
The Fluid Catastrophe
John Reid

The Fluid Catastrophe

The Fluid Catastrophe

By

John Reid

Cambridge
Scholars
Publishing

The Fluid Catastrophe

By John Reid

This book first published 2019

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-3206-2
ISBN (13): 978-1-5275-3206-9

Contents

Foreword

We arrived at Buckles Bay, Macquarie Island on MV Nella Dan in December 1963. We clambered down the ship's side on a rope ladder into an amphibious vehicle and were driven ashore. As the vehicle ground its way up the steep pebble beach, I recall being not overly impressed by my new surroundings. Four ton elephant seals clashed their chests in mock battle and needed to be persuaded to get out of the way of our vehicle. Despite the brisk, cold wind whipping through the tussock grass, the place stank to high heaven. The smell came from the piles of rotting bull kelp strewn along the beaches. This was going to be my home for the next 12 months. It neither looked nor smelled hospitable.

The southern aurora (Aurora Australis) is most active at the latitude of Macquarie Island, which lies 1200 km SE of Tasmania, Australia. There is a band of activity known as the Auroral Oval which encircles the South Magnetic Pole located at the edge of Antarctica south of Tasmania. I had taken a job with the Australian Antarctic Division as Auroral Physicist. This entailed operating various semi-automatic instruments which measured various aspects of the aurora. Long nights of visual observation were no longer part of the job description. There was an "all-sky-camera" which took a ten second exposure of the whole sky once per minute during the hours of darkness, a spectrophotometer which measured auroral colours and an instrument called a "riometer" which measured radio wave absorption, which occurred when the aurora was active.

All of this high-tech equipment was new. I was one of the first of a new breed of auroral physicists being sent south. First we had to build the auroral lab, then install the new gear, and then keep it operational. The prefabricated lab was erected behind a small hill to shield the instruments from stray light from the base.

I spent the whole year in a frantic attempt to keep it all working. The station experienced frequent power failures due to contaminated diesel fuel, and these led, in turn, to failures in the lab DC power supply which supplied all the instruments. Solid-state electronics was very new back then. I had barely even seen a transistor before I arrived. Eventually I redesigned the lab power supply to make it sufficiently robust to handle the power failures and brown outs. I also developed about 5 km of all-sky-camera film. It was a hectic year but I coped and in the end stayed on an extra three months to show the new guy the ropes. I had no time to do any research. I was a maintenance technician.

But this wild and windy place had got into my blood. In some way, I grew up there.

It teemed with wildlife. Several species of penguin, fur seals and elephant seals, skuas, giant petrels and other sea birds were all to be found near the main base at Buckles Bay at the northern end of the island. The wandering albatross nested on the coastal plain and could be approached without their attempting flight. They are the world's largest flying bird – so large that they only become airborne with difficulty. The beautiful, light-mantled sooty albatross nested among the giant grass tussocks which covered the steep slopes. Occasionally, on the plateau among cushion plants and rocks, a feral cat would be glimpsed from afar. They seemed larger than domestic cats and were generally ginger in colour.

On the coastal plains of grass and feather-bed bog were rabbits and wekas, introduced as a food source by sealers in the 19th century. The Stewart Island weka is a flightless rail similar to the kiwi. They are amusing little creatures, like fluffy, chestnut bantams. They occupied small territories with fiercely contested, but invisible, boundary lines. On the beaches and around the base, numerous fat elephant seals were moulting, mating or fighting, depending on the time of the year. It was like living in a big zoo.

The riometer is really a radio telescope with a very wide beam antenna pointing upwards and operating at the lowest frequency at which the ionosphere is nearly transparent. It receives radio noise from the Milky Way as it passes overhead, giving a smooth curve on the paper charts. During times of auroral activity, radio waves are absorbed and big dips appear in the daily curve. It was a reliable and sensitive instrument and, unlike optical methods, operated well in daylight and on cloudy nights. I became interested in the fine

structure in the riometer absorption records. It was very tantalizing; certain rough or "noisy" patches seemed to correspond to the X-ray "microbursts" which had recently been observed by means of high altitude balloons launched from the island. The trouble was, the riometer did not have enough resolution to isolate the individual pulses; what was needed was a faster riometer.

In late 1966 I returned to Macquarie Island with the world's first operational "fast-response riometer" designed using a state-of-the-art, low-noise radio receiver developed at the University of Tasmania's Physics Department. It had 50 times the resolution of a conventional riometer. This was my PhD project. Over the next three months I recorded the first observations of cosmic noise absorption pulsations. They had an unexpected sawtooth shape. Their slower decay was a measure of electron-ion recombination in the D-region. I returned home in March 1967 and published my discovery in a Letter to Nature. Later I was able to show that other auroral-zone phenomena, Pi1 micropulsations, were also asymmetrical and caused by the same mechanism.

This was all, of course, entirely useless; the purest of pure research. There were no practical reasons for studying the ionosphere following the advent of communication satellites. No-one today has the slightest interest in cosmic noise absorption pulsations nor in Pi1 micropulsations apart from a few researchers in closely related fields. This may change if there is a nuclear war and the issue of EMP (electromagnetic pulse) becomes important.

I learned a lot. I learned how important it is to have instruments that work properly and are properly calibrated. I learned that auroral physics is not about plugging a computer into the sky and pressing the ON button. Science is about interacting with the real world; you have to manipulate things in order to understand them, and you have to understand them in order to manipulate them. That is how understanding, the end-product of science, is built up. The idea that you can sit indoors at a computer terminal and determine how the real world *must* work is nonsense.

I remember what my supervisor, Prof G.R.A. (Bill) Ellis, told me before I went south with my gear. He had said:

When you first use an instrument with much higher resolution than before, not only do you see the same things bigger, you see new things.

This is what happened to Galileo. When he first looked at the

planet Jupiter in 1610, he not only saw a sphere on a larger scale, he saw four of Jupiter's moons as well. This was the first observation of the satellites of another planet. It strongly supported the new heliocentric theory and ultimately led to the development of modern astronomy and physics.

When I operated my fast riometer for the first time, I too saw new things; I saw sawtooth pulsations with a decay time related to the electron density 80 km above the earth.

That is what scientific research is about; it is about exploring the Universe, about broadening understanding, about seeing new things. It formalizes the natural human tendency to explore. It involves a sense of wonder. But we also have another, conflicting tendency which inhibits our understanding; we have a tendency to kowtow to authority and to suppress ideas which threaten that authority. We live in social hierarchies based on shared dogma. Science can be subverted by such dogma.

The famous story of King Canute commanding the waves is told by chronicler Henry of Huntingdon, who lived within 60 years of the death of Canute (1035 AD). When at the summit of his power, Canute ordered a seat to be placed for him on the sea-shore when the tide was coming in. Then, before a large group of his flattering courtiers, he spoke to the rising sea, saying,"Thou, too, art subject to my command, for the land on which I am seated is mine, and no one has ever resisted my commands with impunity. I command you, then, o waters, not to flow over my land, nor presume to wet the feet and the robe of your lord." The tide, however, continued to rise as usual, dashing over his feet and legs without respect to his royal person. Then the King leaped backwards, saying: "Let all men know how empty and worthless is the power of kings, for there is none worthy of the name, but He whom heaven, earth, and sea obey by eternal laws."

Canute was the most powerful monarch ever to rule England. He was, at once, King of England, Denmark, Norway and Scotland. He was a Viking, a Christian and an educated man. He performed an experiment by which he demonstrated to his superstitious courtiers the objective reality of natural laws and the limitations of human agency. To him the idea that a human being could override God's "eternal laws" was blasphemy. The belief in the existence of laws of nature is a consequence of monotheism. It led ultimately to the scientific revolution which accompanied the Renaissance.

For the first time in a millennium, scholars now question the ex-

istence of natural laws. Evidently Canute was wrong and science is merely a social construct. From now on, global temperature will be controlled politically, by decree. The Scientific Method is to be abandoned, it seems.

This book is intended as a robust refutation of this fashionable tendency. It is written in praise of the Scientific Method as a means of comprehending and manipulating the natural world.

In Part I, Popper's description of the Scientific Method is used to address and rectify a persistent myth: the myth of the continuum, the rationalist belief in the universal applicability of differential calculus to fluid processes.

In so doing, emphasis is placed on time series rather than continuous functions, and a methodology for dealing with time series is developed, *Discrete Time Spectral Analysis*; this is discussed in Part II. It allows rigorous methods of statistical inference to be applied to time series and to the time series of global average temperature in particular. Most of this mathematical development could be heavy going for the non-specialist and can be skipped; it is the implications which are important (page 85).

In Part III, we abandon mathematical rationalism and adopt an empirical approach, in order to speculate about a variety of natural processes such as the recurrence of Ice Age Terminations, the growth of wind-seas, the effect of subaqueous volcanism on ocean circulation and the dynamics of the Earth's interior.

If a movie were to be made of the workings of the Universe and then shown backwards, the only Law of Physics that would not be obeyed is the Second Law of Thermodynamics. It is the only Law in which the direction of time is important. In the reversed movie, the motions of the planets in their orbits would look much the same. Planetary motions are determined solely by Newton's differential equations and do not involve the Second Law. On the other hand, a breaking wave would look very wrong indeed when viewed backwards in time. Waves never "unbreak". Breaking waves involve the Second Law and cannot be adequately described by differential equations. That is what this book is about.

Part I

The Myth of the Continuum

Chapter 1

The Scientific Method

The Scientific Method

The fundamental things we know about the physical world are either hard-wired into our brains or we found them out by experience. Watch toddlers playing with kitchen utensils. They manipulate objects, they experiment. They find out by experience that big things don't fit inside small things. As we grow up, the results of those early experiments are automatically understood but the actual experiments have been forgotten. The same applies to all learned behaviour about cause and effect, about the nature of flowing water and the danger of fire. As we go through life, we continue to find out about the world around us by experiencing it, by living in it, by manipulating it, by carrying out millions of informal "experiments", few of which we remember except for one or two epiphanies.

However, as humans we also have the gift of language so that our knowledge of the world is far broader than our personal experience. The vast majority of things people know about the world are learned from other people. Humans have a strong desire to learn through communication: firstly from our parents, then from our teachers, then from our playground friends and peers. In addition, we learn from various media: books, television, social media, and so on. Nevertheless, "facts" are often communicated and remembered as summary principles, tendencies, or trends rather than as observed details, which would be far too numerous to remember.

But there is a problem. It is a very big problem.

Beliefs about the world, the most general and important ones, are also political banners which both unite and divide people. It is human nature to bond with those who share a belief and to reject those who do not as alien, as foreign, as other. At best we regard people whose beliefs conflict with our own as unsound, eccentric or downright crazy. If we are told, by those in authority, that such-and-such is the truth, then to publicly air doubts can be seen as a mad or traitorous act. No doubt this socially unifying tendency of belief once served an evolutionary purpose in uniting believers against a common enemy – "survival of the loyal". At this level, a belief becomes a religion or an ideology.

Such unifying beliefs are usually concerned with politics and value judgements: "Henry Tudor is the rightful King of England", "Communism is evil", "There should be an equal number of women and men on company boards" and so on. However that is not always the case. For example, the belief that the Earth is the centre of the Universe is value free, but in 1633 Galileo was forced to recant under threat of torture when he proposed otherwise. At that time the Christian Church in Europe was the absolute authority on the legitimacy of a belief; validity or otherwise was determined solely by the authority of the Church which, in turn, regarded the word of the Bible as the ultimate arbiter, subject to its own interpretation of course.

Francis Bacon, a contemporary of Shakespeare, was the foremost exponent of the Scientific Method of the early modern era. A great legacy of Bacon was the description, in his *Novum Organum* (1620), of "Idols of the Mind": beliefs which commonly obstruct the path to correct scientific reasoning. These are:

1. Idols of the Tribe: *The human understanding is of its own nature prone to suppose the existence of more order and regularity in the world than it finds ... The human understanding when it has once adopted an opinion draws all things else to support and agree with it.*

2. Idols of the Cave: *The Idols of the Cave are the idols of the individual man. ... men look for sciences in their own lesser worlds, and not in the greater or common world.*

3. Idols of the Marketplace: *names of things which do not exist and names of things which exist, but yet confused and ill-defined, and hastily and irregularly derived from realities.*

4. Idols of the Theatre: *... in the plays of this philosophical theatre you may observe the same thing which is found in the theatre of the poets, that stories invented for the stage are more compact and elegant, and more as one would wish them to be, than true stories out of history. Idols which have immigrated into men's minds from the various dogmas of philosophies.*

In modern language these are equivalent to:

1. *To a man with a hammer, every problem is a nail.*
2. *I've made up my mind. Don't confuse me with facts.*
3. *How many devils can sit on the head of a pin?*
4. *Never spoil a good story for the sake of the truth.*

In the ensuing centuries, natural philosophers – scientists – have striven to purge science of these conceptual errors. Some are still there, embedded in the fabric of sciences such as Fluid Dynamics and the environmental sciences.

In fluid dynamics, an Idol of the Tribe is the widespread belief in the universal applicability of differential calculus in describing the behaviour of fluids. According to fluid dynamicists, every fluid is a *continuum* and so is continuous and differentiable almost everywhere. This despite the overwhelming evidence in support of the atomic theory from nuclear physics, and in support of the granularity of action space from Quantum Mechanics.

The mystique of Chaos Theory now dominates some fields despite this granularity. Mathematical Chaos is a an Idol of the Marketplace whereby a thing may exist (chaos) but is "yet confused and ill-defined and hastily and irregularly derived from realities". Mathematical Chaos bears no resemblance to reality nor to the $\chi\alpha o\zeta$ of the Greeks. It is a fashionable buzzword intended to mislead us into believing that the pathological behaviour of differential equations somehow provides a profound insight into the nature of the physical world. By its very existence, Chaos Theory misdirects attention from stochastic methods which provide a more fruitful approach.

By the 16th century, the invention of the printing press meant that ideas could be disseminated widely and rapidly. This was the Church's problem with Martin Luther. A century later, other inventions such as the telescope enabled people to discover the nature of reality for

themselves, and what they saw often contradicted Church authority. This was the Church's problem with Galileo. Furthermore, Europe was in a state of religious ferment as Christianity fragmented into various splinter groups.

A handful of thinkers in England found a way out of this mess. It was to draw a strong distinction between science and religion and to separate religion from the secular aspects of human inquiry. In 1663, the "The Royal Society of London for Improving Natural Knowledge" was founded. Its motto, "Nullius in verba", broadly translates as "take nobody's word for it". Thinkers like Newton and Priestley were often intensely religious people, but they were careful to keep their religion separate from their scientific work.[1]

This separation of modes of thought is an aspect of civilization. The great sagas of the past, such as the Iliad, pre-dated writing and were part of an oral tradition. They were written as poetry to be more easily remembered. As such they were a mixture of poetry, religion, history and technology. Parts of the Finnish saga, the Kalevala, appear to have been a handbook for finding and refining bog iron. Another confluence of myth, poetry and technology is perhaps the legend of King Arthur and the sword Excalibur. It may well have been an Iron Age myth about making weapons from iron meteorites. "He drew the sword out of the stone"; red hot iron is "drawn" by beating it with a hammer because the melting point of iron is too high for primitive furnaces.

How can we know about the world?

We can continue to experiment and to see for ourselves as we did as children, but it is a very impractical solution and beyond the reach of even the best-resourced individual. The recent observation of gravitational waves by LIGO cost one third of a billion dollars and involved hundreds of scientists and technicians.

We are inevitably forced to accept the beliefs of other individuals or groups of people who have done the experiments themselves. We have to take their word for it. But how are we to know that these beliefs are based, ultimately, on experiment and observation and are not the expression of some tribally unifying ideology to which humans are so prone?

[1] Priestley was first to isolate an interesting gas which he called "dephlogisticated air". It was later renamed "Oxygen" by Lavoisier. Priestley emigrated to America, was a personal friend of Jefferson and, like him, passionately believed in the separation of church and state.

We do not know. We have to trust them. There is no other way.

The people we must trust are the scientists. It is their task to draw conclusions from observations uninfluenced by their personal ideologies in the tradition of Newton and Priestley and the 17th century Royal Society. The stock-in-trade of scientists is *understanding*. Understanding is their product. As with music, once a new idea has been seen or heard or understood, it is impossible to go back and un-hear it or un-understand it. This makes it difficult to adequately reward creative scientists and for them to protect their work. Musicians have copyright; scientists have peer-reviewed papers. As with music, much that is published may turn out to be dross, but that which is not can last indefinitely. As a civilized society we pride ourselves on our symphony orchestras. Our scientific institutions should have a similar role.

Our trust need not be absolute. We can, to some extent, assess for ourselves whether a particular piece of research is cutting-edge or fairly ordinary or even suspect, but in order to do so, we too must put ideology aside. We too must be conscientious and do our homework. We can judge to some extent by the way the expert talks about what they have done. It is similar to talking to one's doctor or motor mechanic: we may not understand the detail, but we can judge the validity of their insights by their general demeanour, by whether the things they say measure up to what we do already know, and, ultimately, by whether they actually work. If we find out that our G.P. is a closet naturopath, we may start to have doubts about what he/she is telling us. Likewise, if we discover that a scientist is a passionate Green, we may have doubts about his/her prognoses on species numbers or climate change unless he/she has made some obvious effort to guard against his/her ideological predilections.

There are two major streams in European thought: Rationalism and Empiricism. In the case of the former, the power of human reason is assumed to be so great that, starting with a few general principles that are obvious to everyone, it is possible to sit at a desk in a closed room and by reason alone deduce the nature of the Universe and all its workings. The Ancient Greeks were Rationalists who believed that their geometry was an exploration of the properties of space. The great modern Rationalist, Immanuel Kant, talked about the *synthetic a priori* postulate, a statement about the real world that we somehow just *know* to be true. Examples are cause and effect (i.e. *every effect must have a cause which precedes it*) and the axioms of geometry.

Empiricists on the other hand believe that all knowledge is based on experience derived from the senses. In order to understand the world, we must observe it, and observation is paramount. It is a fundamental part of the scientific method that all hypotheses and theories must, ultimately, be tested against observation.

Once again, the big breakthrough came in the 17th and 18th centuries with the Empiricist philosophers, Locke, Berkeley and Hume. With it came the Enlightenment and the rise of the Scientific Method put into practice by early scientists such as Newton, Halley, Galileo, Hooke and Boyle. These people did not completely abandon Rationalism but tempered it with empirical observation. Many people think that physicists such as Newton and Einstein produced their great unifying theories out of thin air without recourse to observation, but that is not the case. Newton experimented extensively with pendulums and Einstein recognized the need for observation in verifying his theories of relativity. For Einstein, time *is* that thing which is measured with clocks, and space *is* that thing which is measured with measuring rods.

When I was a student, a post-apocalyptic science fiction novel was very popular. It was *Earth Abides* by George R. Stewart. As I recall, the protagonist helps his people as they descend into savagery by teaching them how to make and use the bow and arrow, so saving them thousands of years of technical evolution. The idea has stayed with me. In a similar situation I would attempt to teach them the Scientific Method for similar reasons.

This interaction between ideas and observations is complex. Young research scientists spend many years learning the skills of their trade. There is no simple formula, no button to press: you have to learn how to do it under the tutelage of skilled practitioners, much like a musician.

The Scientific Method set out by Bacon in the early 17th century was further refined by Newton and others and set out by Popper (1962) in the form of his Seven Principles:

1. *It is easy to obtain confirmations, or verifications, for nearly every theory if we look for confirmations.*

2. *Confirmations should count only if they are the result of risky predictions; that is to say, if, unenlightened by the theory in question, we should have expected an event which was incom-*

patible with the theory, an event which would have refuted the theory.

3. *Every "good" scientific theory is a prohibition: it forbids certain things to happen. The more a theory forbids, the better it is.*

4. *A theory which is not refutable by any conceivable event is non-scientific. Irrefutability is not a virtue of a theory (as people often think) but a vice.*

5. *Every genuine test of a theory is an attempt to falsify it, or to refute it. Testability is falsifiability; but there are degrees of testability: some theories are more testable, more exposed to refutation, than others; they take, as it were, greater risks.*

6. *Confirming evidence should not count except when it is the result of a genuine test of the theory; and this means that it can be presented as a serious but unsuccessful attempt to falsify the theory.*

7. *Some genuinely testable theories, when found to be false, are still upheld by their admirers for example by introducing ad hoc some auxiliary assumption, or by reinterpreting the theory ad hoc in such a way that it escapes refutation. Such a procedure is always possible, but it rescues the theory from refutation only at the price of destroying, or at least lowering, its scientific status.*

These principles are descriptive not prescriptive and pre-date Popper by three centuries. Popper was not stating *how* Science *ought* to be done but how it actually *is* done. Most working scientists would recognize these ideas and would support their application in their own research field. Popper describes how Science frees itself from Bacon's Idols.

The remarkable advances in science and technology witnessed in the modern era are largely the result of the meticulous application of the Scientific Method. When a theory is tested against observation and fails the test, new insights into the underlying reality are gained, whereas clinging tenaciously to a "correct" theory can only lead to a sterile absolutism. This is the fundamental difference between Physics and Applied Mathematics. It is also the fundamental difference between science and superstition.

In recent times some aspects of Physics itself appear to have abandoned the Scientific Method as described by Popper. An example is superstring theory of Theoretical Physics:

> The possible existence of, say, 10^{500} consistent different vacuum states for superstring theory probably destroys the hope of using the theory to predict anything. If one picks among this large set just those states whose properties agree with present experimental observations, it is likely there still will be such a large number of these that one can get just about whatever value one wants for the results of any new observation.[2]

According to Popper then, superstring theory, the most advanced (and glamorous) form of contemporary theoretical physics, is not even science. It is incapable of making risky predictions about the real world and so is evidently a form of pure mathematics, and not science at all.

There is another so-called "science" which has abandoned the empirical and lost touch with reality. That is Fluid Dynamics or at least those aspects of Fluid Dynamics which have ignored observation in favour of a mathematical ideal.

[2]Woit (2006) p122.

Chapter 2

Numbers and Entropy

Numbers

> God made the natural numbers. Everything else is Man's handiwork.[1]
>
> *Leopold Kronecker*

> All the mathematical sciences are founded on relations between physical laws and laws of numbers, so that the aim of exact science is to reduce the problems of nature to the determination of quantities by operations with numbers.
>
> *James C. Maxwell*

> I believe that mathematical reality lies outside us, that our function is to discover and observe it, and that theorems which we prove, and which we describe grandiloquently as our "creations" are simply the notes of our observations.
>
> *G.H.Hardy*

There are different sorts of numbers. Sometimes numbers are just codes, like telephone numbers, but mostly numbers represent quantities, i.e. things in the real world such as the number of potatoes in a shopping bag or the radius of the earth. Some numbers, like π and e,

[1] Die ganzen Zahlen hat der liebe Gott gemacht; alles andere ist Menschenwerk.

come out of mathematics itself as if the mathematical world has its own objective reality as Hardy's quote above suggests.

In fact numbers have evolved. It started with the natural numbers. The natural number "5" is the property of any set of objects that can be put into one to one correspondence with the fingers of my right hand. Add a new member to this set and it then has the number property "6" which is called the "successor" of "5". Every natural number has a successor, but there is a natural number that is not a successor to anything. It is called "1".

Later on the number zero was invented. Zero started out as a place-holder when the Arabic system replaced the Roman numerals. Arithmetic operations such as multiplication became a lot easier. Then came negative numbers, which were a boon to accountants. Having negative money meant that you owed money. The negative numbers, zero and the natural numbers are called "integers".

Then came fractions. You have 2 acres of land to split equally between 3 sons so they each get 2/3 of an acre. These are called rational numbers. A rational number is an "ordered pair" of integers, (p, q), which is usually written as p/q or $p \div q$ or $N\frac{p}{q}$ where N is also an integer. There are rules for manipulating rational numbers which we called "doing fractions" in primary school.

Then around 300 BCE came the first scientific catastrophe which we will call "The Rational Number Catastrophe", commonly known as "Euclid's proof that the square root of two is not a rational number".

The proof is easy to follow, even for a non-mathematician. It works by showing that the assumption that there is a rational number (i.e. a fraction) whose square is 2 leads to a contradiction. It goes as follows: Suppose

$$\frac{p}{q} = \sqrt{2} \tag{2.1}$$

where p and q are integers with no common factor, i.e. p/q is in its simplest form. Then

$$\frac{p^2}{q^2} = 2 \tag{2.2}$$

Therefore

$$p^2 = 2q^2 \tag{2.3}$$

so that p^2 is an even number. If p^2 is an even number, then p must

also be an even number, i.e.

$$p = 2a \tag{2.4}$$

where a is an integer. Then

$$4a^2 = p^2 = 2q^2 \tag{2.5}$$

and so

$$2a^2 = q^2 \tag{2.6}$$

Therefore q^2 is even which implies that q itself must also be even. Therefore both p and q are even numbers which contradicts our original assumption.

More than two millennia later, the square-root-of-two-problem is still with us. $\sqrt{2}$ was termed "irrational". Other numbers such as π and e were even worse; they are not only irrational, they are "transcendental".

From the point of view of mathematics, the problem of irrational numbers was solved in the late 19th century by Dedekind, Wieirstrasse and others who devised the "real numbers". This was done using the concept of limits. Think of the set of all rational numbers which are less than $\sqrt{2}$. The upper limit of this set is called the "supremum". Now think of the set of all rational numbers which are greater than $\sqrt{2}$. The lower limit of this set is called the "infimum". It can be shown that the supremum and infimum are the same number. That number is the definition of the real number, $\sqrt{2}$. Real numbers make mathematics self-consistent.

From a physics point of view, however, this is fairyland. The mathematicians have defined a set of numbers which scientists are unable to use to perform calculations. Real numbers are not computable. Real numbers are useless for dealing with the real world. All the calculations done by scientists involve rational numbers not real numbers. Computers use rational numbers like p/q, where $q = 2^{64}$.

Real numbers are not real. What happened was that mathematics became more and more refined and elegant until it ceased to be useful. As Einstein once said, "Elegance is for tailors". Belief in the relevance of *real* numbers to the natural world is an Idol of the Tribe.

Entropy

Contrary to popular belief, James Watt did not invent the steam engine. The Newcommen steam engine had been around for seventy years. It was used for pumping water out of mines to prevent them flooding. The Newcommen engine used steam to force a piston along a cylinder. The hard part was getting the piston back to its original position. This had involved cooling the piston itself with water before reheating it again for the "power stroke". As a result more heat was used to reheat the cylinder than was converted into mechanical energy. Watt's invention allowed the steam to be cooled outside the cylinder in a separate apparatus called a *condenser*. The condenser sucked the steam out of the cylinder leaving the cylinder hot and ready for the next stroke. There was a huge increase in efficiency of steam engines; much less fuel was required to pump the same amount of water. Imagine a Cornish tin miner having to import from Wales, all the coal to drive his pumps. To him, a more efficient engine meant lower costs.

Following this breakthrough there were numerous incremental improvements, such as converting reciprocal motion to rotary motion to drive lathes, presses and the like. Steam engines enabled mass-production and so revolutionized manufacturing. Stevenson's locomotive in 1829 revolutionized transport as well.

Fuel costs money, and enormous effort went into attempts to improve efficiency. Clearly heat was a form of energy (the First Law of Thermodynamics). People asked why couldn't *all* the available heat be converted to mechanical energy?

In 1850, the German physicist Rudolf Clausius published a paper in which he proposed the Second Law of Thermodynamics:

It is impossible to construct a device which operates in a cycle and produces no other effect than the transfer of heat from a cooler body to a hotter body.

which is equivalent to:

It is impossible to construct a device which operates in a cycle and produces no other effect than the transfer of heat from a single body in order to produce work.

At this time an important thought experiment was developed to help gain an understanding of the implications of the Second Law. It is called the *Carnot Cycle* and comprises a piston in a cylinder oper-

ating between two heat reservoirs, rather like a steam engine, except that the cylinder contains an ideal gas rather than steam. The heat reservoirs have absolute temperatures T_H and T_C and heat is passed between the cylinder and the reservoirs in a four phase cycle. It is done very slowly or "reversibly" so that the gas is always in equilibrium and friction can be ignored. It turns out that the efficiency, η, is given by

$$\eta = 1 - \frac{T_C}{T_H} \tag{2.7}$$

i.e. efficiency depends solely on the ratio of the absolute temperatures of the reservoirs and it is always less than one. Furthermore it can be shown that this is the best case. There is no other heat engine that can convert heat to work more efficiently than the Carnot Cycle. It is the perfect heat engine.

The work, W, done by the Carnot Cycle is succinctly described by the equation:

$$W = (T_H - T_C)(S_B - S_A) \tag{2.8}$$

where S_B and S_A are the initial and final *entropy* of the two reservoirs. The change in entropy of a reservoir, ΔS, is defined by

$$\Delta S = \int_A^B \frac{dQ}{T} \tag{2.9}$$

where a quantity of heat, dQ, is transferred to the reservoir at absolute temperature, T.

This equation (2.9) is the thermodynamic definition of entropy or, at least, of entropy change. It is a macroscopic, observable quantity which is measurable using thermometers, calorimeters and the like.

But what does it mean at a molecular level? Heat is the total kinetic energy of molecules in a gas, temperature is related to the average kinetic energy of the gas molecules while pressure is the sum of forces per unit area when molecules collide with the boundaries of the container.

But what is entropy?

There have been a number of definitions, starting with Boltzmann's famous:

$$S_B = k_B \ln(W) \tag{2.10}$$

where S_B is the entropy and k_B is Boltzmann's constant and ln() is the natural logarithm.

Boltzmann's paradigm was an ideal gas of N identical particles, of which N_1, N_2, etc. are the numbers of particles in various microscopic conditions of position and momentum. Using the formula for permutations

$$W = \frac{N!}{N_1! \times N_2! \times \cdots} \tag{2.11}$$

where ! denotes factorial,[2] W is the number of microstates associated with a particular observed macrostate. Although more difficult to understand than Einstein's famous $E = mc^2$, Boltzmann's entropy formula is equally profound. It is engraved on his tombstone in Vienna. It is the first formula in physics to relate a measurable, supposedly continuous, physical quantity to probability.

An easy way to understand macrostates and microstates is to consider the "entropy" of a deck of cards.

State 1

Suppose the top 26 cards are all red and the bottom 26 cards are all black. These represent two microstates, M1 and M2, say. The number of possible permutations of both M1 and M2 are $26 \times 25 \times ... \times 2 \times 1$, i.e. factorial 26 (written 26!). Likewise the number of permutations of the whole deck is 52! so that (2.11) becomes

$$W = \frac{52!}{26! \times 26!} \tag{2.12}$$

for this particular case. Substituting in (2.10) and assuming $k_B = 1$ for cards gives $S_B = 0.042$.

State 2

Now suppose the top 13 cards are hearts and the bottom 39 are the other three suits so that W becomes

$$W = \frac{52!}{13! \times 39!} \tag{2.13}$$

Now $S_B = 0.065$. The entropy is greater because this is a less orderly arrangement of the cards.

A transition from State 1 to State 2 results in an entropy increase of $\Delta S_B = .065 - .042 = .023$.

Suppose we use two decks of cards so that W for State 1 is given by

$$W = \frac{104!}{52! \times 52!} \tag{2.14}$$

[2] e.g. 6!=6 X 5 X 4 X 3 X 2 X 1 = 720

and for State 2 by

$$W = \frac{104!}{26! \times 78!} \tag{2.15}$$

This time the entropy change is $\Delta S_B = .024 - .016 = .008$. i.e. less than the value of .023 in the single deck case.

This example indicates that entropy change depends on the number of entities being shuffled – it depends on the *granularity* of the system under investigation. It depends on the total number of grains in the system. It does not work for a continuum.

There are other statistical definitions of entropy, such as those of Gibbs and (in a different context) Shannon:[3]

$$S_B = -k_B \sum_i p_i \ln(p_i) \tag{2.16}$$

If we made a movie of the Universe and showed it backwards, all the laws of physics would still be true except for the Second Law of Thermodynamics. This is the only Law in which time has a direction. It says something about the nature of time itself; time cannot go backwards, entropy can only increase with time.

The three entropy equations, (2.8), (2.9) and (2.10), indicate that thermodynamic systems are coarse-grained. They are coarse grained because of quantum physics. Not only is matter itself coarse-grained according to the atomic theory, but dynamical systems for which a Hamiltonian exists must also be coarse-grained or quantized in "action space"[4].

The Second Law of Thermodynamics is perhaps the most profound of all the Laws of Physics; it arose from a desire to make mining more profitable. The Second Law leads to the concept of entropy. Entropy is a measure of how energy is ordered in a stochastic, granular system. It has no meaning in a deterministic continuous system. The idea that deterministic equations relating continuous physical quantities can provide a comprehensive picture of physical reality is an Idol of the Tribe. It is false because such a model cannot account for the Second Law of Thermodynamics.

[3]The two definitions, (2.10) and (2.16), are not as different as they may first appear because of an approximation for the logarithm of a factorial known as Stirling's Theorem.

[4]*Action* is a physical quantity which has the units of energy×time or momentum×length.

