Scientific Uncertainty: A User's Guide

Seamus Bradley

Department of Philosophy, Logic and Scientific Method
London School of Economics and Political Science
Houghton Street
WC2A 2AE
London
lse.ac.uk/philosophy
seamusbradley.net

August 9, 2011

Abstract

There are different kinds of uncertainty. I outline some of the various ways that uncertainty enters science, focusing on uncertainty in climate science and weather prediction. I then show how we cope with some of these sources of error through sophisticated modelling techniques. I show how we maintain confidence in the face of error.

Contents

> On two occasions I have been asked, "Pray, Mr. Babbage, if you put into the machine wrong figures, will the right answers come out?" I am not able rightly to apprehend the kind of confusion of ideas that could provoke such a question
>
> Charles Babbage

1 Two motivating quotes

How do error and uncertainty enter science? What can we do about it? This paper explores these questions in the context of scientific modelling, specifically climate modelling. While my main interest is in models of the climate, I think a lot of what I say will translate into other disciplines. I want to start this paper with two quotes that highlight two aspects of the modelling process.

1.1 Laplace's demon: ideal scientist

In trying to characterise what determinism was, Laplace described his famous "demon" who was a kind of ideal scientist.

> 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. The perfection that the human mind has been able to give to astronomy affords but a feeble outline of such an intelligence. Discoveries in mechanics and geometry, coupled with those in universal gravitation, have brought the mind within reach of comprehending in the same analytical formula the past and the future state of the system of the world. All of the mind's efforts in the search for truth tend to approximate the intelligence we have just imagined, although it will forever remain infinitely remote from such an intelligence.
>
> Laplace (1951 [1816])[1]

This demon knows everything there is to know: for him, there is no uncertainty. We of course fall far short of Laplace's ideal scientist, and our actual

[1] This is actually a slightly different translation to the one cited, given in Smith (2007).

science is full of uncertainties of various kinds. My aim here is to characterise what kinds of uncertainty arise in science, and point out some of the ways we cope with them.

Laplace attributed three important capacities to his demon.[2] First it must know "all the forces that set nature in motion": it must be using the same equations as Nature is. Second, it must have perfect knowledge of the initial conditions: "all positions of which nature is composed". Laplace makes reference only to positions, but let's grant that he was aware his demon would need to know various other particulars of the initial conditions — momentum, charge... — let's pretend Laplace was talking about *position in state space*. The third capacity that Laplace grants his demon is that it be "vast enough to submit these data to analysis". This translates roughly as infinite computational power. I will have less to say about this third capacity, at least directly.

Laplace was operating in a purely Newtonian world, without quantum effects, without relativistic worries. In this paper, I will do the same. On the whole, climate scientists make the assumption that it is safe to model the climate system as a Newtonian, deterministic system.

1.2 On Exactitude in Science

My second motivating quote is pulling in the opposite direction. Laplace is thinking about the ultimate in accurate prediction, this story from Jorge Luis Borges points to the limits inherent in modelling.

> ...In that Empire, the Art of Cartography attained such Perfection that the map of a single Province occupied the entirety of a City, and the map of the Empire, the entirety of a Province. In time, those Unconscionable Maps no longer satisfied, and the Cartographers Guilds struck a Map of the Empire whose size was that of the Empire, and which coincided point for point with it. The following Generations, who were not so fond of the Study of Cartography as their Forebears had been, saw that that vast Map was Useless, and not without some Pitilessness was it, that they delivered it up to the Inclemencies of Sun and Winters. In the Deserts of the West, still today, there are Tattered Ruins of that Map, inhabited by Animals and Beggars; in all the Land there is no other Relic of the Disciplines of Geography.
>
> Borges (1999)

Borges wasn't the first to consider this idea[3], but his expression of it is characteristically evocative. The map here is serving as a metaphor for the modelling process. A map that is too big is like a model that is too slow. And

[2] Smith (2007) characterises Laplace's demon in this way, and I follow his exposition.

[3] Lewis Carroll toyed with the concept in *Sylvie and Bruno Concluded*

the point I want to make is that modelling is *inherently*, *inescapably* a process of abstraction, idealisation, ignoring of detail. In a way, at the very heart of the whole project of modelling a phenomenon is the need to introduce error: an exact replica (a 1:1 map) of the phenomenon of interest wouldn't tell you anything. This isn't strictly true. If we had a perfect 1:1 replica of the climate we could run various different scenarios on it and explore the system more in the way of a lab experiment. Given the time taken to actually run that sort of model, it would not be a useful *predictive* tool. So what follows isn't a criticism of scientific modelling, but rather marvelling at how much we can throw away and still have informative useful models.

The question we always have to ask ourselves when looking at complex models is: "what is the value added by incorporating these extra processes?" Does the extra insight gained from adding in these extra feedbacks balance out making the model slower and more complex? Does the extra detail justify making the map bigger?

I'm not making any claims that what I say is specific to computer simulation. I more or less agree with Frigg and Reiss (2009) that simulation brings out no new philosophical issues over and above those of other kinds of modelling.[4] Simulation does however emphasise different elements of the issues. Given my interest in climate science specifically, a focus on simulation seems reasonable.

1.3 Characterisations of uncertainty

I should mention something about my use of the word "uncertainty". I am using this as a catch-all term for all the ways one might fail to be certain about things. Economists and decision theorists sometimes make a distinction between "risk" where the probabilities are known, and "uncertainty" where they are not. This distinction is traced back to the work of Frank Knight (Knight 1921). Unfortunately, physicists have the opposite view: "uncertainty" connotes a situation where you know the probabilities and "ambiguity" is the term for a situation where the probabilities are unknown. I use uncertainty to cover both of these possibilities.[5]

This is not the first attempt to classify kinds of uncertainty. For example Walker et al. (2003) offer a framework for understanding uncertainty when using evidence from models to make decisions. They identify three distinct "axes" of uncertainty of relevance for "model-based decision support". These are:

Location Where exactly the error enters into the process

[4] I believe it is up to someone who thinks there is a relevant difference to fill in exactly how this distinction — between simulation and non-computational modelling — can be made.

