is the range of lifetimes of sunspots?
(b) What is the time of one solar rotation?

4.14 (a) How are convective processes affected by sunspots?
(b) How does convection affect the sunspots?

4.15 What three different basic models are there for explaining solar activity?

4.16 What does the butterfly diagram convey?

4.17 (a) What are bipolar regions?
(b) What features are associated with these?

4.18 Explain the two basic sunspot models (theories).

4.19 Explain the vortex model proposed by Villhelm Bjerknes.

4.20 Explain the magnetic cooling models.

4.21 What is *flux expulsion*?

4.22 What is the origin of the solar magnetic field?

4.23 What role may radially different rotation rates play in terms of sunspots?

4.24 (a) What is the formulae for the Wolf sunspot number (R_z)?
(b) What do the symbols represent?

4.25 What is *metadata*?

4.26 (a) What is the *Hale cycle*?
(b) What basic models are used to explain this phenomenon?

4.27 (a) List the different common measures for solar activity level.
(b) Exactly, what do they (presumably) indicate?

4.28 What does an east–west asymmetry in sunspot observations imply?

4.29 How is TSI affected by the 11-year solar cycle?

4.30 What solar phenomena are thought to affect the UV radiation?

4.31 What wavelength of light varies most with the 11-year solar cycle?

4.32 What four basic models are there relating the TSI to the 11-year solar cycle?

4.33 Explain why SCL may provide a measure of the level of solar activity.

4.34 What three steps are required for the reconstruction of the TSI?

4.35 What are (a) *flares*, (b) *prominences*, (c) *faculae*, and (d) *corpuscular clouds*?

CHAPTER 5

5.1 (a) What is the principal driver for weather?
(b) What is the evidence of (poleward) heat transport?

5.2 What is the *butterfly effect*?

5.3 (a) What should temperature measurements represent?
(b) How should the measurements be made?

5.4 What are *inversions*?

5.5 (a) What are urban heat islands?
(b) How can this effect be accounted for?
(c) What other factors may introduce 'spurious' changes in long-term trends?

5.6 (a) How are SSTs measured?
(b) What do the different measurements (ship, satellite) represent?

5.7 (a) What does the SLP really measure?
(b) What units are commonly used for SLP?

5.8 (a) How is precipitation usually measured?
(b)What may affect the readings?

5.9 What are the differences between polar orbiting and geostationary satellites?

5.10 (a) What is a *radiosonde*?
(b) What is SMMR, and what does it measure?
(c) What is SSM/I?
(d) What physical entities are really measured by satellite-borne instruments?
(e) How are quantities such as air temperature inferred?

5.11 (a) What is MSU?
(b) What does it measure?
(c) What instrument does it consist of and what channels does it involve?

5.12 (a) What do tree rings give an indication of?
(b) What geological proxies can provide clues about past climatic conditions?
(c) What is archeological data?

5.13 (a) Describe the important conservation laws for Earth's climate.
(b) Discuss one fundamental implication of the equation of continuity.

5.14 (a) Describe the climatological 'heat engine'.
(b) In what main forms is energy present on Earth.

5.15 (a) What is *entropy*? State the first and second laws of thermodynamics.

5.16 (a) What is meant by *hydrostatic equilibrium*?
(b) What does it imply?

5.17 (a) What is *vorticity*?
(b) What is *potential vorticity*?

5.18 (a) What role does the Coriolis force play for cyclones?
(b) What happens to a volume of air as it is forced over mountain ranges such as the Rockies?

5.19 (a) What is *albedo*?
(b) Estimate Earth's emission temperature for the different values for planetary albedo: [0.3, 0, 0.5, and 0.7]

5.20 (a) Why is temperature not very sensitive to variations in TSI?
(b) What other factors affect the temperature response?

5.21 (a) What is *nocturnal jet*?
(b) How does it arise?

5.22 (a) What is *eccentricity*?
(b) What is it's present value for the Earth?

5.23 (a) Compare the difference in the Earth emission temperature at min. and max. distances, assuming that the albedo is uniform and constant.
(b) What is *perihelion*, *aphelion*, *apoge* and *apoapsis*?
5.24 Explain the greenhouse effect.
5.25 (a) What role do clouds play in the radiation budget?
(b) Is there a difference between low and high clouds, and if so, what?
5.26 (a) What are the names of the different vertical regions in the atmosphere?
(b) What characterises these and how do they differ?
(c) What is *lapse rate*?
5.27 (a) What is meant by *advection*?
(b) How does it affect the temperature at high latitudes?
(c) Describe the Hadley cell.
5.28 (a) What are *CCN*?
(b) What is the difference between warm and cold initiation?
(c) Explain how cloud drops grow and what processes are involved.
(d) How may electric fields be involved.
5.29 What cloud parameters are considered to be important in climatological context?
5.30 What causes hydrostatic stability in the stratosphere?
5.31 Outline which properties oceans have that affect the climate.
5.32 What is *Ekman drift* and *Ekman pumping*?
5.33 What types of ocean waves are there?
5.34 What role do the oceans play in the carbon cycle?
5.35 Describe two ways in which solar activity may affect oceans.
5.36 Describe properties of the cryosphere important for climate.
5.37 Describe the *thermohaline circulation*.
5.38 How may biological activity influence climate?
5.39 (a) What are *feedback processes*?
(b) Explain two different definitions.
5.40 Describe the *Stefan-Boltzmann feedback*.
5.41 How does the *water vapour feedback* work?
5.42 Discuss the snow–albedo feedback.
5.43 What two effects may *cloud feedback* have on climate?
5.44 (a) What four properties are important for climatic differences between the planets in the solar system?
(b) How do the inner planets differ from the outer ones?

CHAPTER 6

6.1 Discuss three ways solar activity may affect climate involving the stratosphere.
6.2 Discuss the relative and absolute variations in short wave solar radiation.
6.3 How does enhanced UV-levels affect ozone photo-chemistry and the stratosphere?
6.4 (a) What is *XUV*?

(b) What is *ozone*?
(c) Discuss differences between tropospheric and stratospheric ozone.

6.5 (a) Recite the equations on which the Chapman theory is based.
(b) What is a *catalyst*?
(c) What happens to ozone when exposed to UV light?

6.6 (a) What is the *ozone hole*?
(b) What role do CFCs play?

6.7 (a) How may the Hadley Cell be affected by stratospheric ozone?
(b) What implications may this have in terms of solar variability?

6.8 (a) What are *planetary waves*?
(b) How may these be involved in a calimate response to solar variability?