Chapter 3

Two Catastrophes

At first pass the world appears deterministic. Every event has a cause which is itself an event which must also have a cause, and so on. So go the first and second arguments for the existence of God (by St Thomas Aquinas, 1225–1274 AD). However, the more deeply physicists looks into the natural world, the more it appears to be random. At a subatomic level, there are the random events of radioactive decay and of wave and particle interactions. At human scales, phenomena such wave breaking, vortex shedding and fluid turbulence have a random or "stochastic" character.[1]

In the introduction to his 1814 *Essai philosophique sur les probabilités*, Pierre-Simon Laplace had this to say:

> We may regard the present state of the universe as the effect of its past and the cause of its future. An intellect which at a certain moment would know all forces that set nature in motion, and all positions of all items of which nature is composed, if this intellect were also vast enough to submit these data to analysis, it would embrace in a single formula the movements of the greatest bodies of the universe and those of the tiniest atom; for such an intellect nothing would be uncertain and the future just like the past would be present before its eyes.[2]

[1] *Random* and *stochastic* mean the same thing

[2] *Nous devons donc envisager l'état présent de l'universe comme l'effet de son état antérieur, et comme la cause de celui qui va suivre. Une intelligence qui pour un instant donné connaîtrait toutes les forces dont la nature est animée et la situation respective des êtres qui la composent, si d'ailleurs elle était assez vaste pour soumettre ces données à l'analyse, embrasserait dans la même formule les mouvements des plus grands corps de l'universe et ceux du plus léger atome;*

This is a statement of absolute determinism. It is not something to be learned about the universe by observation, and it can never be disproved. It is not a scientific theory. It can perhaps best be regarded as a secular substitute for an omniscient God with perfect foreknowledge. To record the position of even a single particle would require knowledge of a quantity with an infinite number of decimal places, a *real number*. Computations with such numbers would take an infinite amount of computer time. Instead we have to use rational numbers and live with the imprecision brought about by round-off errors. Despite these difficulties there are many who work in the physical sciences who still, deep down, agree with Laplace, particularly those who deal with systems of partial differential equations which are, by their nature, deterministic. As Einstein supposedly said: *I, at any rate, am convinced that [God] does not throw dice.*[3] It is hard to comprehend how someone who won a Nobel Prize for his contribution to Quantum Mechanics could have made such a statement. Perhaps it has been taken out of context or otherwise misunderstood.

The calculus was devised by Newton and Liebniz to deal with the motions of solid bodies – planets, satellites and artillery shells – at which it proved to be brilliantly successful. Halley's accurate prediction of the return of the eponymous comet was an early example.

Stochastic variables are often expressed as functions of past states rather than as functions of time. They include an additional, random component. As a consequence, when we use a stochastic model to predict further and further into the future, the predictions become less and less reliable because randomness accumulates from one predicted state to the next. This leads to the idea of a *prediction horizon* described by Koutsoyiannis (2010). For example, meteorological models that predict the weather have a prediction horizon of about a week.

The mid-19th century was the heyday of determinism in the application of Newton's differential calculus to the physical world. Here we examine two cases where this Newtonian determinism broke down: Maxwell's electromagnetic field equations and the Navier–Stokes equations of fluid dynamics. These sets of partial differential equations are written out in full in Appendix 16. In the former case, the purely deterministic description had clearly failed by the turn of the century whereas in the second case, this inadequacy was glossed over, and a

rien ne serait incertain pour elle, et l'avenir comme le passé serait présent a ses yeux.

[3] *Jedenfalls bin ich überzeugt, daßder nicht würfelt.*

deterministic description is still in use today.

The Ultraviolet Catastrophe

At the beginning of the 20th century, two revolutions occurred in physics. One was Relativity, the other, Quantum Mechanics. Without doubt it was the second which has had the greater impact, both on technology and on the way we think about the natural world.

By the mid-19th Century, Maxwell had summarized Faraday's electrical experiments with two elegant equations, (A.6) and (A.7), which described electromagnetism. Maxwell's Equations are partial differential equations and as such are deterministic.

At the time it was thought that these equations described the behaviour of an underlying *aether*. Their form led to the possibility of the existence of waves which would be carried by the aether in much the same way that solids, liquids and gases carry sound waves. The measured properties of the aether, its permittivity and permeability, were then used to determine the aether wave velocity in much the same way that density and elasticity are used to compute the velocity of sound. When this was done, the aether wave velocity turned out to be very close to the measured velocity of light. Light was clearly electromagnetic.

The whole Universe was considered to be bathed in aether, and attempts were made to find out how fast the Earth was moving through this aether pond. Michelson and Morley found that wherever the Earth was located in its orbit around the Sun, its velocity through the aether was always zero. This experiment was the basis of Relativity. The Relativity theory could well be called "the non-existence of aether theory". The theory of aether was replaced with a theory of fields. Maxwell's equations remained formally the same and became known as the electromagnetic field equations.

Now came new questions: why does a hot object emit light, and why does the spectrum[4] of this light have a particular shape depending on the temperature? In simple terms, why does an object such as an iron bar change colour as its temperature increases? The problem was attacked by assuming the hot object is a cavity containing electromagnetic waves. This was called a "black body" because colour

[4]The concept of *spectrum* is dealt with more fully in Part II.

from reflected light is ignored. The electromagnetic waves bounce around inside the cavity interacting with molecules at random and bouncing off the walls. Think of sound waves bouncing back and forth in a woodwind instrument – a fundamental frequency is generated, together with numerous harmonics which give the instrument its particular timbre. In a similar manner, electromagnetic waves are assumed to bounce around inside the black body. Instead of particular harmonics or modes being determined by the shape of the cavity, the total energy is partitioned *equally* among all the modes. This is because energy is transferred from mode to mode internally by wave-molecule interactions rather than reflections from the walls. These assumptions lead to the Rayleigh–Jeans formula for B_λ, the spectral radiance of black-body radiation:

$$B_\lambda(T) = \frac{2ck_BT}{\lambda^4}$$

where c is the velocity of light, T is the absolute temperature, λ is the wavelength and k_B is Boltzmann's constant.

The formula is a direct consequence of Maxwell's Equations. It works very well at long wavelengths, i.e. when λ is large. However it breaks down at short wavelengths. It is obviously wrong because it tends to infinity as λ tends to zero. There would have to be an infinite amount of energy radiated at zero wavelength! In fact, it already goes badly wrong even in the ultra-violet part of the spectrum.

This is *the ultraviolet catastrophe*: Maxwells's equations cannot predict the distribution of radiated black-body energy with wavelength in the ultraviolet and beyond.

In November 1900, German physicist Max Planck postulated that electromagnetic energy could be emitted only in quantized form, in other words, the energy could only be a multiple of an elementary unit:

$$E = h\nu = hc/\lambda$$

where h is Planck's constant and ν is the frequency of the radiation. Physicists now call these quanta photons, and a photon of frequency ν will have its own specific and unique energy. The total energy at that frequency is then equal to $h\nu$ multiplied by the number of photons at that frequency. This restricts the amount of energy distributed to shorter wavelengths and leads to Planck's Law:

$$B_\lambda(T) = \frac{2ck_BT}{\lambda^4}\frac{1}{e^{hc/(\lambda k_BT)} - 1}$$

Planck's Law describes black-body radiation very well indeed.

Later Einstein and Bose postulated that Planck's quanta were real physical particles that we now call photons, not just a mathematical fiction. They modified statistical mechanics in the style of Boltzmann to an ensemble of photons. Einstein's photon had an energy proportional to its frequency, and this also explained the photoelectric effect. For this work Einstein was awarded the Nobel Prize in Physics in 1921.

The unsung hero in all of this was Boltzmann. Planck used Boltzmann's statistical interpretation of the second law of thermodynamics as a way of gaining a more fundamental understanding of the principles behind his radiation law. Thus Maxwell's deterministic equations break down at high frequencies and are replaced by Planck's Law which is based on Boltzmann's probabilistic world-view.

Boltzmann's kinetic theory of gases presupposed the reality of atoms and molecules, but almost all German philosophers and many scientists disbelieved their existence. Furthermore, many defenders of "pure thermodynamics" were trying hard to refute the kinetic theory of gases and statistical mechanics because of Boltzmann's assumptions about atoms and molecules and especially his statistical interpretation of the second law of thermodynamics. He had a long-running dispute with the editor of the preeminent German physics journal of his day, who refused to let Boltzmann refer to atoms and molecules as anything other than convenient theoretical constructs. Only a couple of years after Boltzmann's death, Einstein's theoretical description of molecular diffusion and subsequent experiments confirmed the values of Avogadro's number and Boltzmann's constant, and convinced the world that atoms really do exist. There are few, if any, scientists today who would doubt this.

The Fluid Dynamic Catastrophe

Today the physics of the very large (Cosmology) and of the very small (Particle Physics) are well established. Strangely the physics of the mesoscale covering the interval from laboratory to planetary scales is less well understood. This is the scale in which fluids predominate. Fluid dynamics is the field of applied mathematicians who harbour a deep suspicion of statistical inference and the experimental method. They proceed according to a sort of mathematical rationalism, an

axiomatic approach whereby the fundamentals are given, *a priori*, and all that is required is mathematical rigour in dealing with the partial differential equations describing "ideal" fluids.

In the mid 19th century, the equations embodying Newton's laws of mechanics were adapted to deal with fluids as well as solid objects. Out of this came the Navier–Stokes equations (Navier, 1831, Stokes, 1845), which embody Newton's Laws and the conservation of mass in the form of differential equations. These equations are based on the assumption that the fluid, at the scale of interest, is a *continuum* – in other words, it is not made up of discrete particles but rather is a continuous substance. Another necessary assumption is that all the fields of interest like pressure, flow velocity, density, and temperature are differentiable, i.e. smooth at every scale.

These assumptions, although necessary for the derivation, have some unfortunate consequences, viz.:

1. The Navier–Stokes equations are in direct conflict with atomic theory and so cannot account for the Brownian motion,

2. The Navier–Stokes equations do not predict turbulence.

3. The Navier–Stokes equations do not allow entropy to increase.

The first people to come to grips with turbulence were Taylor (1935) and Kolmogorov (1941). In sharp contrast to the deterministic approach of Navier–Stokes, both Taylor and Kolmogorov treated the problem as a stochastic one, which, using dimensional arguments, led to Kolmogorov's spectrum of the turbulent energy in the *inertial range* of scales where viscosity is unimportant:

$$E(k) = C\epsilon^{\frac{2}{3}}k^{-\frac{5}{3}}$$

where C is a constant, ϵ is the rate of energy dissipation and $k(=1/\lambda)$ is the wavenumber.

Like Planck's formula for electromagnetic black-body radiation, Kolomogorov's spectrum is based on probabilistic assumptions. It appears then that both electromagnetism and fluid dynamics depend on two different theories: a deterministic theory, which applies only to low energy densities and a stochastic theory, which applies high energy densities.

Just as Maxwell's Equations work very well indeed at low frequencies for non-ionizing radiation, the Navier–Stokes equations work very

well indeed for low Reynolds Number, *streamline* flows in which there is no turbulence. It is only at high energy densities and shorter time scales that both systems of equations break down. In neither case do these deterministic equations make provision for increases in entropy as described by Boltzmann's equation, (2.10).

In 2000 the Clay Mathematics Institute of Cambridge, Mass., U.S., designated the Navier–Stokes equations a "Millennium Problem", one of seven mathematical problems selected for a special award worth $1 million. Unfortunately the Navier–Stokes Millennium Problem is couched in terminology that precludes anyone other than a mathematician from claiming the prize. The problem with these equations is not mathematical; it has more to do with the fact that they do not describe real fluids.

The Navier–Stokes equations may work well for streamline flow at laboratory scales, but they break down badly at environmental scales where sub-grid-scale turbulence becomes a major issue. In these circumstances they cannot be solved explicitly and we must resort to *finite difference methods*, i.e. the Navier–Stokes equations must be dragooned into iterative form in order to be useful. This is done using Taylor's Theorem, viz.: for each partial derivative, $y'(a) = \frac{\partial y}{\partial z}|_a$ in (A.10), (A.11) and (A.12), we use the following type of approximations for the various partial derivatives:

$$\frac{\partial y}{\partial z}\Big|_a \approx \frac{y(a+\Delta z) - y(a)}{\Delta z} \tag{3.1}$$

Various difference schemes are used, each with its own advantages and drawbacks, but all are based on the assumption that derivatives such as $\frac{\partial y}{\partial z}$ actually do exist, in other words, that the fluid under consideration is a *continuum*, i.e. that it is continuous and differentiable everywhere. By definition

$$\frac{dy}{dz} = \lim_{\Delta z \to 0} \frac{\Delta y}{\Delta z} \tag{3.2}$$

which requires that $\Delta y \to 0$ as $\Delta z \to 0$ in order for the ratio to remain finite. Given that fluids are granular rather than continuous and differentiable, this assumption is unwarranted.

It might be argued that this is nitpicking and that (3.1) is good enough for practical purposes. However, problems arise with finite difference models under numerical conditions which closely resemble

those giving rise to turbulent phenomena in the real world. Because of the deterministic character of both the Navier–Stokes equations and their finite difference approximations, it is not possible for the stochastic character of such turbulence to be properly represented. Instead, numerical instability ensues and the model "blows up", i.e. some variables increase in magnitude so rapidly that the computer word length is exceeded within a few time steps and an error condition is raised.

In order to avoid such instabilities, various fudges are introduced. These include:

1. arbitrarily smoothing the bathymetry or topography at the boundaries of the model so that there are no cliffs or spikes where vortex generation would occur in the real world, and

2. arbitrarily increasing frictional parameters such as the viscosity, η, in (A.9).

In order to satisfy #2, new parameters are introduced into (A.9) to remove mechanical energy from potential instabilities while making the model appear truly "physical". One such new parameter is the "eddy viscosity" which, in numerical models, is often set an order of magnitude higher than observed values. There is some justification for including eddy viscosity because it is related to the apparent viscous drag which comes about when turbulent eddies cross the boundary between neighbouring flows. In the real world, eddy viscosity is only important where eddies are found, i.e. where there is turbulence. In numerical models it is assigned a high value throughout the model in order to keep things stable.

In short, numerical models based on the Navier–Stokes equations only work if the tendency of moving fluids to become turbulent is artificially suppressed. As a result, deterministic numerical models based on these equations are not accurate emulations of physical reality. They rapidly part company with the system they purport to emulate. Because of excessively high viscosity values needed for stability, they are often over-damped and so underestimate the spatial and temporal variability of the real world.

As an example, look at the variances of sea surface temperature (SST) predicted by the HadCM3 model and based on the Navier–Stokes equations. These are shown in Figure 3.1. SST variances derived from the HadSST2 data set are shown Figure 3.2 for comparison. Over most of the ocean, the observed variances are noticeably

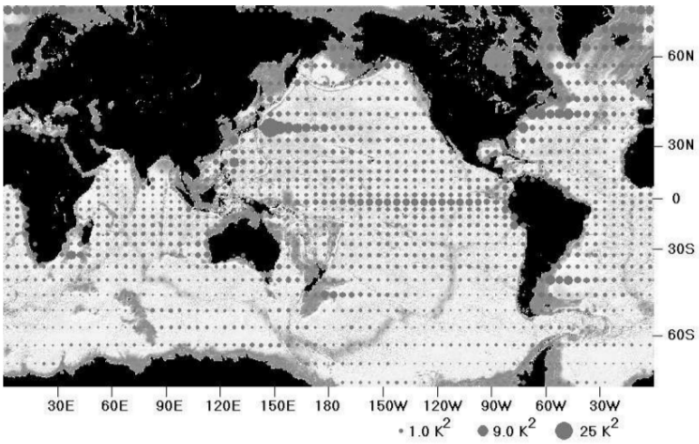


Figure 3.1: Sea Surface Temperature variances from the HadCM3 Climate Model.

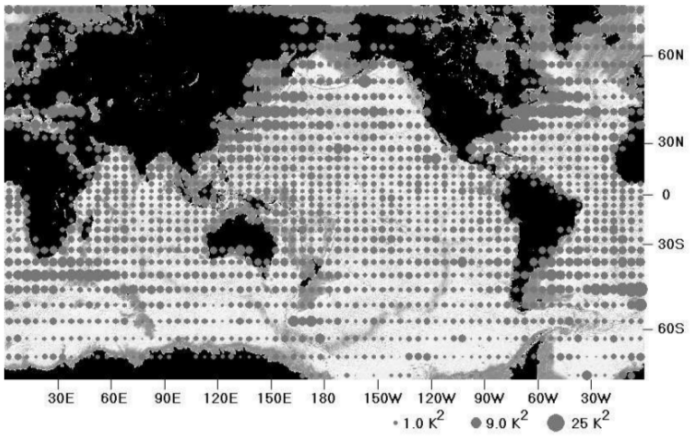


Figure 3.2: Sea Surface Temperature variances from the HadSST2 data set. (Variances along the Equatorial Pacific are artificially low for data processing reasons.)

greater than model output even to the naked eye. This was tested statistically by testing the ratio of observed and model variances using Fisher's F-test. The null hypothesis that model output only differed from the observations by chance can be rejected at a very high level of significance.

Furthermore, model output is different in character from the observations; it *looks* as if it has been generated by a machine. The big warm-core eddies which spin off the Cape of Good Hope and Cape Horn were not captured at all.

The continuum - an Idol of the Tribe

The problem with Maxwell's equations and the Navier–Stokes equations is that they are Newtonian – they are partial differential equations and, as such, deterministic and continuous. They do not allow for the granularity and stochastic behaviour of the systems they purport to describe. This is particularly true of the Navier–Stokes equations which do not even accommodate the atomic theory of Democritus but apply instead to some mythical "continuum" which has never been observed. Neither do they accommodate the granularity of action space, which is why they cannot deal with turbulence. A turbulent region of a fluid is a region which is less ordered than streamline flow, i.e. where the entropy density is higher than that of streamline flow. Turbulent phenomena in fluids such as Kolmogorov spectra and the Law of the Wall were not derived from the Navier–Stokes equations but from dimensional arguments and stochastic assumptions.

This is the *fluid dynamic catastrophe*. The Navier–Stokes equations cannot predict the distribution of turbulent energy with wave number. Newton's differential calculus does not work for fluids. Stochastic methods and iterative models are more appropriate for dealing with fluid phenomena.

The belief that every fluid is a continuum, describable by the Navier–Stokes equations, is an Idol of the Tribe. It is perhaps the most well-established, most widespread and most destructive of Bacon's Idols in contemporary science. It has held back fluid dynamics for more than a century. Almost all of the apparent gains in this field in its application to hydraulics, aeronautics and marine science have been the result of painstaking empirical observations involving wave tanks, wind tunnels and marine hydrography.

When the limitations of Laplacian Determinism are recognised and the mythical nature of the continuum acknowledged, it may become possible, once again, to develop meso-scale physics empirically. Clearly much of the real world is not continuous, differentiable and deterministic as the mathematicians would have us suppose. A real quantity is not a "real number", which is a mathematical ideal and unmanageable in practice. Rather the real world is granular and stochastic. It requires a calculus which accommodates this fact and emphasises stochastic time series over deterministic continuous functions. The spectral analysis of time series comprises such a calculus and is discussed in Part II. In Part III some hypotheses and speculations based on this approach are proposed. It is an approach freed from the strictures of Newtonian Rationalism and more compatible with the established methods of empirical science and statistical inference.

Part II which follows is intended for the specialist. The lay reader should skip to Implications (page 85).

Part II

The Spectral Analysis of Time Series

Chapter 4

Spectral Analysis

This Part is intended for the specialist. The lay reader should skip to Implications (page 85).

The term *spectrum* refers to the distribution of energy or power with frequency or wavelength. It began when Newton passed a beam of white light through a glass prism and saw how it split into its component colours. This happened because different colours have different velocities in glass and so were refracted at different angles. The rainbow effect produced was termed the spectrum after the Latin word for ghost. The splitting of light into its component parts became a very powerful technique, particularly in astronomy where the spectral character of an object is pretty much the only information available.

As every blacksmith or potter knows, the temperature of a hot object can be judged by its colour. As an object is heated it first becomes dull red, then bright red, then orange, then yellow then white, the colour of the Sun's surface. Apart from temperature, the spectrum can also reveal great detail about the chemical composition of the object, but that need not concern us here. The "Quantum Revolution" which overturned classical physics a century ago started with the consideration of the spectra of hot objects as discussed in the previous chapter.

Clearly the break-down of energy into components associated with different scales is of major interest in the analysis of a physical system because it provides insights into underlying physical processes. In the analogue world, various gadgets such as prisms, diffraction gratings and electronic circuitry are used to perform this task. In this digital age, numerical methods are appropriate.

Scientific measurements of real world quantities do not commonly come singly or in small groups but in very large data sets which may contain, quite literally, millions of values. This is particularly so since the advent of electronic storage methods. One particularly common form that data takes is the *time series*, which a sequence of measurements taken at equal intervals of time or averaged over equal intervals of time. An example would be the Weather Bureau's three hourly readings of the temperature at a particular location.

Drawing conclusions from time series is more complicated than one might expect. Most people like to think they have an intuitive grasp of time series, particularly with regard to the weather: "It was much hotter when I was a kid" and so on. Often these intuitions are quite wrong even when made by supposedly skilled observers. What is needed is a formal discipline for dealing with time series in the Popperian manner. Such a discipline evolved from signal processing theory and Econometrics.

The spectrum is an essential concept in dealing with time series. Rather than the power spectrum of the analogue world it is the variance spectrum which is used. Variance, the mean square of a set of numbers, is the digital analogue of power or energy. Hence it is the *variance spectrum* which must be calculated or estimated. Like the power spectrum, it can be a powerful tool for revealing underlying physical processes. The method is known as *spectral analysis*.

Spectral analysis co-evolved with the rapid development of digital computers in a number of different disciplines. As a consequence, some myths and misconceptions crept in and have become established as Baconian Idols – for example, the belief that the periodogram is not a consistent statistical estimator and must always be smoothed by "windowing" in order to be useful.

Here we develop a rigorous approach to the spectral analysis of time series, one which accommodates conventional methods of statistical inference.

Chapter 5

Statistical Inference

Popper's principles listed in Chapter 1 are well served by methods of statistical inference developed by Fisher (1934) and others in the early part of the 20th century. These methods formalize and quantify the idea of falsifiability by introducing the concept of the Null Hypothesis, that is, an hypothesis which has been set up to be deliberately falsified using the methods of mathematical statistics. This technique is known as hypothesis testing.

Here is a very simple example:

- **Experiment**: A man tosses a coin 4 times and gets Heads every time. Is there something strange going on, such as a biased coin being tossed?
- **Null Hypothesis**: There is nothing strange going on; it is just due to chance
- **Assumed distribution**: Probability of heads=0.5, probability of tails=0.5.
- **Calculation**: Probability of outcome (4 successive heads) under Null Hypothesis ("It's just chance") $= 0.5^4 = 1/16 = .0625 = 6.25\%$.
- **Conclusion**: There is 1 chance in 16 or a probability of six percent that this could happen by chance so perhaps it did. We cannot reject the Null Hypothesis.

Now consider a second example involving ten tosses rather than four:

- **Experiment**: A man tosses a coin 10 times and gets Heads every time. Is there something strange going on, such as a biased coin being tossed?
- **Null Hypothesis**: There is nothing strange going on; it is just due to chance
- **Assumed distribution**: Probability of heads=0.5, probability of tails=0.5.
- **Calculation**: Probability of outcome (10 successive heads) under the Null Hypothesis $= 0.5^{10} = 1/1024 = .00975$ which is less than one percent.
- **Conclusion**: There is less than 1 chance in a 100 that this could happen by chance so it looks like something strange is going on and that the assumed distribution is wrong. We can reject the Null Hypothesis.

As with all statistical tests, the devil is in the detail. This must be the *only* set of 10 tosses. Obviously if we keep on tossing the coin thousands of times until, sooner or later we do get 10 Heads in a row, then we would have biased our sample to favour our conclusion, and the conclusion would not be valid. Selecting data to fit the hypothesis being tested in this way is termed "cherry picking". We have to trust the experimenters not to cherry-pick.

Note how well Fisher's hypothesis-testing methodology fits Popper's Principles. We don't actually *prove* that the coin is biased; we cannot even be sure that the coin *is* biased – we may be dealing with sleight of hand. We can only say that something highly unlikely has happened which is worthy of further investigation. The only difference between these examples and a real scientific experiment lies in the simplicity of the assumed distribution.

In these examples we make two important assumptions. We assume

1. that the outcome of a normal coin toss is random in that we cannot predict the outcome with certainty, and

2. that successive coin tosses are independent, i.e. that the probability of a coin toss being a Head is 0.5 whatever the outcomes of the preceding tosses may have been.

More precisely, the null hypothesis was that the coin tossing was an independent, random process from a binomial distribution with equal probabilities.

These were examples of statistical inference using "frequentist" statistics and hypothesis testing as a practical example of Popper's ideas. Other statistical methods include "Bayesian" statistics, which have more to do with optimizing strategies when applying established theories to real world problems.

In discussing these ideas, we must distinguish between a *sample* and the *population* or *ensemble* from which it is drawn. The terms population and ensemble designate the same thing: the former comes from the biological sciences, the latter from signal processing.

An ensemble is a set of parallel universes, each with identical statistical properties but differing in random detail. Think of an electronic gadget, a "noise generator", sitting on a laboratory bench making a hissing noise. Now imagine thousands of identical gadgets sitting side by side all making a hissing noise. The hissing noises may all sound the same, but if you make graphs of them with an oscilloscope or a chart recorder, the graphs would all be different; it is only *statistical* properties that we hear.

The average output of all the electronic gadgets at a given time is called the "ensemble", the "expectation" average or "the average across the ensemble". Generally we are not dealing with electronic gadgets nor are we dealing with multiple universes, but we still retain the concept of the ensemble average; it is a useful idea. If x_t is the value of the time series at time t then its ensemble average is written $\mathbb{E}(x_t)$. Note that the ensemble average has the subscript t indicating that it depends on time. We often assume that all ensemble averages of a particular quantity are the same whatever the time, i.e. that

$$\mathbb{E}(x_1) = \mathbb{E}(x_2) = \cdots = \text{constant} \tag{5.1}$$

If ensemble averages all satisfy this assumption, the time series is termed *stationary*. That is the definition of stationary. Stationarity is an assumption. It is not possible to prove that a time series is stationary, but hypothesis testing can be used to prove that it is not.

In reality we must deal with a *sample* time series. Generally there is only one time series, the sample time series, $x_1, x_2, x_3 ... x_N$, and there is only one type of sample average that can be calculated, the time average. This is written $< x >$ or $\bar{x}$, which is the sum of all the

values of the time series divided by the total number of values, i.e. $(x_1 + x_2 + x_3 + ... + x_N)/N$.

For a stationary time series, time averages become closer and closer to ensemble averages as the number of quantities averaged (the "sample size") gets larger. Although intuitively obvious, this is by no means trivial. It is known as the Ergodic Theorem. It forms the basis of statistical inference about time series. Excepting for artificially generated time series, we can never *know* the values of ensemble quantities; we can only estimate them from the data. For example, the sample mean is an estimate of the ensemble mean because of the Ergodic Theorem, i.e. $< x >$ is an estimator of $\mathrm{E}(x)$. Much of statistical theory is concerned with knowing how "good" such an estimate is. When the t-test is used to place 95% confidence limits on a value, we are implying that there is only 1 chance in 20 that the null hypothesis is false, the null hypothesis being that the ensemble mean lies inside the range of the confidence limits.

When estimating ensemble values from sample quantities, there are two important concepts to be considered. They are whether the estimate is *biased* and whether the estimate is *consistent*. An estimate is biased when its expectation value differs from the ensemble value it is supposed to estimate. A consistent estimate is one which converges as the sample size gets larger. This is true if and only if the variance of the estimate tends to zero as the sample size increases.

For notational simplicity only simple averages have been discussed above, and now we shall assume $\mathbb{E}(x_t) \equiv 0$. The same arguments apply to averages of combinations of values. Of particular importance is the product of values which are separated in time by a fixed amount, i.e. $\mathbb{E}(x_t x_{t-L})$ where L is the lag in time between two values of the time series. Its ensemble average, Φ_L, is known as the *covariance* at lag L. When $L = 0$ it is called the *variance*.

The covariance function, Φ, completely summarizes the important statistical properties of a stationary time series. It is a "population parameter", i.e. it is a property of the ensemble. Like the ensemble mean, it can be estimated from the sample data, but placing confidence limits on it is more complicated.

The covariance function does not display well. Covariance functions from entirely different processes can look remarkably similar. Fortunately there is a way around this. The Fourier Transform of the covariance function $S(f)$ displays very well indeed. It shows how variance is distributed with frequency. It has visible features such as

peaks, troughs and power law slopes which correspond to underlying physical processes such as blurring, resonance and integration. It is called the *variance density spectral estimate*, the *variance density spectrum*, sometimes the *power spectrum* but usually just the *spectrum*.

Like the covariance function, the ensemble spectrum comprises a set of population parameters which are estimated from the data. There are various methods for estimating spectra which are discussed further below.

Variance density spectra defined in this way are the numerical analogues of the energy density spectra of physics, such as black body radiation spectra or turbulent energy wavenumber spectra mentioned above. Like them, it is crucially dependent on underlying physical processes and so is a powerful analytical tool for exploring such processes.

The Periodogram

The squared modulus of the Fourier Transform of the time series is the simplest, most straightforward method of estimating its spectrum. This estimate is called the *periodogram* and is defined by equations (7.1) and (7.5).

Periodogram spectra tend to be noisy compared with other methods of spectral estimation. In the 1950s, when digital computers first started to be used for the spectral analysis of time series, it was widely believed that this was because the periodogram was not a consistent estimator of the ensemble variance density, see, for example, Hannan (1960). This is a fallacy. The length, N, of a sample time series is *not* the sample size. The sample size is 1, and each ordinate has an F-distribution with 2 and $N-1$ degrees of freedom. In this context, the length, N, is the dimension and not the sample size in a statistical sense. A proof that the periodogram is a consistent estimator for a well-behaved (ARMA) process is given in Chapter 7.

This widespread misconception has led to a variety of methods for smoothing or 'windowing' the data, such as the 'Hanning Window' and so on, as in Blackman and Tukey (1958), for example. These may sometimes be convenient for display purposes but are unnecessary and frequently misleading. The noisiness of the periodogram spectral estimate can be attributed to the small number of degrees of freedom,

namely 2, associated with each ordinate; the N degrees of freedom of the sample time series are spread out among $N/2$ spectral values. Despite this noisiness, useful features of the periodogram can still be identified.

A narrow, statistically significant peak at frequency f_0 in the periodogram indicates cyclic behaviour in the underlying time series at that frequency. This in turn implies that a deterministic cycle is present in the data, perhaps of astronomical origin. A broad peak in the spectral estimate implies resonant, stochastic behaviour. A flat or 'white' spectrum implies that the time series is unselfcorrelated. So et al. (1999) have shown that the unwindowed, unsmoothed periodogram is the most sensitive method of detecting sinusoids in noisy time series.

Widespread features of time series of physical quantities are power law spectra which are a consequence of natural processes such as integration, i.e.

$$S = Af^{\nu} \tag{5.2}$$

where S is the variance density at frequency f and A and ν are constants. For this reason it is convenient to plot $\log(S)$ against $\log(f)$ when displaying spectra. Thus (1) becomes

$$\log(S) = \nu \log(f) + log(A) \tag{5.3}$$

so that power law relationships appear as straight lines when spectra are plotted using logarithmic scales. Log-log spectra have the additional advantage of conveniently displaying behaviour over a wide range of scales and diminishing the apparent noisiness of the periodogram. Temperature time series often have a power law spectrum with $\nu = -2$ because temperature is commonly dependent on the integration of heat or other forms of energy as discussed in Chapter 9.

Of interest is the spectrum obtained by (5.2) when $\nu = 0$ and $\hat{S}$ =const. In this event, by analogy with light, the spectrum is termed 'white'. It corresponds to a population covariance function which has a positive value at zero lag and is zero elsewhere, i.e. the time series is unselfcorrelated. While this can never be demonstrated conclusively in practice for any particular time series, it is a powerful null hypothesis. If the sample time series is assumed unselfcorrelated then it can be shown that the spectral ordinates are independent of each other and, furthermore, that each ordinate divided by the spectral mean,

$\sum \hat{S}/N$, has an F-distribution with 2 and $N-1$ degrees of freedom. This allows confidence limits to be placed on spectral ordinates so that spectral peaks can be tested for significance. Figure 11.2b is an example. This topic is discussed in more detail in Chapter 7.

The null hypothesis, that the given time series is the outcome of a white noise process with zero mean, can be rejected when either one or more peaks exceed the confidence level or when the slope is significantly less than zero. There may be other situations in which the null hypothesis can be rejected, e.g. the presence of band-limited noise. The periodogram is a powerful tool for dealing with the behaviour of a time series in the frequency domain. There is no requirement that the data be windowed or filtered in any way. In fact such windowing methods preclude the use of the periodogram as a well-defined statistic.

Maximum Entropy Spectra

The idea of maximizing entropy in order to improve the resolution of an image or spectrum first arose in radio astronomy (Gull and Daniell, 1978). This Maximum Entropy Method (MEM) was rapidly adopted in time series analysis as a more sophisticated method of reducing the inherent noisiness of spectral estimates present in the periodogram of short time series, see, for example, Reid (1979) and references therein. At the time computers were very slow and various tricks were used to speed up calculations. In spectral analysis, one such trick was to compute the sample covariance of the time series out to some maximum lag, N, and then to take the Fourier Transform of this truncated covariance estimate as the spectral estimate.

The resulting spectral estimate has rather low frequency resolution because truncation in the time domain resulted in convolution in the frequency domain. The new Maximum Entropy Method (MEM) improved spectral resolution considerably, or, at least, it appeared to.