[5] What it would mean to "know the probabilities" in this context isn't clear, in fact. Single case probabilities are tricky beasts, and our models of the climate are deterministic, so what is the actual chance, the real probability of some climate outcome?

Level How much the uncertainty affects the model-relevant predictions

Nature Whether the uncertainty is due to a deficiency in the modelling, or due to unavoidable natural variation

I am particularly interested in the first of these.

Regan, Colyvan, and Burgman (2002) offer a typology of uncertainty for ecologists. They define two broad kinds of uncertainty "Epistemic" and "Linguistic". I am particularly interested in the epistemic uncertainties, and my characterisation of epistemic uncertainties is not all that different from theirs.

What I aim to do differently with this paper is discuss kinds of uncertainty with a view to what coping strategies we adopt to deal with them. Understanding the nature and diversity of scientific uncertainty is important for policy making. Oversimplifying scientific evidence, or downplaying the uncertainties involved is to the detriment of model based decision making (Stirling 2010).

1.4 Outline

This paper is primarily aimed at philosophers of science and others who have an indirect interest in uncertainty in science (policy makers, research funders...). Most of what I have to say will be obvious to working scientists, although perhaps the classification and overview might be useful to them as well.

My examples and much of the discussion will be motivated by climate modelling. However, I will keep things on a fairly "conceptual" and idealised level: much of the detail will not reflect actual practices of working scientists. What I am aiming for is to explore the basic sources of errors and the basic processes for mitigating them.

We are nowhere near Laplace's demon in our capacities. The next two sections explore ways in which we fail to live up to Laplace's dream. First I look at uncertainty in our measurements of initial conditions (section 2). Second I look at uncertainty in our models (section 3). Then I turn to looking at the coping strategies we have developed to deal with uncertainty (section 4), bearing in mind the cautions of Borges' story. I finish with a discussion of realism and instrumentalism about climate models.

2 Data gathering

Scientists gather a lot of information about the phenomena under study. These data are uncertain in a number of ways. For example, the data we collect are finite strings of integers, whereas the quantities we measure can presumably be infinitely precise (2.1). Our measurement instruments might cause the data to be inaccurate (2.2). Or worse, we might not be measuring the quantities we thought we were (2.3).

2.1 Truncated data: imprecision

Let's say we're measuring temperature at a weather station. The thermometer gives a reading of, say 13.45 °C, but in fact the temperature is more like 13.452 934 °C. Our data of the system's current state is *truncated*. This is one type of uncertainty. It is a fairly mild form of uncertainty: the number is right as far as it goes.

One might complain that temperature is a derived concept, and it doesn't really make sense once you start trying to measure the temperature on smaller and smaller scales. But even granting that, our temperature data is still less precise than it *could* be. I treat this issue at more length in section 2.4.

What are the reasons for truncated data? Often, the differences in value are not important: for most purposes, all of the temperatures that round to 13.45 °C are indistinguishable, so the truncation doesn't matter. However, if we have reason to believe the system of interest behaves chaotically,[6] then even these tiny differences *can* have an effect. This will be important later.

Second, our measuring devices have only a limited accuracy, so maybe even where the differences do matter, we might not be able to detect them. Our capacity to store data is also limited, so even if we could actually make arbitrarily fine measurements, we wouldn't be able to store all that data.

Third, the longer the strings of data we collect and feed into our models, the longer the calculations take. So given our limited computational capabilities, even if we could collect more fine-grained data, maybe we would truncate it before feeding it into the computer, in the interests of computational tractability. If predicting Thursday's weather takes until Friday, the forecast is useless.

There is a trade off between precision of the data fed into the model (and the precision of the model itself) and how long that model takes to run.

2.2 Noisy data: inaccuracy

Sometimes our measurements are not perfect. To stick with the weather example, imagine that the thermometer registered 13.47 °C rather than 13.45 °C. This isn't just a case of truncation: this is an error. This isn't imprecision: this is inaccuracy. This is to assume, of course, that there is a fact of the matter about what the value *should have* been that is different from what it was. Hardcore instrumentalists about quantities might object. I defer a discussion of the metaphysical assumptions underlying my project until section 5.

There are many reasons why this sort of error might creep in. For example, the equipment might be faulty, or slightly miscalibrated. The more accurate we attempt to make our predictions, the more significant figures we try to determine, the more danger there is that noise — random fluctuations in the

[6] This awkward locution is necessary, since it is effectively *unknowable* whether some physical system is mathematically chaotic or not. Indeed, a physical system's being chaotic isn't even well defined: chaos is a property of mathematical equations.

system — might overwhelm the signal. For most purposes these errors will be small. But again, if we're worried about non-linear systems, then these small differences can lead to big differences somewhere down the line.

2.3 Deeper errors

Measuring physical quantities is a tricky business. It is not some straightforward pre-scientific enterprise that is, somehow, a *precursor* to scientific enquiry. Working out how to measure certain quantities is a long, involved process that makes use of theory and evolves in tandem with it. Chang (2004) charts the fraught history of thermometry, for instance. This suggests another way our measuring of the initial conditions can go wrong.

Imagine if we did not know that there is a connection between temperature and atmospheric pressure. We would not then be controlling for differences in atmospheric pressure, and this would cause our measurements of temperature to be faulty. We would falsely attribute changes in measurement reading to changes in temperature, where actually the change is due to change in pressure. That is, we would be implicitly making an assumption about our thermometer readings not being sensitive to air pressure. If this implicit assumption turns out to be important (and importantly false[7]) then our readings would not be readings of temperature, but readings of some combination of temperature and pressure. To put it another way, we wouldn't be measuring what we thought we were. There is nothing wrong with the measurement *per se*, but with our understanding of what we are getting from the measurement.