6.9 (a) What is the *QBO*?
(b) What is its preferred timescale?

6.10 How is the QBO thought to be affected by solar activity?

6.11 (a) What is the AO?
(b) What is the usual definition of the mode?

6.12 (a) Discuss the link between stratospheric phenomena and the surface climate.
(b) Mention more than one mechanism through which the troposphere is affected by the stratospheric response to solar variability.

6.13 (a) Discuss how volcanic eruptions may affect climate.
(b) What is the usual time scale of the climatic response.

6.14 Discuss the differences between small and large particles (aerosols) in the stratosphere.

CHAPTER 7

7.1 (a) What are the northern lights?
(b) When are they most active?

7.2 (a) How does the solar activity affect the geomagnstic field?
(b) What are geomagnetic storms and what causes them?

7.3 (a) What is the *solar wind*?
(b) How does it affect Earth's magnetic field?

7.4 What is the *magnetosphere*?

7.5 (a) What are the *Van Allen belts*?
(b) How is it related to the southern Atlantic anomaly?

7.6 (a) Describe the atmospheric electric field.
(b) What role does lightening play for the potential difference between Earth's surface and the ionosphere?

7.7 What is meant by *charge separation*?

7.8 (a) What are *galactic cosmic rays*?
(b) What other type of cosmic rays are there (in terms of source)?
(c) How were these discovered?

7.9 (a) What do GCR do to the air?
(b) What type of GCR-related particles are there in the atmosphere?

7.10 (a) Describe *airglow*, *sprites* and *elves*.
(b) What phenomenon are sprites and elves associated with?

7.11 Who is known to be the first scholar who proposed that northern lights were associated with the geomagnetic field?

7.12 (a) Is auroral activity changing with season, and if so, what time of year is most active?
(b) How does the auroral activity relate to the level of solar activity?

7.13 (a) Describe the two types of auroras.
(b) What are the main differences?

7.14 (a) How do northern lights arise?
(b) Explain why it is usually seen at high latitudes.

7.15 What is the *IMF*?

7.16 (a) What was the *Little Ice Age*?
(b) What is meant by the *Maunder Minimum*?

7.17 a) How can ^{14}C and ^{10}Be be used to infer levels of solar activity?
(b) Outline the physical mechanisms.

7.18 Discuss the two main mechanisms through which GCR theoretically can affect cloud formation.

7.19 How do low and high clouds affect climate and what are the main differences between these two cloud types?

7.20 (a) What is the Köler curve?
(b) How are GCR usually measured?

7.21 Discuss the mechanism proposed by Svensmark on how GCR may affect Earth's temperatures.

7.22 Why is a decrease in the GCR flux level expected to give a stronger day-time response (as opposed to the night-time)?

7.23 What are Forbush decreases?

7.24 Describe the geodynamo.

CHAPTER 8

8.1 Discuss the *scientific criteria*.

8.2 What are *parametric tests*, *null-hypotheses*, and *null-distributions*?

8.3 Describe the basics and purpose of Monte-Carlo simulations.

8.4 Give an account of the first ideas about feedback mechanisms in solar–terrestrial relationships (Blandford)?

8.5 (a) Describe the early relationships established between sunspots and temperature on Earth (before 1910).
(b) How do these compare with current knowledge?

8.6 What was the established view of the relationship between TSI and sunspots before Abbot and Fowle (1913)?

8.7 Discuss the inter-decadal cycle proposed by Brückner.

8.8 (a) What is normally meant by *climatology*?
(b) Why does the analysis often focus on anomalies?

8.9 What is meant by *standardisation*?
8.10 How can trends and filters affect correlation estimates?
8.11 Explain how filters may affect the degrees of freedom.
8.12 What is *bootstrapping*?
8.13 How may aerosols affect climate?
8.14 What is *ANOVA*?
8.15 What is the proper way of validating (evaluing) a model?
8.16 List 7 different proxies for solar activity (directly observable, or derived thereof, i.e., not palaeo records).
8.17 What is the *Gleissberg cycle*?
8.18 What are *EBMs* and what are their shortcomings?
8.19 What steps are involved in reconstucting past irradiance?
8.20 (a) Outline the stellar evolution.
(b) What is the faint young Sun–warm Earth paradox?
8.21 Outline the *snowball Earth hypothesis*.

CHAPTER 9

9.1 What is meant by *modes of natural variability*?
9.2 (a) Outline the main features of ENSO.
(b) What main models account for the variability?
9.3 Describe the mean state along the equatorial Pacific, including sea levels, temperatures, winds and currents.
9.4 (a) What is *Ekman drift*?
(b) What is the *cold tongue*?
(c) How are these related?
9.5 Outline the *delayed oscillator model*.
9.6 What is the difference between Rossby and Kelvin waves?
9.7 What is the *mixed layer*?
9.8 What is meant by *ocean–atmosphere (air–sea)* coupling?
9.9 What is the *MJO*?
9.10 Is solar activity required to account for inter-annual variability associated with ENSO, and if so, what mechanism can explain the connection?
9.11 Discuss how the Hadley Cell may be a link between solar activity and ENSO.
9.12 What is the *SOI*?
9.13 Explain the mian features of the south Asian monsoon.
9.14 What evidence is there pointing to a connection between solar activity and the monsoon.
9.15 What is the *NAO*?
9.16 Discuss plausible explanations, whereby solar activity might affect the NAO.
9.17 (a) Describe the Gulf Stream.
(b) What is its role in the climate system?
9.18 What are the difficulties associated with a hypothesised link between solar activity and the Gulf Stream position?

9.19 (a) What is the *PDO*?
(b) Are there any indications of an association between the PDO and solar activity?

9.20 Describe the greenhouse effect.

9.21 What are the Milankovitch cycles?

9.22 (a) How does the moon affect the oceans and the atmosphere?
(b) What timescales are involved?