MEM assumes that the sample time series $\{x_t,\, t=1,...,N\}$ is the realization of a random vector $\boldsymbol{X}$; the $p+1$ ensemble variances and covariances, $\mathbb{E}(X_i X_j)\ i,j \leq p$, are fixed; and that the other covariances at higher lags are simply unknown and undetermined. The Gibbs entropy given by (2.16) is generalized to the continuous H-function of

the probability density distribution of the ensemble.

$$H = -\oint f(\boldsymbol{X}) \ln(f(\boldsymbol{X})) \boldsymbol{dX} \tag{5.4}$$

This is then maximized under the constraint that the $N+1$ ensemble covariances are fixed at their sample values

$$\oint \boldsymbol{X}\boldsymbol{X}^T f(\boldsymbol{X}) \boldsymbol{dX} = [< x_i x_j >] \tag{5.5}$$

and

$$\oint f(\boldsymbol{X}) \boldsymbol{dX} - 1 = 0 \tag{5.6}$$

where (assuming zero mean) $[< x_i x_j >]$ is the matrix of sample covariances.

The details of the derivation need not concern us here, but it turns out that:

1. the resulting joint distribution is Gaussian and
2. it is an autoregressive process of order p.

MEM spectral analysis was very popular for a while because it gives spectacularly sharp spectral peaks in contrast to the degraded resolution of other spectral estimators. However there are some problems with it.

The number of peaks in the spectral estimate depends on the order chosen for the analysis; there are usually about $p/2$ spectral peaks. This means that choosing the order, p, appropriately is crucial. For this purpose, a parameter known as the Akaike Information Criterion (AIC) is used (Akaike, 1970). The AIC estimates of the relative quality of a statistical model for a given set of data and so provides a means for model selection. The AIC does not provide a test of a model in the sense of testing a null hypothesis. It tells us nothing about the absolute quality of a model, only the quality relative to other models.

But the worst aspect of MEM must surely be the false sense of achievement it gives the researcher. MEM spectra have razor sharp peaks with elegantly sloped skirts; they *look* precise. A researcher enamoured of this method would be well advised to generate a synthetic time series, i.e. a sequence of random numbers of the same length as

his/her data, subject it to MEM analysis with the same maximum lag and then compare the results. It is commonly the case that the MEM spectrum of the random numbers looks equally as convincing as that of the data.

These problems arise from the fact that MEM is a purely autoregressive process; it is an *all-pole model*. In cases where a moving average component is present, the assumption that the variance and first p covariances are the only constraints on the data are incorrect; for example, for a pure moving average (MA) process of order q, the covariances from $q+1$ to infinity are all zero. Experience shows that MA components are frequently present in real world data. They are often a consequence of the convolution or "blurring" of data due to the inevitable imperfections of the measurement technique.

The ARMA Spectrum

The Autoregressive Moving Average or $ARMA$ model for spectral estimation specified by equation (6.13) is discussed in the following chapters. It is the best model for spectral estimation. It is the most general linear model of a stochastic time series.

The ARMA model has two advantages over other spectral techniques:

1. The drift parameter estimate, $\hat{c}$, can be used to test whether there is a significant linear trend in the time series.

2. For given order, (p,q), the model itself can be tested for significance.

When the time series exhibits a trend, i.e. when $\hat{c} \neq 0$ in (6.13), the process is non-stationary and the ensemble spectrum has a singularity at the origin. Such situations are dealt with either by using an ARIMA ("Autoregressive Integrated Moving Average") model, by forming a new time series of first differences or by removing the trend from the data by the usual methods of ordinary least squares linear regression.

As discussed above, the covariance function, Φ, as defined by (7.14), encapsulates the important statistical properties of a stationary time series. The sequence $\{\epsilon_t\}$ in (6.13) is assumed to be "white" (i.e. unselfcorrelated) in order that the two sequences of coefficients,

$\{a_1, ..., a_p\}$ and $\{b_1, ..., b_q\}$ be estimated from the data. The sequence, $\{\epsilon_t\}$, can be found by iteratively substituting the coefficient estimates $\{\hat{a}_1, ..., \hat{a}_p\}$ and $\{\hat{b}_1, ..., \hat{b}_q\}$ and the original data back into (6.13), when $\{\epsilon_t\}$ becomes known as the "sequence of residuals".

If the estimation has not been effective, for example because p and q were chosen to be too small, then the sequence of residuals will be self-correlated, i.e. be non-white. Intuitively speaking, information remains behind in the residuals. As it happens there are powerful statistical tests for assessing whether such a sequence of residuals is indeed white. The *Ljung–Box Test* (Ljung and Box, 1978) is one such test.

Most statistical packages return the sequence of residuals when fitting an ARMA model. These can then be input into the package's Ljung–Box procedure. This returns the Ljung–Box Q-parameter which has a χ^2-distribution. Hence a probabilty, α, can be found for the Q-parameter as a function of lag. Should this probability lie below some given threshold at some lag, say $\alpha < 0.05$, then the null hypothesis that the residuals are unselfcorrelated can be rejected.

This is a powerful test of a model, far more powerful than the Akaike Information Criterion mentioned aboved. In practice it is a simple matter to run the ARMA model-fitting procedure with different values of p and q until a fit which satisfies the Ljung–Box test is found. In practice either a number of pairs of values (p, q) are found or none at all. If the former, then the smallest values should be chosen.

The process of finding an ARMA model is similar to factorizing an algebraic expression by inspection: sometimes factors can be found and sometimes not. Indeed it may well be that a linear algebraic solution may be found for factorizing the sample covariance matrix. Sometimes no ARMA model can be found that fits a given sample time series even though it appears stationary. For example, this happens when the sample is a strongly non-linear function of an underlying ARMA process, but there may well be other reasons.

Chapter 6

Iterative Processes

The famous Fibonacci series is perhaps the simplest example of an iterative, deterministic process. It goes:

1, 1, 2, 3, 5, 8, 13, 21, 34, ...

i.e. each term is the sum of the preceding two terms. Mathematically

$$x_t = x_{t-1} + x_{t-2} \tag{6.1}$$

It is very interesting. It grows exponentially, and the ratio of successive terms converges to the Golden Ratio beloved of the Ancient Greeks.

It can be generalized:

$$x_t = a_1 x_{t-1} + a_2 x_{t-2} \tag{6.2}$$

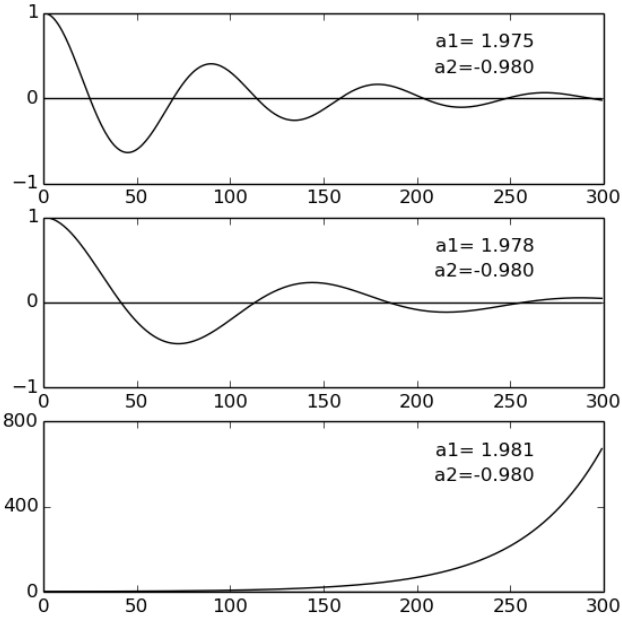
so that (6.1) is just the special case when $a_1 = a_2 = 1$.

Consider what happens if we make $a_1 = 1.183$ and $a_2 = -1$. Starting with (1,1) as before, the sequence goes

1.00, 1.00, 0.18, -0.78, -1.11, -0.53, 0.48, 1.10, 0.82, -0.13, -0.98, -1.02, -0.23, 0.75, 1.12, 0.57, -0.44, -1.09, -0.854, 0.08 ...

This time it doesn't go to infinity; it is cyclic.

The gross behaviour of the generated sequence can be strongly dependent on the values of the coefficients a_1 and a_2. The figure shows the 300-long sequences which resulted from three numerical experiments. In each case the second coefficient was set at $a_2 = -0.98$,



Three sequences of the form (6.2) showing the effect of changing the first coefficient by a small amount.

and the first coefficient varies slightly from run to run, i.e. $a_1 = 1.975, 1.978$ and 1.981. In the first two cases (top two panels), the sequences are damped sinusoids with slightly different frequencies, but in the third case (bottom panel), the sequence increases exponentially like the original (1,1) Fibonacci sequence. A very small change in one coefficient has led to a massive change in the generated sequence.

It turns out that the behaviour of the sequence depends on the *characteristic equation*. Equation (6.2) can be written using a "shift operator", Z:

$$x_t = a_1 Z x_t + a_2 Z^2 x_t \tag{6.3}$$

i.e.

$$(a_2 Z^2 + a_1 Z - 1) x_t = 0 \tag{6.4}$$

where, by definition, $Zx_t = x_{t-1}$. The characteristic equation has the same form as (6.4) in which the shift operator, Z, becomes a complex number, z:

$$a_2 z^2 + a_1 z - 1 = 0 \tag{6.5}$$

which has roots $z_{1,2} = \frac{-a_1 \pm \sqrt{a_1^2 + 4a_2}}{2a_2}$

When (6.5) has complex roots, i.e. when $a_1^2 + 4a_2 < 0$, the generated series is cyclic; otherwise it is exponential. This explains the radical difference between the sequences depicted in the two lower panels of the figure. The characteristic equation (6.5) has two roots, z_1 and z_2 where

$$z_1 + z_2 = a_1 \tag{6.6}$$

and

$$z_1 z_2 = -1 \tag{6.7}$$

When the parabola described by the LHS of 6.5 does not cross the zero axis, then z_1 and z_2 are complex and $z_2 = z_1^*$. When it does cross the zero axis, z_1 and z_2 are both real.

Equation (6.2) can be generalized to any finite order, p, viz.:

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} \tag{6.8}$$

where $\{a_i, i = 1, ..., p\}$ are fixed coefficients.

The shift operator equation is

$$(\sum_{i=1}^{p} a_i Z^i - 1) x_t = 0 \tag{6.9}$$

and the characteristic equation is

$$\sum_{i=1}^{p} a_i z^i - 1 = 0 \tag{6.10}$$

The expression on the LHS of (6.10) is known as the *z-transform* of the model.

Iterative stochastic processes – ARMA

Equation (6.8) can be generalized still further to

$$x_t = \sum_{i=1}^{p} a_i x_{t-i} + \epsilon_t \tag{6.11}$$

where $\{\epsilon_t\}$ is a sequence of random numbers with zero mean. They are assumed to be independent, which implies that they are unselfcorrelated. They are commonly assumed to be Gaussian and identically distributed, but that is not essential.

$\{\epsilon_t\}$ is known as *the innovation.*

As before, (6.11) can be used to generate sequences iteratively using a random number generator to provide $\{\epsilon_t\}$. There is, however, a major difference. For the same set of starting conditions, every sequence generated by (6.8) will be the same whereas every sequence generated by (6.11) will be different. Equation (6.11) describes a *stochastic* process.

An important feature of (6.11) is that it is also a regression equation, i.e.

$$X_t = \sum_{i=1}^{p} a_i x_{t-i} + \epsilon_t \tag{6.12}$$

where at time τ, where $t-1 < \tau < t$, X_t and ϵ_t are random variables, and the x_{t-i} are past fixed sample values. This is the standard regression model with the p coefficients $\{a_i\}$ being the regression coefficients. It means that, given a sequence $\{x_t\}$ and assuming a given order, p, the values of the coefficients can be estimated from the data. For this reason, (6.11) is referred to as an *autoregression* or AR model since the sequence is regressed on itself, i.e. on its own past values.

The most general linear model of a stationary, stochastic process is the *Autoregressive Moving Average* process or *ARMA* process/model. The output of an ARMA(p,q) process at time t is defined as:

$$\xi_t = c + \epsilon_t + \sum_{i=1}^{p} a_i x_{t-i} + \sum_{j=1}^{q} b_j \epsilon_{t-j} \tag{6.13}$$

where ξ_t is the random variable output by the process, the x_{t-i} are past realizations of ξ_t, $\{a_i, i = 1,...,p\}$ are known as the autoregression coefficients, the $\{b_j, j = 1,...,q\}$ are known as the moving average

coefficients , $\{\epsilon_t\}$ is the zero mean, unselfcorrelated innovation, and c is a constant known as the *drift term* because, when non-zero, it represents an underlying drift in the series values over time. The ordered pair (p,q) is known as the *order* of the process.

Given (p,q), estimates of the $p+q+1$ parameters $\hat{a}_1$, $\hat{a}_2$, ... $\hat{a}_p$, $\hat{b}_1$, $\hat{b}_2$, ... ,$\hat{b}_q$ and $\hat{c}$ can be made from the data. In fact they are estimated from the variance/covariance function of the sample time series. How this is done need not concern us here; suitable packages for ARMA estimation are available in high-level programming languages such as *R, Matlab* and *Python*. ARMA functions in such packages not only return parameter estimates but also complete time series of the *residuals*, i.e. the time series $\{\epsilon_t\}$ obtained when the parameter estimates, $\{\hat{a}_i\}$, $\{\hat{b}_j\}$ and $\hat{c}$, and the sample time series $\{x_t\}$ are substituted in (6.13). Useful statistics of $\{\epsilon_t\}$ such as its variance σ_ϵ^2 are also returned.

In the definition (6.13) above, the predicted value of the process, ξ_t, is a random variable. The innovation, ϵ_t, is also a random variable. Consequently the coefficient estimates $\hat{a}_1$ etc. are realizations of random variables with known distributions. Hence confidence limits can be placed on these estimates. This is particularly important with regard to the drift term, $\hat{c}$. All real time series exhibit an apparent long-term trend, either positive or negative. At issue is whether such apparent trends are statistically significant, i.e. whether $\hat{c}$ is significantly different from zero and the trend cannot be attributed to chance alone.

In this way, with an ARMA model, the drift in a sample time series can be tested for significance using accepted methods of statistical inference. While it can never be proven that there is no drift in a population, at least we can show when there is such a drift and at what level of significance. When the drift is so small as to be insignificant, then Occam's Razor allows us to treat the drift as if it were zero.

Chapter 7

Discrete Time Spectral Analysis

Introduction

A thorough description of Fourier Transforms and Spectral Analysis is given by Bracewell (1986). Hamilton (1994) offers an equally thorough description of time series methods.

In the present development, the variance density spectrum is defined as the ensemble average of the periodogram. The periodogram of the sample is thus a statistic. For an ARMA process, it is a consistent estimator of the ensemble variance density spectrum.

In this way, some key theorems are reproduced without recourse to continuous functions. In practice any real time series occupies a range of time scales, $< \Delta t, N\Delta t >$, where Δt is the sampling interval and N is the number of terms in the series. The behaviour of the underlying continuous process, should there be one, at time scales shorter than the sampling period, Δt, is of no concern. If there is evidence of aliasing such as, for example, a concentration of variance at the high-frequency end of the spectrum, then it is the responsibility of the researcher to change the sampling regime to something more suitable. Likewise the length, N, and span, $N\Delta t$, of the time series are assumed to be finite. These assumptions allow key results to be developed while avoiding the pitfalls of spectral methods based on deterministic, continuous functions of infinite extent.

The Periodogram

We start with a finite sequence of real numbers, $\{x_t, t = 0, ..., N-1\}$, which is assumed to comprise samples of a varying physical quantity

Name	Symbol	Domain	Sample or Random Variable
t.s. value	x_n	time	sample
t.s. value	y_n	time	sample
F.T of x	X_j	frequency	sample
F.T of y	Y_j	frequency	sample
estimate	$\hat{P}_k$	frequency	sample
t.s. variable	ξ_n	time	random variable
F.T. of ξ	Ξ_j	frequency	random variable
estimator	Π_k	frequency	random variable

Table 7.1: Symbol conventions

separated by equal intervals of time or covering equal intervals of time, i.e. a 'time series'. We are concerned with $\{X_k, k = 0, ..., N-1\}$, the discrete Fourier transform (DFT) of the time series $\{x_n\}$ given by

$$X_k = \sum_{t=0}^{N-1} x_t . e^{-2\pi i k t/N} \ , \ k = 0, 1...N-1 \tag{7.1}$$

where $i = \sqrt{-1}$. In general, the X_k are complex numbers and, because the x_t are real

$$X_{-k} = X_{N-k} = X_k^* \tag{7.2}$$

where X_k^* is the complex conjugate of X_k. Because of the cyclical character of the DFT, X_j can be defined for arbitrary k since for any integer m

$$X_{k+mN} = X_k \tag{7.3}$$

The underlying concept of spectral analysis is encapsulated by Parseval's theorem which, for the finite, discrete case under discussion here, can be expressed as:

$$\sum_{t=0}^{N-1} x_t^2 = \frac{1}{N} \sum_{k=0}^{N-1} \mid X_k \mid^2 . \tag{7.4}$$

We define a new sequence, the *periodogram*, $\{\hat{P}_k\}$, where

$$\hat{P}_k = 2 \mid X_k \mid^2 / N \ , \ k = 1, 2...N/2 - 1 \tag{7.5}$$

and, if N is even,

$$\hat{P}_{N/2} = \mid X_{N/2} \mid^2 / N.$$

After substituting from (7.2), (7.4) becomes

$$s^2 = \frac{1}{N}\sum_{n=0}^{N-1} x_n^2 = \sum_{k=1}^{N/2} \hat{P}_k. \tag{7.6}$$

since $\hat{P}_0 = 0$.

For notational convenience (and commonly in practice), the data sequence $\{x_n\}$ has zero mean, and so s^2 is the sample variance. Equation (7.6) can be regarded as a conservation law which describes variance density as a function of frequency, $\hat{P}_k$ being that component of the variance associated with the frequency $k\Delta f = k/N\Delta t$. Note that $\hat{P}_k$ has the same units as s^2; to make it a variance density, it must be divided by Δf. In the following development, we assume $\Delta t = 1$ without loss of generality.

The ensemble periodogram is given by the equivalent equation to (7.5), viz.:

$$\Pi_k = 2 \mid \Xi_k \mid^2 /N \ , \ k = 1,2...N/2-1 \tag{7.7}$$

and, for even N,

$$\Pi_{N/2} = \mid \Xi_{N/2} \mid^2 /N.$$

where

$$\Xi_k = \sum_{t=0}^{N-1} \xi_t . e^{-2\pi i k t/N} \ , \ k = 0,1...N/2 \tag{7.8}$$

The Variance Density Spectrum

The variance density spectrum S_k is defined as the ensemble average of the periodogram, viz.:

$$S_k = \mathbb{E}(\Pi_k) \quad \text{for} \quad k = 0,1,...N/2 \tag{7.9}$$

Note that, while this development is discrete in time, it is not, necessarily, discrete in frequency. The discrete frequency forms of (7.1) and (7.8) could equally well have been written as continuous functions of frequency, f, viz.:

$$X(f) = \sum_{t=0}^{N-1} x_t . e^{-2\pi i f t/f_s} \ , \ 0 < f < f_s \tag{7.10}$$

and, for the ensemble:

$$\Xi(f) = \sum_{t=0}^{N-1} \xi_t . e^{-2\pi i f t/f_s} \ , \ 0 < f < f_s \tag{7.11}$$

where $f_s = 1/\Delta t$, Δt being the sampling interval, usually defined to be one. This allows a variance density spectrum, $S(f)$, to be specified as a continuous function of frequency. This is sometimes desirable for display purposes whereas the discrete frequency forms are more easily calculated using the Fast Fourier Transform.

The covariance function

From (7.7), (7.8) and (7.9) and using $|\Xi_k|^2 = \Xi_k \Xi_k^*$:

$$S_k = \frac{1}{N} \mathbb{E} \left[\left(\sum_{m=0}^{N-1} \xi_m . e^{-2\pi i k m/N} \right) \left(\sum_{n=0}^{N-1} \xi_n . e^{2\pi i k n/N} \right) \right] \tag{7.12}$$

Thus

$$S_k = \frac{1}{N} \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \mathbb{E}(\xi_m \xi_n) e^{-2\pi i k (m-n)/N} \tag{7.13}$$

For a wide-sense stationary process, the ensemble covariance function is defined by

$$\Phi(m-n) = \mathbb{E}(\xi_m \xi_n) = \Phi(n-m) \tag{7.14}$$

and (7.13) becomes

$$S_k = \frac{1}{N} \sum_{L=-N}^{N} (N - |L|) \Phi(L) e^{-2\pi i k L/N} \tag{7.15}$$

so that S is the DFT of the function $(N-|L|)\Phi(L)/N$ on the domain $-N \leq L \leq N$.

Note that, as defined here, the variance density spectrum is the Fourier Transform not of the covariance function, as is commonly assumed, but of the covariance function multiplied by the triangle function. This comes about because of the finite length of the time series. Consequently the variance spectrum of any finite time series has, in effect, been convoluted with the Fourier Transform of the triangle function, i.e. with the $sinc^2$ function. This is important when sinusoids are involved as discussed further below.

When the time series is unselfcorrelated, i.e. when

$$\mathbb{E}(\xi_m \xi_n) = \Phi(m-n) = \sigma^2 \delta_{mn} \tag{7.16}$$

where σ^2 is the variance of ξ_n and δ_{mn} is the Kronecker delta, i.e. $\delta_{nn} = 0$ and $\delta_{mn} = 0$ when $m \neq n$, then

$$S_k = \frac{1}{N}\Phi(0) = \frac{\sigma^2}{N} = \text{const.} \tag{7.17}$$

and the time series is said to be "white" by analogy with the energy spectrum of visible light.

The periodogram of a white time series

The given time series, $\{x_n\}$, can be considered one realization of an ensemble of random sequences, $\{\xi_n\}$. Likewise $\{X_k\}$, the DFT of $\{x_n\}$, is the corresponding realization of the random sequence $\{\Xi_k\}$, the DFT of $\{\xi_n\}$; and the computed periodogram, $\{P_k\}$, is the realization of the ensemble periodogram, $\{\Pi_k\}$.

As a null hypothesis we assume that the time series has zero mean and is unselfcorrelated. Thus

$$\mathbb{E}(\xi_n) = 0 \tag{7.18}$$

and

$$\mathbb{E}(\xi_m \xi_n) = \sigma^2 \delta_{mn} \tag{7.19}$$

The random variable Ξ_k defined by (7.8) can be written in terms of its real and imaginary components, viz.:

$$\Xi_k = \Re_k + \Im_k i \tag{7.20}$$

i.e.

$$|\,\Xi_k^2\,|^2 = \Xi_k \Xi_k^* = \Re^2 + \Im^2 \tag{7.21}$$

where

$$\Re_k = \sum_{n=1}^{N-1} \xi_n \cos(2\pi k n/N) \tag{7.22}$$

and

$$\Im_k = \sum_{n=1}^{N-1} \xi_n \sin(2\pi k n/N) \tag{7.23}$$

For large N, by the central limit theorem, both $\Re_k$ and $\Im_k$ must be close to Gaussian whatever the distribution of the ξ_n may be. They are also unselfcorrelated, uncorrelated with each other and have zero mean. The variance of $\Re_k$ is given by

$$\mathbb{E}(\Re_k^2) = \sigma^2 \sum_{m=1}^{N}\sum_{n=1}^{N} \delta_{mn} \cos(2\pi km/N)\cos(2\pi kn/N) \tag{7.24}$$

$$= \frac{\sigma^2}{2}\left[N + \sum_{n=1}^{N} \cos(4\pi kn/N)\right] \tag{7.25}$$

$$= \frac{\sigma^2}{2}\left[N + \frac{1}{2}\sum_{n=1}^{N}(z^n + z^{*n})\right] \tag{7.26}$$

$$= \frac{\sigma^2}{2}\left[N + \frac{1}{2}\left(\frac{z(1-z^N)}{1-z} + \frac{z^*(1-z^{*N})}{1-z^*}\right)\right] \tag{7.27}$$

where $z = e^{4\pi ik/N}$ so that

$$z^N = z^{*N} = 1 \tag{7.28}$$

and

$$\mathbb{E}(\Re_k^2) = \frac{\sigma^2 N}{2} \tag{7.29}$$

Similarly

$$\mathbb{E}(\Im_k^2) = \frac{\sigma^2 N}{2} \tag{7.30}$$

It follows that the rescaled random variables $\sqrt{\frac{2}{N\sigma^2}}\Re_k$ and $\sqrt{\frac{2}{N\sigma^2}}\Im_k$ are both normal(0,1). Hence, from (7.7) the random variable

$$\frac{N\Pi_k}{\sigma^2} = \frac{2\,|\,\Xi_k\,|^2}{N\sigma^2} = \frac{2\Re^2}{N\sigma^2} + \frac{2\Im^2}{N\sigma^2} \tag{7.31}$$

has a χ^2-distribution with two degrees of freedom, and so the scaled spectral estimate of an unselfcorrelated time series, $N\hat{P}_k/\sigma^2$, has a χ^2-distribution with two degrees of freedom.

In practice we do not know σ^2 in (7.31), which is a population parameter. We only know its estimate $\hat{\sigma}^2$, where

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{k=1}^{N} x_k^2 \tag{7.32}$$

since $< x_k >= 0$. Assuming that the $\{x_k\}$ are the realization of a sequence of unselfcorrelated random variables $\{\xi_k\}$ which are normally distributed (0,σ), then

$$\frac{\hat{\sigma}^2}{\sigma^2} = \frac{1}{N-1}\sum_{k=1}^{N}\frac{x_k^2}{\sigma^2} \tag{7.33}$$

has a χ^2-distribution with $N-1$ degrees of freedom. Hence the quantity

$$\frac{N\hat{P}_k}{\hat{\sigma}^2} = \frac{N\Pi_k}{\sigma^2}\Big/\frac{\hat{\sigma}^2}{\sigma^2} \tag{7.34}$$

is the realization of a random variable which is the ratio of two random variables having χ^2-distributions with 2 and $N-1$ degrees of freedom, respectively.

If X_1^2 and X_2^2 are independent random variables following chi-square distributions with ν_1 and ν_2 degrees of freedom respectively then the distribution $F = \frac{X_1^2/\nu_1}{X_2^2/\nu_2}$ is said to follow the variance ratio or F-distribution with ν_1 and ν_2 degrees of freedom (see: Abramowitz and Stegun (1965), section **26.6.2**). Inserting degrees of freedom of 2 and $N-1$ into (7.34) gives an F statistic for each value of the periodogram, viz.:

$$F_k = \frac{N(N-1)\hat{P}_k}{2\hat{\sigma}^2} = \frac{N\Pi_k}{2\sigma^2}\Big/\frac{\hat{\sigma}^2}{(N-1)\sigma^2} \tag{7.35}$$

and so F_k has an F-distribution with 2 and $N-1$ degrees of freedom.

The p-values for the F-distribution are provided in most statistical software packages. These can be used to test each ordinate of the periodogram for whiteness using the F_k statistic defined by (7.35). In practice the F-distribution with these degrees of freedom is barely distinguishable from the χ^2-distribution with 2 degrees of freedom for $N > 100$.

Averaged white periodograms

The resolution of periodograms may be improved by averaging a number of them (Bartlett, 1948). In order that sample statistics be calculable, it is essential that the periodograms come from time series that are of the same length and were generated by the same processes or

by identical processes and that they be independent. In the present case of time series which are assumed unselfcorrelated under a null hypothesis, it is sufficient that they not overlap in time.

The kth ordinate of the mean of the periodograms of M, N-long time series with ensemble variance σ^2 is given by

$$\bar{\hat{P}}_k = \frac{1}{M}\sum_{m=1}^{M} P_k^m \quad \text{for } k = 1 \cdots N/2 \tag{7.36}$$

Multiplying both sides by N/σ^2 gives

$$\frac{N\bar{\hat{P}}_k}{\sigma^2} = \frac{1}{M}\sum_{m=1}^{M} \frac{NP_k^m}{\sigma^2} \quad \text{for } k = 1 \cdots N/2 \tag{7.37}$$

Since each of the terms in the sum has a χ^2-distribution with 2 degrees of freedom, then the sum itself must have a χ^2-distribution with $2M$ degrees of freedom. Hence

$$\text{Var}\left(\frac{N\bar{\hat{P}}_k}{\sigma^2}\right) = \frac{1}{M^2}\text{Var}\left(\sum_{m=1}^{M} \frac{NP_k^m}{\sigma^2}\right) = \frac{2}{M} \tag{7.38}$$

Hence $\text{Var}\left(\frac{N\bar{\hat{P}}_k}{\sigma^2}\right) \to 0$ as $M \to \infty$, and P_k of a white time series is a consistent estimator of S_k as defined by (7.9).

Fourier Transform of a Convolution

Measurements are often contaminated or blurred by the physical characteristics of the measuring device. Thermometers have thermal mass and so do not respond instantaneously to temperature changes. Lenses are never perfect and all photographic images are blurred to some degree. Prior to measurement, physical quantities may themselves be blurred by diffusion and mixing. Numerically such blurring is expressed as *convolution*.

Consider an "output" time series $\{y_t, t = 0, ..., N-q-1\}$ formed by convolving an "input" time series $\{x_t, t = 0, .., N-1\}$ with a "filter" function, $\{\beta_j, j = 0, ..., q\}$, i.e.

$$y_t = \sum_{j=0}^{q} \beta_j x_{t-j} \tag{7.39}$$

Let Y_k, X_k and B_k be the Fourier Transforms of $\{y_t\}$, $\{x_t\}$ and $\{\beta_j\}$, respectively; then from (7.1),

$$\begin{aligned}
\sum_{t=0}^{N-1} y_t e^{-2\pi ikt/N} &= \sum_{t=0}^{N-1} \left[\sum_{j=0}^{q} \beta_j x_{t-j} \right] e^{-2\pi ikt/N} \\
&= \sum_{j=0}^{q} \beta_j \left[\sum_{t=0}^{N-1} x_{t-j} e^{-2\pi ik(t-j)/N} \right] e^{-2\pi ikj/N} \\
&= \sum_{j=0}^{q} \beta_j \left[\sum_{t-j=0}^{N-1} x_{t-j} e^{-2\pi ik(t-j)/N} \right] e^{-2\pi ikj/N} \\
&= \sum_{j=0}^{q} \beta_j X_k e^{-2\pi ikj/N} \\
&= \left[\sum_{j=0}^{q} \beta_j e^{-2\pi ikj/N} \right] X_k
\end{aligned}$$

i.e.

$$Y_k = B_k X_k \tag{7.40}$$

and so the Fourier Transform of convoluted functions is the product of the Fourier Transforms of those functions. The change from t to $t-j$ in the sum is allowable because the kernel, $e^{-2\pi ikt/N}$, is cyclic over the domain of the dummy variable.

Periodogram of an ARMA process

For notational convenience we assume the sampling interval, Δt, to be 1 time unit and the periodogram frequency resolution, $1/N\Delta t$, to be $1/N$ frequency units.

An ARMA(p,q) process is defined by

$$y_t = x_t + \sum_{i=1}^{p} a_i y_{t-i} + \sum_{j=1}^{q} b_j x_{t-j} \tag{7.41}$$

where $\{x_t\}$ is an unselfcorrelated sequence. Equation (7.41) can be written

$$y_t - \sum_{i=1}^{p} a_i y_{t-i} = x_t + \sum_{j=1}^{q} b_j x_{t-j} \tag{7.42}$$

We define two new vectors γ and β where

$$\gamma_0 = 1 \quad \text{and} \quad \gamma_i = -a_i \quad \text{for} \quad i > 0 \tag{7.43}$$

and

$$\beta_0 = 1 \quad \text{and} \quad \beta_j = b_j \quad \text{for} \quad j > 1 \tag{7.44}$$

Hence

$$\sum_{i=0}^{p} \gamma_i y_{t-i} = \sum_{j=0}^{q} \beta_j x_{t-j} \quad \text{for } t = 1 \cdots N \tag{7.45}$$

For the ensemble

$$\sum_{i=0}^{p} \gamma_i \upsilon_{t-i} = \sum_{j=0}^{q} \beta_j \xi_{t-j} \quad \text{for } t = 1 \cdots N \tag{7.46}$$

where x_t and y_t are assumed to be realizations of random variable sequences ξ_t and υ_t respectively. Taking the (continuous) Fourier Transform of both sides of (7.45) and using (7.40) gives

$$Y(f)\Gamma(f) = X(f)B(f) \tag{7.47}$$

where $Y(f)$, $\Gamma(f)$, $X(f)$ and $B(f)$ are the Fourier Transforms of y_t, γ_t, x_t and β_t, respectively. Rearranging, taking the squared modulus of both sides and replacing f with integer value k gives

$$|Y_k|^2 = |X_k|^2 \left| \frac{B_k}{\Gamma_k} \right|^2 \tag{7.48}$$

By analogy with (7.5), let Q_k be the periodogram of the ARMA time series, $\{y_k\}$, so that (7.48) becomes

$$Q_k = P_k \left| \frac{B_k}{\Gamma_k} \right|^2 \tag{7.49}$$

Now let M independent periodograms, Q_k^m, be averaged, i.e.

$$\bar{\hat{Q}}_k = \frac{1}{M} \sum_{m=1}^{M} Q_k^m \quad \text{for } k = 1 \cdots N/2 \tag{7.50}$$

Then from (7.49),

$$\bar{\hat{Q}}_k = \bar{\hat{P}}_k \left| \frac{B_k}{\Gamma_k} \right|^2 \tag{7.51}$$

where Γ_k and B_k are functions of population parameters such as a_i and b_j, and so are independent of M.