Given the indirect way we make a lot of measurements in climate science, this is a live issue. What exactly studies of ice cores and tree rings show us about the paleoclimate is a contested issue, so this is another example of where our measurement techniques are not unequivocal (see Schiermeier 2010). How one calibrates size of tree rings with estimates of the temperature at the time that ring was growing involves theory, and a great deal of knowledge of different areas of science. For example, size of tree rings could be correlated with rainfall as well as temperature. So if we made an assumption that rainfall can be ignored then this might lead to discrepancies.

Indeed, we don't have as many actual measurements of data as we would like. We mostly only have readings for surface level land values. We don't have as good information about what's going on out at sea or up in the atmosphere (Parker 2006, p.353). A technique known as "assimilation" is used to "generate missing data" using a particular type of climate model (Talagrand 1997).

To summarise: our measurement techniques might not be measuring what we thought they were. This could be due to some physical relationship among the quantities that we are not modelling, or perhaps due to some artefact of the measuring process.

[7] If the air pressure where we are measuring doesn't change, then this assumption — strictly speaking still false — would not cause misleading results.

2.4 A digression on the meaningfulness of derived quantities

Temperature is a shorthand for "mean kinetic energy". So "temperature at a point" is meaningless. Temperature is a quantity defined for some volume, and is defined by the average energy of the particles in that volume. Imagine shrinking the volume of interest until it is of a size with the atoms. Now imagine that at a particular time, that volume is empty.[8] But say at some later time, a particle whizzes through the volume. The mean kinetic energy of the box would suddenly jump from 0 to k where k is the kinetic energy of that particle, and then it would jump instantaneously back to 0 when the particle leaves the volume. On this scale, it should be obvious that temperature is not a useful quantity. In short mean kinetic energy is just not a stable, meaningful, useful concept on scales smaller than the mean free path of the atoms.

A similar meaninglessness is apparent if we consider the millionth decimal place of a temperature in a "normal sized" volume. The millionth decimal place might well change radically as particles enter and leave the volume. So at this resolution, again, the quantity is useless: it is pure noise.

But the above discussion does not undermine temperature as a meaningful, useful quantity on the appropriate scales. Of course, at some level, at some degree of precision, the whole concept of temperature breaks down; but we are not near that limit in practice. Even if the thousandth decimal place of a temperature reading might well be meaningless, there are meaningful significant figures that we aren't measuring. It does not mean that striving for more accuracy is wasted effort: it simply means that there is a limit.

The limit isn't an absolute limit on what science is capable of: if we were working on the tiny volume scale, we simply wouldn't be using quantities like temperature. We would be using the resources of quantum mechanics, rather than classical statistical mechanics and thermodynamics. So while arbitrarily precise temperature readings don't make sense as the ultimate goal of better measurement, it is still the case that there are meaningful significant figures we aren't capturing (again assuming there is a fact of the matter).

3 Model building

Now that we have gathered our data, what do we do with it? How do we turn our data into predictions? We will typically have data in the form of "time series".[9] A time series is a series of measurements of a quantity (for example temperature) indexed by when the measurement was taken. We will have time series of several quantities — temperature, pressure, precipitation, wind speed... — from a variety of locations. This gives us a rough summary of the climactic conditions at those times, in those locations.

[8] Let's pretend we've never heard of vacuum energy: we are being good Newtonians...

[9] Other kinds of data aren't all that different. I focus on time series since talking about "evolution in time" is easy and intuitive.

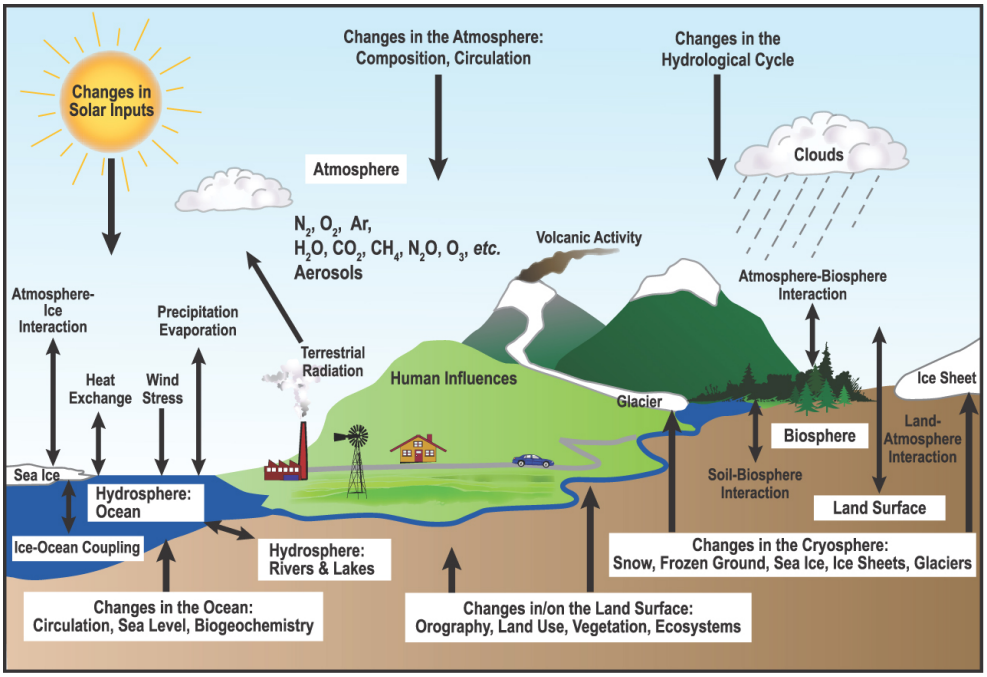

Figure 1: The climate system is tremendously complex

Using a variety of statistical tools — regression analysis, radial basis functions, autocorrelation functions and so on — we can explore how these variables change in time, and in response to changes in each other. We can also use our knowledge of the basic physics to inform how we think the relationships among the variables *should* behave. For example, using the Navier-Stokes equation,[10] the basic equation of fluid mechanics, we can guess at how wind speed in adjacent cells should evolve in response to each other over time. Lloyd (2009) claims that this sort of support from basic science gives us confidence in our models. I return to this point later.