Index

Colour plates

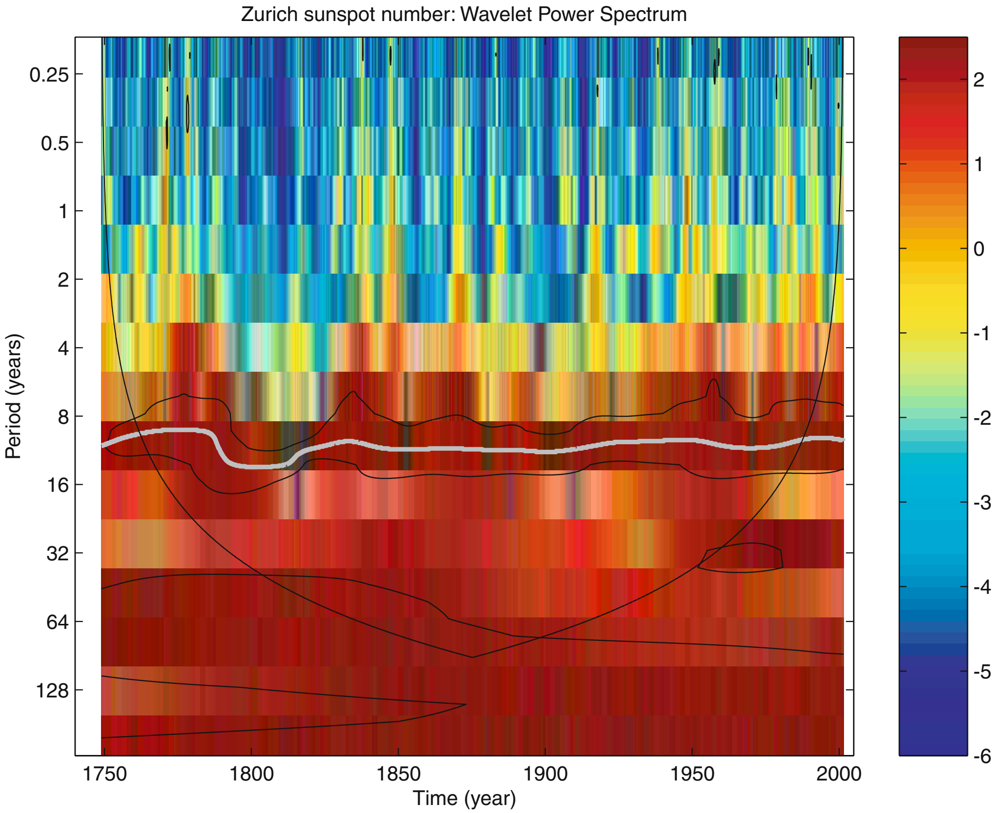


Figure 4.9. Wavelet analysis of the Zurich sunspot number showing how the power spectrum estimate evolves over time (Morlet wavelet). The "11-year" cycle is marked with a thick light grey curve. There were some interesting fluctuations in the periodicity of the Schwabe cycle at the end of the 18th century. The wavelet analysis also suggests that there are cycles with timescales of the order of 100 years which may have experienced a slight trend towards longer periodicities. There are also hints of an envelope in the high-frequency (0.5–6 years timescale) variations. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

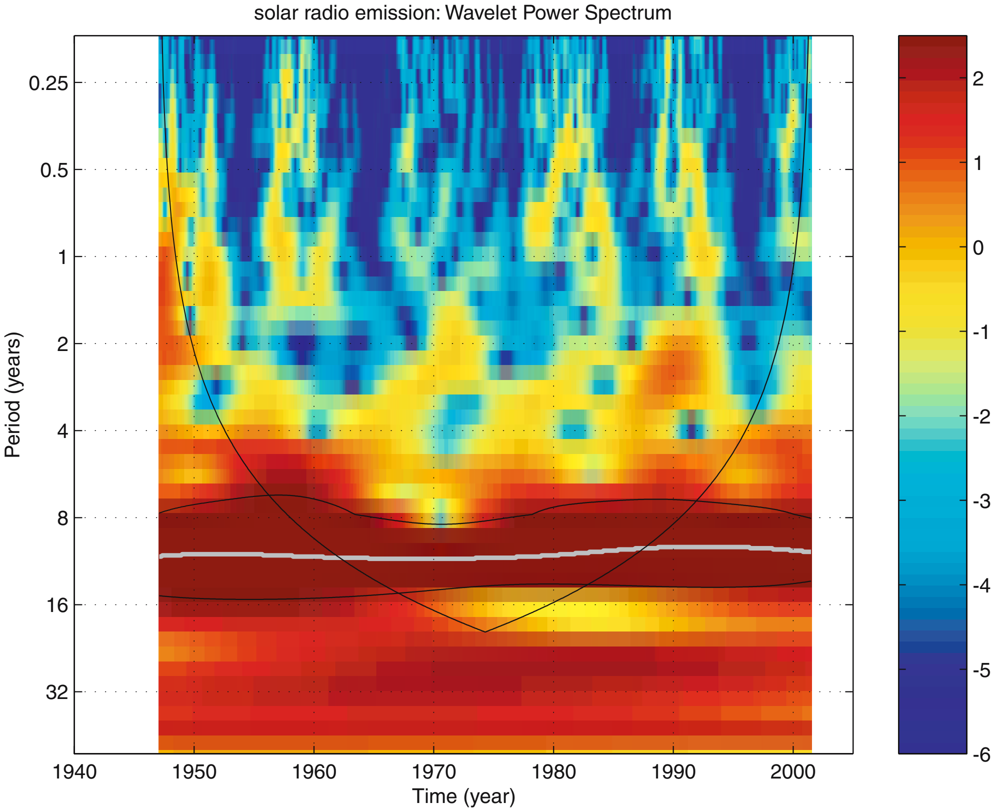


Figure 4.12. Results from a wavelet analysis of the solar radio emission. The 11-year cycle is prominent, and the modulation of the high-frequency variations can be seen as alternating blue and yellow shading in the frequency band 0.25–2.00 years. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

Figure 5.4. One complication associated with some satellite measurements is determining which retrieval algorithm is most appropriate for twilight conditions as reflected solar glare may contaminate the atmospheric signal.