By (7.38), the variance of $\bar{\hat{P}}_k$ converges as $M \to \infty$. Therefore, by (7.49), $\bar{\hat{Q}}_k$, the variance of the periodogram of the ARMA time series, $\{y_k\}$, also converges as $M \to \infty$.

Hence the periodogram of an ARMA time series is a consistent estimator of its ensemble spectrum.

The spectral estimate of an ARMA process

From (7.49),

$$\mathbb{E}(Q_k) = \mathbb{E}(P_k) \left| \frac{B_k}{\Gamma_k} \right|^2 \tag{7.52}$$

Hence, using (7.9) and (7.17), the variance density spectrum, S_k, of an ARMA time series is given by

$$S_k = \frac{\sigma^2}{N} \left| \frac{B_k}{\Gamma_k} \right|^2 \tag{7.53}$$

and the spectral estimate, $\hat{S}_k$, by

$$\hat{S}_k = \frac{\hat{\sigma}^2}{N} \left| \frac{\hat{B}_k}{\hat{\Gamma}_k} \right|^2 \tag{7.54}$$

where $\hat{\sigma}^2$ is the variance of the residuals. $\hat{B}_k$ and $\hat{\Gamma}_k$ are given by

$$\hat{\Gamma}_k = 1 - \sum_{j=1}^{p} \hat{a}_i z^j \tag{7.55}$$

and

$$\hat{B}_k = 1 + \sum_{j=1}^{q} \hat{b}_j z^j \tag{7.56}$$

where $z = e^{-2\pi i k/N}$, and $\{\hat{a}_i\}$ and $\{\hat{b}_j\}$ are estimated from the data.

The periodogram of a sinusoid

The periodogram of a sinusoid exhibits a narrow peak. The peak is not a δ-function for a time series of finite length as discussed in reference to (7.15).

Consider a time series $\{c_j, j = 0, ..., N-1\}$, defined explicitly by a cosine function with arbitrary frequency, f_0, and phase, ϕ:

$$c_j = A\cos(2\pi j f_0/N + \phi) = x_j + y_j \tag{7.57}$$

where

$$x_j = Ae^{2\pi i j f_0/N + \phi i)}/2 \tag{7.58}$$

and

$$y_j = Ae^{-2\pi i j f_0/N + \phi i)}/2 \tag{7.59}$$

Substituting x_j into (7.10) gives

$$X(f) = \frac{A}{2N}\sum_{j=0}^{N-1} e^{-2\pi i j(f-f_0)/N)+\phi i} \tag{7.60}$$

After multiplying both sides by $e^{-2\pi i(k-f_0)/N}$

$$X(f)e^{-2\pi i(f-f_0)/N} = \frac{A}{2N}\sum_{j=0}^{N-1} e^{-2\pi i(j+1)(f-f_0)/N+\phi i} \tag{7.61}$$

Subtracting (7.61) from (7.60) and transposing gives

$$X(f) = \frac{Ae^{\phi i}\left[1 - e^{2\pi i(f-f_0)}\right]}{2N\left[1 - e^{2\pi i(f-f_0)/N}\right]} \tag{7.62}$$

$$= \frac{Ae^{\pi i(k-f_0)+\phi i}\left[e^{-\pi i(k-f_0)} - e^{\pi i(k-f_0)}\right]}{2Ne^{\pi i(k-f_0)/N}\left[e^{-\pi i(k-f_0)/N} - e^{\pi i(k-f_0)/N}\right]} \tag{7.63}$$

i.e.

$$X_k = \frac{Ae^{\pi i(f-f_0)+\phi i}\sin\{-\pi(f-f_0)\}}{2Ne^{\pi i(f-f_0)/N}\sin\{-\pi(f-f_0)/N\}} \tag{7.64}$$

Likewise, $\{Y(f)\}$, the DFT of $\{y_j\}$ is given by

$$Y(f) = \frac{Ae^{\pi i(f+f_0)+\phi i}\sin\{-\pi(f+f_0)\}}{2Ne^{\pi i(f+f_0)/N}\sin\{-\pi(f+f_0)/N\}} \tag{7.65}$$

where, from (7.57) the squared modulus of the DFT of $\{c_j\}$ is given by

$$|\,C(f)\,|^2 = X(f)X(f)^* + X(f)Y(f)^* + X(f)^*Y(f) + X(f)^*Y(f)^* \tag{7.66}$$

From (7.64) and (7.65),

$$X(f)X(f)^* = \frac{A^2 \sin^2\{\pi(f-f_0)\}}{4N^2 \sin^2\{\pi(f-f_0)/N\}} \tag{7.67}$$

$$X(f)Y(f)^* = \frac{A^2 e^{2\pi i f_0 + 2\phi i} \sin\{\pi(f+f_0)\} \sin\{\pi(f-f_0)\}}{4N^2 e^{2\pi i f_0/N} \sin\{\pi(f+f_0)/N\} \sin\{\pi(f-f_0)/N\}} \tag{7.68}$$

$$X(f)^*Y(f) = \frac{A^2 e^{-2\pi i f_0 - 2\phi i} \sin\{\pi(f+f_0)\} \sin\{\pi(f-f_0)\}}{4N^2 e^{-2\pi i f_0/N} \sin\{\pi(f+f_0)/N\} \sin\{\pi(f-f_0)/N\}} \tag{7.69}$$

and

$$Y(f)Y(f)^* = \frac{A^2 \sin^2\{\pi(f+f_0)\}}{4N^2 \sin^2\{\pi(f+f_0)/N\}} \tag{7.70}$$

Adding the cross-modulation terms gives an error term $E(f)$,

$$E(f) = X(f)Y(f)^* + X(f)^*Y(f) \tag{7.71}$$

Thus

$$E(f) = \frac{A^2 \cos\{2\pi(1-1/N)f_0 + 2\phi\} \sin\{\pi(f+f_0)\} \sin\{\pi(f-f_0)\}}{2N^2 \sin\{\pi(f+f_0)/N\} \sin\{\pi(f-f_0)/N\}} \tag{7.72}$$

Substituting (7.2) in (7.65) gives

$$Y(f)Y(f)^* = X(f)X(f)^* \tag{7.73}$$

and the periodogram of a cosine function of arbitrary frequency and phase becomes:

$$Q(f) = \frac{A^2 \sin^2\{\pi(f-f_0)\}}{2N^2 \sin^2\{\pi(f-f_0)/N\}} + E(f) \tag{7.74}$$

Note that the error term, $E(f)$, is dependent on the phase, ϕ. Its absolute value $|E(f)|$ has a maximum value of $0.025A^2$ when $f_0 \approx 1$ and $k = 1$ and is bounded above by $\mid 1.5A^2/2N \sin(2\pi f/N) \mid$.

When N is large, (7.74) becomes

$$Q(f) \approx \frac{A^2 \sin^2\{\pi(f-f_0)\}}{2\{\pi(f-f_0)\}^2} \tag{7.75}$$

i.e. the $sinc^2$ function. The discrete frequency periodogram is

$$Q_k \approx \frac{A^2 \sin^2\{\pi(k-f_0)\}}{2\{\pi(k-f_0)\}^2} \tag{7.76}$$

The Spectrum of a Random Walk

Many environmental time series are comprised of temperature measurements or depend in some way on temperature, particularly proxy time series such as isotope ratios in ocean sediment and ice. Temperatures are, in turn, dependent on the transport of heat and other forms of energy by turbulent stochastic processes (Hasselmann, 1976). The relationship between temperature and heat flux is given by Fourier's heat equation:

$$\rho C \frac{\partial T}{\partial t} = -K_0 \frac{\partial T}{\partial x} + Q(x,t) \tag{7.77}$$

where ρ and C are the density and specific heat of the body being heated, T is its temperature, t is the time, x is a spatial variable with the units of length, K is the conductivity and Q is the flux of heat or other form of energy. In general terms and in discrete form, (7.77) becomes

$$\rho C \frac{\Delta T}{\Delta t} = -K \frac{\Delta T}{\Delta x} + Q \tag{7.78}$$

As previously we choose a time scale for which $\Delta t = 1$ gives

$$\rho C (T_t - T_{t-1}) = -\frac{K}{\Delta x}(T_{t-1} - T_{res}) + Q \tag{7.79}$$

where T_{res} is the temperature of some relatively stable reservoir to and from which heat is being transferred. This becomes

$$T_t = q_t + a_1 T_{t-1} \tag{7.80}$$

where

$$q_t = Q + \frac{K}{\rho C \Delta x} T_{res} \tag{7.81}$$

and

$$a_1 = 1 - \frac{K}{\rho C \Delta x} \tag{7.82}$$

$$= 1 - \alpha \tag{7.83}$$

where $0 \leq \alpha << 1$.

At sufficiently large time scales, Δt, q_t can be assumed to have zero mean and to be unselfcorrelated, in which case (7.80) describes an ARMA(1,0) process for temperature with variance density spectrum given by:

$$S(f) = \frac{\sigma_q^2}{N|1 - a_1 z|^2} \tag{7.84}$$

where σ_q^2 is the variance of the q_t white noise heating/cooling process,

$$z = e^{-2\pi i f/N} \tag{7.85}$$

and N is the length of the time series of temperature values. This is a *single-pole model*; there is a single pole (i.e. singularity) in the z-plane. It is on the real axis outside the unit circle at $z_1 = 1/a_1$.

Setting the discrete time angular frequency, ω as $\omega = 2\pi f/N$, (7.84) can be rewritten

$$\begin{aligned} S(f) &= \frac{\sigma_q^2}{N(1-2a_1\cos\omega + a_1^2)} \\ &= \frac{\sigma_q^2}{N((1-2a_1+a_1^2)+2(a_1-\cos\omega))} \\ &= \frac{\sigma_q^2}{N((1-a_1)^2+4a_1\sin^2\frac{\omega}{2})} \end{aligned}$$

Hence for small values of f, $\omega << 1$, $\sin\omega = \omega$ and (7.84) becomes

$$S(f) = \frac{\sigma_q^2}{N(\alpha^2 + a_1\omega^2)} \approx \frac{\sigma_q^2}{N(\alpha^2+\omega^2)} \tag{7.86}$$

When $\omega > \alpha$, $S(f)$ is inversely related to ω^2 and so inversely related to f^2.

When $\alpha = 0$ there is a singularity at $S(f) = 0$, and the time series is said to be a *random walk*. Since the variance of a random walk increases with time, (7.14) does not hold, the time series is not stationary and a variance density spectrum cannot be defined. However, when α is small but non-zero, the time series is stationary, and a variance density spectrum can be defined. Such a time series is called a *centrally biased random walk* in statistics and *low-pass filtered noise* or *red noise* in signal processing. We will use the term *red noise*.

Chapter 8

Spurious Regression

Introduction

In statistics, *regression* is the process of fitting a straight line to a collection of points on a graph. The sample correlation coefficient, r, is a measure of "goodness of fit", i.e. of how well the fitted line fits the data points. When the line fits perfectly then $r = \pm 1$. The sign of r indicates the slope of the regression line; a negative value indicates the line slopes downwards, i.e. that the values tend to decrease with increases in the independent variable.

If a system is deterministic then its variables are all single-valued functions of time. Experimental observations of dynamical variables are commonly displayed as functions of time and a regression line fitted to the observations to display the trend or rate of change with time. This is commonplace, something most researchers learned in school.

However, there can be serious problems with this methodology when the system under investigation is stochastic. Nelson and Kang (1984) demonstrated that, for certain stochastic processes such as a *random walk*, the use of time as the explanatory variable can lead to the appearance of a trend even though none was present in the original data. An observed trend obtained by regressing a physical quantity on time may or may not be real, depending on the deterministic or stochastic nature of the system under investigation. A similarly misleading effect is observed when two random walk time series are compared. This was noted nearly a century ago by the then President of the Royal Statistical Society, George Udny Yule and

termed "Nonsense-Correlation" by him (Yule, 1926). More recently Granger and Newbold (1974) have described it as *Spurious Regression*. Although not widely known outside the field of Econometrics, the implications of these papers cannot be overestimated. They are little known in the physical sciences.

It is widely assumed that this effect is solely a consequence of the time series being a random walk and so non-stationary, but that is not the case. Here we show that the effect is the result of a high concentration of variance at the low frequency end of the spectrum, i.e. at periods which are long compared with the length of the sample time series, such as the red noise time series described in the previous section. Such time series are common in nature, particularly temperature time series.

Time as the independent variable

A useful feature of the ARMA approach is that it provides means for generating synthetic time series with specific statistical properties. A random walk time series can be considered an ARMA(1,0) process with unit autoregression coefficient, i.e. the process whose realization is $\{y_t\}$ given by

$$y_t = a_1 y_{t-1} + x_t \tag{8.1}$$

where $a_1 = 1$ and x_t is unselfcorrelated.

Of interest is the case for which $a_1 \neq 1$ but lies close to 1. Let

$$a_1 = 1 - \alpha \tag{8.2}$$

where $0 \leq \alpha \leq 1$. Variance spectra for four such time series are shown in Figure 8.1.

A numerical experiment was performed in which a realization of a 2000-long synthetic time series was generated using equation (8.1) with $\alpha = .003$. It is shown graphically in the upper diagram of Figure 8.2. In the diagram, the time series has been subdivided into 20 non-overlapping intervals, each of length 100. In each interval the regression against elapsed time was found. The regression lines for each interval are plotted in the upper diagram, and the values of the correlation coefficient for each interval are plotted in the lower diagram. It can be seen that most of the regression lines fit the data very well indeed and that there is a wide range of correlation coefficient values ranging from close to -1 to close to +1 resulting from the long "period" of the quasi-cyclic behaviour of the time series.

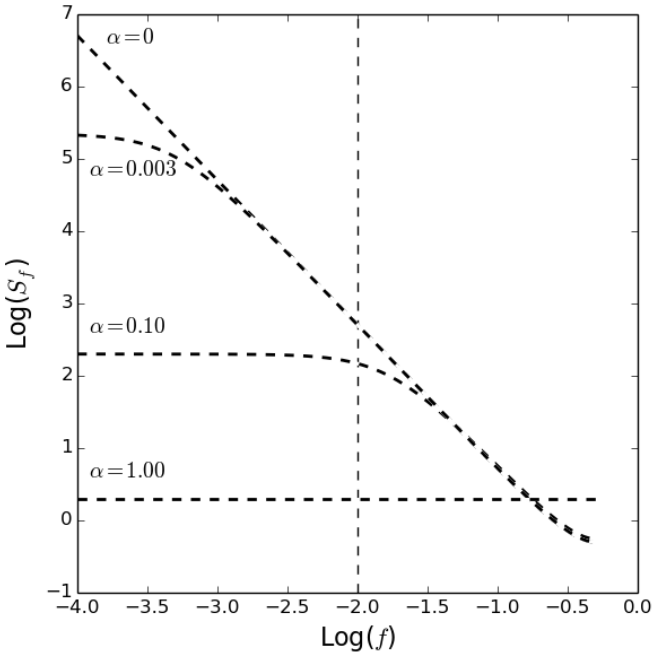


Figure 8.1: Spectra of four synthetic time series for various values of α

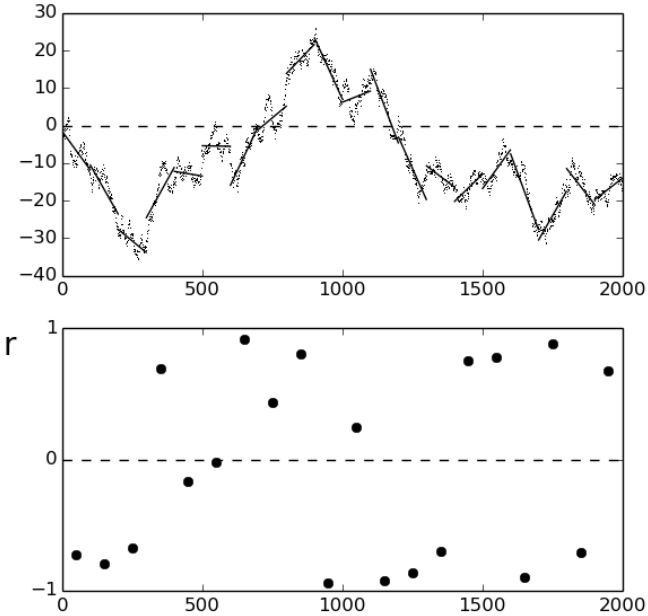


Figure 8.2: Synthetic time series of an ARMA(1,0) process with $\alpha = .003$. The time series was subdivided into twenty, 100-long, non-overlapping intervals. In each interval the correlation between value and elapsed time was evaluated. The corresponding correlation coefficients are plotted in the lower diagram.

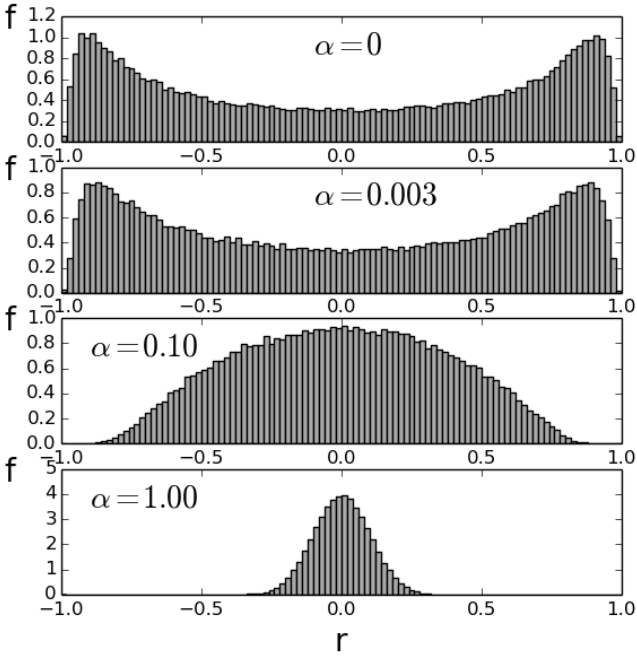


Figure 8.3: Probability distributions of the sample correlation coefficient, r, for the four time series whose spectra are shown in Figure 8.1.

In fact there can be no true correlations or regressions because the time series is purely the outcome of equation (8.1). These apparent correlations and regression fits are entirely spurious. Likewise neither is there any periodic or cyclic behaviour. The time series displayed in Figure 8.2 can properly be described as red noise.

It should also be noted that the process described by (8.1) is stationary: all of its statistical properties are independent of the time, t. Clearly the spurious correlations and spurious regressions observed in this case are not the outcome of the non-stationarity of the process. Tests for stationarity are irrelevant in assessing the validity or otherwise of the regression of a time series against the elapsed time.

What is relevant is the variance density at periods longer than

the sample time series under investigation. The reciprocal of the time series length (.01) is shown as the vertical dashed line in Figure 8.1. Intuitively we "know" about the variance spectrum to the right of this line from the data sample. There is no way to tell much about what is happening to the left of the line because this involves periods greater than the length of the data sample. By using stationarity tests such as Dickey–Fuller, we are trying to assess what is happening at infinitesimal frequencies, i.e. at infinite periods. Common sense would suggest that this is not possible.

How then is spectral shape related to spurious correlation? A further numerical experiment was performed in which the process described by equation (8.1) was used to generate 100,000 time series realizations, each 100 time units long, and the values correlated with time in each case. Histograms showing the resulting frequency distributions, f, of the correlation coefficient, r, are plotted in Figure 8.3. In each case the histogram has been normalized by the bin width so that the area under the curve is one, i.e. they are probability distributions.

In Figure 8.3 the distributions for $\alpha = 0$ (i.e. a true random walk) and $\alpha = .003$ are almost identical. They are bimodal with peaks near $|r| = 0.9$, indicating that in these cases a high absolute value of the correlation coefficient is the most likely outcome when a red noise sample time series is regressed on elapsed time. For $\alpha = .01$ the curve is broadened compared with $\alpha = 0$, indicating that high correlations are possible. For $\alpha = 0$ the time series is white and the displayed histogram is approximately Gaussian as expected. It is the Pearson r-distribution.

Correlations between time series

The above discussion concerns the spurious correlation of red noise sample time series on elapsed time. Spurious correlation also occurs when two red noise time series samples are compared. This is demonstrated by Figure 8.4. There is a similar broad scatter of correlation coefficients indicating that spurious correlation occurs in this situation as well. Indeed, it was this aspect of spurious correlation or nonsense-correlation that was first noted by Yule (1926).

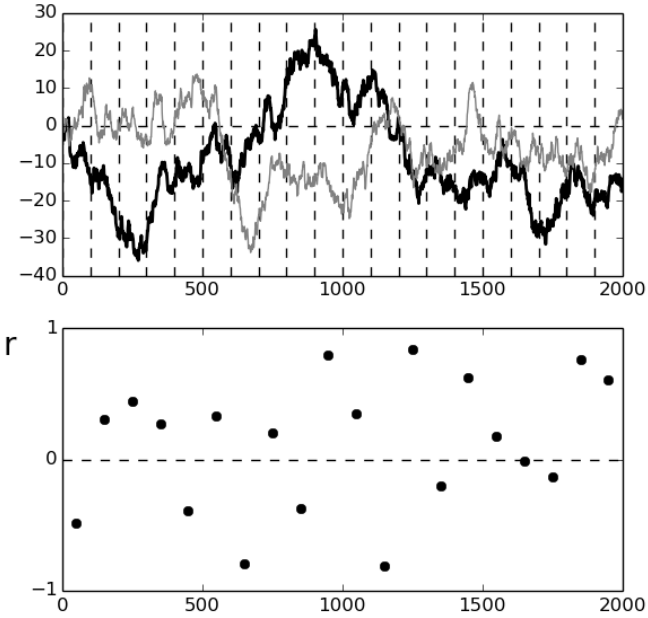


Figure 8.4: Two synthetic time series of ARMA(1,0) processes with $\alpha = .003$. The 2000-long interval was subdivided into twenty, 100-long, non-overlapping intervals shown by the vertical dashed lines. In each interval the correlation between the two time series samples was evaluated. The corresponding correlation coefficients are plotted in the lower diagram.

Chapter 9

Spurious Regression and Climate

This chapter reproduces the author's paper in Energy and Environment, (Reid, 2017).

Introduction

In recent times much has been made of the apparent rising trend in global average temperature commonly attributed to increased greenhouse gas concentrations in the atmosphere. At issue is whether this trend is a real, deterministic trend or whether the observed variations are merely the outcome of a random process.

Hasselmann (1976) proposed a stochastic model of climate variability wherein slow changes in climate are explained as the integral response to continuous random excitation by short period 'weather' disturbances. Thus intrinsic quantities such as temperature are the outcome of the integration by natural processes of quasi-random, extrinsic quantities such as heat. As a consequence, such measurements can be regarded as the outcome of a stochastic process and can be expected to exhibit a power law spectrum with negative exponent due to such integrating effects. The best known and simplest example of such a process is the 'random walk' obtained when white noise is integrated or summed. It has a power law spectrum with an index of -2 as described by (7.86).

Here we examine the HadCRUT4 data set of 166 annual values of global average temperature from 1850 to 2015 inclusive (Morice

et al., 2012).[1] There are a number of such global temperature data sets available, e.g. those from GISS, NOAA and BEST. Statistically they are almost identical. HadCRUT4 was chosen because it was the longest of these.

The data were analyzed using

1. a deterministic model in which each temperature value is regarded as a deterministic function of elapsed time plus a measurement error, and

2. a stochastic model whereby each temperature value is regarded as a function of preceding values plus an innovation.

The deterministic model

A deterministic model typically comprises a linear function of one or more functions of the explanatory variable plus a random element. In this case the explanatory variable is time. The parameters are estimated by minimizing the sum of squares of the differences between an estimate and the true values of the sample, the residuals. This is the ordinary least squares (OLS) method. It is based on the assumption that a deterministic relationship with the explanatory variable does exist and that the random elements at different times have zero mean and are independent of one another.

The HadCRUT4 time series values were fitted with a straight line by the OLS method of linear regression, i.e. the model

$$y_t = a_0 + a_1 t + \xi_t \tag{9.1}$$

was fitted to the data and is shown as the straight line in Figure 9.1(a)

The fitting of a function by OLS regression requires that the sequence of residuals $\{\xi_t\}$ in (9.1) or $\{\xi'_t\}$ in (9.1) be unselfcorrelated. Clearly that is not the case for (9.1) where a sinusoidal function or 'multidecadal oscillation' would remain after removal of the linear trend. For this reason a sinusoid of arbitrary phase was included in the model of equation (9.2):

[1]Downloaded from `http://www.metoffice.gov.uk/hadobs/hadcrut4/data/current/time_series/` on 12/4/16.

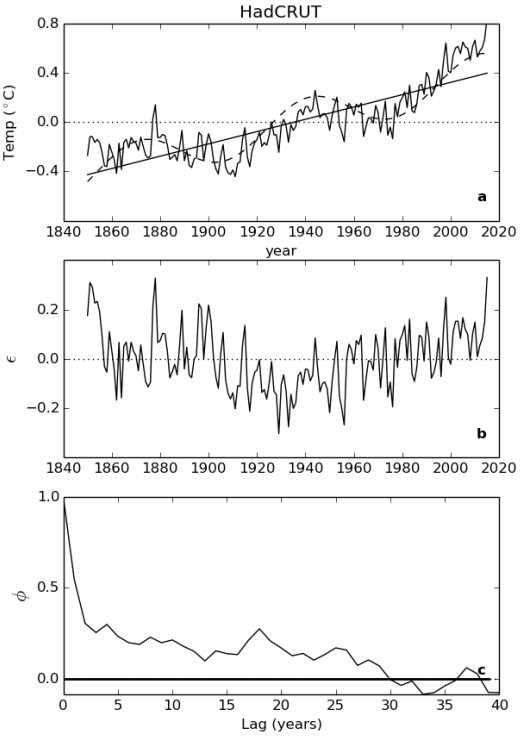


Figure 9.1: The deterministic model: (a) The HadCRUT time series. The straight solid line shows the linear trend of temperature vs. time. The dashed line shows the multiple regression fit of a linear trend plus a sinusoid. (b) Residuals from the time series of the linear trend plus sinusoid. (c) The autocorrelation function of the residuals, ϕ and the p-value of Ljung–Box Q-statistic (thick line). The p-value is zero at all lags.

	coef	std err	t	P>|t|	95.0% Conf. Int.
$\hat{a}_0$	-0.0149	0.010	-1.508	0.134	<-0.035, 0.005>
$\hat{a}_1$	0.0050	0.000	24.394	0.000	<0.005, 0.005>
$\hat{a}_2$	0.1210	0.014	8.841	0.000	<0.094, 0.148>
$\hat{a}_3$	0.0791	0.014	5.569	0.000	<0.051, 0.107>

Table 9.1: Coefficient estimates for the deterministic model of equation (9.2). Standard error, t-value, p-value and 95% confidence intervals are shown for each.

Test		Q-statistic	at lag	probability
Ljung–Box	min	92.665	3	0
Ljung–Box	max	232.052	39	0

Table 9.2: Testing the residuals of regression model of equation (9.2) for self-correlation. Minimum and maximum values of the Ljung–Box test statistics and their corresponding probabilities for a maximum lag of 40.

$$y_t = a_0 + a_1 t + a_2 \cos(\omega t) + a_3 \sin(\omega t) + \xi'_t \tag{9.2}$$

The angular frequency, ω, was chosen by trial and error and corresponded to a period of 70 years. The estimates $\hat{a}_0$, $\hat{a}_1$, $\hat{a}_2$ and $\hat{a}_3$ of the model parameters a_0, a_1, a_2 and a_3 are shown in Table 9.1.

The sequence of residuals, $\{\xi'_t\}$, is shown in Figure 9.1b and its autocorrelation function (ACF)[2] in Figure 9.1c. Even to the naked eye there appears to be a systematic positive tendency in the ACF out to lag 30. There is a statistical test which can be used to determine whether the non-zero values of the ACF at non-zero lags are significant or just due to chance. It is the Ljung–Box test (Ljung and Box, 1978). The results obtained when fitting equation (9.2) to the HadCRUT4 data are shown in Table 9.2.

The probabilities listed in Table 9.2 are so small that we can reject the null hypothesis that the non-zero ACF values are purely random. Equation (9.2) is not a good fit to the data. It can be rejected at a very high level of significance.

[2]The ACF is the covariance function normalized by dividing it by the variance

	coef	std err	z	P>|z|	95.0% Conf. Int.
$\hat{a}_1$	0.9955	0.006	176.629	0.000	<0.984, 1.007>
$\hat{b}_1$	-0.4068	0.074	-5.490	0.000	<-0.552, -0.262>
$\hat{b}_2$	-0.2276	0.067	-3.379	0.001	<-0.360, -0.096>
$\hat{c}$	0.0736	0.351	0.210	0.834	<-0.615, 0.762>

Table 9.3: Coefficient estimates for the stochastic model of equation (9.3). Standard error, z-value, p-value and confidence limits are shown for each.

The stochastic model

An ARMA(1,2) model was fitted to the HadCRUT4 time series using the *Python* statistical package: *statsmodels.tsa.arima-model.ARMA*. The package's *css-mle* option was selected whereby the conditional sum of squares likelihood was maximized and its values used as starting values for the computation of the exact likelihood via a Kalman filter.

The fitted model was thus

$$y_t = a_1 y_{t-1} + \epsilon_t + b_1 \epsilon_{t-1} + b_2 \epsilon_{t-2} + c \tag{9.3}$$

where a_1, b_1, b_2 and c were parameters to be fitted, and the $\{\epsilon_t\}$ were independent, identically distributed random variables with zero mean. The orders, $p = 1$ and $q = 2$, were found by trial and error, i.e. as the smallest values which resulted in unselfcorrelated residuals.

Note that the parameter c is similar to the coefficient a_1 in (9.1) and (9.2). Setting $a_1 = 1$ and the other coefficients in (9.3) to zero for the moment gives

$$y_t = y_{t-n} + nc \tag{9.4}$$

so that y_t becomes a deterministic linear function of the elapsed time, $n\Delta t$. For this reason c is known as the 'drift term'. It is a deterministic element in an otherwise stochastic model.

The estimates $\hat{a}_1$, $\hat{b}_1$, $\hat{b}_2$ and $\hat{c}$ of the model parameters a_1, b_1, b_2 and c are shown in Table 9.3.

The most important feature of Table 9.3 is the small value and large confidence interval of the drift term estimate, $\hat{c}$. It is not significantly different from zero. Unlike the deterministic model, stochastic

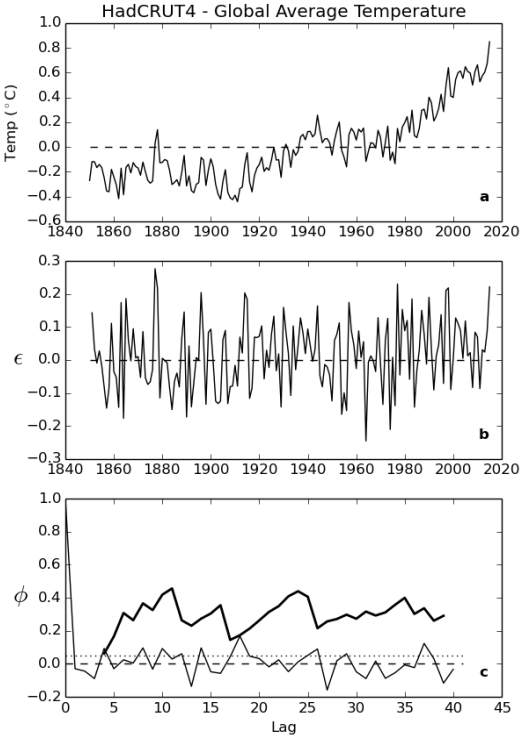


Figure 9.2: The stochastic method: (a) The HadCRUT time series. (b) Residuals from fitting an ARMA(1,2) model, (c) The autocorrelation function of the residuals, ϕ and the p-value of Ljung–Box Q-statistic (thick line).

Test		statistic	at lag	probability
Ljung–Box	min	3.093	3	0.079
Ljung–Box	max	39.843	39	0.345

Table 9.4: Testing the residuals of the ARMA(1,2) model for self-correlation. Minimum and maximum values of the Ljung–Box test statistics and their corresponding probabilities for a maximum lag of 40.

modelling indicates that there is no significant drift in the HadCRUT4 time series of global average temperature.

The sequence of residuals, $\{\epsilon_t\}$, is shown in Figure 9.2b and its autocorrelation function in Figure 9.2c. The ACF values at non-zero lags appear to be randomly distributed on either side of zero. As before, the Ljung–Box test was used to see if the null hypothesis that the residuals are unselfcorrelated can be rejected. The results are shown in Table 9.4.

None of the probabilities listed in Table 9.4 lie below the critical value of 0.05 and so there is no reason to reject the null hypothesis that the non-zero values of the ACF are due entirely to chance. The ARMA(1,2) model is a very good fit to the HadCRUT4 time series.

The Frequency Domain

Figure 9.3 shows the ARMA(1,2) spectral estimate of the HadCRUT4 time series plotted using logarithmic scales (thick dashed line). Also shown is the periodogram of the sample (thin line). The autoregressive coefficient, $\hat{a}_1$, in Table 9.3 is very close to one. In the nomenclature of the previous chapter, $\alpha = 1 - \hat{a}_1 = .0045$ – a small value indicating that the time series is red noise. This then is the reason for the f^{-2} trend at low frequencies (dashed line) and for the apparent regression with elapsed time shown by the Deterministic Model above. The HadCRUT4 time series of global average temperature is red noise, and the apparent regression is spurious.