There are a variety of simplifying assumptions that need to be made here, and none of the techniques used is infallible, so the model building process adds its own sources of error. I outline these sources of error now. The plan is to first list the sources of error, then in section 4 to discuss how we deal with these errors.

What I should make clear before beginning is that I am interested in systems that will often be chaotic or non-linear. This means that even small errors can quickly becoming worrisome. For a more detailed discussion of the particular issues with nonlinear dynamics and predictability, see Smith (2000).

[10] That said, our understanding of the Navier-Stokes equation is not perfect: indeed, we can't even be assured of the existence and uniqueness of solutions (Fefferman n.d.)

3.1 A toy example

To make the discussion more vivid, consider the following toy example. Laplace's demon is trying to predict the evolution of the fish population in my pond. He has observed the past evolution of the system and has arrived at a model of the system. He knows that the population density of fish in my pond in a week's time N_{t+1} depends on the current population density N_t. These numbers are normalised such that the absolute maximum population of fish possible is given by 1. He also knows that the future population is adversely affected by how crowded the pond is (captured by the $1 - N_t$ factor). The fish population dynamics are given by:

$$N_{t+1} = 4N_t(1 - N_t) \tag{Log}$$

This is the famous "logistic map", probably the simplest dynamics that produce chaotic behaviour.

May (1976) discusses this model in detail. Figure 2 shows how nearby initial conditions diverge radically after a few time steps. This is a figure of 9 initial conditions very close together. The solid line represents the "actual" evolution of the fish population, while the dotted lines represent what the model would predict had you got the initial condition slightly wrong in one of 8 distinct ways. As you can see, after about 12 timesteps, you could be very wrong indeed.

3.2 Curve fitting

Let's start with an easy problem. We have a one-dimensional time series of data. These might represent daily temperature readings from a particular weather station, or volume of internet traffic through a particular server, or cases of a disease at a particular hospital, or weekly estimates of the population of fish in my garden pond... The question is given the data, what kind of process could generate that time series? What sort of equation describes the dynamics of how the system evolves?

For any finite data set, any number of different graphs pass through all the points. This is a well known problem with fitting curves to data. If there are n data points, then there is a polynomial of degree $n - 1$ that will go through all of these points.[11] And in fact, there are infinitely many polynomials of higher degree that go through each of these points. On what grounds can you say that the process generating the data is represented by *this* rather than *that* function?

Choosing on the grounds of simplicity might be (at least pragmatically) appealing, but there are problems. First, what counts as a simple equation? For example, consider the following equation:

$$\alpha \sin(\beta x) \tag{Fit}$$

[11] The only case where this is not possible is if two points took the same x value, which is impossible in a time series.

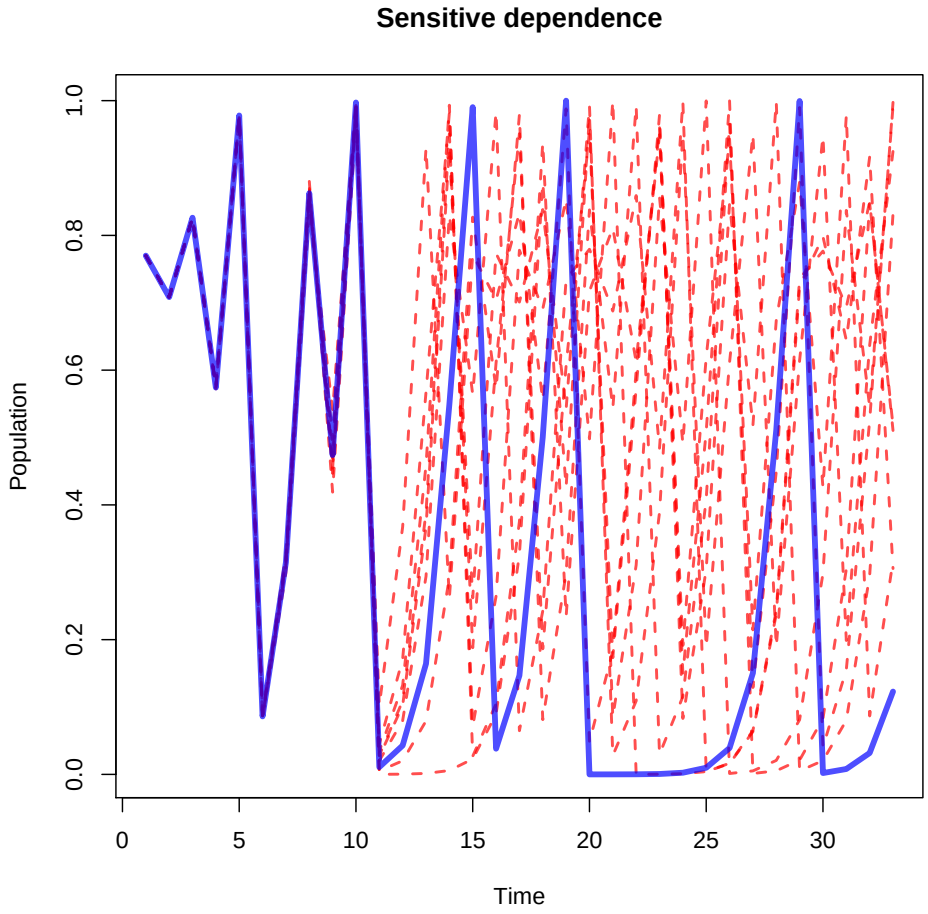

Figure 2: Diverging time series

For any finite data set, there are values of α and β such that the graph of (Fit) passes arbitrarily close to all the points. This equation is, arguably extremely simple: it has only two parameters. On another measure, however, it's a rather complicated equation, since the graphs it gives rise to are normally "unnecessarily wiggly" given the amount of data it needs to fit. That is, one feels like the curve should bend *only if* it is bending towards a data point it would otherwise miss. And what guarantee do we have that simplicity is an indicator of truth of a functional relationship?