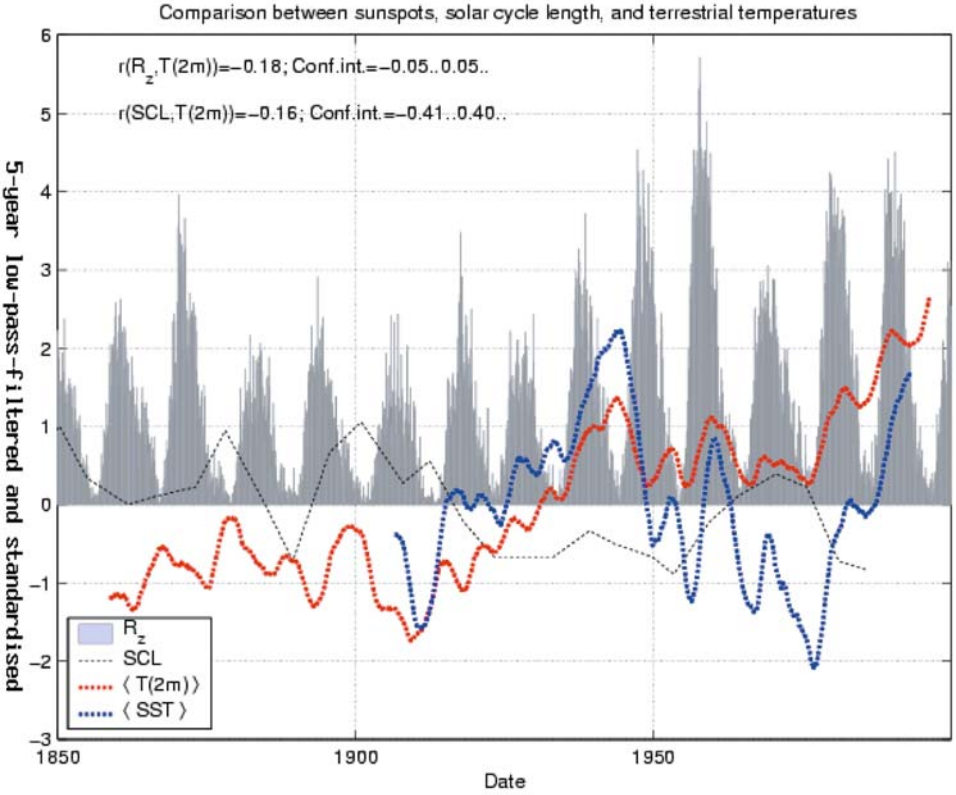


Figure 8.4. The Zurich sunspot number curve (bars, rescaled by dividing with one standard deviation) as well as standardised global mean temperature (red curve), global mean sea surface temperature (blue) and solar cycle length (black). The standardisation allows easy comparison between the various records despite different units and magnitudes. The correlation analysis has been applied to unfiltered sunspot number and 5-year Gaussian low-pass-filtered temperature, and the correlation estimates are: $r(R_z, T[2m]) = -0.18$, $r(\mathrm{SCL}, T[2m]) = -0.16$. Note the negative correlation estimate for R_z and temperature, which can be explained by Figure 8.8. The seasonal cycle has been removed from the temperature record. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA, CRU and the UK Meteorological Office.

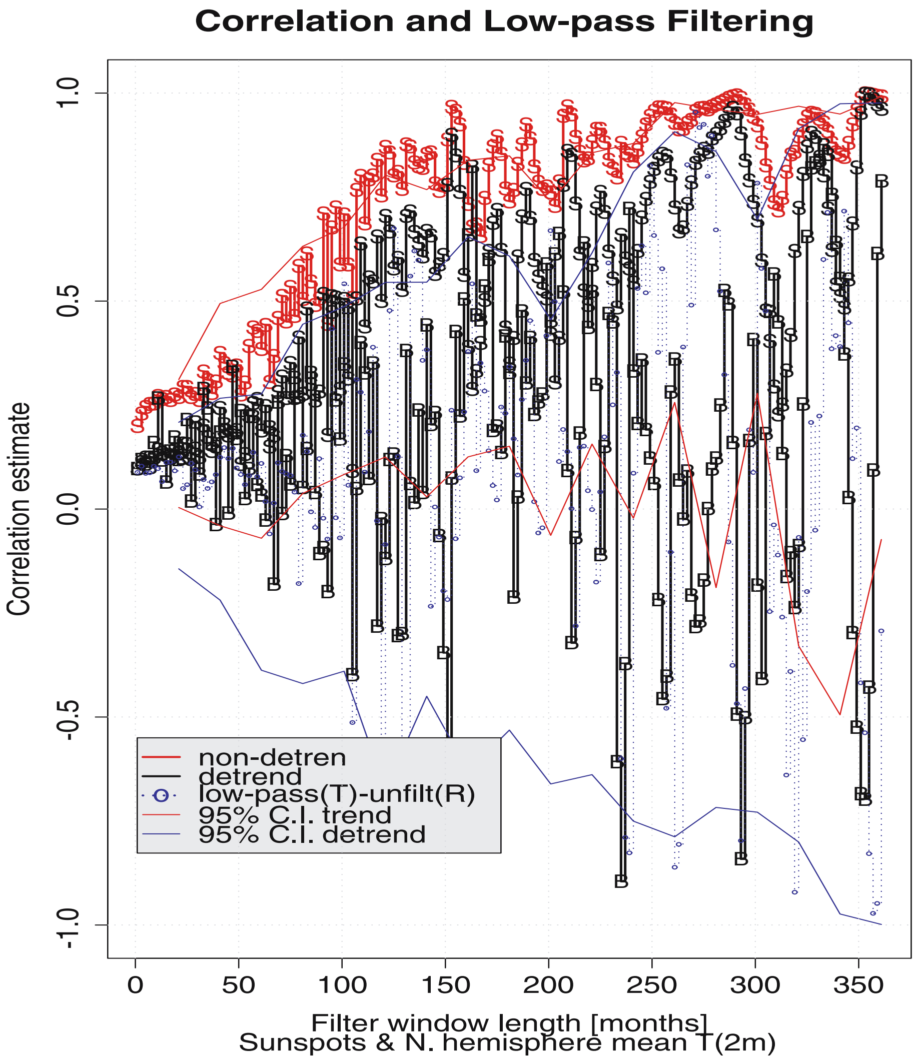


Figure 8.8. The effect of low-pass-filtering on the correlation estimate for the monthly mean sunspot number and northern hemisphere mean temperature for de-trended (red) and non-de-trended (black dashed) series. The two black curves show the results derived using a binomial filter ("B") and a square moving average filter ("S"). The blue curve represents an analysis where both records have been de-trended but only the temperature has been low-pass-filtered. The filter is a box-car type and the *x*-axis represents the filter window width. The greater the width the stronger the smoothing of the curves. The filtered curves have been sub-sampled by selecting values at intervals corresponding to one filter window length. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and CRU.