As a consequence of the above-discussed whiteness tests confirming the absence of self-correlation of the residuals, this spectral estimate is optimal. There can be no peak, trough or trend in the spectrum other than those depicted in Figure 9.2 because this would

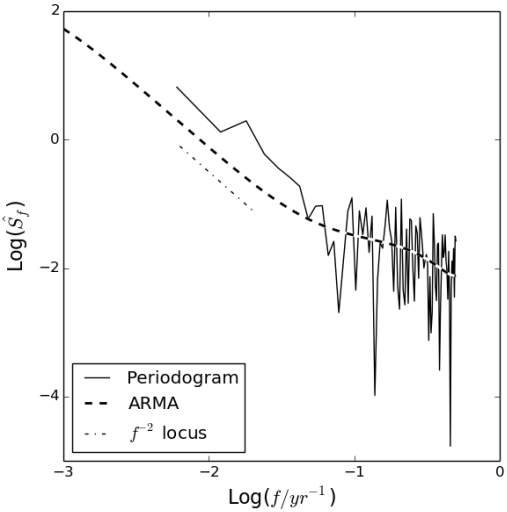


Figure 9.3: The variance density spectral estimate, $\hat{S}_f$, vs. frequency, f, of the ARMA(1,2) model fitted to the HadCRUT time series (thick dashed line). The thin line shows the periodogram of the time series.

require further poles and/or zeros in the z-plane which are not included in the ARMA model. Such extra poles or zeros, if unaccounted for, would inevitably lead to self-correlation of the sequence of residuals which would then have failed the Ljung–Box test; the 70-year "multidecadal oscillation" of equation (9.2) is therefore also spurious.

The *Python* code used to generate Figure 9.2 and Figure 9.3, *arma.py*, can be found in the Appendix.

Chapter 10

Implications

A rigorous, stochastic, discrete time theory of spectral analysis has been developed without making any assumptions about continuity and differentiability. The estimation of the variance density spectrum of a time series only applies to a range of time and frequency scales defined by the sampling period and the span of the sample itself. In this way it is possible to gain insights into underlying processes such as integration and convolution and to detect the significance of sinusoidal components by the accepted hypothesis testing methods of frequentist statistics. In particular the validity of a particular model can itself be assessed by testing the whiteness of the residuals using the Ljung-Box test.

It has been demonstrated (equation 7.38) that the periodogram of a sample which is the outcome of an ARMA process is a consistent estimator of the ensemble spectrum of the process. This contradicts the widespread assumption that the periodogram is not a consistent estimator and renders traditional methods of windowing unnecessary. In fact any such windowing is likely to subvert whiteness testing of the residuals and should be avoided.

The pitfalls of spurious regression against time and the spurious correlation of sequences with one another have been shown to be due to the concentration of variance at the low frequency end of the spectrum rather than to the non-stationarity of the sample, i.e. to the "redness" of the spectrum. Yule's "nonsense-correlations" are a property of red noise.

The process which gives rise to red noise is widely found in engineering and in nature. In electronics it occurs when noise is fed

through an integrator as with the bass control of an audio amplifier. In the natural world it occurs when matter or energy is stored, e.g. when water from random rainfall events is stored in a lake, dam or river catchment. Most importantly, it occurs for temperature whenever heat is stored according to Fourier's equation (7.77) with a spectrum given by equation (7.84). Statistically it is a particular sort of Markov process termed a "centrally biased random walk".

The small increase in global average temperature over the last 166 years is not a trend and it is not likely to continue. It is red noise. So-called "multidecadal oscillations" are also red noise. Any correlation between global average temperature and other environmental quantities is likely to be spurious.

Part III

Speculations

Chapter 11

The Ice Ages

Introduction

The final chapters, while not concerned with spectral analysis as such, demonstrate how abandoning the deterministic, "clockwork" mind-set of Newtonian calculus leads to a fruitful description of environmental phenomena as the outcome of stochastic processes. Some speculations about ocean and mantle dynamics based on this approach are given.

This chapter shows how a model with both stochastic and deterministic components leads to a mechanism which accounts for rapid ice-age terminations.

The idea that the Earth was once much colder than at present has been debated by scientists since the early 18th century. In 1824, Danish-Norwegian geologist, Jens Esmark, appears to have been the first to propose that the Earth had undergone climatic changes and that these were caused by changes in the Earth's orbit. Serbian astrophysicist, Milutin Milankovitch, was the first to "put the numbers in" to Esmark's theory (Milankovitch, 1941), and these orbital variations are now known as "Milankovitch Cycles". In more recent times, Hays et al. (1976) used oxygen isotope ratios in ocean sediment cores to show that Esmark, Milankovitch (and many others) were correct in this proposal. Hays et al. conclude:

1. *Three indices of global climate have been monitored in the past 450,000 years in Southern Hemisphere ocean-floor sediments.*

2. *Over the frequency range* 10^{-4} *to* 10^{-5} *cycles per year, climatic*

variance of these records is concentrated in three spectral peaks at periods of 23,000, 42,000 and approximately 100,000 years. These peaks correspond to the dominant periods of the earth's solar orbit, and contain respectively about 10, 25 and 50 percent of the climatic variance.

3. *The 42,000-year climatic component has the same period variations in the obliquity of the earth's axis and retains a constant phase relationship with it.*

4. *The 23,000-year portion of the variance displays the same periods (about 23,000 and 19,000 years) as the quasi-periodic precession index.*

5. *The dominant, 100,000-year climatic component has has an average period close to, and is in phase with, orbital eccentricity. Unlike the correlations between climate and the higher frequency orbital variations (which can be explained on the assumption that the climate system responds linearly to orbital forcing), an explanation of the correlation between climate and eccentricity probably requires and assumption of non-linearity.*

6. *It is concluded that changes in the earth's orbital geometry are the fundamental cause of the succession of Quaternary ice ages.*

7. *A model of future climate based on the observed orbital-climate relationships, but ignoring anthropogenic effects, predicts that the long-term trend over the next several thousand years is toward Northern Hemisphere glaciation.*

Since that paper was published, there have been numerous incremental improvements in pinning down the time scales of ocean sediment sequences. Most importantly isotope ratios in ice cores, which act as proxies for global temperature, have become available from Greenland and the Antarctic ice cap. These not only tally with the ocean sediment time series but, with one or two exceptions, support the above conclusions.

The 100,000-year spectral peak contains half the variance and certainly dominates graphs of the spectra unless they are plotted on logarithmic scales. Hays et al. were well aware that variations in the eccentricity of the earth's orbit were not sufficient to give rise to such a comparatively strong peak, which is why they appealed to

an unknown non-linear mechanism to account for it. Their use of "approximately" indicates that they were also well aware that visual inspection of their data in the time domain reveals no 100,000-year cycle. Interstadials (i.e. interglacials) are separated by 80,000 years or 120,000 years, multiples of the obliquity period. Others have noted these discrepancies, e.g. Maslin and Ridgwell (2005).

Here we show that this peak is not statistically significant. Its large size is due solely to the integrating effect of Fourier's heat equation (7.77), which results in a power law temperature spectrum with an index of -2 as described by (7.86).

It is also not the case that higher frequency variations "can be explained on the assumption that the climate system responds linearly to orbital forcing": the rapidity of ice age terminations compared with ice age onsets implies that some non-linear mechanism must be involved. Elsewhere in their paper, Hays et al. recognise this but still have difficulty accounting for the 100,000-year peak.

Orbital Parameters

Temperatures, T, derived from EPICA Dome C ice cores (Jouzel et al., 2007) covering the period from 490 kyr to the present were examined. For spectral analysis purposes it was desirable that time series be created comprising a sequence of values sampled at equal intervals of time. However, as a result of ice flow behaviour, more recent ice core strata from near the surface are sampled much more densely in time than are deeper, older strata. Time series were constructed by partitioning the data into sampling intervals of equal length. Values lying within each sampling interval were then averaged and the mean assigned to the mid-time of that sampling interval. The time interval was chosen as 1 kyr with 491 samples spanning the period from 490 kyr BP to the present.

The *Python* code used to generate Figure 11.1 and Figure 11.2, *OPspectra.py*, *OPlist.py* and *DailyInsolation.py*, can be found in the Appendix.

Figure 11.1**a** shows 1 kyr averages of EPICA Dome C Ice Core Deuterium temperature estimates due to Jouzel et al. (2007). Figure 11.1**b** shows the first differences of the sequence shown in Figure 11.1**a**, ΔT. Its mean, m, and $\pm 2\sigma$ levels are shown as dashed

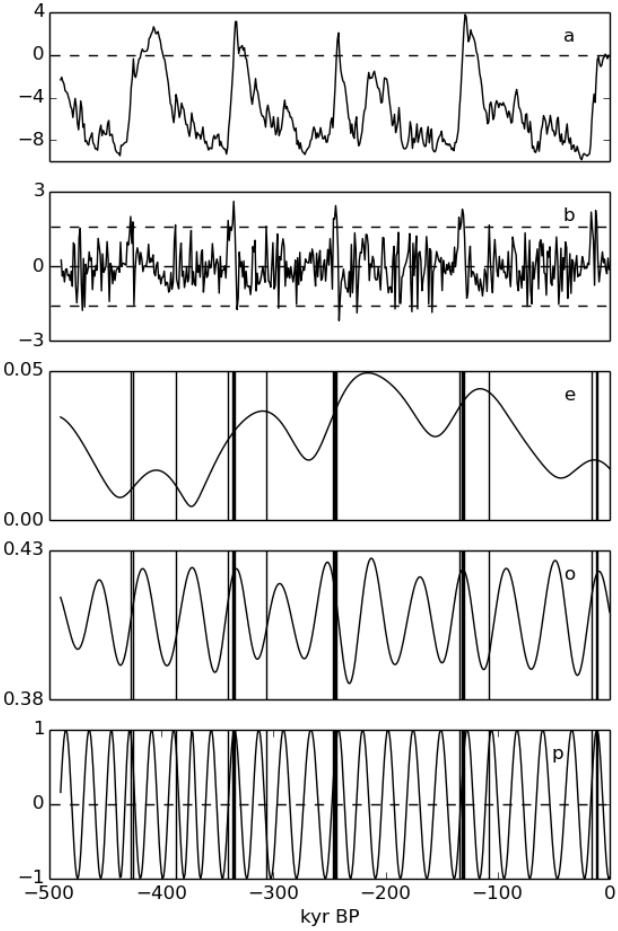


Figure 11.1: **a** 1 kyr averaged temperatures, T, derived from EPICA Dome C ice cores. **b** First differences, ΔT, showing mean, m and upper and lower 2σ values (dashed lines). **e** Calculated eccentricity of the earth's orbit. **o** calculated obliquity of the earth in its orbit. **p** Calculated precession angle of the earth in its orbit. Vertical lines in **e**, **o** and **p** show the times at which ΔT exceeds $m + 2\sigma$.

lines. The temperature differences exceed the upper 2σ level during terminations when the temperature is rapidly increasing.

Also shown are the secular variations in the Earth's orbital parameters derived from the data and the *MATLAB* programs originally due to Berger (1979) and Berger and Loutre (1991) and later used by Huybers (2006) and Huybers and Eisenman (2006). The vertical lines in these three figures represent the time at which the temperature differences of Figure 11.1**b** were more than 2σ above the mean, i.e. the times of terminations as defined here. On four out of five occasions in Figure 11.1**o**, terminations occurred at times of increasing obliquity, but the one near 250 kyr BP did not. On the other hand, in Figure 11.1**p** the three most recent terminations occurred at times of increasing precession, but the earliest two did not.

Figure 11.2 shows spectra in the form of periodograms of the five time series shown in Figure 11.1 with vertical scales displaced arbitrarily for ease of comparison.

Both the obliquity and precession spectra have multiple peaks. In this way they resemble the spectra of amplitude modulated (AM) and frequency modulated (FM) radio signals. Radio engineers refer to the largest peak as the "carrier" and the other peaks as "side-bands".

The spectrum of ΔT in Figure 11.2 is almost flat implying that the time series is unselfcorrelated. If we take as null hypothesis the hypothesis that the population time series of ΔT is indeed unselfcorrelated as discussed above then confidence limits can be placed on each ordinate value. The horizontal dashed lines in Figure 11.2 are the upper and lower 99.9% confidence limits for the ΔT spectrum computed using (7.35).

It can be seen that the carrier peaks in the obliquity and precession spectra give rise to significant peaks in the ΔT spectrum at the 99.9% level. Peaks corresponding to the side-bands are not present. The statistical significance of the obliquity and precession peaks implies that the temperature time series does indeed have deterministic sinusoidal components: it is not purely stochastic. On the other hand, the so-called "eccentricity peak" near $(f) = 10^{-2}$ is not significant and can be regarded as noise, as can other peaks in the ΔT spectrum at higher frequencies.

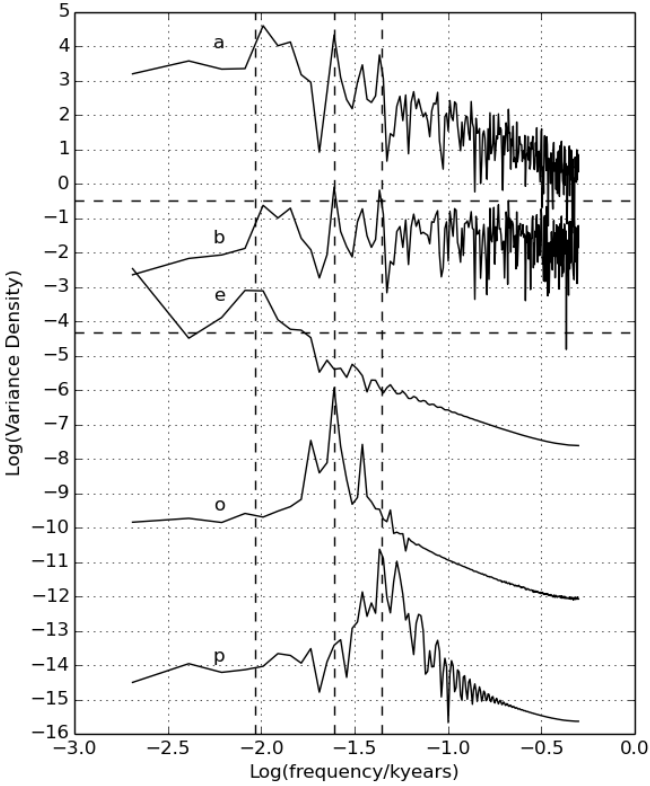


Figure 11.2: Periodogram spectra of the five time series depicted in Figure 11.1: **a** 1 kyr averaged temperatures, T, **b** First differences, ΔT, **e** eccentricity, **o** obliquity, **p** precession. Vertical dashed lines show the frequencies of the major peaks in the orbital parameter spectra at .00922, .0246 and .044 kyr^{-1} respectively. Horizontal dashed lines show the upper and lower 99.9% confidence limits of the ΔT spectrum.

A mechanism for rapid terminations

Huybers (2006) has pointed out that while the intensity of summer insolation is primarily controlled by the .044 kyr^{-1} precession frequency, early Pleistocene glacial cycles occurred at the .0246 kyr^{-1} obliquity frequency because glaciers are sensitive to insolation integrated over the duration of the summer. Nevertheless this cannot be the whole story because, as can be seen in Figure 11.1, Termination III ($\sim$250 kyr BP) occurred at a time of declining NH obliquity but increasing precession. Huybers pointed out that ice sheet ablation is only indirectly related to insolation via air temperature and will not occur when air temperature is below the freezing point of water. For this reason he proposed that insolation only be realistically considered as forcing at a given latitude when it exceeded a threshold sufficient to raise the air temperature above 0°C. Huybers proposed a threshold of 275 W/m^2 at 65°N. Introducing this threshold had the effect of bringing the precession cycle back into play so that ice sheet ablation was no longer dependent on the obliquity cycle alone. This led to an improved fit between glacial cycles and the Earth's orbital parameters.

However, this does not account for the remarkable rapidity of the terminations: ice sheets that took more than 40,000 years to form collapsed within 4,000 years or so as can be seen in Figures 11.1a and 11.1b. Quasi-sinusoidal orbital components, while related to temperature to some degree in both phase (Figure 11.1) and frequency, (Figure 11.2) gave rise to the saw-tooth temperature variations of Figure 11.1a. This requires some sort of positive feedback and a mechanism for storing heat or enthalpy from one summer to the next. Ablation by solar radiation and raised air temperatures alone does not provide such a mechanism. Furthermore, there is a need to account for the fact that terminations did not occur on every occasion when conditions were suitable. There is a random element: increased insolation only seems to increase the probability of ice sheet collapse rather than act as a fully deterministic cause.

Water is an atypical liquid in having its maximum density at 4.0°C – well above its freezing point. This implies that a body of water between 0°C and 4.0°C will behave much like a solar pond in that the upward convection of heat is suppressed. In solar ponds this condition is set up and maintained by adding salt to maintain the density

gradient, but in melt ponds it happens naturally. In solar ponds incoming solar radiation heats most of the top 2 m. There is a surface 'mixed layer' due to wind induced turbulence, and below 2 m heat is preserved. The density gradient suppresses convection, and conduction is negligibly small – $\sim$1 W/m^2 compared with $\sim$100 W/m^2 from incoming radiation. In calm conditions the water heated to 4.0°C by radiation might be expected to sink to the bottom of the pond due to 'inverse convection'. To first order this will commence spontaneously when the Rayleigh number exceeds a critical value of 1700 for which the length scale is $\sim$5 metres. In summer, wind induced turbulence will predominate, and the melt lake will be well mixed to the bottom because density variations will be too small to induce stratification. In winter, a surface cover of ice will form rapidly allowing turbulence to die out and the water to become stratified. At depth the water will remain liquid because its latent heat cannot escape.

On sea ice, the warmed water finally thaws through to the base of the ice and mixes with the ocean beneath so that melt ponds on sea ice only last for a single season. In a melt pond over land, however, radiatively heated water will mix down and melt the ice beneath it until bedrock is reached. From then on extra heat will ablate the edges of the pond, and the pond will become larger. Each summer more heat will be added, and the pond will grow in area in proportion to the solar radiation absorbed which is proportional, in turn, to the pond's current area. Small melt ponds will thus grow exponentially until they join up to form lakes and, conceivably, lakes of continental scale. Ultimately a final ice dam will be broken and the lake's contents will pour into the ocean carrying the remaining ice with it and leaving bare earth behind.

This is the ideal case. It may not always happen like this. Firstly ponds of sufficient depth must form: ponds that are too shallow will lose all their heat by conduction during the winter. In steep terrain, underwater rivers or 'moulins' can form and drain a pond so that any stored heat finishes up in the ocean, so breaking the positive feedback cycle. There is a random component in whether melt ponds will start in the first place, whether any will deepen sufficiently to retain their heat through the first few winters and whether any will develop to be large enough for their collapse to have global consequences.

Here then is a mechanism which accounts for rapid ice age terminations:

- It accounts for storage from summer to summer,
- It has a positive feedback component whereby the area of a melt pond increases exponentially, i.e. at a rate proportional to its current area, and
- There is random component controlling whether sufficient melt ponds will form in the first place.

Water surface absorption

Another feature of melt ponds which makes them relevant to ice sheet collapse is their low albedo compared with surrounding ice and snow, see, for example, Flocco et al. (2016).

When light from a point source shines on a reflective surface at the boundary of two transparent media, part is reflected and part refracted according to Snell's Law, viz.:

$$n_1 sin(\theta_1) = n_2 sin(\theta_2) \tag{11.1}$$

where θ_1 and θ_2 are the angles of incidence and refraction respectively.

The amplitude of the transmitted rays depend solely on the refractive indices of the two media, n_1 and n_2. The Fresnel reflection coefficient for p-polarized and s-polarized light, r_{12p} and r_{12s}, are given by

$$r_{12p} = \frac{tan(\theta_1 - \theta_2)}{tan(\theta_1 + \theta_2)} \tag{11.2}$$

$$r_{12s} = \frac{sin(\theta_1 - \theta_2)}{sin(\theta_1 + \theta_2)} \tag{11.3}$$

The fraction of incident power reflected, R, is given by

$$R = r_{12p}^2 + r_{12s}^2 \tag{11.4}$$

This is, in effect, the albedo of a water surface. For most solar elevation angles it will be less than for the surrounding ice and snow. The fraction of the radiance transmitted and ultimately absorbed is $A = 1 - R$. The solar irradiance per unit horizontal area is $A(h).sin(h)$ (where $h = \pi/2 - \theta_1$ is the solar elevation angle). This can be integrated over time to give the water surface absorbed insolation. A more rigorous discussion of this topic is given by Malinka et al. (2017).

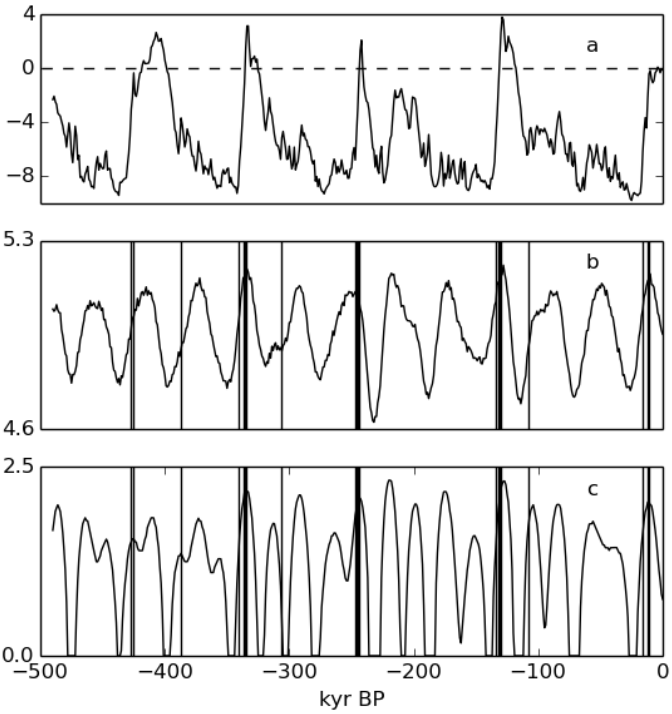


Figure 11.3: **a** Time series of 1 kyr averaged temperatures. **b** Annual insolation at 65° N with threshold $\tau = 275 W/m^2$ (Huybers, 2006). **c** Annual water surface absorbed insolation at 65° N with threshold, $\tau = 470 W/m^2$. Vertical bars show times when ΔT exceeded $m + 2\sigma$.

The *MATLAB* programs of Huybers and Eisenman (2006) for calculating daily and yearly insolation were modified to calculate daily water surface absorption of solar radiation by multiplying the irradiance at a given time and latitude by the function $A(h)$ prior to integrating. Their algorithm for rejecting those days in which the total daily insolation fell below a particular threshold, τ, was retained. The quantity calculated in this way is the water surface absorbed insolation, which was summed over the year to give annual water surface absorbed insolation (AWSAI) in units of J.m^{-2}.

Time series of AWSAI were computed for a number of high northern latitudes and threshold values, τ. These AWSAI time series were not very different from the corresponding insolation time series of Huybers for the same latitudes and thresholds; evidently the shape of the A(h) curve is not particularly important. However, the resulting time series were strongly dependent on the value of the threshold, τ, chosen. Huybers chose a value of $270W/m^2$ below which air temperature would be below the freezing point of water and no ablation of ice could be expected. The threshold value chosen here, $470W/m^2$, is almost double that. This higher threshold is justified by the fact that precipitation must be in the form of rain so that ponds can form on the ice in the first place and then not freeze overnight in the summer.

Figure 11.3c shows the time series for an optimum combination of latitude (65°N) and threshold ($\tau = 470W/m^2$). Time series of 1 kyr averaged temperatures and Huybers (2006) insolation time series are also shown for comparison in Figure 11.3a and Figure 11.3b, respectively. As before, vertical dashed lines indicate times when ΔT exceded $m + 2\sigma$.

Visual examination of Figure 11.3 shows that AWSAI is generally a better fit to the temperature time series than is insolation alone, particularly near the problematic interstadial at ∼250 kyr BP.

More detailed plots of temperature for 60 kyr during and following each of the last five Ice Age Terminations are shown in Figure 11.4. The temperatures plotted are the raw data from the EPICA Dome C ice cores. Rescaled plots of AWSAI are also shown as a series of lobes when the AWSAI was above the 470 W/m^2 threshold (dashed curves).

These plots show some interesting features, viz.:

1. Termination temperature peaks were 1 to 3 kyr wide and followed by a plateau in temperature such as the Holocene.

2. In each case the temperature peaked near the time of maximum AWSAI of the Termination Lobe.

3. Following Terminations I to IV, in each case temperature began to fall more rapidly after the end of the Termination Lobe.

4. Following Terminations I to IV, in each case temperature fell by about 5°C between the end of the Termination Lobe and the beginning of the subsequent lobe.

Melt ponds and ice sheet dynamics

The first part of this chapter is based on a paper submitted to a peer-reviewed journal. The paper was rejected on the advice of the Associate Editor who reasoned as follows:

1. *Less than half of melt ponds in Greenland store water between seasons with the remainder emptying by either vertical or lateral drainage (Selmes et al., 2013).*

2. *Melt ponds on grounded ice do not form randomly. Their location and size is controlled by the topographic depressions in which they form. These depressions, in turn, are controlled by the topography of the underlying bedrock (Echelmeyer et al., 1991).*

3. *Since ice flows, any ice which has been subject to enhanced ablation due to the presence of the melt-pond is advected downstream. On Greenland, the ice on which a lake sits is renewed approximately every 10 years. This, together with (2) means that it is impossible for exponential lake growth to occur through enhanced ablation on grounded ice.*

This view, that the ice sheet dynamics of present day Greenland provides a suitable model for the behaviour of the Ice Sheets during the last five Terminations, is not justified; they are very different scenarios. In present day Greenland, large outlet glaciers, which are restricted tongues of the ice sheet, move through bordering valleys around the periphery of Greenland to calve off into the ocean. The best known of these outlet glaciers is the Jakobshavn Glacier discussed

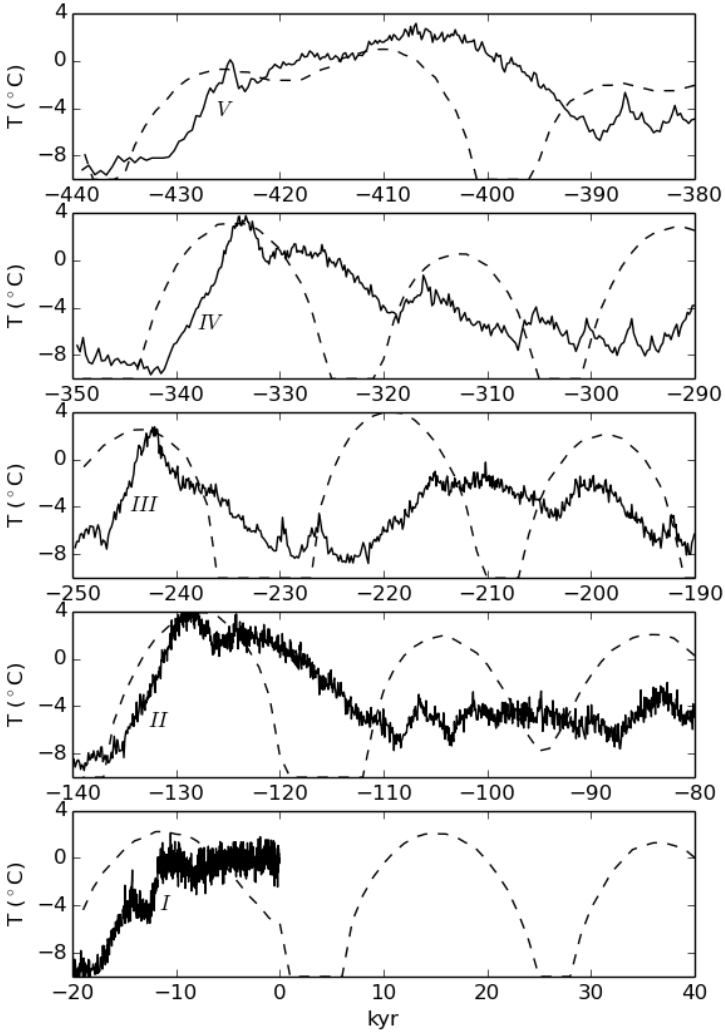


Figure 11.4: Epica Dome C Ice Core Deuterium temperature estimates (solid line) and annual water surface absorbed insolation at 65° N (dashed line) for the intervals including and following the five last Ice Age Terminations (V, IV, III, II and I).

by Echelmeyer et al., which, at its terminus, flows at speeds of 7000 m/yr. A comprehensive study of ice streams by Margold et al. (2015) has shown that flow speeds over most of the area of the Laurentide Ice Sheet were close to zero at the time of the Last Glacial Maximum (Margold et al., Figure 2). Given the accepted dynamics of ice flow, these regions must have been almost flat and horizontal in contrast to the steep Greenland ice flows.

Furthermore, the present day is a time of decreasing solar declination and water surface absorption. In contrast, Terminations occurred at times of increasing solar declination when old, static, horizontal ice was increasingly exposed to solar radiation, so allowing new melt ponds to form when rain fell during the summer months.

Conclusions

The catastrophic collapse of ice sheet melt lakes appears to have first been proposed by Bretz (1923) who attributed the scouring of the Scablands of Washington State to the draining of Lake Missoula in western Montana. Since then many other topographic features have been seen to be associated with such dramatic collapses, see, for example, Patton et al. (2017).

That melt ponds and melt lakes play a major role in amplifying the effects of increased Northern Hemisphere insolation has hitherto gone unnoticed. These catastrophic climate disruptions are surely the most important feature of the Late Pleistocene and, in a sense, define it.

The succession of Ice Ages and Interglacials, of Stadials and Interstadials, is a deterministically forced stochastic process, very much like present-day weather but at longer time scales. There is little evidence of other deterministic cycles forcing climate between the annual cycle and the precession cycle, i.e. between one year and 23,000 years in period: a span of four orders of magnitude. Supposed "multidecadal cycles" are nothing more than red noise.

According to Figure 11.4, under this model, on a time scale of millennia, the next Ice Age has already commenced. The present "Holocene" lobe of AWSAI comes to an end in 1200 years time, and the subsequent lobe commences 5000 years after that. Based on previous interglacials, we can speculate that global average temperature is likely to be 5°C cooler in 6000 years time.

Software

The *Python* programs used to generate the diagrams found in this chapter, *OPspectra.py, OPlist.py* and *DailyInsolation.py*, are listed in the Appendix.

Chapter 12

Ocean Waves

On Christmas Eve 2001, 400 angry people and their vehicles were left stranded on the wharf at Georgetown, Tasmania. They had been waiting to board the ferry to Melbourne, Victoria on the other side of Bass Straight, a six hour voyage on the new wave-piercing catamaran, *Devilcat*, designed and built in Tasmania, and, at the time, one of the world's fastest and most economical vehicular ferries. *Devilcat* had arrived from Victoria and discharged its passengers and cargo but could not make the return trip.

Why not? There were no mechanical or personnel problems, and the sea was "like a millpond" according to one eye-witness who had just arrived on the vessel.

According to Australian Marine Safety Authority (AMSA) regulations, the ferry was not permitted to sail if the *Significant Wave Height*[1] in Bass Strait was greater than two metres. Since there was no wave measuring equipment deployed in Bass Strait at the time, the Significant Wave Height was calculated by means of a numerical wave model from meterorological wind velocities. The world wave model in question was run by NOAA, the US National Oceanic and Atmospheric Administration in Boulder, Colorado. According to the NOAA model, the Significant Wave Height in Bass Strait on 24 December 2001 was greater than the AMSA threshold value of two metres and so *Devilcat* was not allowed to sail. Direct observations by several hundred passengers and crew were irrelevant. They were not qualified observers. A computer on the other side of the world knew better.

[1]The Significant Wave Height is defined by $H_s = 4\sigma_z$, where σ_z is the standard deviation of sea surface height.

Not long after this, unsurprisingly, the *Devilcat* service went out of business. The vessel was sold to Bay Ferries of Canada and operated on the Bay of Fundy for many years.

The numerical prediction of ocean waves has always been problematic. This is largely because there is not a good theory of wave generation. The propagation of swells is well understood and can be described with great precision, but the development of wind-sea is not quite so straightforward.[2]

Pierson and Moskowitz revisited

In a classic case of the theoretical tail wagging the observational dog, the breakthrough came with Pierson and Moskowitz (1964) in a paper entitled "A proposed spectral form for fully developed wind seas based on the similarity theory of S. A. Kitaigorodskii". This paper became the *sine qua non* of all future discussion of ocean wave spectra. It interpreted the observations of wave spectra at a variety of wind speeds made by Moskowitz (1964) in the light of a new theory of waves proposed by the Russian mathematician who made the assumption that the variance density spectrum of a wind-generated sea is a function of only four variables, viz.:

$$S(f) = F(f, g, U_+, X) \tag{12.1}$$

where f is the frequency, g is the acceleration due to gravity, U_+ is the Charnock–Ellison friction velocity (of the wind) and X is the fetch. Pierson and Moskowitz used the concept of a *fully developed sea* to eliminate the dependence on fetch, X, from their equations to derive a dimensionless spectrum, $S'(f)$, as a function of dimensionless frequency, f', viz.:

$$S'(f) = \frac{S(f)g^3}{U^5} \tag{12.2}$$

$$f' = \frac{Uf}{g} \tag{12.3}$$

where U is the measured wind velocity. Moskowitz's observed spectra are replotted in dimensionless form in Figure 12.1.