This is a very general worry about science: how do we know that our theory is getting things right about the world?

3.3 Structure error

The standard practice when trying to fit a curve to some data is to start with a parametrised family of functions. We can then work out which member of this family most closely fits the data: we can work out which parameters fit the data best. But who's to know that that first step (picking a family of functions) is valid? What assumptions can underwrite that sort of choice? Sometimes we have basic physics evidence that suggests the functional form is, say linear. But other times, we don't have this kind of information. When considering the climate system, for example, there are any number of feedbacks all of them interacting with each other; so we don't really know what model structure is appropriate.

What interpretation can we give to this process? Fitting a curve from a family of functions would make sense if we knew that that data had been generated by some process that shares that functional form. That is, if we knew that data were generated by some function in that parametrised family, then this would be the obvious way to approach that problem. But it is almost never the case that we are justified in making that kind of assumption. So what does the choice of family mean? What justifies it?

This sort of worry about structure error arises from trying to solve the above uncertainty (3.2). I discuss this concern at more length once I have outlined what strategies are employed to overcome those worries.

3.4 Missing physics

There are processes in the climate that we know we aren't modelling. We have improved greatly (see Figure 3) but there are still many aspects of the climate system left out of our models. In the early 90s, we weren't modelling ocean currents: convection in the ocean wasn't in the model. So we in fact *don't* want our models to fit too closely to the data, because they aren't modelling all the things that are going on in the world.

There are still processes that aren't making it into our models. Some of these processes will have a negligible effect on the climate, but we can't be sure which. Given the nonlinear character of our models, tiny differences due to apparently negligible processes could cause the models to differ significantly from the target system on the sort of timescales we are interested in.

3.5 Overfitting

As well as structure error, we can be uncertain of the parameters we fix. If there is noise or inaccuracy in the data, then there is a danger of "overfitting". Using Lagrange interpolation to find the polynomial of degree $n-1$ that fits your n points exactly is obviously a crazy idea if you *know* that some of those points are inaccurate! All that you could be doing here is "fitting the noise" which clearly leads to a functional relationship that has very little relation to the underlying process. This is discussed in Hitchcock and Sober (2004). In their own words:

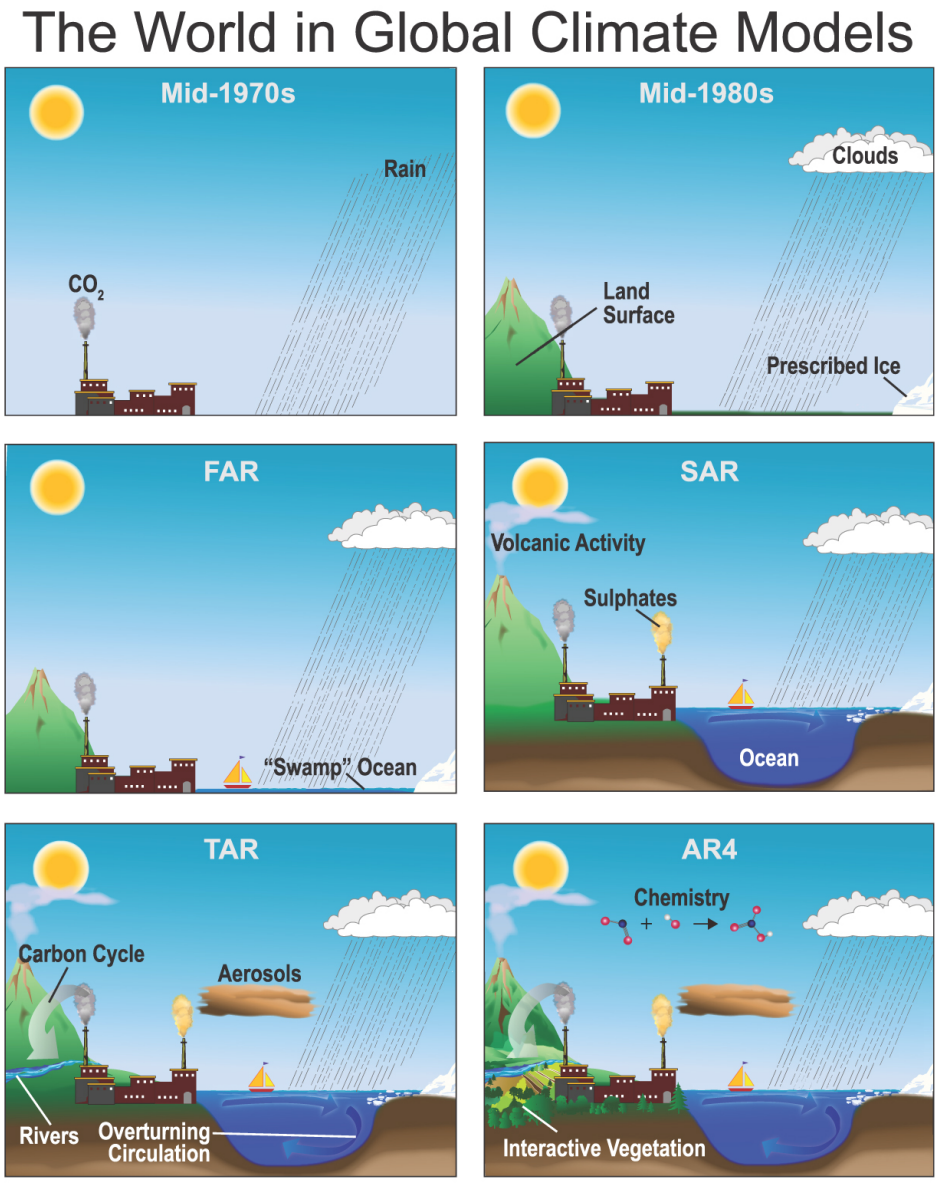

Figure 3: How climate models have improved

> The data D are bound to contain a certain amount of noise in addition to the information they carry about the underlying relationship between [the variables]. By constructing a relatively complex curve that exactly fits [all the data, the scientist] is bound to *overfit* the data. That is, she is bound to propose a theory which is too sensitive to idiosyncrasies in the data set D that are unlikely to recur in further samples drawn from the same underlying distribution.
>
> Hitchcock and Sober (2004, p.11)

So the question that needs to be asked whenever there is a choice between a complex, better fitted model and a simpler model with less good fit, is: does the extra model complexity really afford an increase in *predictive* accuracy, rather than just an increase in capacity to accommodate past data? What is the value added of the more complex model? Answering this question depends on how good the models are at prediction, but we don't really have that information for climate modelling.

This is a problem that worries climate scientists. Obviously, they want to build the best models they can: they want to build models based on the best, richest data available. They want the models to fit the data: if the models didn't fit the data, they wouldn't be very good models! But this data will inevitably contain noise. So how do they avoid the problem of overfitting? I defer answering this question until section 4.5.

3.6 Discretisation

Another issue is that most of the parameters and quantities we are interested in in atmospheric science are more or less continuous.[12] However, the data we have are discrete: hourly values of particular measurements at particular locations, for example. Quantities of interest vary continuously in time and space. We only have particular measurements from particular places. The computer implementations of our simulations are also discrete: computers are finite state machines, they can't really do continuity. So we have spatial and temporal discretisation problems.

This problem leads to "rounding errors". Normally, these aren't all that problematic, but for iterated procedures, chaotic dynamics and the like, these small errors can compound themselves and quickly become serious.

This raises issues about how to go about accurately representing the continuous system discretely, and how to evolve a discrete system so as to track the evolution of the continuous target system.

In fact, there is a deeper problem here. Continuous and discrete systems of roughly the same dynamics can have very different overall properties.[13]

[12] "More or less" because of the concerns about temperature and the like being derived quantities built up out of discrete entities, but see section 2.4

[13] Thanks to Charlotte Werndl for pointing me to this issue.

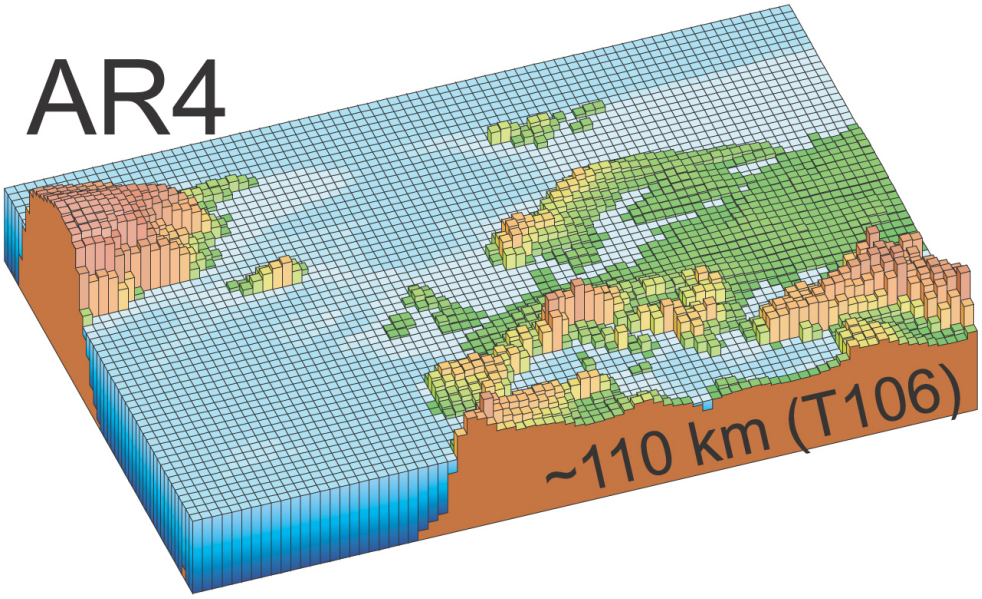

Figure 4: Europe as viewed by the models in IPCC AR4 (2007)

The logistic map (which I discussed in section 3.1) is chaotic, but the logistic equation, the "continuous version" is non-chaotic, and in fact fairly tame. This "continuous versions" is given by the differential equation $\frac{d}{dt}x(t) = 4x(t)(1 - x(t))$. The similarity should be clear.

Indeed, this must be the case, since one dimensional continuous flow cannot be "chaotic" in any sense, whereas the logistic equation is one-dimensional. Discrete dynamics, however, can exhibit chaos in one dimension, as the logistic map demonstrates.

3.7 Model resolution

There is some minimum resolution to the simulation: the world is split into grid squares 100 km on a side, say. See Figure 4 for a taste of what Europe looks like at this resolution. Then processes that happen at smaller scales aren't in the model. For example, clouds are an important factor in climate models. However, typical climate models will have too low a resolution to "resolve" clouds properly, so the effects of clouds have to be put in "by hand". How this is done will be discussed later.