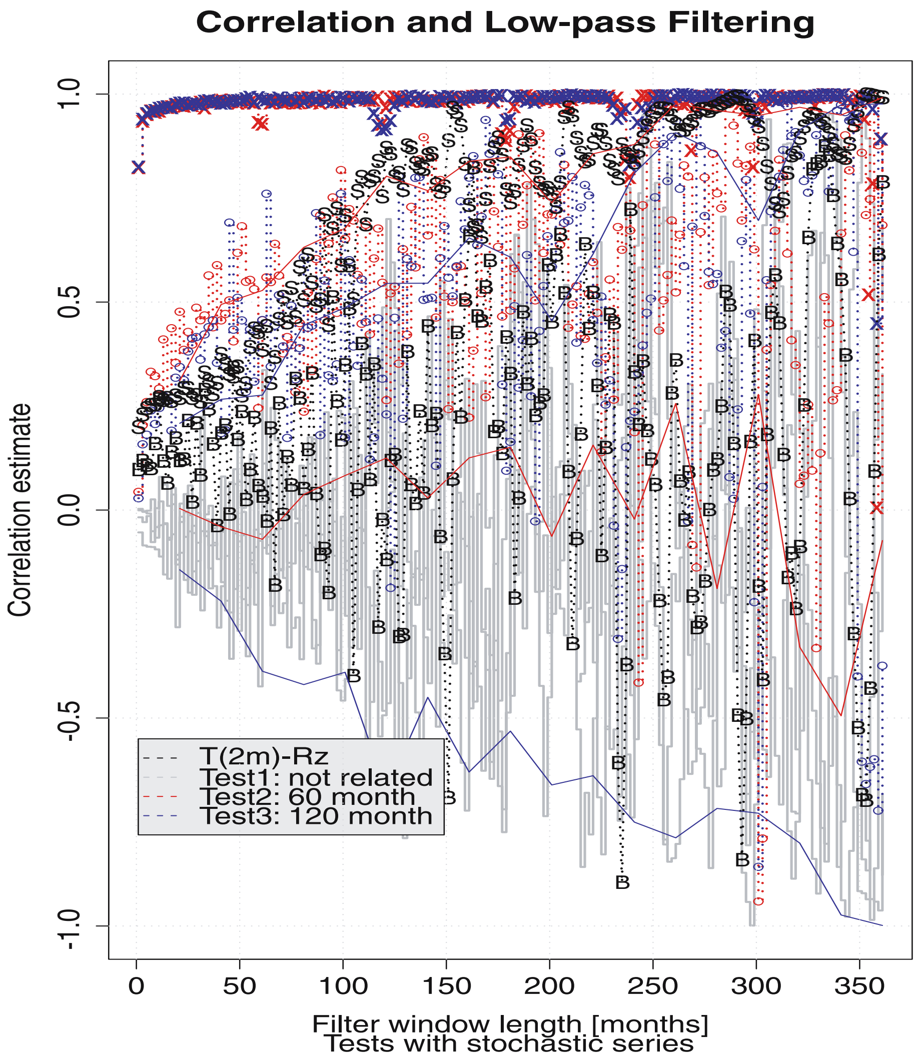


Figure 8.9. Test where similar analysis as in Figure 8.8 is applied to stochastic series (grey) and constructed series with imposed "weak" ($x_1 = 0.25 \times x_2 + 0.75\times$ white noise) and "strong" relationships ($x_1 = 0.75 \times x_2 + 0.25\times$ white noise). The strongly related data give correlations close to unity.