[2] *Swell* describes waves propagated from distant storms in which white-capping has ceased. *Wind-sea* describes shorter wavelength waves being generated by wind and accompanied by white-capping.

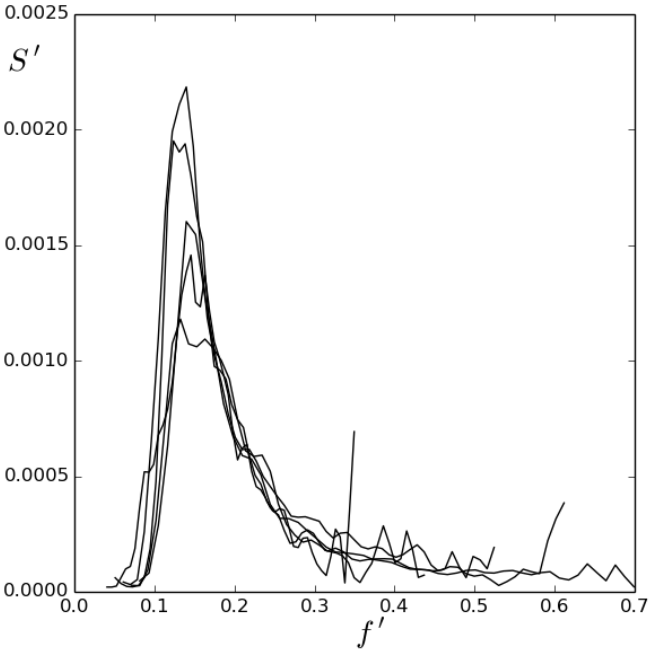


Figure 12.1: Mean dimensionless spectra at wind speeds of 20, 25, 30, 35 and 40 knots, listed by Moskowitz (1964)

Certainly Moskowitz's spectra proved to be remarkably similar when displayed in this way. Pierson and Moskowitz fitted a curve of the form

$$F(f') = \frac{ABe^{-B/f}}{f^5} \tag{12.4}$$

and this, the *Pierson and Moskowitz spectrum*, became the accepted canonical form with some minor additions and modifications from time to time.

Some issues

There are, however, some issues with the Pierson and Moskowitz spectrum and its variations. Chief among them is the assumption made in (12.4) that variance density in the high frequency tail of the spectrum is inversely proportional to the fifth power of the frequency, i.e. that it is a power law spectrum with an index of -5. As they state in their 1964 paper:

> There is increasing theoretical support for the appearance of f'^5 in the denominator. The original theory of Phillips (1958) suggests that the waves, if high and if having a spectrum with $f'^{4.5}$ or f'^4 in the denominator would break and dissipate energy. The analysis of Kitaigorodskii supports this.

As illustrated by equation (5.3), a convenient method of finding power law trends in spectra empirically is simply to plot them using logarithmic axes. The spectra of Figure 12.1 are replotted in Figure 12.2. A strong power law trend is clearly evident in the range $-0.8 < logf' < -0.5$.

But the index is not -5. The index is -3.

For more than fifty years, fluid dynamicists have continued to believe that wind-seas exhibit a fifth power spectrum when, in fact, the very data on which this theory is based clearly exhibits an inverse cube spectrum. Once again theory trumps experiment in Fluid Dynamics.

There are other issues. Kitaigorodskii's scale free wave spectra are based on the idea of "similarity" whereby, like fractals, waves are assumed to look the same at every scale. This is true of some waves, for example swells when there is no white-capping. But white-capping is an ever-present phenomenon in the case of wind-seas and white-capping does in fact vary according to scale. For large waves,

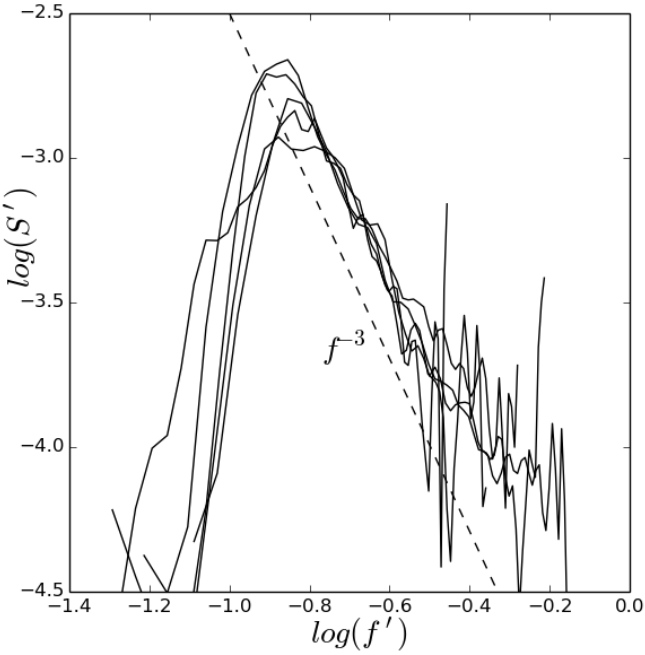


Figure 12.2: The mean dimensionless spectra of Figure 12.1 replotted on logarithmic scales

white-capping looks very different from that of small waves because of the spume that is formed. Individual bubbles can be seen in the spume from small waves, but spume at the scale of larger waves looks more like a white continuous fluid. This is well known to cinematographers attempting to film storms at sea using model ships in wave tanks. The larger (and more expensive) the model, the more realistic is the shoot. At smaller scales molecular forces such as surface tension become important and alter the appearance of the spume in a way that is obvious to a cinema audience.

In 1991 I presented a paper (Reid, 1992) to the IUTAM Symposium on Breaking Waves held in Sydney, Australia.[3] I related details of an experiment I had carried out using the wave tank at the Australian Maritime College in Launceston, Tasmania. The experiment demonstrated unequivocally that when waves of different frequencies interact to cause white-capping, energy and momentum are lost only from the higher frequency wave. As this process continues, it leads to the frequency down-shifting of the spectral peak commonly observed in wind-seas in the open ocean. Furthermore, it leads to similar spectra to those observed in the open ocean.

Although this experiment was performed in a wave tank in the absence of wind, it illustrated quite clearly how wind waves grow. This can be observed with the naked eye in an enclosed body of water such as a pond or reservoir. At first the wind generates ripples with a wavelength of about 15 cm. The whole of the pond is covered in ripples of similar amplitude and wavelength. As the wind increases, some ripples start to break. Once this happens the downwind ripples become larger and longer in wavelength than the upwind ripples. Frequency down-shifting occurs, but it does not occur until after white-capping has commenced. Wave-breaking is fundamental to the growth of a "wind sea" even on a pond.

The paper was not well received and was indignantly refuted by a leading wave modeller, who was given special permission by the Chair to take the platform following my presentation. He argued that my conclusion, that frequency down-shifting occurs because of wave breaking, could not possibly be true. If it were, it would mean that decades of work on ocean wave modelling had been wasted. To many of those present, this was clearly unthinkable.

Conclusion

Similarity theory rests on an assumption of continuity and the absence of wave breaking. As Phillips pointed out in the above quote, waves having a spectrum with a power law index of -4.5 or more will "break and dissipate energy". That is precisely the point. That is what happens in a wind-sea. Waves regularly break and dissipate the

[3]Downloadable from `https://apps.dtic.mil/dtic/tr/fulltext/u2/a264196.pdf`.

energy that they have gained from the wind. Wave dynamicists are in denial about the role of wave-breaking in the formation of ocean wave spectra.

Wave breaking involves changes in entropy and so cannot be handled by the continuous equations of fluid dynamics. Its role in the development of wind-seas is therefore ignored.

This is the Fluid Catastrophe in action.

Chapter 13

Volcanoes under the ocean

Ocean currents are driven by tides, by wind and by variations in water density caused by variations in salinity and temperature, the *thermohaline* circulation. Their paths are also controlled by the Earth's rotation in the same way that winds in the atmosphere rotate about low pressure systems. These circular current paths are called gyres.

The Gulf Stream, the Kurashio (Japan) and the Aghulas (South Africa) are well known ocean currents. The flow of ocean currents are measured in Sverdrups, one Sverdrup (Sv) being a million cubic meters of water passing a fixed point in one second. The flow of the Amazon is about 0.5 Sv. The biggest ocean current, the Antarctic Circumpolar Current has a flow of 100 Sv where it runs through Drake Passage between South America and Antarctica. The flow of the ACC is equal to that of two hundred Amazons.

Ocean currents often determine local climate. Deserts, such as the Atacama Desert of Peru, tend to be flanked by cold currents. Offshore currents flowing towards the equator are generally colder and give rise to less evaporation and less precipitation, whereas warm, pole-ward currents lead to enhanced evaporation and higher rainfall. On the other hand, pole-ward currents are generally poor in life-supporting nutrients compared with equator-ward currents. The Labrador current flows between Greenland and Canada and carries nutrient southward to nourish the cod fishery of the Grand Banks of the North Western Atlantic.

It is commonly assumed by oceanographers that world ocean currents have remained fundamentally unchanged since the last ice-age termination 11,000 years ago. Historical and archaeological evidence

indicates that this was not the case.

Greenland and Easter Island

Isotope analysis of bone samples from a Greenland Viking colony from the late 10th to mid 15th century show that their diet changed dramatically over this time. At the time of Eric the Red around 1000 AD they ate predominantly terrestrial food, but by the end of the settlement period around 1450 AD their diet comprised predominantly marine food (Arneborg et al., 1999). This may reflect changing climatic conditions with the onset of the Little Ice Age. It could also indicate that fewer fish were available in the early years of settlement. Both phenomena are attributable to changes in the circulation of the North Atlantic.

A similar puzzle concerns the consumption of porpoise by the inhabitants of Easter Island. According to Jared Diamond in his book *Collapse* (Diamond, 2005):

> Comparison of early garbage deposits with late prehistoric ones or conditions on modern Easter Island reveals big changes in ... initially bountiful food sources. Porpoises and deep ocean fish like tuna, virtually disappeared from the islander's diet.

According to J. Loret in Loret (2003):

> Throughout Polynesia archaeologists have found that 90 percent of middens consist of fish or shellfish. On Easter Island however, from the period of 900–1300 A.D. one third of the bones were of porpoises. Yet nowhere in Polynesia do the bones of porpoises account for more than one percent. These marine mammals had to be hunted far offshore indicating that their craft had to be sufficiently seaworthy and constructed from large trees.

Both authors are reaffirming the accepted narrative that Easter Islanders destroyed their own livelihood by cutting down all the big trees, from which they constructed their canoes.

Maps of the present-day distribution of the four major species of porpoise indicate that there are no porpoise within thousands of kilometres of Easter Island. It seems unlikely that early Easter Islanders could have travelled so far to hunt these animals; there must have been porpoise closer to the island. Today the ocean around Easter Island is very poor in nutrients. There is no marine ecosystem sufficiently large to support a top predator such as porpoise; there is insufficient nutrient available to generate such an ecosystem.

Once again this implies that ocean circulation between 900 and 1300 AD differed from that of the present day. These events both preceded a time of warm weather in Europe known as the Mediaeval Warm Period (MWP) and suggest that changes in ocean circulation accompanied and perhaps even caused the MWP.

Finally, consider the Maya, a civilisation which flourished on the Yucatan peninsula of present day Mexico from about 400 to 900 AD. Oxygen isotope ratios and sulphur abundances in lake sediments from Yucatan suggest that the cause of the Mayan Collapse was sustained drought.

The Mayan drought immediately preceded the MWP.

Subaqueous volcanism

But what caused these changes? A promising candidate is subaqueous volcanism.

Subaqueous volcanism refers to volcanic activity beneath the surface of the ocean, as distinct from sub-aerial volcanism, which refers to volcanic activity on land. It is believed that subaqueous volcanism accounts for eighty five percent of all volcanic activity, sub-aerial fifteen percent (White et al., 2003) so that volcanic activity is five times more common under the sea than on land. The rapid formation of steam when water flash-boils on contact with hot magma is the reason for the explosive power of sub-aerial volcanoes. On the other hand, the large pressure at the bottom of the ocean prevents water from boiling and forming steam. As a consequence ocean floor volcanoes are much less likely to be explosive than are those on land: five percent compared with eighty percent.

Instead, although some subaqueous volcanos are explosive, they mostly take the form of hydrothermal vents (HTVs). These were first discovered in 1977. Also known as “black smokers”, they are found

on mid-ocean ridges and above volcanic hotspots at depths of two to three kilometres. They are formed when seawater penetrates the ocean floor and comes into contact with hot magma. The heated water then exits through vents as black plumes which rise 100 m or so above the ocean floor. The vents are often surrounded by strange life forms – life forms which utilise the free energy of the chemicals in the plume and are not dependent on sunlight and photosynthesis.

Instead of boiling, the circulating seawater becomes "supercritical" when it is heated above 374°C at high pressure. Supercritical water is neither a true liquid nor a true gas and contains nascent oxygen and hydrogen ions. It is a hot oxidising acid and highly corrosive. It is so corrosive that it tears rocks apart chemically. As a result, HTV plumes contain heavy metals and hydrogen sulphide, which recombine as the plume cools and precipitate onto the ocean floor as metal sulphides. These precipitated metal sulphides make very rich ores. The Mount Isa ore body is comprised of sea floor sulphides created in this way. Some mining companies such as Nautilus Minerals are mining HTV sulphides directly from the ocean floor.

Known HTVs release 17 terawatts of power into the ocean as heat – about the same as global human usage of energy (Elderfield and Schultz, 1996). Each year they pump into the ocean

- 500 tonnes of arsenic,
- 1500 tonnes of lead,
- 50,000 tonnes of copper,
- 140,000 tonnes of zinc

and many other metals (Tivey, 2015). They have been doing this for hundreds of millions of years. These metals are rapidly precipitated as sulphides, which explains why their concentrations have not built up in sea water. In fact, HTVs may have the net effect of actually cleansing sea water of dissolved metals by this sulphide precipitation reaction.

These quantities dwarf any human contribution to oceanic pollution by heavy metals. In addition, Uranium and its radioactive daughter elements are also released copiously by HTVs, so much so that the life forms surrounding the vents are often themselves radioactive.

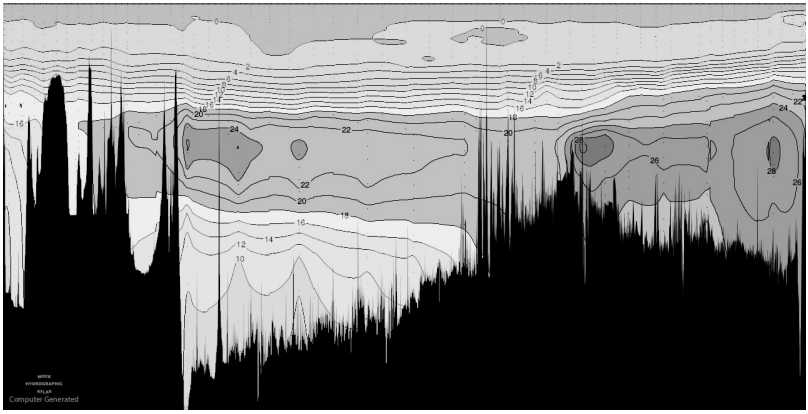


Figure 13.1: WOCE oceanographic section P06 showing two ^{3}He plumes at 2000m depth at latitude 32°S in the South Pacific. The plume on the left with a maximum at 178°W is close to the Kermadec Ridge north of NZ. The plume on the right with maximum at 111°W is close to the East Pacific Rise near Easter Island (27°S, 109°W) (From the WOCE Pacific Ocean Atlas).

Helium-3

Natural Helium occurs as two isotopes, Helium-4 (^{4}He) and Helium-3 (^{3}He). Typically, Helium-4 is more abundant in Earth's crust, whereas Helium-3 is more abundant in the mantle below. When water is heated by mantle magma within hydrothermal vents, the helium that diffuses into the water has the characteristics of mantle Helium, i.e. it has a higher ^{3}He/^{4}He ratio than Helium that comes from the crust. Furthermore, it has been found that this ratio (displayed as δ^3He) varies in proportion to the heat added to the ocean and therefore provides a useful trace of hydrothermal heating of the deep ocean.

Figure 13.1 shows the WOCE[1] oceanographic section of the South Pacific at latitude 32°S from 154°E to 72°W. There are two large plumes of ^{3}He extending laterally for thousands of kilometres from 1500m to 3000m in depth. In each case the ^{3}He concentrations are highest at the westernmost end implying that the source is at the

[1]World Ocean Circulation Experiment

westernmost end. The western plume has the volcanically active Kermadec Ridge as its source, and the other plume has the East Pacific Rise near Easter Island. There are similar plumes in the North Pacific.

The plume volumes and concentrations can be used to estimate the total heat added to the Pacific Ocean from hydrothermal sources and the thermal expansion of the ocean which results. The temperature inside the plumes is about one degree warmer because of the hydrothermal contribution. As a result, because of thermal expansion, the sea level of the world's oceans is about half a metre greater than it otherwise would have been. These figures indicate that hydrothermal heating most likely makes some contribution to global mean sea level variations, but more detailed modelling is not possible without further information concerning the vertical diffusion rates of both heat and ^{3}He. More detailed knowledge of the variance spectrum of the thermal contribution is also needed.

All that can be said is that Figure 13.1 does not fit the traditional "conveyor belt" paradigm of steady-state abyssal ocean circulation and supports the notion that hydrothermal heating may be sufficiently large to contribute to the ocean's thermohaline circulation.

Megaplumes

Volcanism on land is intermittent, random and explosive. Although subaqueous volcanism is less likely to be explosive, there is no reason to assume that it is not similarly intermittent and random. Major sub-aerial volcanic events of the magnitude of Krakatoa are once-in-a-century events. It is likely then that subaqueous events of similar energy to Krakatoa occur four or five times per century.

Detailed examination of HTVs is expensive. It requires the deployment of remotely operated vehicles from specially equipped research vessels. Only a small fraction of the 50,000 km of mid-ocean ridges has been explored, and even then sites of interest are only visited occasionally. The long term variability of HTV fields remains almost a closed book. In the absence of observations, the best we can do is to assume that subaqueous volcanism varies over time in a similar way to sub-aerial volcanism. We can expect that, from time to time, subaqueous volcanic events will occur which gives rise to plumes much larger than those routinely observed. In fact remnants of such plumes, called "megaplumes", have been observed as anomalous chemical sig-

natures high in the water column. Small patches of high sea surface temperature indicative of megaplumes have been recorded by satellites.

There is an aspect of megaplume formation which has hitherto been ignored. HTV plumes, which typically have an exit temperature of 360°C, are rapidly cooled by the entrainment of surrounding sea water until they lose buoyancy and spread horizontally much like chimney smoke on a frosty night. However plume dynamics indicates that for a large plume the temperature of the plume interior could remain sufficiently elevated for the water to boil, as the hydrostatic pressure decreases with decreasing depth. Indeed one HTV field, the Lucky Strike field near the Azores, is sufficiently shallow at 1700 m depth for effluent to be close to boiling point as it exits the vents.

Once boiling commences, plume dynamics alters radically. The formation of steam bubbles greatly increases the buoyancy of the plume causing vertical acceleration of the effluent stream leading to increased further boiling, and so on. In effect, the plume becomes a geyser.

Using buoyancy calculations alone and neglecting heat losses, one litre of vent effluent at 360°C, when it boils in a megaplume, will create sufficient buoyancy to lift 30 tonnes of cold sea water to the surface. A sufficiently powerful HTV field can conceivably generate enough buoyancy to turn an entire ocean basin upside down.

The most powerful HTV field so far discovered is the TAG field on the Mid-Atlantic Ridge. The power output of the TAG field is around 6 GW. It lies at 26°N, the same latitude as the Yucatan Peninsula, one-time home of the Mayan Civilisation. We can speculate that the TAG Field is the remnant of a major subaqueous event with a megaplume large enough to trigger an Atlantic Meridional Overturning event . The overturning brought large volumes of cool deep ocean water into the tropical mixed layer so cooling the sea surface. leading to decreased ocean evaporation and drought in Yucatan.

The Maya were proceeded by the Olmecs. The Olmec civilization came to an abrupt end around 400 BC, perhaps for the same reason.

Upwelling

Seventy percent of the Earth's surface is ocean and ninety percent of the ocean is "desert"; it sustains very little life. Most of the top 50 to 100 metres, the *mixed layer*, contains very little nutrient, particularly the deep ocean well away from land. Most of the nutrient and hence most of the plankton and other living organisms are found on continental shelves where nutrient is continually replenished by river run-off from the land. In many parts of the deep ocean, nutrients such as nitrate, phosphorus and silicate are in plentiful supply at depths of 1000 m or so but support no life there because essential sunlight cannot penetrate to this depth. This is true of the South Atlantic and North Pacific, for example. In some places deep ocean water is forced to the surface. When this deep water contains sufficient nutrient, plankton blooms and marine ecosystems form. Fifty percent of the world's commercial ocean fish come from such nutrient-rich *upwellings*.

Ocean upwellings are caused by the deflection of ocean currents by ridges, seamounts and coastlines. However, a megaplume from an HTV field can have a similar effect. Easter Island lies close to the southern edge of the North Pacific nutrient mass and also on the North Pacific Rise, the world's fastest spreading mid-ocean ridge and one of the most active. Perhaps a megaplume from an HTV field on the East Pacific Rise once supported a mid-ocean marine ecosystem during the period in which Easter Islanders dined on porpoise.

Carbon dioxide

In *The Voyage of the Beagle*, Darwin opens his chapter of on the Galapagos Islands with the following remark:

> Considering that these islands are placed directly under the equator, the climate is far from being excessively hot ... this seems chiefly caused by the singularly low temperature of the surrounding water, brought here by the great southern Polar current.

In fact three ocean currents converge near the Galapagos, the current to which Darwin referred, now known as the Humboldt Current,

its Northern Hemisphere mirror image, the California Current and the Cromwell Current also known as the Equatorial Counter Current. The first two currents cause deep, nutrient-rich, cold water to come to the surface with the result that the waters around the Galapagos are relatively cool and well-endowed with marine life. The presence of iron in the upwelled water suggests that HTVs above the Galapagos' volcanic hot-spot contribute to this upwelling.

The last-named current, the Cromwell Current, while poor in nutrients, is rich in dissolved inorganic carbon. As a result, when this current comes to the surface near the Galapagos, excess carbon dioxide is released into the atmosphere . A simple calculation shows that three gigatonnes (i.e 3 Pg = 3 petagrams) of carbon dioxide are out-gassed per year by this current alone, equal to about half the calculated human contribution per annum.

The degree of out-gassing of CO_2 by ocean upwellings depends on the nutrient concentration and the final temperature achieved by the initially cold, upwelled water. Like most gases, CO_2 is much more soluble in cold water than in warm water. In more nutrient-rich temperate and polar regions, upwelling generally leads to the net absorption of CO_2 from the atmosphere due to photosynthesis by phytoplankton. The atmospheric carbon taken up by the biomass is ultimately returned to the deep ocean in the form of faecal and skeletal material.

The concentration of CO_2 in the atmosphere has been monitored continuously since 1958 when Charles Keeling established a monitoring station on Mauna Loa in Hawaii. Since that time a number of other, similar stations have been established. All such stations show a steady increase in atmospheric CO_2 from year to year once seasonal variations are excluded. It is widely believed that this is due to human activities such as the combustion of fossil fuels and that emissions from natural sources such as the ocean and volcanoes play a relatively minor role. Increases in global average surface temperature observed during the 20th century have been attributed to this increase because, as a triatomic molecule, CO_2 is a strong absorber of infra-red radiation. A number of complex computer models predict the degree of global warming according to various atmospheric CO_2 scenarios.

The most relevant aspect of the impact of CO_2 on the atmosphere is the length of time for which it remains in the atmosphere. Over the previous half century an experiment has been under way which illuminates the relationship between CO_2 stored in the atmosphere and its

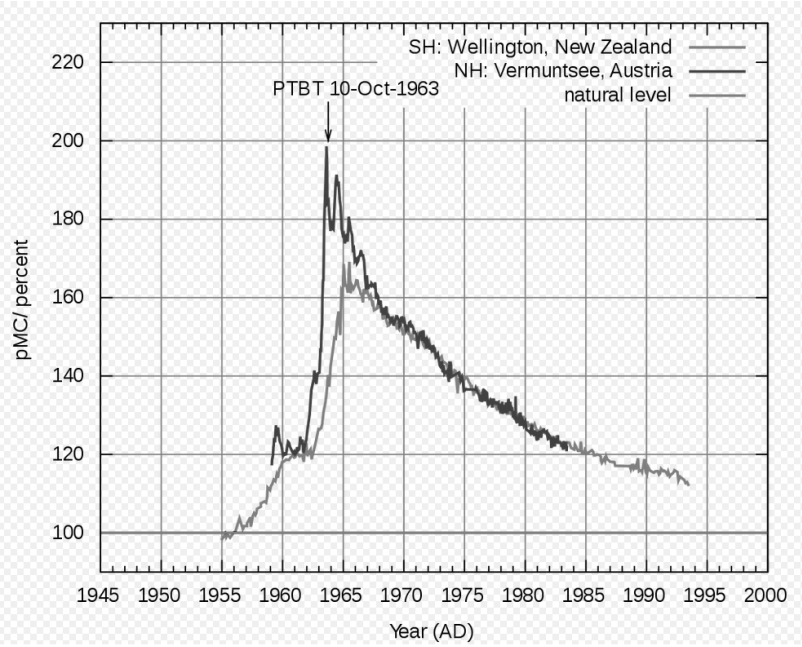


Figure 13.2: Atmospheric ^{14}C, New Zealand and Austria. The New Zealand curve is representative for the Southern Hemisphere; the Austrian curve is representative for the Northern Hemisphere. Atmospheric nuclear weapons tests almost doubled the concentration of ^{14}C in the Northern Hemisphere. The Partial Test Ban Treaty is the abbreviated name of the 1963 treaty banning atmospheric testing which formally went into effect on 10 October 1963. (Diagram due to Hakonomono.)

interchange with other potential reservoirs. That experiment involved the injection of a pulse of radioactive carbon, ^{14}C, into the atmosphere as an unintended consequence of the nuclear weapons testing that took place in the 1950s and 60s. The circumstances were entirely serendipitous, but it would be difficult to design a better experiment with which to investigate the carbon cycle as it pertains to the atmosphere. The observed decay in atmospheric concentrations of ^{14}C which followed the cessation of the tests is known as the *bomb-test curve*.

The bomb-test curve is shown in Figure 13.2. The following features are immediately evident:

1. Atmospheric ^{14}C exhibits exponential decay
2. The exponential decay has a single time constant of 14 years giving a residence half-time of 10 years
3. After 50 years the curve has returned to its pre-bomb-test background value.

The bomb-test curve is a direct observation of the impulse-response of the atmospheric reservoir to a perturbation in its CO_2 content. The above listed facts lead to the following conclusions:

1. The exponential decay implies that the removal of atmospheric CO_2 from the atmosphere is dominated by a first order differential equation, a diffusion equation, whereby the rate of diffusion from the atmosphere reservoir is proportional to the concentration.
2. The single time constant implies that the CO_2 diffuses into a single reservoir (or multiple reservoirs with similar time constants).
3. The asymptotic approach to a new equilibrium which is indistinguishable from the pre-bomb background value implies that the second reservoir into which the atmospheric $^{14}CO_2$ has diffused must be very much larger than the atmospheric reservoir itself.

These facts and deductions form the basis for a simple linear model of ocean–atmosphere CO_2 exchange. We assume that there are regions of the ocean surface where atmospheric CO_2 diffuses into the deep ocean, which constitutes the second reservoir, and that there

are other regions of the ocean surface where CO_2 out-gasses from the ocean into the atmosphere. This out-gassing occurs, not by a diffusion process, but by pumping. Deep ocean CO_2 at 4°C is brought into the surface mixed layer by cold, upwelling ocean currents. As the upwelled seawater from beneath the thermocline is warmed by sunlight, it becomes supersaturated with CO_2, which then comes out of solution. This happens because, like most gases, the solubility of CO_2 is inversely related to temperature.

The simplest mathematical description is the following first order differential equation:

$$\frac{\mathrm{d}K}{\mathrm{d}t} + \frac{K}{\tau} = a(t) + p(t) \tag{13.1}$$

where $K = K(t)$ is the atmospheric concentration of CO_2 at time t (the "Keeling curve"), τ is the diffusion time constant estimated from the bomb-test curve, $a(t)$ is the production rate of anthropogenic CO_2 and $p(t)$ is the rate at which CO_2 is being pumped out of the ocean. Three out of four of the terms in equation (13.1) can be measured or calculated. Only $p(t)$ is not known, and it can be calculated from the other three terms, viz.:

$$p(t) = \frac{\mathrm{d}K}{\mathrm{d}t} + \frac{K}{\tau} - a(t) \tag{13.2}$$

i.e. the rate at which CO_2 is being pumped into the atmosphere by upwelling, $p(t)$, is equal to the rate of increase of the Keeling curve, $\mathrm{d}K/\mathrm{d}t$, plus the rate of loss by diffusion back into the ocean, K/τ, less the anthropogenic component, $a(t)$.

The solution is shown in Figure 13.3. The rapid rate of diffusion loss, due to a surprisingly small value of τ, means that the rate of diffusion into the ocean dominates the process. Since concentration in the atmosphere is not diminished as a result, this implies, in turn, that atmospheric CO_2 is replenished by a comparably high rate of pumping, $p(t)$, of approximately 50 Pg/year in 2010, so dwarfing the anthropogenic production rate, $a(t)$, of 10 Pg/year.

Utilising the bomb-test curve as a measure of the impulse response function of the atmospheric CO_2 reservoir, and assuming this reservoir is replenished by pumping rather than diffusion, leads to the conclusions that the residence time of CO_2 in the atmosphere is 10 years and that less than twenty percent of the observed recent increase in atmospheric CO_2 concentration is anthropogenic in origin.

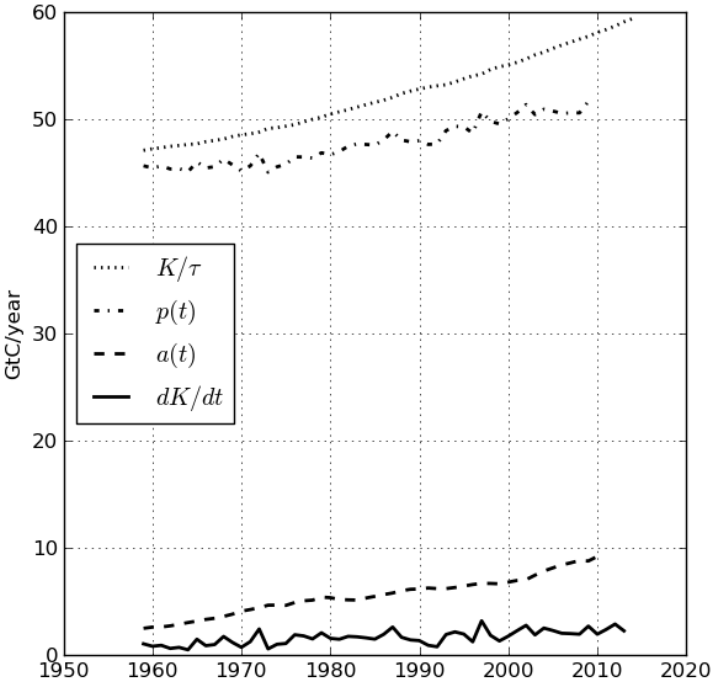


Figure 13.3: Solution of equation (13.2) for $p(t)$, the rate at which CO_2 is pumped from the ocean (dot-dash). Also shown are K/τ, the atmosphere to ocean diffusion rate, (dots), $a(t)$, the anthropogenic production rate, (dashes), and $\mathrm{d}K/\mathrm{d}t$, the measured rate of increase of atmospheric CO_2 concentration, (solid).

There is a vast amount of CO_2 stored in the ocean: 38,000 Pg compared with 380 Pg generated by human activity since the beginning of the industrial revolution. It is doubtful whether humankind's modest one percent contribution has made very much difference.

Chapter 14

Liquid-in-solid Convection

A model is proposed in which upward-moving, liquid-in-solid convection cells are the primary mechanism by which the Earth's interior is cooled. Convection cells form in the liquid core when heat from radioactive sources is convected outward. Each cell penetrates the mantle by melting the rock above it due to the extra heat transported by the cell. In this way each convection cell propagates upwards through the otherwise solid mantle as the convected heat melts rock above the cell and solidifies magma below it. Because the cells are initiated in the core spontaneously and at random, the model accounts for the observed random nature of many geophysical phenomena. A laboratory experiment is proposed whereby a low melting point solid such as wax is subjected to a high temperature gradient in order to promote liquid-in-solid convection and so emulate aspects of mantle dynamics.

Introduction

Spectral analysis of observational data by Pelletier (2002) indicates that the geomagnetic field has a power law variance spectrum with no sharp peaks that would indicate deterministic, periodic behaviour. The observed random occurrence of geomagnetic reversals is described by such a power law spectrum. A more profound implication is that processes which dominate the Earth's interior are primarily random or stochastic, i.e. governed by the laws of probability. Geothermal processes are evidently not steady-state and cannot be accounted for by deterministic models.

Conventionally the mantle is assumed to be sufficiently plastic to allow convection but sufficiently rigid to support seismic S-waves; see, for example Turcotte and Oxburgh (1972) and Richter (1984) and references therein. The problem is that such plastic mantle convection is very slow, with turnover times of the order of 1×10^9 years, whereas Pelletier's spectrum follows its power law locus to periods shorter than 1000 years.