Discretisation and model resolution are obviously linked. A particular discretisation method brings with it a particular scale, a particular resolution.

3.8 Implementation

Computer programs — which is what simulations are, after all — will inevitably contain bugs. Typical General Circulation Models (GCM) run to

hundreds of thousands of lines of code. It would be surprising if there were no mistakes at all in all that code. For example, perhaps a particular function takes "temperature" as a variable, but it wants a relative temperature. If it's used in some module where it is fed an absolute temperature, then this will lead to problems. Imagine if a process were expecting a temperature in Celsius and got one in Farenheit. The process would just receive a number, and would manipulate it as normal, but the mistake would propagate through subsequent processes and invalidate the results.

There will be initial conditions where they won't do what we expect them to. What ways do we have to know that *the predictions we use them for* aren't the places where they will break?

At an even lower level, hardware can be faulty or buggy (Muldoon 2007). For example, the Pentium FDIV floating point unit bug caused some calculations to go wrong by as much as 61 parts in a million. Hardly a big deal in general, but if you are dealing with potentially chaotic systems, these things can cause noticeable errors (after a few iterations). Recall Figure 2: nearby initial conditions diverge.

Different chip architectures implement arithmetic differently, so running the exact same program on different chips can lead to different results, even without hardware faults. Even different compilers, compiling the same code on the same chipset can lead to differences in output.[14] So this is an extra source of problems.

4 Coping with uncertainty

This has so far seemed to be a fairly negative discussion. However in this section, I want to outline why things aren't as bad as they might be. I want to highlight some ways we have of maintaining confidence in our models in spite of all the above sources of error.

Before starting I want to point out that these ways of maintaining confidence come in two varieties. One variety are the methodological ways of maintaining confidence. These are methods we can use to cope with uncertainty. For example, derivation from theory (section 4.2) and ensemble forecasting (section 4.4) are such methods.

The other category is harder to name. These aren't *methods* we can follow, but rather, aspects of the output that give us confidence. Robustness (section 4.6) and past success (section 4.7) are of this sort. These properties of output aren't really things we can aim for, but they are things that warrant confidence if they turn out to hold.

4.1 Make better measurements!

An obvious response to at least some of the issues raised above is that we need to make better measurements. This is a reasonable point and there are

[14] L. Smith in conversation

certainly cases where better measurements could help, but it certainly won't solve all our problems. While better measurements might mitigate against some of the problems, it won't really *solve* any of them. Better measurements would give us more significant figures in our data, but it wouldn't really stop it being the case that the data are truncated.

While mitigation of error is obviously still a good thing — and it's normally all we can do — I want to point out that it doesn't solve the problems. For practical purposes, good mitigation is as good as solving the problem. How good the mitigation has to be depends on the practical purpose in question. How many significant figures are needed such that the remaining error in initial conditions doesn't lead to an unacceptable level of error in the output depends on what counts as "acceptable" error in the output. Parker (2009) discusses this notion of "adequacy for purpose". The point is that it is important to be clear what you want out of your model, since this affects what you can put in it and still be confident of the results.

There's a question of whether better measurements are a cost-effective way of increasing adequacy. For example, we could put thousands and thousands of buoys in the seas to get more detailed readings of temperature and so on across the ocean. Would this extra data (gathered at considerable cost) lead to better models? Would other uses of that money have led to *even better* models?

This relates to Karl Popper's this idea of "accountable" predictions (Popper 1982). John Earman discusses "Popper's demon" as a weakening of Laplace's demon who is limited to only use finitely precise initial conditions (Earman 1986). This demon can, however, calculate how accurate its predictions are given the initial data. The idea is, for any given desired precision of output, the demon knows how accurate its initial conditions would need to be. So, given the model the demon has, it knows how quickly nearby initial conditions diverge. From this, it can work out how accurate its initial data have to be in order for the predictions to be adequately precise. That is, in order for the trajectories to have not spread out too much yet. This idea is related to those discussed in section 4.3.

4.2 Derivation from theory

One thing that helps with confidence is to know that our models and our parametrisations have a sound basis in basic physics. This is something Lloyd (2009) mentions as a source of "confirmation" for climate models.[15] If a model bases its heat transport processes on basic thermal physics principles, then that is good reason to think it will predict well. It's clearly not sufficient to guarantee good predictions that a model be based on theory, but it is a good start!

Here is an example of how a simple model of the climate can be derived

[15] It seems Lloyd is using "confirmation" in a non-standard way: to mean something like "giving us confidence".

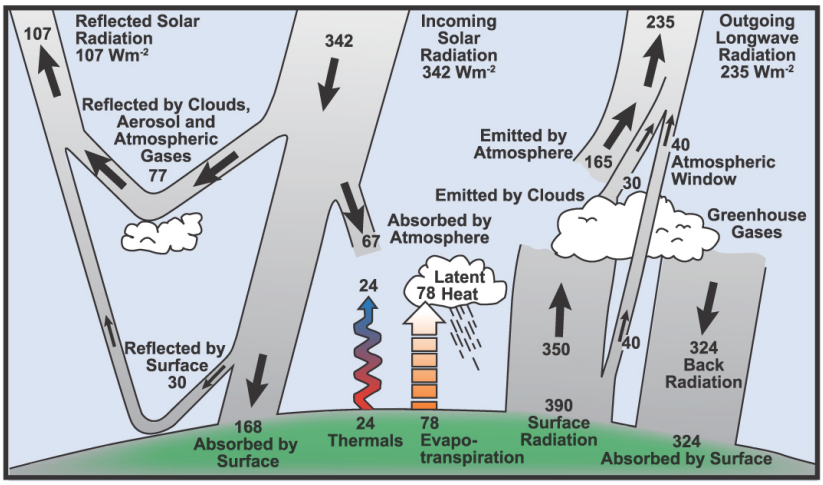

Figure 5: Basic energy balance of the Earth

from some simple physics. Simple one-dimensional energy balance models derive their heat-transport equations direct from principles of thermodynamics. McGuffie and Henderson-Sellers (2005, p.82 ff) derive a simple 0-dimensional "energy balance model" (EBM) from basic physics with the help of some geometry. The basic processes involved in these simple models are summarised in Figure 5.