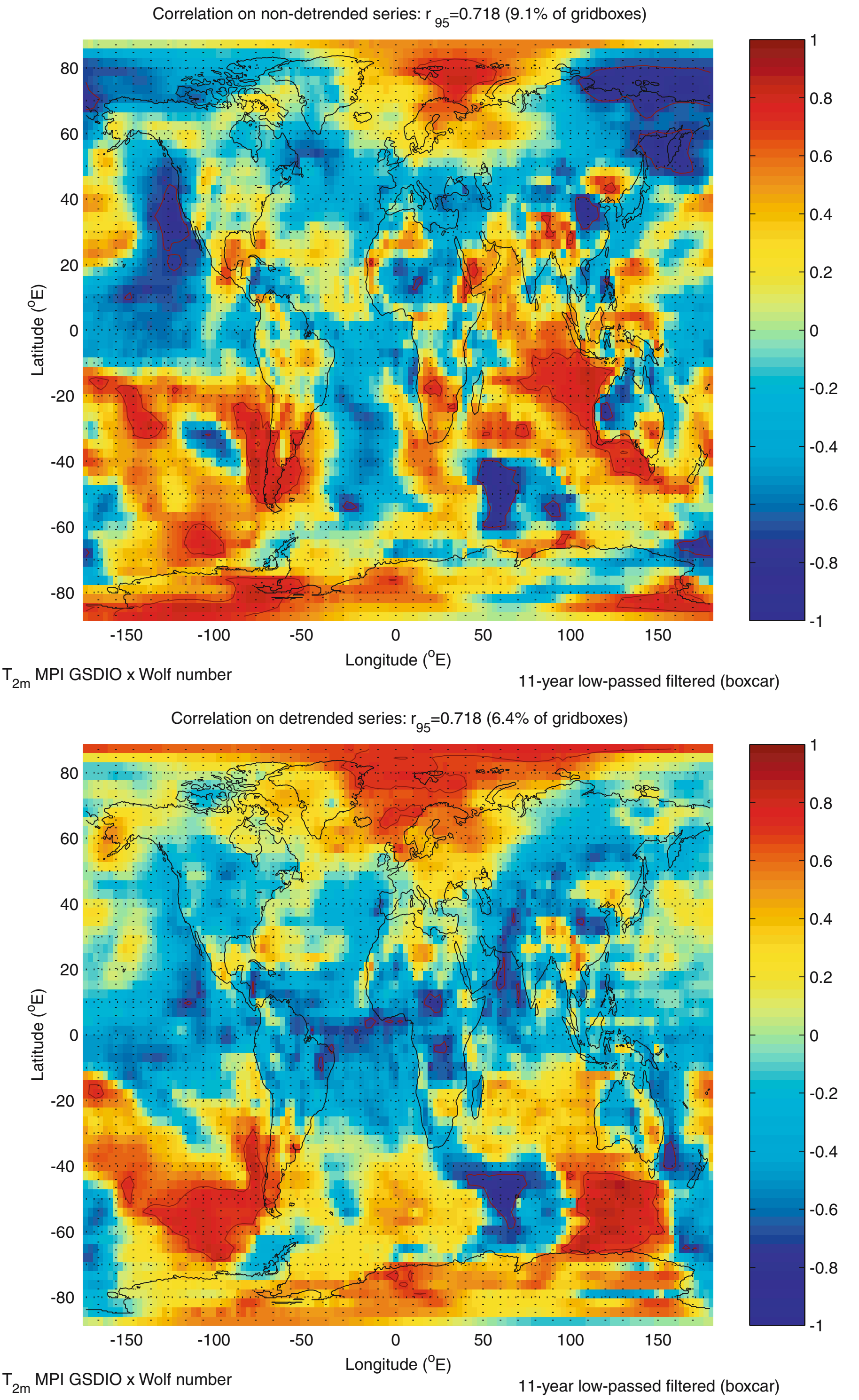

Correlation on non-detrended series: r $_{95}$=0.718 (9.1% of gridboxes)
Latitude (^{o}E)
80
60
40
20
0
-20
-40
-60
-80
-150
-100
-50
0
50
100
150
Longitude (^{o}E)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
T$_{2m}$ MPI GSDIO x Wolf number
11-year low-passed filtered (boxcar)
Correlation on detrended series: r$_{95}$=0.718 (6.4% of gridboxes)
Latitude (^{o}E)
80
60
40
20
0
-20
-40
-60
-80
-150
-100
-50
0
50
100
150
Longitude (^{o}E)
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
T$_{2m}$ MPI GSDIO x Wolf number
11-year low-passed filtered (boxcar)

Figure 8.17. Correlation maps between the SCL and the 2-meter temperature from Max-Plancks-Institute's climate model (ECHAM4/OPYC3, GSDIO). In the upper panel the data have not been de-trended, whereas the lower panel shows the correlation map for the de-trended data. The temperatures have been subject to a 11-year low-pass filter, but the temperature series has been sub-sampled so that only those data points coinciding with the SCL (mid-point of each solar cycle) have been used in the analysis.

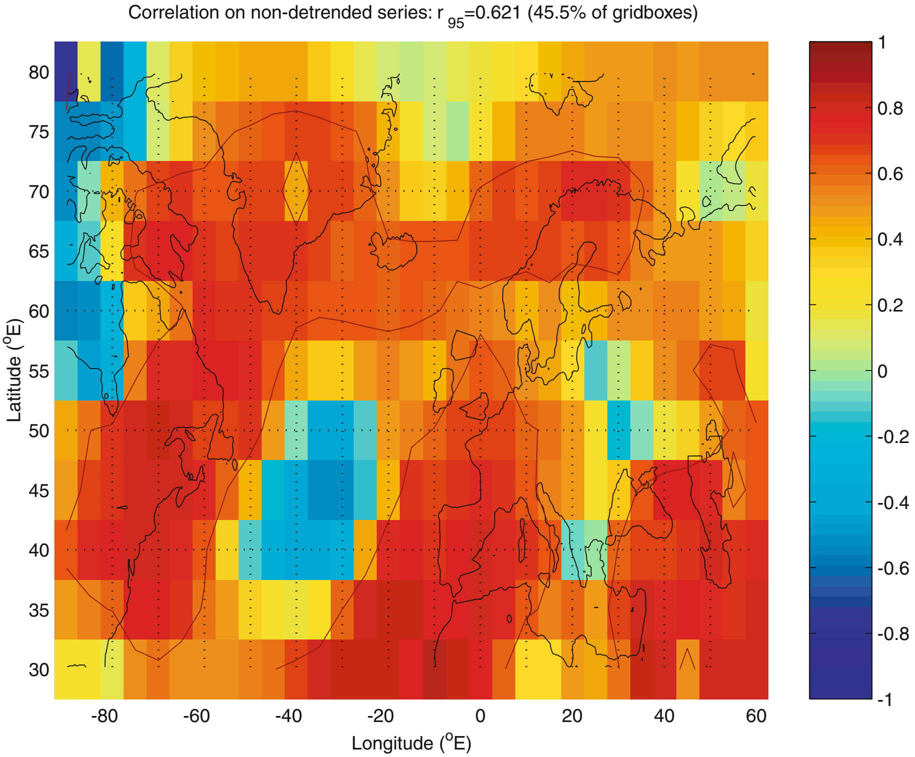


Figure 8.18(a). Map showing the correlation between the SCL and 11-year low-pass-filtered $T(2m)$ reconstruction using non-de-trended data. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and Benestad (2000a).

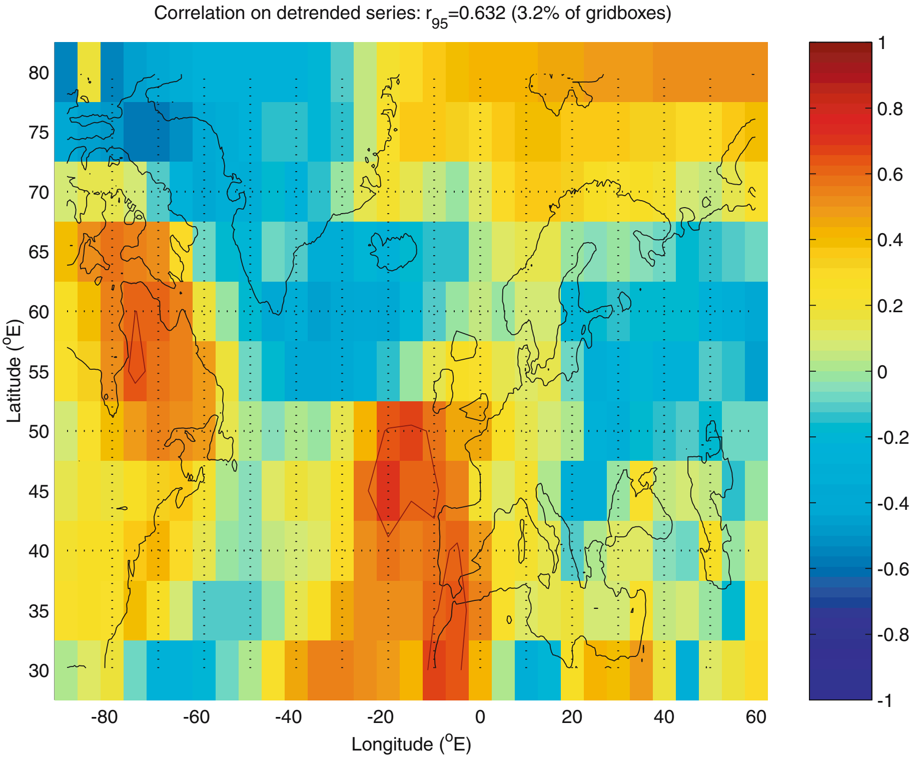


Figure 8.18(b). Map showing the correlation between the SCL and 11-year low-pass-filtered T(2m) reconstruction using de-trended data. Data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA and Benestad (2000a).