It might be argued that geomagnetism is generated entirely in the core – for example, the deterministic model of Glatzmaier and Roberts (1995). This implies a spectrum which is independent of mantle dynamics. Even so questions concerning the random nature of geomagnetism and the relatively short time scales of mantle processes remain unresolved.

A simple, one dimensional model in which conduction is the only form of heat transport predicts a fluid core with a size and temperature far in excess of those deduced from observation. For instance assuming uniform radiogenic heating, $H_r = 1.4 \times 10^{-8}$ W m^{-3} (Pollack et al., 1993), and uniform conductivity ($\kappa = 1.0$ W.m^{-1}K^{-1}) (Tang et al., 2014) gives a core temperature of $27,000\,^{\circ}$C and a core radius of 6100 km. In the present epoch the liquid core radius is observed to be only 3500 km, and so conduction alone cannot account for the observed structure of the Earth's interior. It follows that some mechanism of heat transport other than conduction and visco-plastic turnover must play a major role in cooling the Earth.

Liquid-in-solid convection

We propose an alternative, as depicted in idealized form in Figure 14.1. In Figure 14.1, convection cells form in the liquid outer core. Rayleigh–Bernard convection is a stochastic process. It occurs when a body of liquid lies in a vertical temperature gradient at length scales sufficiently large for the critical Rayleigh number to be exceeded. As a consequence of the formation of these cells, regions of the lower mantle, M, which are directly above each Rayleigh–Bernard convection cells will attain a higher temperature than surrounding regions because of the extra heat transported by the cell. These regions then melt allowing the cell to penetrate the mantle. Ultimately the cell becomes completely enclosed by the solid mantle as the mantle solidifies behind it in region S. This happens because the increased heat trans-

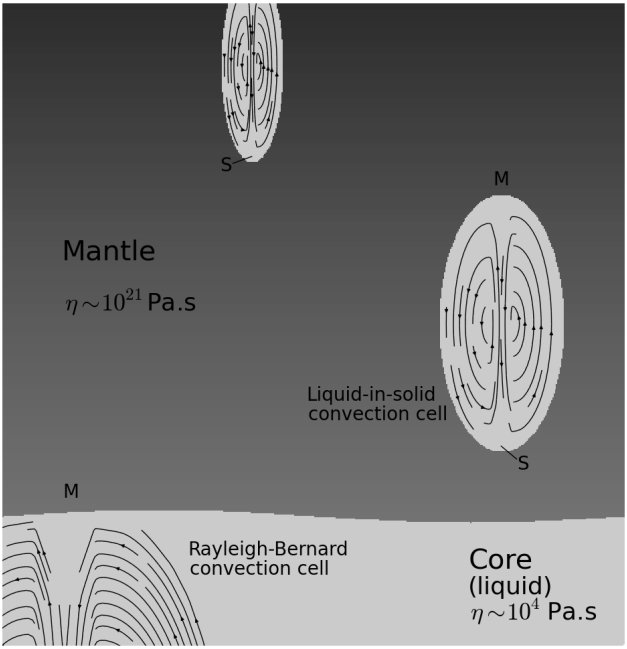


Figure 14.1: A conceptualization of liquid-in-solid convection cells formed in the lower mantle from Rayleigh–Bernard convection cells in the core. M: regions of melting due to increasing temperature. S: regions of solidification due to cooling and increasing pressure. Note that this diagram is highly idealized and does not represent numerical model output. The approximate dynamical viscosities, η, are also shown.

Table 14.1: Values of the Rayleigh Number, R_a, as a function of length scale and viscosity.

L km	η =1 Pa.s	$\eta = 1 \times 10^{11}$ Pa.s
0.001	10×10^{-3}	10×10^{-14}
0.010	10×10^{1}	10×10^{-10}
0.100	10×10^{5}	10×10^{-6}
1.000	10×10^{9}	10×10^{-2}
10.000	10×10^{13}	10×10^{2}
100.000	10×10^{17}	10×10^{6}
1000.000	10×10^{21}	10×10^{10}

port promoted by the cell causes the mantle below it to be cooled.

This process will be designated LISC, for "liquid-in-solid convection". Rayleigh number calculations indicate that such LISC cells are feasible. Their existence or otherwise depends on whether the Rayleigh number, R_a, for such a cell exceeds the critical Rayleigh number, R_{ac}. This in turn depends largely on their scale size, viscosity and other parameters and is given by

$$R_a = \frac{\rho\alpha(T_{bottom} - T_{top})gL^3C_p}{\eta k} \tag{14.1}$$

where ρ is the density, α is the coefficient of thermal expansion, T is the temperature, L is the length scale, C_p is the heat capacity, η is the viscosity, and k is the thermal conductivity. This can be written

$$R_a = \frac{\rho\alpha\Delta T g L^4 C_p}{\eta k} \tag{14.2}$$

where ΔT is the temperature gradient.

One of the factors in (2), the viscosity, η, is not well known. Estimates of its value in the lower mantle vary from 1 to 1×10^{11} Pa.s. A similar variation is found in near-surface magmas depending on their composition. Despite this, meaningful values of R_a can be found because it is such a strong function of L in (2). Table 1 shows some values at different length scales for both low and high viscosity values.

Convection can only occur when R_a exceeds the critical Rayleigh Number, R_{ac}. A nominal value for the latter is 1707, but it varies

slightly according to the geometry. Hence, according to Table 1, the length scale above which LISC is possible lies somewhere between 300 m and 30 km depending on the assumed viscosity.

Each cell melts the solid mantle above it because of the extra heat being convected outward from the hot core. In this way each convection cell propagates upwards through the otherwise solid mantle at speeds determined by the solution of the Stefan problem[1] for a liquid-solid boundary. The upward-moving, LISC cells formed in this way are proposed as the primary mechanism by which the interior of the Earth is cooled.

The LISC cells shown in Figure 1 are highly idealised. In reality LISC cells are unlikely to be neat ellipsoids. The solution of the Stefan problem for a moving phase transition within a non-uniform medium will be difficult or intractable, but intuitively it seems likely that rising cells will follow paths where the melting point is lower or the solid is hotter and avoid places where it is cooler. Thus strings of ascending LISC cells can be expected to form wherein each cell follows the path of a preceding cell. Such a string of LISC cells could well be described as a "plume" when it is in the mantle. An individual LISC cell becomes a "diaper" when it interacts with the crust.

Some implications of a random model

The initial generation of convection cells is a random process. It is impossible to predict when or where a new cell will start up or how big it will be because turbulent convection itself is a random process. All physical quantities associated with LISC will therefore also have a random component and have power law spectra. These include ascent speed, volume, magnetic field and temperature.

In this model of the Earth's interior, specific heat, thermal conductivity, radiogenic heating and density are all assumed to be constant. Despite this simplicity, complex behaviour occurs as a consequence of convection. Rayleigh–Bernard convection cells form spontaneously and at random in the outer liquid core as heat builds up from radioactive sources. The model assumes that the primary distinction between mantle and core is due to phase alone: one is solid, the other liquid.

[1]The Stefan problem is the mathematical description of a phase boundary which moves with time.

(This does preclude compositional differences. The concentration of heavier elements such as iron can be expected to be greater nearer the centre of the Earth.)

- LISC cells will have a large range of scales, ranging from the Rayleigh scale to scales comparable to the dimensions of the core itself, from volcanic eruptions to Large Igneous Provinces.
- The total volume of the Earth is the sum of the volumes of the core, the solid mantle and all the LISC cells in the mantle. It must also be a quantity which varies over time in a random manner according to the proportion of molten material in the core and mantle.
- Hence sometimes the earth expands and at other times it contracts. When the Earth is expanding, new crust forms at mid-ocean ridges (Fig. 5)
- When the Earth contracts, the crust is forced to crumple and mountain building occurs. Such cooling and crumpling would result from, and be modified by, extensive diapirism.
- Sea floor spreading occurs when the Earth expands – recent magnetic stripes average 6 km across indicating a total expansion of $\sim$36 km in the Earth's circumference, i.e. roughly 0.1 percent variation in circumference during each expansion phase.
- The geomagnetic field is the sum of the self-exciting magneto-hydrodynamic (MHD) dynamo fields of all of the LISC and Rayleigh–Bernard convection cells in the mantle and core respectively. It is therefore a random quantity which varies over time in a random manner.
- Glatzmaier and Roberts (1995) account for isolated field reversals but not the variance spectrum of the geomagnetic field observed by Pelletier. Their deterministic model could be upgraded to include random forcing as described here.

A wax block experiment

Numerical modeling is deterministic and so cannot fully emulate the random aspects of convection. However, modelling of individual LISC cells under various viscosity and temperature gradient scenarios could provide important insights. In particular, local solutions of the Stefan Problem under physical conditions pertaining in the mantle could be used to provide realistic parametrizations for use in larger scale models.

However, the reality of liquid-in-solid convection should first be verified at laboratory scale. This can be done, for example, by subjecting a block of wax to a vertical temperature gradient. Such an experiment could provide great insight into plume formation and diaperism.

Once again, the relevant parameter is the Rayleigh Number. For, say, a ~ 1 m^3 block of paraffin wax on an electric hot plate at, say, 100 deg C, cooled at the top with iced water, and assuming $g = 9.8$ m.s^{-2}, $k = 0.25$ W.m^{-1}K^{-1}, $C_p = 2500$ JK^{-1}.kg^{-1}, $\alpha = .0007$ K^{-1}, $\rho =$ 900 kg.m^{-2} $\eta = 3.0$ Pa.s, $\Delta T = 100.0$ K, $R = 1707$, substituting into (14.2) and solving for x gives a critical Rayleigh scale of $x \approx .01$ m. Thus molten LISC cells should form at scales of 1 cm or more in solid wax under these conditions.

An interesting aspect of this experiment is whether permanent liquid layers would form both at the base and just below the cold top crust, so corresponding to the core and lithosphere of the Earth. If so, the interaction of ascending LISC cells with the lithosphere and crust layers might emulate some of the effects of diaperism discussed by Carey (1999).

Chapter 15

Earthquakes and the South Atlantic Anomaly

For an interval of 18 months in 1967 and 1968 a swarm of submarine earthquakes occurred in the South Atlantic with locations that were remarkably collinear. To the north-east of the Mid-Atlantic Ridge, the line of earthquakes followed the chain of volcanic activity known as the Cameroon Line. On the south-western side, the ocean floor is relatively featureless. The line showed no discernible kink or deviation where it crossed the Mid-Atlantic Ridge. At the time of the earthquakes, the South Atlantic Magnetic Anomaly had begun expanding rapidly towards the south-east, crossing the line of the earthquakes, and it seems likely that the two phenomena were associated. It is difficult to account for these empirical observations under conventional theories of crust or mantle dynamics.

Introduction

The ready availability of high quality data sets on-line, together with improvements in programming languages means that significant features can come to light which may have been missed when the data were first acquired. This is largely due to better graphical display techniques. The human eye is very good at spotting patterns in data when it is displayed in graphical form. In the 1960s graphical display was difficult and time consuming; nowadays it is almost effortless. The pattern of collinear earthquakes discussed here was noticed

only recently, by accident when downloading data for entirely different purposes. That it had gone unnoticed until then can be attributed to these factors.

The Data

Earthquake data were downloaded from the IRIS Wilber 3 global dataset (available online at `http://ds.iris.edu/wilber3/find_event`, accessed 10/18/2017). Earthquake locations were mapped using the *Python* programming language and its *basemap* software package. Submarine earthquakes of magnitude ≥ 4.0 which occurred between January 1960 and October 2017 were mapped initially. Despite the large number of earthquakes ($\sim 400,000$), a diagonal line of earthquakes across the South Atlantic could be clearly seen. Further investigation revealed that most of these occurred between 31/1/1967 and 15/9/1968. For greater clarity only earthquakes occurring in 1967 and 1968 are shown in Figure 15.1. In this figure, the collinear swarm of earthquakes forms a diagonal line from near Tierra del Fuego (37°S, 35°E) to the African coast near Cameroon (3°N, 8°W). The earthquakes were highly collinear, even more collinear than those occurring along the Mid-Atlantic Ridge (MAR) which can also be seen in the diagram. A depth of 33 km was recorded for all 62 earthquakes, and their magnitudes ranged from 4.2 to 5.9. The line of earthquakes did not fit a great circle, and their successive occurrence along the line appeared to be sporadic and random, i.e. they did not start at one end and progress regularly from east to west or vice versa. Plots of the locations of all submarine earthquakes with magnitude ≥ 4.0 for the whole Earth were generated for each year from 1960 to 2017 and examined carefully. No similar event could be discerned.

The collinearity of the earthquakes is truly remarkable; nothing similar has been observed before. This suggests that they were not real and perhaps due to an artefact of the recording network. They were all recorded in the early years of the modern seismic network and were all reported at the same depth of 33 km. This is certainly disconcerting. However depth resolution at the time was poor, most similar submarine earthquakes being recorded at depths of either 10 km or 33 km.

The strongest evidence for the reality of the collinear swarm is that the north-eastern end of the line corresponds to the well-known

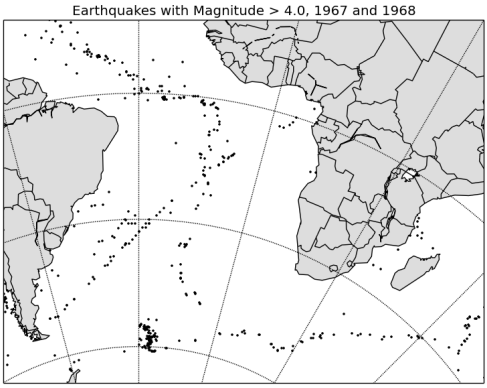


Figure 15.1: The dots show the locations of all submarine earthquakes with magnitude greater than or equal to 4.0 which occurred in the displayed area during 1967 and 1968.

Cameroon Line of volcanic activity. In effect the swarm represents an extension of the Cameroon Line beyond the Mid-Atlantic Ridge (MAR). If the collinear swarm were indeed an artefact, this would be a remarkable coincidence.

Association with the South Atlantic Anomaly

The reality of the collinear swarm would be more convincing if it were associated with some other geodynamic phenomenon. One such phenomenon is the South Atlantic Anomaly (SAA), a region of the Earth's surface where the geomagnetic field intensity is anomalously low. The relation of the collinear swarm to the expanding SAA is depicted graphically in Figure 15.2. The total geomagnetic field strength, F, at the earth's surface derived from the $igrf12$ spherical harmonic expansion of the geomagnetic field (Thébault et al., 2015) is shown in Figure 15.2. The line of the 1967/68 collinear earthquake swarm is also shown. It is a quadratic fit to the 62 collinear earthquakes depicted in Figure 15.1.

The intensity anomaly and the area of the SAA both increased considerably during the 20th century. It also become more elongated

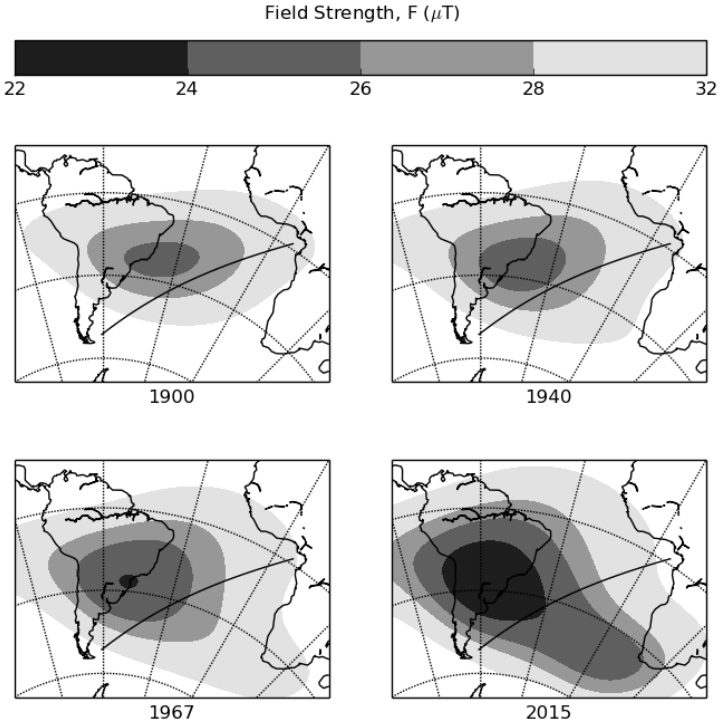


Figure 15.2: Maps of the geomagnetic field strength, F, at sea level computed using *igrf12* for the years 1900, 1940, 1967 and 2015. This region of lower field strength is the South Atlantic Anomaly. The quadratic line of best fit to the 1967/68 collinear earthquake swarm is also shown in each map.

in a south-easterly direction. The elongation process commenced around the time of the collinear swarm in 1967.

Evidence of LISC in the Mantle?

A remarkable feature of the collinear swarm is that the line intersects the MAR with no discernible kink or discontinuity. Such a discontinuity might be expected because the MAR is spreading at an average rate of 2.5 cm/year and the age of the Cameroon Line is 65 million years (Fitton and Dunlop, 1985). If the line was a pre-existing feature of the South Atlantic crust, a break of the order of 1600 km should be evident. In reality no discontinuity is resolved and so, if one exists, it must be less than about one degree of latitude (100 km). This implies that either the extension of the Cameroon Line beyond the MAR is a new feature of the crust (i.e. less than 4 million years old) or it is not a feature of the crust at all but a feature of the upper mantle. The association with the south-eastward expansion of the South Atlantic Anomaly suggests the latter.

A recent paper by Bendick and Bilham (2017) showed that earthquakes of magnitude ≥ 7.0 occur synchronously when appropriately parametrized, i.e. earthquakes with similar properties occur at similar times even though only loosely coupled spatially. One such property was the "renewal interval" calculated "using a semi-empirical relation between event magnitude and mean slip divided by the product of the local tectonic loading rate determined from the REVEL global plate motion model and a local value for seismic coupling". The present group of earthquakes do not lie on any known plate boundary and have magnitudes much less than 7.0. As such they lie outside the domain of interest of that paper. The paper also suggests that stresses in the crust related to decreased rotation rate are the ultimate cause of such earthquake clusters. In the present case, the year 1967 was a time of increasing length-of-day and increasing M$\geq$ 7.0 earthquake activity, which peaked 3 to 5 years later. Since then length-of-day has not exceeded its 1967 value, and such a prime cause cannot be completely ruled out. On the other hand, the recorded depth of 33 km and the association with the recent rapid elongation of the South Atlantic Anomaly imply a Mantle origin.

We can speculate that both phenomena are the result of a large LISC cell rising in the Mantle below the South Atlantic. Accordingly

the South Atlantic Anomaly is due to the magneto-hydrodynamic (MHD) field of the LISC cell, and the collinear earthquakes are due to new stresses induced in the Crust as the cell rises in the Mantle.

Part IV

Chapter 16

Conclusions

Implications for "Climate Change"

1. There is no significant trend in global average temperature. The apparent trend is due to spurious regression (Chapters 8 and 9).

2. The apparent correlation between global average temperature and atmospheric CO_2 concentration is also spurious. (Chapter 8)

3. Items 1 and 2 are both a consequence of global average temperature being a centrally biased random walk with a red-noise variance spectrum. (Chapter 7)

4. The bomb-test curve shows that about 80 percent of recent increases in atmospheric CO_2 concentration are due to ocean upwelling and only 20 percent are due to human activity. The ratio of the total contribution of anthropogenic CO_2 to the total in the ocean-atmosphere system since the beginning of the Industrial Revolution is only 1 percent. This would have no measurable effect. (Chapter 12)

5. Predictions of temperature increases based on numerical coupled ocean-atmosphere general circulation models (i.e. climate models) are meaningless because such deterministic models cannot account for turbulence, which is stochastic. Because of this, they include unrealistically large values of parameters such as

eddy viscosity in order to remain stable and so can never faithfully emulate reality. (Chapter 3, Figures 3.1 and 3.2)

6. Climate modellers ignore the effect of subaqueous volcanic activity on ocean circulation despite that fact that 85 percent of volcanic activity occurs beneath the ocean and that heating from a major oceanic eruption would dwarf all other ocean processes. This aspect of ocean circulation and climate has not simply been forgotten; it is conscientiously avoided. It is the elephant in the room. (Chapter 13)

7. Based on the above we can conclude that, at multidecadal time scales, unexplained variations in global temperature and in global mean sea level may be attributed to subaqueous volcanism and are unrelated to human activity.

A return to Empiricism

Over the last four centuries the Newtonian description of the physical world in terms of differential equations has been brilliantly successful. It began with precise and accurate descriptions of the motion of the planets and their satellites and found further application in the design of the machinery which became the basis of industrial civilization. At first General Relativity appeared to undermine Newton's ideas, but in effect it amounted to only a minor refinement of the Newtonian system.

A more profound threat to the Newtonian hegemony was the quantum theory which followed the "Ultraviolet Catastrophe": the realization that some physical quantities are granular and require a radically different approach. At about the same time, the Second Law of Thermodynamics was proposed leading to the concept of entropy which describes how matter and energy are ordered and which requires the Universe be granular rather than continuous.

Newton's ideas were successful when applied to the dynamics of planetary motions because friction and turbulence are negligibly small at planetary scales. Planetary motions are, to all intents and purposes, deterministic and amenable to a description in terms of differential calculus. Machinery is intentionally designed to minimise friction and turbulence and to be amenable to a deterministic description. This

even applies to semi-conductor design where so called "race conditions" are eliminated in order to preclude any possibility of stochastic behaviour in electronic components.

But it is not true of fluids. Stochastic behaviour in the form of turbulence clearly plays a major role in the dynamics of fluids. The Navier–Stokes equations, which describe fluids in purely Newtonian terms, fail in high Reynolds Number regimes where turbulence is generated. This is the "Fluid Catastrophe". In effect the Quantum Revolution bypassed Fluid Dynamics, whose practitioners still cling to the 19th century idea of the continuum.

The belief that any real fluid can only be dealt with as a deterministic, Newtonian continuum has had a stifling effect on development. Fluid Dynamics has become the province of Applied Mathematicians who are skilled in the manipulation of partial differential equations but in very little else. They are not trained to perform experiments. They do not have an empirical, "Popperian" outlook. They are mathematical Rationalists who only pay lip service to the scientific method. They are not really scientists at all but they they think they are.

Fluid dynamicists might argue that there is no alternative, that only a Newtonian approach to fluids will work. The way to break the stranglehold of determinism is to regard the mathematical expressions of physical laws, such as the Navier–Stokes equations, not as immutable, deterministic Laws but, rather, as constraints. The future state of a fluid is then predicted as that state which maximises the entropy within constraints determined by conservation of mass, momentum and energy. Because of our imperfect knowledge of the initial boundary conditions and the existence of bifurcation points in the evolving trajectory of the system, future states can never be known with certainty but, rather, a most likely state is predicted and its variance estimated. Russian physicists and mathematicians such as Kolmogorov, Monin, Obukhov and Kitaigorodskii laid the groundwork for this approach, but their work has been overshadowed by the advent of the digital computer which favoured fully deterministic numerical models. These now dominate the discipline. The current fad for running "ensembles" of deterministic models with randomized boundary conditions is naive and simplistic. Internally such models are kept stable only by means of unrealistic parameter values and smoothed topography and bathymetry. Garbage in, garbage out.

Science is not a collection of given "truths" like religion; it is a work in progress. This applies as much to Fluid Dynamics as to any

other branch of science. What is needed is for Fluid Dynamics to become, once again, the empirical science that flowered in the 19th century. The wax block experiment described in Chapter 14 would empirically test the validity and relevance of the liquid-in-solid convection mechanism proposed here. If successful, new insights into Mantle dynamics would be gained. The wave-wave interaction algorithms, presently used in numerical wave models, were initially borrowed, second-hand, from Quantum Mechanics. This accounts for the poor performance of these models. What is needed are empirical studies of those wave-wave interactions which really do occur among gravity waves on a fluid surface and which involve entropy increases due to wave-breaking. Initially such studies could be carried out with the wave tanks and hydraulic basins presently used by naval architects. This would open a new field similar to the field of sub-atomic particle interactions which followed the development of the cyclotron. These experiments are not expensive; they are within the reach of a small university department.

It is time for Fluid Dynamics to recognize its "catastrophe", to put its house in order and to become an empirical science once again. Until that happens, its predictions of future states of the Earth's atmosphere must be taken with a large grain of salt.

This book has challenged contemporary Fluid Dynamics. It has examined the myth, the Baconian Idol, on which it is based: that a fluid is a continuum which can be fully described by the deterministic equations of differential calculus. This idea is at odds with the stochastic assumption underlying statistical inference and with the post-quantum conception of entropy. It is a relic of a bygone era when the natural world was regarded as a perfect machine. Once the world was Paley's timepiece, created and wound up by God at the Creation and left ticking steadily for us to examine and to marvel at. Perhaps this was appropriate in the early 19th Century but there has since been a vast increase in the range of phenomena accessible to science. Wherever we look we see a Universe that is chaotic and unpredictable; more like stock market than timepiece.

Furthermore it is not the pristine perfection of a Laplacian universe that matters, but rather its imperfections. Without occasional, random imperfections in nucleic acids, life could never have become more complex than the virus. Turbulence is everywhere, in cumulus clouds, in breaking waves, in the sound of a clarinet and yet it is inaccessible to the elegant equations of 19th century physics. By pre-

supposing an underlying perfection, we blind ourselves to the amazing realities of the world around us. Only by experiment and observation can we truly see.

Acknowledgments

I am indebted to William Hempstone for Francis Bacon and his Idols, to Terence Mills for Yule and nonsense-correlations and to John Colman for his advice about Rationalism.

Appendices

Maxwell's Equations

The behaviour of electric and magnetic fields, $\boldsymbol{E}$ and $\boldsymbol{B}$, in free space are described by Maxwell's Equations:

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = \frac{\rho}{\epsilon_0} \tag{A.1}$$

$$\frac{\partial B_z}{\partial y} - \frac{\partial B_y}{\partial z} - \frac{1}{\mu_0 \epsilon_0} \frac{\partial E_x}{\partial t} = \mu_0 J_x \tag{A.2}$$

$$\frac{\partial B_x}{\partial x} - \frac{\partial B_z}{\partial x} - \frac{1}{\mu_0 \epsilon_0} \frac{\partial E_y}{\partial t} = \mu_0 J_y \tag{A.3}$$

$$\frac{\partial B_y}{\partial y} - \frac{\partial B_x}{\partial y} - \frac{1}{\mu_0 \epsilon_0} \frac{\partial E_z}{\partial t} = \mu_0 J_z \tag{A.4}$$

where $\boldsymbol{E} = (E_x, E_y, E_z)$, $\boldsymbol{B} = (B_x, B_y, B_z)$, ρ is the electric charge density, $\boldsymbol{J} = (J_x, J_y, J_z)$ is the current density and ϵ_0 and μ_0 are measurable physical quantities known as the permittivity and permeability of free space, respectively. It turns out that

$$\mu_0 \epsilon_0 = c^2 \tag{A.5}$$

where c is the velocity of light.

Using vector operators and (A.5), these reduce to the two equations:

$$\nabla \cdot \boldsymbol{E} = \frac{\rho}{\epsilon_0} \tag{A.6}$$

$$\nabla \times \boldsymbol{B} - \frac{1}{c^2} \frac{\partial \boldsymbol{E}}{\partial t} = \mu_0 \boldsymbol{J} \tag{A.7}$$

The first equation, (A.6), is the inverse square law for the electric field strength and the second, (A.7), is the electromagnetic wave equation.

The Navier–Stokes Equations

The behaviour of a fluid continuum is described by the Navier–Stokes equations:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0 \tag{A.8}$$

$$\rho \left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \cdot \nabla \mathbf{u} \right) = -\nabla P + \eta \nabla^2 \mathbf{u} \tag{A.9}$$

where η is the viscosity, $\mathbf{u}$ is the velocity of the fluid parcel, P is the pressure and ρ is the fluid density.

Written out in full, (A.9) becomes:

$$\rho \left(\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} \right) = -\frac{\partial P}{\partial x} + \eta \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right) \tag{A.10}$$

$$\rho \left(\frac{\partial v}{\partial t} + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} + w\frac{\partial v}{\partial z} \right) = -\frac{\partial P}{\partial y} + \eta \left(\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right) \tag{A.11}$$

$$\rho \left(\frac{\partial w}{\partial t} + u\frac{\partial w}{\partial x} + v\frac{\partial w}{\partial y} + w\frac{\partial w}{\partial z} \right) = -\frac{\partial P}{\partial z} + \eta \left(\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right) \tag{A.12}$$