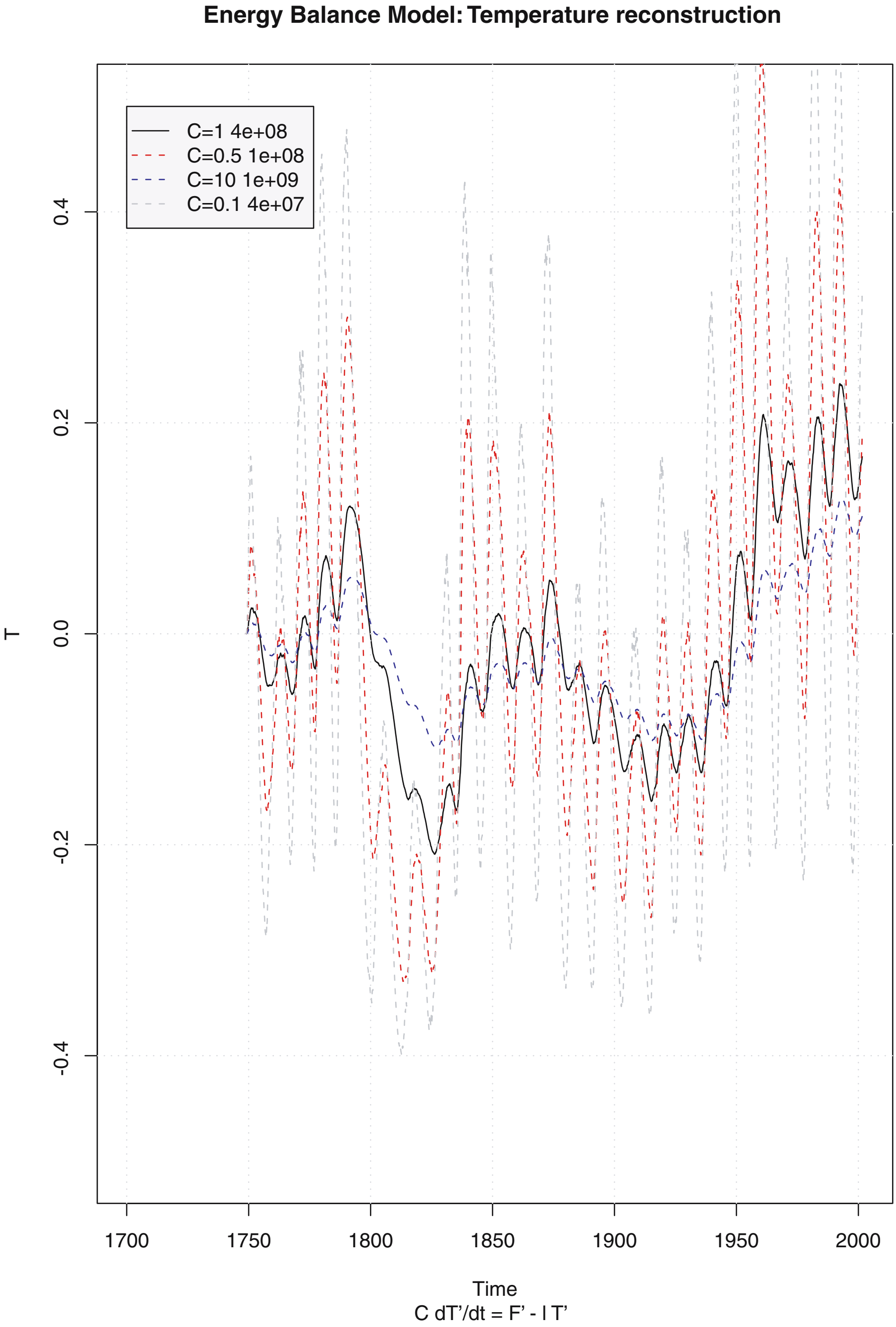


Figure 8.23(a). Temperature reconstruction based on the simple energy balance model described by equation (8.6) and using TSI reconstruction based on equation (4.14). This model has been linearised ($\lambda \Delta T$). Sunspot data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

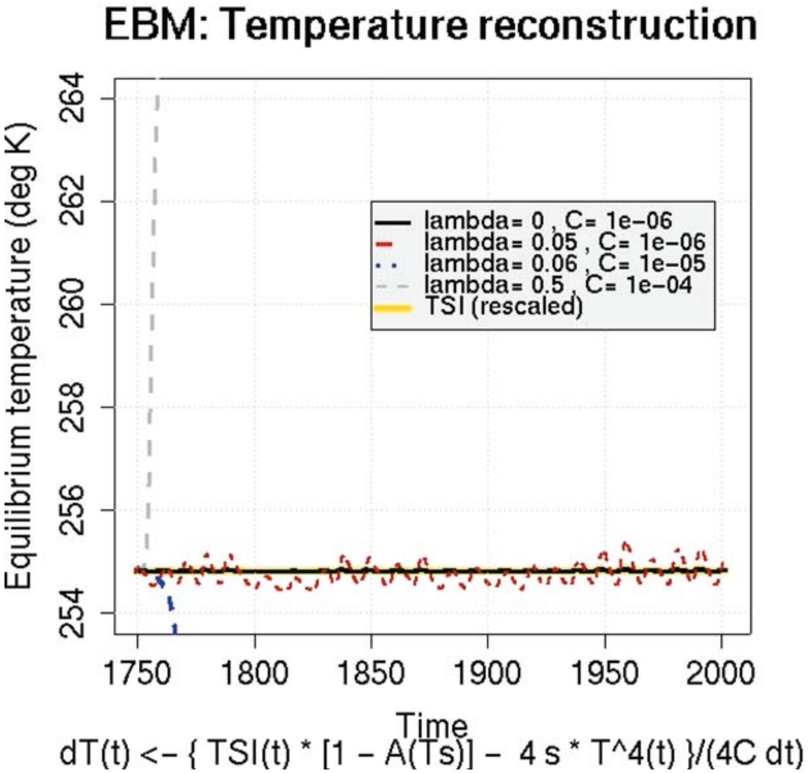
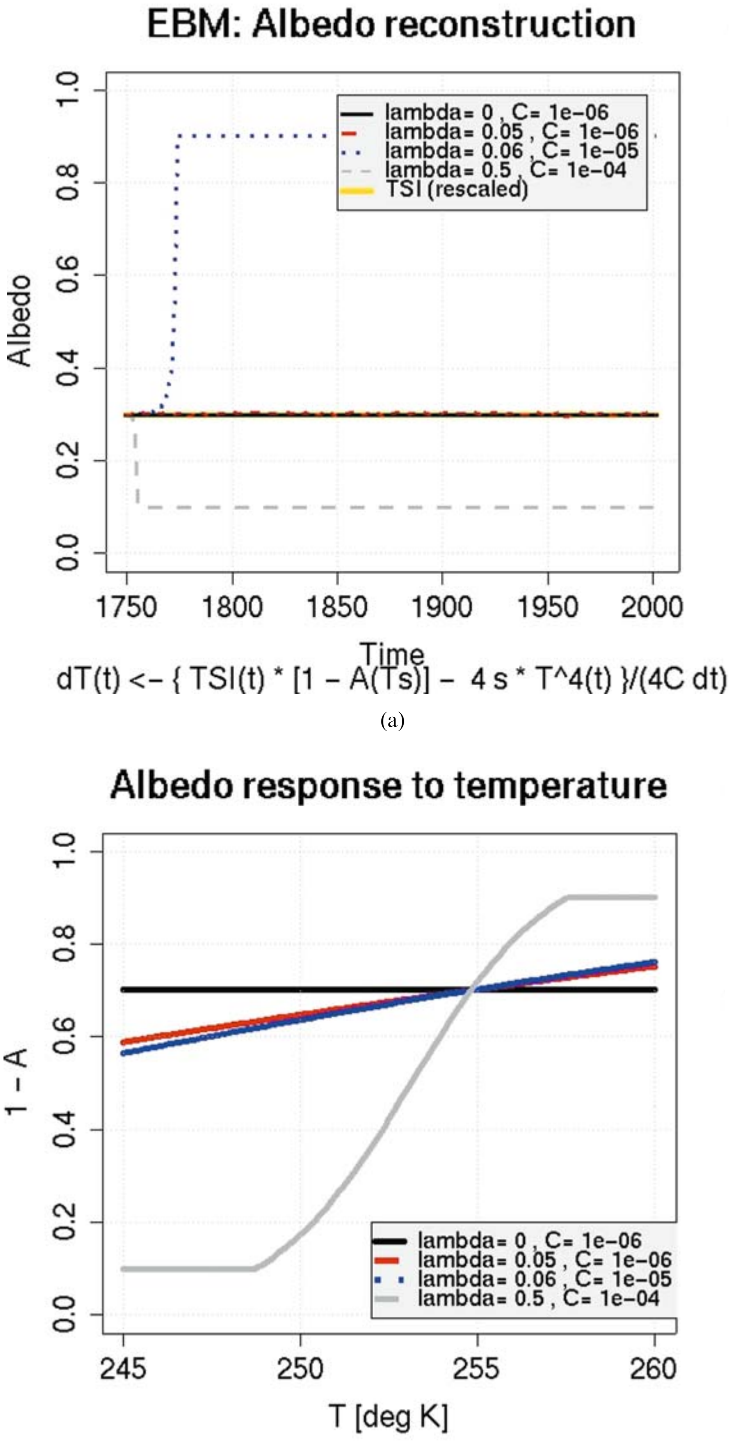


Figure 8.23(b). As (a) but incorporates an albedo feedback. This model includes the full quadratic response in temperature ($\lambda \Delta T^4$). Sunspot data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.



(a)

(b)

Figure 8.24. The albedo derived from the EBM model in Figure 8.23(b) (a) and the sensitivity of the theoretical albedo model to temperature (b). The albedo is calculated according to $A = 1/1 + \exp((T - T_0)l)$, but the minimum and maximum values are restrained to 0.1 and 0.9 respectively. Sunspot data from: ftp://ftp.ngdc.noaa.gov/STP/SOLAR_DATA.

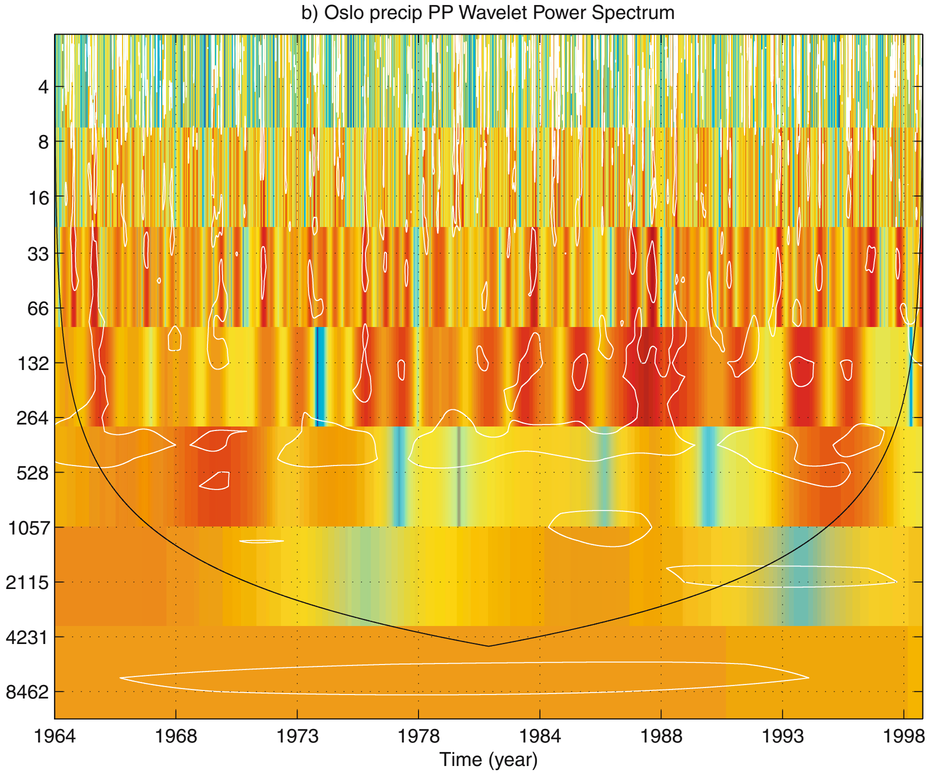


Figure 8.25. Wavelet analysis of the daily precipitation in Oslo bringing out the annual cycle, and possibly a weak cycle with a timescale of 7000–8000 days (19–22 years). There are also hints of variability with preferred timescales of 1000 days (2–3 years) and 2000 days (5 years). Data from the Norwegian Meteorological Institute.

Printing: Mercedes-Druck, Berlin
Binding: Stein + Lehmann, Berlin