arma.py

```
""" arma.py fits an ARMA(p,q) model to the time series specified
        by the input function. Values of p and q are optimized by
        trial and error by minimizing the residual variance or the
        Ljung--Box pmin value. These are appended to a 'Summary'
        output file.
        Spectra and time domain graphs are saved in 'spectrum'
        and 'resid' folders.
        Periodograms are saved in numerical form in a
        'periodogram' folder.
        Statistics are saved in a 'stats' folder.
        These four subfolders must be created before running arma.py.
        A text file 'Summary' is created and appended after each run.
        Call: python arma.py p q where (p,q) is the order.
"""
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import gridspec
from scipy.fftpack import fft
from scipy import stats
import statsmodels.tsa.stattools as sts
import statsmodels.tsa.arima_model as sta
import sys

def gat():
        #
        #   read in temperature data
        #
        Ns = 166
        f = open('HadCRUT.4.4.0.0.annual_ns_avg.txt', 'r')
        title = 'HadCRUT'
        x = np.zeros(Ns,float)
        y = np.zeros(Ns,float)
        dy = np.zeros(Ns,float)
        i = 0
        for line in f:
                myline = line.split()
                x[i] = float(myline[0])
                y[i] = float(myline[1])
                i+=1
        f.close()
        dt = 1.
        ylabel=r'Temp ($^\circ$C)'
        xlabel='year'
        title = 'HadCRUT4 - Global Average Temperature'
        """
        fig=plt.figure(num=99,figsize=(6.4,3.2))
        plt.plot(x,y,color='k',linewidth=1)
```

```
        plt.hlines(0,x[0],x[-1],colors='k',linestyles='dashed')
        plt.title(title)
        plt.ylabel(ylabel)
        plt.savefig('rawdata')
        plt.show()
        """
        return x,y,Ns,dt,xlabel,ylabel,title

if len(sys.argv)<>3:
        print 'Wrong number of arguments:'
        print 'Should be: arma.py p q'
        quit()
p = int(sys.argv[1])
q = int(sys.argv[2])
outputcode='{0:1d}{1:1d}'.format(p,q)
ab = [0 for i in range(p+q)]
print "ab = ",ab

x,y,Ns,dt,xlabel,ylabel,title = gat()

df = 1./(Ns*dt)
fs = 1./dt                          #sampling frequency
Nf = 0.5/dt                          #Nyquist frequency
xbar = np.mean(x)
ybar = np.mean(y)
y = y-ybar
yvar = np.var(y)

print title
print 'Ns: {0:4d}, dt:{1:4.0f}, Log(Nf):{2:6.2f}'\
.format(Ns,dt,np.log(Nf)/np.log(10.))
#
#
#
#
fig=plt.figure(num=1,figsize=(6.4,6.4))
gs = gridspec.GridSpec(1, 1, height_ratios=[1])
#
#   prepare log-log periodogram
#
Y = fft(y)
X = np.linspace(0,Nf,Ns/2+1)
Y2 = 2*(np.abs(Y[:Ns/2+1])/Ns)**2
XX = X[1:]
Pgm = Y2[1:]/df
Pgm[Ns/2-1]=0.5*Pgm[Ns/2-1]
lX = np.log(XX)/np.log(10.)
lP = np.log(Pgm)/np.log(10.)

ax=plt.subplot(gs[0])
```

```
plt.plot(lX,lP,linewidth='1',color='k',label='Periodogram')
plt.xlabel(r'Log($f/yr^{-1}$)',fontsize=16)
plt.ylabel(r'Log($\hat{S_f}$)',fontsize=16)
#
#        Compute ARMA coefficients using statsmodels
#
model2 = sta.ARMA(y,(p,q))
results2 = model2.fit(start_params=ab, trend='nc', disp = False)
#
resultspq='stats/results{0:s}'.format(outputcode)
f=open(resultspq,'w')
f.write(str(results2.summary(0.05)))
f.close()
#
epsvar = results2.sigma2
#
#        save parameters
#
paramspq ='stats/{0:s}'.format(outputcode)
f = open(paramspq,'w')
for num in results2.arparams: f.write('{0:f} '.format(num))
for num in results2.maparams: f.write('{0:f} '.format(num))
f.close()
print results2.params
#
#        plot ARMA population spectrum, S, with confidence limits.
#
import cmath
Np = 1000
lf = np.zeros(Np,float)
S = np.zeros(Np,float)
i = 0
deltaf = Nf/Np
integral = 0.
dof = 2
lNf=np.log(Nf)/np.log(10)                #log(Nyquist frequency)
lfarray = np.linspace(-4,lNf, num=Np)
for lf in lfarray:
        f=10.**lf
        phi = 2*np.pi*f/fs
        z = complex(np.cos(phi),np.sin(phi))
        den = 1.
        for ip in range(p):
                den-=results2.arparams[ip]*z**(ip+1)
        num = 1.
        for iq in range(q):
                num+=results2.maparams[iq]*z**(iq+1)
        S[i] = 2*dt*epsvar*(abs(num/den))**2
        integral += S[i]*deltaf
        i+=1
```

```
print'yvar, integral = ',yvar, integral
lS = np.log(S)/np.log(10.)
print lfarray.shape,S.shape
plt.plot(lfarray,lS,linewidth=2,color='w')
plt.plot(lfarray,lS,linewidth=2,color='w',linestyle = '-')
plt.plot(lfarray,lS,linewidth=2,color='k',\
linestyle = '--',label='ARMA')
#
#        plot power law locus
#
pllx=np.zeros(2,float)
plly=np.zeros(2,float)
pllx[0]=-2.2
plly[0]= -0.1
pllx[1]=-1.7
plly[1]=-1.1
plt.plot(pllx,plly,color='k',linestyle='-.',label=r'$f^{-2}$ locus')
#
#        Save spectral details
#
periodogrampq='periodograms/{0:s}'.format(outputcode)
f = open(periodogrampq,'w')
f.write('   log f    log P        f        T    log S\n')
for i in range(XX.size):
        phi = 2*np.pi*XX[i]/fs
        z = complex(np.cos(phi),np.sin(phi))
        den = 1.
        for ip in range(p):
                den-=results2.arparams[ip]*z**(ip+1)
        num = 1.
        for iq in range(q):
                num+=results2.maparams[iq]*z**(iq+1)
        S[i] = 2*dt*epsvar*(abs(num/den))**2
        lS[i] = np.log(S[i])/np.log(10.)
        f.write('{0:8.3f} {1:8.3f} {2:8.5f} {3:8.4f} {4:8.4f}\n'\
.format(lX[i],lP[i],XX[i],1/XX[i],lS[i]))
f.close()
#
#
plt.legend(loc='lower left')
plt.xlim(-3,0)
plt.xticks([-3,-2,-1,0])
plt.ylim(-5,2)
plt.yticks([-4,-2,0,2])
plt.savefig('spectrum/{0:s}'.format(outputcode))
plt.show()
#.................................................
#
#        residuals
#
```

```
fig=plt.figure(num=2,figsize=(6.4,9.6))
gs = gridspec.GridSpec(3, 1, height_ratios=[1,1,1])
#
#
#
ax=plt.subplot(gs[0])
plt.plot(x,y,color='k',linewidth=1)
plt.hlines(0,x[0],x[-1],colors='k',linestyles='dashed')
plt.title(title)
plt.ylabel(ylabel)
ax.annotate('a',xy=(.95,.1),xycoords='axes fraction', \
xytext=(.95,.1), textcoords='axes fraction',weight='bold')
#
#   find ARMA residuals.
#
ax = plt.subplot(gs[1])
eps = results2.resid
plt.plot(x[1:],eps[1:],color='k')
plt.hlines(0,x[0],x[-1],colors='k',linestyles='dashed')
string = r'$\epsilon$'
ax.annotate(string,xy=(-.12,.48),xycoords='axes fraction', \
xytext=(-.12,.48), textcoords='axes fraction',weight='bold',\
fontsize='20')
ax.annotate('b',xy=(.95,.1),xycoords='axes fraction', \
xytext=(.95,.1), textcoords='axes fraction',weight='bold')
#
#   get acf of residuals
#
nlags = 40
ax=plt.subplot(gs[2])
epsacf,lbvalue,pvalue = sts.acf(eps, unbiased=False, \
nlags=nlags, qstat=True)
iepsacf = range(len(epsacf))
plt.plot(iepsacf,epsacf,linewidth='1',color='k')
plt.hlines(0,0,len(epsacf),linestyles="--")
plt.hlines(0.05,0,len(epsacf),linestyles=":")
plt.xlabel("Lag")
ax.annotate('c',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.95,.1), textcoords='axes fraction',weight='bold')
#
string = r'$\phi$'
ax.annotate(string,xy=(-.12,.48),xycoords='axes fraction', \
xytext=(-.12,.48), textcoords='axes fraction',weight='bold',\
fontsize='20')
#
#                Ljung--Box
#
truep = np.zeros(nlags,float)
pmin= 1.1
pmax = -1
```

```
lagmin=0
ip = range(nlags)
aip = np.array(ip)
for lag in range(p+q,nlags):
        Q = lbvalue[lag]
        tpl=1-stats.chi2.cdf(Q,lag-(p+q))
        if tpl<pmin:
                pmin=tpl
                Qpmin=Q
                lagmin=lag
        if tpl>pmax:
                pmax=tpl
                lagmax=lag
        truep[lag] = tpl
plt.plot(aip[p+q:],truep[p+q:], color='k', linewidth='2')
print 'pmin = {0:5.3f} at lag {1:5d}'.format(pmin,lagmin)
print 'pmax = {0:5.3f} at lag {1:5d}'.format(pmax,lagmax)
plt.savefig('resid/{0:s}'.format(outputcode))
plt.show()
#
#        Create Summary file if none exists and append
#
if True:
        Smmry = 'Summary'
        Exists=True
        try:
                f = open(Smmry,'r')
        except IOError:
                Exists = False
        else:
                f.close()
        if not Exists:
                f = open(Smmry,'w')
                f.write('{0}{1}'.format(title,'\n\n'))
                header1 = '                          Ljung--Box\n'
                f.write(header1)
                header2 = ' pq      resid var         Q      pmin\n'
                f.write(header2)
                f.close()
        f = open(Smmry,'a')
        f.write(' {0:s}  {1:13.3e}{2:10.4f}{3:10.4f}\n'\
        .format(outputcode,epsvar,Qpmin,pmin))
        f.close()
print
print
```

OPspectra.py

```
""" OPspectra.py - reads temperature, eccentricity, epsilon and
    omega for the last 490 kyear listed in OP491 prepared by
    opList.py. Plots time series and spectra of T, dT, ecc,
    epsilon and omega.
    Shows times when dT exceeds 2 standard deviations from the mean.
    Shows 99.9 percent confidence limits for dT spectrum.
"""
import numpy as np
import matplotlib.pyplot as plt
from matplotlib import gridspec
from scipy.fftpack import fft, ifft
from scipy import stats
#
ipfile='OP491'
Ns = len(open(ipfile).readlines())-1
print 'Ns = ',Ns
f = open(ipfile, 'r')

kyear = np.zeros(Ns,float)
kym = np.zeros(Ns-1,float)
T = np.zeros(Ns,float)
dT = np.zeros(Ns-1,float)
ecc = np.zeros(Ns,float)
epsilon = np.zeros(Ns,float)
omega = np.zeros(Ns,float)
i = 0
j = 0
for line in f:
        if i>0:
                myline = line.split()
                kyear[j] = -float(myline[0])
                T[j] = float(myline[1])
                ecc[j] = float(myline[2])
                epsilon[j] = float(myline[3])
                omega[j] = float(myline[4])
                j+=1
        i+=1
f.close()
prec=np.sin(omega)

for i in range(Ns-1):
        dT[i]=T[i]-T[i+1]
        kym[i]=.5*(kyear[i]+kyear[i+1])

#
figname=ipfile
#
```

```
#
fig=plt.figure(num=1,figsize=(6.4,9.6))
gs = gridspec.GridSpec(5, 1, height_ratios=[1,1,1,1,1])
#
ax=plt.subplot(gs[0])
plt.plot(kyear,T,color='k')
#plt.title(figname)
plt.ylim(-10.0,4.0)
plt.yticks([-8,-4,0,4])
plt.xlim(-500,0)
plt.xticks([])
plt.hlines(0,-499,0,linestyle='--')
ax.annotate('a',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.92,.8), textcoords='axes fraction')

ax=plt.subplot(gs[1])
plt.plot(kym,dT,color='k')
plt.xlim(-500,0)
plt.xticks([])
plt.ylim(-3,3)
plt.yticks([-3,0,3])
mean=np.mean(dT)
std=np.std(dT)
plt.hlines(mean+2*std,-499,0,linestyle='--')
plt.hlines(mean,-499,0,linestyle='--')
plt.hlines(mean-2*std,-499,0,linestyle='--')
ax.annotate('b',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.92,.8), textcoords='axes fraction')

ax=plt.subplot(gs[2])
plt.plot(kyear,ecc,color='k')
plt.ylim(0,.05)
plt.xlim(-500,0)
plt.xticks([])
plt.yticks([0,.05])
for i in range(Ns-2):
        if dT[i]>mean+2*std:
                plt.vlines(kym[i],0,.05)
ax.annotate('e',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.92,.8), textcoords='axes fraction')

ax=plt.subplot(gs[3])
plt.plot(kyear,epsilon,color='k')
plt.xlim(-500,0)
plt.xticks([])
plt.ylim(.38,.43)
plt.yticks([.38,.43])
for i in range(Ns-2):
        if dT[i]>mean+2*std:
                plt.vlines(kym[i],.38,.43)
```

```
ax.annotate('o',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.92,.8), textcoords='axes fraction')

ax=plt.subplot(gs[4])
plt.plot(kyear,prec,color='k')
plt.ylim(-1,1)
plt.xlim(-500,0)
plt.xlabel('kyr BP')
plt.yticks([-1,0,1])
plt.hlines(0,-499,0,linestyle='--')
for i in range(Ns-2):
        if dT[i]>mean+2*std:
                plt.vlines(kym[i],-1,1)
ax.annotate('p',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.9,.8), textcoords='axes fraction')

plt.savefig(figname)
plt.show()
#
#        Normalize all time series to zero mean and unit variance
#
T=(T-np.mean(T))/np.std(T)
dT=(dT-np.mean(dT))/np.std(dT)
#
figname="OPspectra"
f=open('OPPgrams','w')
fig=plt.figure(num=1,figsize=(6.4,8.0))
gs = gridspec.GridSpec(1, 1, height_ratios=[1])
ax=plt.subplot(gs[0])
#
X = np.linspace(0,0.5,(Ns)/2)
XX = X[1:]
lX = np.log(XX)/np.log(10.)
#
Y = fft(T)
S = np.abs(Y[0:(Ns)/2])**2
SS = S[1:]
lS = np.log(SS)/np.log(10.)
plt.plot(lX,lS,color='k')
ax.annotate('a',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.25,.95), textcoords='axes fraction')
f.write('\n a    log f    log P        P        f        T   \n')
for i in range(31):
        f.write(\
        '{0:2d} {1:8.3f} {2:8.3f} {3:8.0f} {4:8.5f} {5:8.4f}\n'.\
        format(i,lX[i],lS[i],SS[i],XX[i],1/XX[i]))
#
Y = fft(dT)
S = np.abs(Y[0:(Ns)/2])**2
SS = S[1:]
```

```
lS = np.log(SS)/np.log(10.)
plt.plot(lX,lS-4,color='k')
#
#        Prepare confidence limits using F-test
#
Sbar=np.mean(S)
lSbar=np.log(Sbar)/np.log(10.)
dfn = 2
dfd = Ns-1
cl001 = np.log(Sbar*stats.f.ppf(.001,dfn,dfd))/np.log(10.)
cl999 = np.log(Sbar*stats.f.ppf(.999,dfn,dfd))/np.log(10.)
plt.hlines(cl999-4,-3.,0.,color='k',linestyle='--')
plt.hlines(cl001-4,-3.,0.,color='k',linestyle='--')
ax.annotate('b',xy=(.7,.1),xycoords='axes fraction',\
xytext=(.25,.68), textcoords='axes fraction')
f.write('\n b    log f    log P        P        f        T   \n')
for i in range(31):
        f.write(\
        '{0:2d} {1:8.3f} {2:8.3f} {3:8.0f} {4:8.5f} {5:8.4f}\n'.\
        format(i,lX[i],lS[i],SS[i],XX[i],1/XX[i]))
#
Y = fft(ecc)
S = np.abs(Y[0:(Ns)/2])**2
SS = S[1:]
lS = np.log(SS)/np.log(10.)
plt.plot(lX,lS-3.5,color='k')
ax.annotate('e',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.25,.6), textcoords='axes fraction')
f.write('\n c    log f    log P        P        f        T   \n')
for i in range(31):
        f.write(\
        '{0:2d} {1:8.3f} {2:8.3f} {3:8.0f} {4:8.5f} {5:8.4f}\n'.\
        format(i,lX[i],lS[i],SS[i]*10000,XX[i],1/XX[i]))
#
Y = fft(epsilon)
S = np.abs(Y[0:(Ns)/2])**2
SS = S[1:]
lS = np.log(SS)/np.log(10.)
plt.plot(lX,lS-7.1,color='k')
ax.annotate('o',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.25,.31), textcoords='axes fraction')
f.write('\n d    log f    log P        P        f        T   \n')
for i in range(31):
        f.write(\
        '{0:2d} {1:8.3f} {2:8.3f} {3:8.0f} {4:8.5f} {5:8.4f}\n'.\
        format(i,lX[i],lS[i],SS[i]*10000,XX[i],1/XX[i]))
#
Y = fft(prec)
S = np.abs(Y[0:(Ns)/2])**2
SS = S[1:]
```

```
lS = np.log(SS)/np.log(10.)
plt.plot(lX,lS-15,color='k')
ax.annotate('p',xy=(.95,.1),xycoords='axes fraction',\
xytext=(.25,.1), textcoords='axes fraction')
f.write('\n e    log f    log P        P        f        T   \n')
for i in range(31):
        f.write(\
        '{0:2d} {1:8.3f} {2:8.3f} {3:8.0f} {4:8.5f} {5:8.4f}\n'.\
        format(i,lX[i],lS[i],SS[i]*1000,XX[i],1/XX[i]))
#
f.close()
plt.ylim(-16,5)
plt.yticks([-16,-15,-14,-13,-12,-11,-10,-9,-8,-7,-6,-5,-4,\
-3,-2,-1,0,1,2,3,4,5])
plt.grid()
plt.xlabel("Log(frequency/kyears)")
plt.ylabel("Log(Variance Density)")
#
#        Plot key frequencies.
#
keyf=np.array([.00922,.0246,.044])
lkeyf=np.log(keyf)/np.log(10.)
for lkeyf1 in lkeyf:
        plt.vlines(lkeyf1,-16,5,color='k', linestyle='--')
#
plt.savefig(figname)
plt.show()
```

OPlist.py

```
""" OPlist - writes list of orbital parameters vs kyear
"""
import numpy as np
import scipy.io as spio
import matplotlib.pyplot as plt
from matplotlib import gridspec
import DailyInsolation as DI
#
#
#
# DomeCdeuterium
#
#   read in raw data
#
Nr = 5159
f = open('DomeCdeuterium.txt', 'r')
title = 'EPICA Dome C Deuterium'
rx = np.zeros(Nr,float)
ry = np.zeros(Nr,float)
i = 1
j = 0
for line in f:
    if i>104:
        myline = line.split()
        #print myline, len(myline)
        if len(myline)>4:
            rx[j] = float(myline[2])/1000.
            ry[j] = float(myline[4])
            if rx[j]>490:
                break
            j+=1
    i+=1
f.close()
#
#   prepare time series
#
tmin = 0
#tmax=5
tmax = 491.
#Ns = 5
Ns = 491
dt = (tmax-tmin)/Ns
df = 1/(Ns*dt)
fN= 0.5/dt              #Nyquist frequency
x = np.zeros(Ns,float)
y = np.zeros(Ns,float)
Nempty = 0
```

```
for j in range(Ns):
    x[j] = tmin+j*dt
    t1 = tmin+j*dt-dt/2
    t2 = tmin+j*dt+dt/2
    nj = 0
    sj = 0
    for i in range(Nr):
        if t1<rx[i] and rx[i]<=t2:
            nj+=1
            sj+=ry[i]
    if nj>0:
        y[j]=sj/nj
    else:
        if j>0:
            y[j]=y[j-1]
            Nempty += 1
        else:
            y[0]=ry[0]
        print "no data in <",t1,",",t2,">"
print Nempty, " empty intervals."
#print x
#
#
#
kyear=-x.copy()
ecc,epsilon,omega=DI.orbital_parameters(-kyear)
opfile='OP{0:03d}'.format(Ns)
f=open(opfile,'w')
header='kyr          T        ecc    epsilon        omega\n'
f.write(header)
for iy,kyear1 in enumerate(kyear):
        string='{0:3.0f} {1:10.6f} {2:10.6f} {3:10.6f} {4:12.6f}\n'\
.format(-kyear1,y[iy],ecc[iy],epsilon[iy],omega[iy])
        f.write(string)
f.close()
```

DailyInsolation.py

```
""" DailyInsolation.py - Functions for computing Daily Insolation

 References:
   Berger A. and Loutre M.F. (1991). Insolation values for the
   climate of the last 10 million years.
   Quaternary Science Reviews, 10(4), 297-317.
   Berger A. (1978). Long-term variations of daily insolation and
   Quaternary climatic changes.
   Journal of Atmospheric Science, 35(12), 2362-2367.

 Authors:
 MATLAB
   Ian Eisenman and Peter Huybers, Harvard University, August 2006
   eisenman@fas.harvard.edu
   This file is available online at
   http://deas.harvard.edu/~eisenman/downloads
 PYTHON
   John Reid
   johnsinclairreid@gmail.com
"""
import numpy as np
import scipy.io as spio

def daily_insolation(*args):
        kyear=args[0]
        lat=args[1]
        day=args[2]
        n1=1.                #refractive index of air
        n2=1.333        #refractive index of Water
        tiny=1.4e-45
        [ecc,epsilon,omega]=orbital_parameters(-kyear)
        obliquity=epsilon*180/np.pi
        long_perh=omega*180/np.pi
        lat=lat*np.pi/180         # latitude
        delta_lambda_m=(day-80)*2*np.pi/365.2422
        beta=(1-ecc**2)**(1/2)
        lambda_m0=-2*((1/2*ecc+1/8*ecc**3)*(1+beta)*np.sin(-omega)\
        -1/4*ecc**2*(1/2+beta)*np.sin(-2*omega)\
        +1/8*ecc**3*(1/3+beta)*(np.sin(-3*omega)))
        lambda_m=lambda_m0+delta_lambda_m
        lambda1=lambda_m+(2*ecc-1/4*ecc**3)*np.sin(lambda_m-omega)
        lambda1+=(5/4)*ecc**2*np.sin(2*(lambda_m-omega))
        lambda1+=(13/12)*ecc**3*np.sin(3*(lambda_m-omega))
        So=1365
        # solar constant (W/m**2)
        delta=np.arcsin(np.sin(epsilon)*np.sin(lambda1))
        # declination of the sun
```

```
        cnst=So/np.pi*(1+ecc*np.cos(lambda1-omega))**2/(1-ecc**2)**2
        Ho=np.empty_like(delta)
        Fsum=np.empty_like(delta)
        Asum=np.empty_like(delta)
        for index,delta1 in np.ndenumerate(delta):
                if(abs(lat)>=np.pi/2-abs(delta1)) and (lat*delta1>0):
                        Ho[index]=np.pi
                elif(abs(lat)>=np.pi/2-abs(delta1)) \
                and (lat*delta1 <= 0):
                        Ho[index]=0
                else:
                        Ho[index]=np.arccos(-np.tan(lat)\
                        *np.tan(delta1))        \
                        #hour angle at sunrise/set
                num=20
                dh=Ho[index]/num
                Fsum[index]=0
                Asum[index]=0
                for h in np.linspace(0,Ho[index],num):
                        sinh=(np.sin(lat)*np.sin(delta1)\
                         + np.cos(lat)*np.cos(delta1)*np.cos(h))*dh
                        Fsum[index]+=sinh
                        theta1 = np.pi/2-h
                        theta2 = np.arcsin(n1*np.sin(theta1)/n2)
                        plus=theta1+theta2
                        if abs(plus)<tiny:
                                plus=tiny
                        minus=theta1-theta2
                        if abs(minus)<tiny:
                                minus=tiny
                        r12p=np.tan(minus)/np.tan(plus)
                        r12s=np.sin(minus)/np.sin(plus)
                        FracReflected=0.5*(r12p**2+r12s**2)
                        FracAbsorbed = 1.0-FracReflected
                        Asum[index]+=sinh*FracAbsorbed
        Fsw=cnst*(Ho*np.sin(lat)*np.sin(delta) \
        + np.cos(lat)*np.cos(delta)*np.sin(Ho))
        Ftest=cnst*Fsum
        WSAbs=cnst*Asum
        return Fsw, ecc, obliquity, long_perh, Ftest, WSAbs

def orbital_parameters(year):
        tmax=5001                        # max years before present
        mat = spio.loadmat('orbital_parameter_data.mat'\
        ,squeeze_me=True)
        m=np.array(mat['m'])
        kyear0=m[:tmax,0]
        # kyears before present for data (kyear0>=0)
        ecc0=m[:tmax,1]                 # eccentricity
        # add 180 degrees to omega
```

```
        #(see lambda definition, Berger 1978 Appendix)
        omega0=m[:tmax,2]+180
        # longitude of perihelion (precession angle)
        omega0=np.unwrap(omega0*np.pi/180)*180/np.pi\
# remove discontinuities (360 degree jumps)
        epsilon0=m[:tmax,3]         # obliquity angle
        iyear=year.astype(int)
        frac=year-iyear
        ecc=ecc0[iyear]+frac*(ecc0[iyear+1]-ecc0[iyear])
        omega=(omega0[iyear]+frac*(omega0[iyear+1]\
        -omega0[iyear]))*np.pi/180
        epsilon=(epsilon0[iyear]+frac*(epsilon0[iyear+1]\
        -epsilon0[iyear]))*np.pi/180
        return ecc,epsilon,omega
```

mandala.py

```
#! /usr/bin/env python
""" mandala.py

        John Reid        2004

        Press Escape to exit.
"""
import numpy as np
import time
import threading
import pygame
from pygame.locals import *

stopEvent=threading.Event()
swidth=950
sheight=1250
sw2=swidth/2
sh2=sheight/2
a = np.zeros((sheight,swidth,3),np.uint8)
b = np.zeros((sheight,swidth,3),np.uint8)

pygame.init()
windowSurface = pygame.display.set_mode((sheight,swidth),FULLSCREEN)

sinceblack=0
blackcount=9
sincecircle=0
circlecount=4
pygame.mouse.set_visible(False)
while not stopEvent.isSet():
        ran=np.random.rand(5)
        rw2=np.int(ran[0]*sw2)
        rh2=np.int(ran[1]*sh2)
        if sinceblack>1:
                R=ran[2]*255
                G=ran[3]*255
                B=ran[4]*255
        else:
                R=0
                G=0
                B=0
        if sincecircle==circlecount:
                sincecircle=0
                for i in range(swidth):
                        x = i-sw2
                        for j in range(sheight):
                                y = j-sh2
```

```
                                if x*x+y*y<0.9*rh2*rh2:
                                        b[j,i,0]=R
                                        b[j,i,1]=G
                                        b[j,i,2]=B
        else:
                for i in range(sw2-rw2,sw2+rw2):
                        for j in range(sh2-rh2,sh2+rh2):
                                b[j,i,0]=R
                                b[j,i,1]=G
                                b[j,i,2]=B
        a=np.bitwise_xor(a,b)
        pygame.surfarray.blit_array(windowSurface, a)
        b=np.bitwise_xor(b,b)
        pygame.display.flip()
        time.sleep(2)
        sinceblack+=1
        sincecircle+=1
        if sinceblack>=blackcount:
                sinceblack=0
#                time.sleep(2)
                a=np.right_shift(a,[1])
        for event in pygame.event.get():
                if event.type == pygame.KEYDOWN:
                        if event.key == pygame.K_ESCAPE:
                                pygame.mouse.set_visible(True)
                                quit()
                                break
```

Bibliography

Abramowitz, M. and I. A. Stegun (1965). *HANDBOOK OF MATHEMATICAL FUNCTIONS*. 180 Varick Street, New York, N.Y. 10014: Dover Publications, Inc.

Akaike, H. (1970). Statistical predictor identification. *Ann. Inst. Statis. Math. 22*, 233–217.

Arneborg, J., J. Heinemeier, N. Lynnerup, H. L. Nielsen, N. Rud, and A. E. Sveinbjornsdottir (1999). Changes of diet of the greenland vikings determined from stable carbon isotope analysis and 14c dating of their bones. *Radiocarbon 41*, 157–168.

Bartlett, M. S. (1948). Smoothing periodograms from time-series with continuous spectra. *Nature 161*, 686–687.

Bendick, R. and R. Bilham (2017). Do weak global stresses synchronize earthquakes? *Geophys. Res. Lett. 44*, 8320–8327.

Berger, A. L. (1979). Long-term variations of daily insolation and quaternary climatic changes. *Journal of the Atmospheric Sciences 35*, 2362–2367.

Berger, A. L. and M. F. Loutre (1991). Insolation values for the climate of the last 10 million years. *Quaternary Science Reviews 10*(4), 297–317.

Blackman, R. B. and J. W. Tukey (1958). *THE MEASUREMENT OF POWER SPECTRA*. 180 Varick Street, New York, N.Y. 10014: Dover Publications, Inc.

Bracewell, R. N. (1986). *THE FOURIER TRANSFORM AND ITS APPLICATIONS*. McGraw-Hill Book Co-Singapore: McGraw-Hill, Inc.

Bretz, J. H. (1923). The channeled scabland of the columbia plateau. *Journal of Geology 31*, 617–649.

Carey, S. W. (1999). *Earth, Universe, Cosmos*. University of Tasmania.

Diamond, J. M. (2005). *Collapse: how societies choose to fail or succeed*. New York: Viking.

Echelmeyer, K., T. Clarke, and W. Harrison (1991). Surficial glaciology of jakobshavns isbræ, west greenland: Part i. surface morphology. *Journal of Glaciology 37*(127), 368–382.

Elderfield, H. and A. Schultz (1996). Mid-ocean ridge hydrothermal heat fluxes and the chemical composition of the ocean. *Annu. Rev. Earth Planet. Sci. 24*(127), 191–244.

Fisher, R. (1934). *Statistical Methods for Research Workers*. Tweeddale Court, Edinburgh: Oliver and Boyd.

Fitton, J. and H. Dunlop (1985). The cameroon line, west africa, and its bearing on the origin of oceanic and continental alkali basalt. *Earth Plan. Sci. Letters*.

Flocco, D., D. L. Feltham, D. Schroeder, and M. Tsamados (2016). Impact of refreezing melt ponds on arctic sea ice basal growth. *The Cryosphere Discuss.*.

Glatzmaier, G. and P. H. Roberts (1995). A three-dimensional self-consistent computer simu- lation of a geomagnetic field reversal. *Nature 377*, 203–209.

Granger, C. W. J. and P. Newbold (1974). Spurious regression in econometrics. *Journal of Econometrics 2*, 111–120.

Gull, S. F. and G. J. Daniell (1978). Image reconstruction from incomplete and noisy data. *Nature 272*, 686–690.

Hamilton, E. J. (1994). *Time Series Analysis*. Princeton, New Jersey: Princeton University Press.

Hannan, E. J. (1960). *TIME SERIES ANALYSIS*. London: Methuen and Co.

Hasselmann, K. (1976). Stochastic climate models, part I, theory. *Tellus XXVIII*(6), 473–485.

Hays, J. D., J. Imbrie, and N. J. Shackleton (1976). Variations in the earth's orbit: Pacemaker of the ice ages. *Science 194*(4270), 1121–1132.

Huybers, P. (2006). Early pleistocene glacial cycles and the integrated summer insolation forcing. *Science 313*(5786), 508–511.

Huybers, P. and I. Eisenman (2006). Integrated summer insolation calculations. *IGBP PAGES/World Data Center for Paleoclimatology, Data Contribution Series 2006-079. NOAA/NCDC Paleoclimatology Program, Boulder CO, USA*.

Jouzel, J., V. Masson-Delmotte, O. Cattani, G. Dreyfus, S. Falourd, G. Hoffmann, B. Minster, J. Nouet, J. Barnola, J. Chappellaz, H. Fischer, J. Gallet, S. Johnsen, M. Leuenberger, L. Loulergue, D. Luethi, H. Oerter, F. Parrenin, G. Raisbeck, D. Raynaud, A. Schilt, J. Schwander, E. Selmo, R. Souchez, R. Spahni, B. Stauffer, J. P. Steffensen, B. Stenni, T. F. Stocker, J. L. Tison, M. Werner, and E. W. Wolff (2007). Orbital and millennial antarctic climate variability over the past 800,000 years. *Science 317*(5839), 793–797.

Koutsoyiannis, D. (2010). A random walk on water. *Hydrol. Earth Syst. Sci. 14*, 585–601.

Ljung, G. M. and G. E. P. Box (1978). On a measure of the lack of fit in time series models. *Biometrika 65*, 297–303.

Loret, J. (2003). A cultural icon: Scientific exploration into the world's environmental problems in microcosm. In J. Loret and J. T. Tancredi (Eds.), *Easter Island*, pp. 19–28. Springer US.

Malinka, A., E. Zege, L. Istomena, G. Heygster, G. Spreen, D. Perovich, and C. Polashenski (2017). Reflective properties of melt ponds on sea ice. *The Cryosphere Discuss.*.

Margold, M., C. R. Stokes, and C. D. Clark (2015). Ice streams in the laurentide ice sheet:f identification, characteristics and comparison to modern ice sheets. *Earth Science Reviews 143*, 117–146.

Maslin, M. and A. Ridgwell (2005). Mid-pleistocene revolution and the "eccentricity myth". *Geological Society, London, Special Publications 247*, 19–34.

Milankovitch, M. (1941). *K. Serb. Akad. Beogr. Spec. Publ. 132*.

Morice, C. P., J. J. Kennedy, N. A. Rayner, and P. D. Jones (2012). Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The hadcrut4 dataset. *Journal of Geophysical Research 117*, D08101.

Moskowitz, L. (1964, December). Estimates of the Power Spectrums for Fully Developed Seas for Wind Speeds of 20 to 40 Knots. *J. Geophys. Res. 69*, 5161–5179.

Nelson, C. R. and H. Kang (1984). Pitfalls in the use of time as an explanatory variable in regression. *Journal of Business and Economic Statistics 2*, 73–82.

Patton, H., A. H. A., K. Andreassen, A. Auriac, P. L. Whitehouse, A. P. Stroevenc, C. Shackleton, M. Winsborrow, J. Heyman, and A. M. Hall (2017). Deglaciation of the eurasian ice sheet complex. *Quaternary Science Reviews 169*, 148–172.

Pelletier, J. D. (2002). Natural variability of atmospheric temperatures and geomagnetic intensity over a wide range of time scales. *PNAS 99 (Suppl 9)*(S1), 2546–2553.

Phillips, O. M. (1958). The equilibrium range in the spectrum of wind-generated waves. *Journal of Fluid Mechanics 4*(4), 426–434.

Pierson, Jr., W. J. and L. Moskowitz (1964). A proposed spectral form for fully developed wind seas based on the similarity theory of S. A. Kitaigorodskii. *J. Geophys. Res. 69*, 5181–5190.

Pollack, H. N., S. J. Hurter, and J. R. Johnson (1993). Heat flow from the earth's interior:analysis of global data set. *Rev. Geophys. 31*, 267–280.

Popper, K. R. (1962). *CONJECTURES AND REFUTATIONS*. New York: Basic Books.

Reid, J. (2017). There is no significant trend in global average temperature. *Energy & Environment 28*(3), 302–315.

Reid, J. S. (1979). Confidence limits and maximum entropy spectra. *J. Geophys. Res. 84*(A9), 5289–5301.

Reid, J. S. (1992). The sideband instability and the onset of wave breaking. In M. L. Banner and R. H. J. Grimshaw (Eds.), *Breaking Waves*, Berlin, Heidelberg, pp. 155–159. Springer Berlin Heidelberg.

Richter, F. M. (1984). Time and space scales of mantle convection. In H. H.D. and A. Trendall (Eds.), *Patterns of Change in Earth Evolution*, Volume 5, pp. 271–289. Berlin, Heidelberg: Springer.

Selmes, N., T. Murray, and T. D. James (2013, February). Characterizing supraglacial lake drainage and freezing on the Greenland Ice Sheet. *The Cryosphere Discussions 7*, 475–505.

So, H. C., Y. T. Chan, Q. Ma, and P. C. Ching (1999). Comparison of various periodograms for sinusoid detection and frequency estimation. *IEEE Trans. Aerospace Electronic Systems 35*(3), 945–951.

Tang, X., M. C. Ntam, J. Dong, E. S. G.Rainey, and A. Kavner (2014). The thermal conductivity of earth's lower mantle. *Geophys. Res. Lett. 41*(3), 2746–2752.

Thébault, E., C. Finlay, C. Beggan, P. Alken, J. Aubert, O. Barrois, ..., and T. Zvereva (2015). International geomagnetic reference field: the 12th generation. *Earth, Planets and Space 67:79.*

Tivey, M. K. (2015). Generation of seafloor hydrothermal vent fluids and associated mineral deposits. *Oceanography 20*, 50–65.

Turcotte, D. L. and E. R. Oxburgh (1972). Mantle convection and the new global tectonics. *Review of Fluid Mechanics 4*, 33–66.

White, J. D. L., J. L. Smellie, and D. A. Clague (2003). Introduction. In J. D. L. White, J. L. Smellie, and D. A. Clague (Eds.), *Explosive Subaqueous Volcanism, Geophysical Monograph 140*, pp. 1–23. American Geophysical Union.

Woit, P. (2006). *Not Even Wrong: The Failure of String Theory and the Search for Unity in Physical Law.* New York: Basic Books.

Yule, G. U. (1926). Why do we sometimes get nonsense-correlations between time-series?–a study in sampling and the nature of time series. *J. Royal Statistical Society 89*, 1–63.

Index