From thermodynamics we know that on average, the energy input and the energy output of the Earth must balance.[16] The Sun provides effectively the only external source of energy. How much sunlight the Earth's surface reflects (its albedo) will be an important factor in this process, as will how much of the energy radiated from the Earth gets through the atmosphere. The "solar constant" S is a measure of the amount of energy per unit area that the Sun provides. The amount of energy received by the Earth is $\pi R^2 S$ where R is the radius of the earth. That is, the energy received is the energy per unit area provided by the Sun times the area of the Earth "visible" to the sun at a time. This will be a circle of radius R. The total surface of the Earth is $4\pi R^2$, so the time-averaged energy input is $S/4$.

Thermodynamics now gives us an equation for the effective temperature of the Earth:

$$(1-\alpha)S/4 = \sigma T^4 \qquad \text{(Temp)}$$

Where T is the effective temperature and σ is the Stefan-Boltzmann constant.

[16] In reality, given the uptake of heat by the deep ocean and similar processes, there is a genuine about whether this claim is true for any given timescale (D. Stainforth in conversation). This is, however, the basic assumption that underlies this derivation.

The surface temperature and the effective temperature will differ depending on how efficiently the atmosphere absorbs the radiation emitted from the Earth. From (Temp) and an estimate of this efficiency, we can calculate what the temperature of the Earth is, and how it would change if the atmosphere changes. From some simple well confirmed physics we have derived a simple model of the climate system. More complex models require more work, but the same general pattern can be seen in them as well.

4.3 Interval predictions

Returning to our fish population model; if the demon knows the exact initial conditions — current population — he can predict the future population of fish with perfect accuracy. But consider a different level of description. The demon is now trying to predict the future population based, not on an exact initial condition, but on some *set* of initial conditions, some "patch" of the initial phase space. This might correspond to some measurement of initial conditions with finite precision (truncated data). In the logistic map example this would be making predictions given some small interval of initial conditions.

Obviously the demon cannot give pinpoint predictions for this sort of initial condition, since different initial points in the initial condition patch end up at different places once subjected to the dynamics. The demon can, however tell you exactly what the patch of the final phase space is that houses all and only those initial conditions that started in that initial patch.[17]

The problem with this "patch prediction" is that for something like the logistic map, it quickly becomes uninformative. Because of the chaotic nature of the logistic map, the initial error grows rather quickly, so after a few iterations the interval you predict is rather big, and it soon covers the whole interval. This is obviously not a very helpful prediction. But it is interesting as a measure of the predictability. That is, if the interval blows up quickly, this indicates that the system is hard to predict. This relates to the discussion of accountability above. For more on accountability and its relation to interval forecasting and ensemble forecasting see Smith (1996).

Patch prediction can help against imprecision (section 2.1) but not much else. This "exploring the uncertainty" is an important idea in climate science. While not directly useful in improving the precision of our predictions, this is a good way of learning about our models. We learn what interactions lead to big changes by seeing which small differences in parameters have big effects.

This is all taking place at a fairly abstract level. How the comments in this section (and those following it) relate to actual practice is complicated. But my claim is that something roughly like these procedures underlies the (much more complex) practices of working scientists. What scientists actually

[17] Recall this demon has unlimited computational capacity: it can complete the supertask of putting every point in the uncountable set through the dynamics

do will be far more sophisticated (and in some ways more constricted), but the basic idea of exploring the uncertainty through ensembles is the same.

4.4 Ensemble forecasting

Say instead of a patch of initial conditions, we give the demon some probability distribution over the possible initial conditions. For example, a "noise distribution" peaked around an observed value. The demon could evolve this distribution through the dynamics and tell you what the distribution of final conditions looks like.

Now, what if the demon weren't quite vast enough to submit all these data to analysis? There are, after all, uncountably many initial conditions in any patch or distribution, to subject to the dynamics. And each initial condition is of infinite precision. So that's an awful lot of computation to do. So, how would a demon of unreasonably large, but still finite computational resources fare? What he might do is the demon might sample from the noise distribution, say 1024 initial conditions of finite precision. That is, he would take his measured initial conditions, and take some probability distribution centred on the measured values. This distribution accounts for his understanding that the measured initial conditions could be slightly wrong. He then samples from this distribution to give him his ensemble of initial conditions. He could then subject each of these initial conditions to the dynamics, and take the distribution of model outcomes as representing the distribution of possible system outcomes.[18] What the demon does is pick out a set of possible initial conditions — sets of possible values for current temperature, precipitation, wind speed and so on — and put each of these possible descriptions of the current climate through the model dynamics. The demon can then explore what future evolutions of the system are possible in the model.

Figure 6 shows an example of an ensemble forecast of the logistic map at various lead times. As you can see, the error spreads out fairly quickly.

Why is ensemble forecasting better than "best-guess" forecasting? Before ensembles, the standard practice in weather modelling was to take the best available data and the best available model and to run one simulation in as much detail as possible. Now, we sample from a noise distribution and run each ensemble member through some model. If there are 1024 ensemble members, then this takes 1024 times as much computing time as one model run, so we need to use a simpler model than if we just had the one member. So how does this lead to better outcomes? What is the value added by the more complex ensembles approach?

The problem with "best-guess" forecasting is that initial conditions very close to the best guess end up looking very different (recall Figure 2). Loosely

[18] A common practice is to take some ensemble of initial conditions and adopt a uniform prior probability over them. This is more or less the same as the sort of procedure I have been describing.

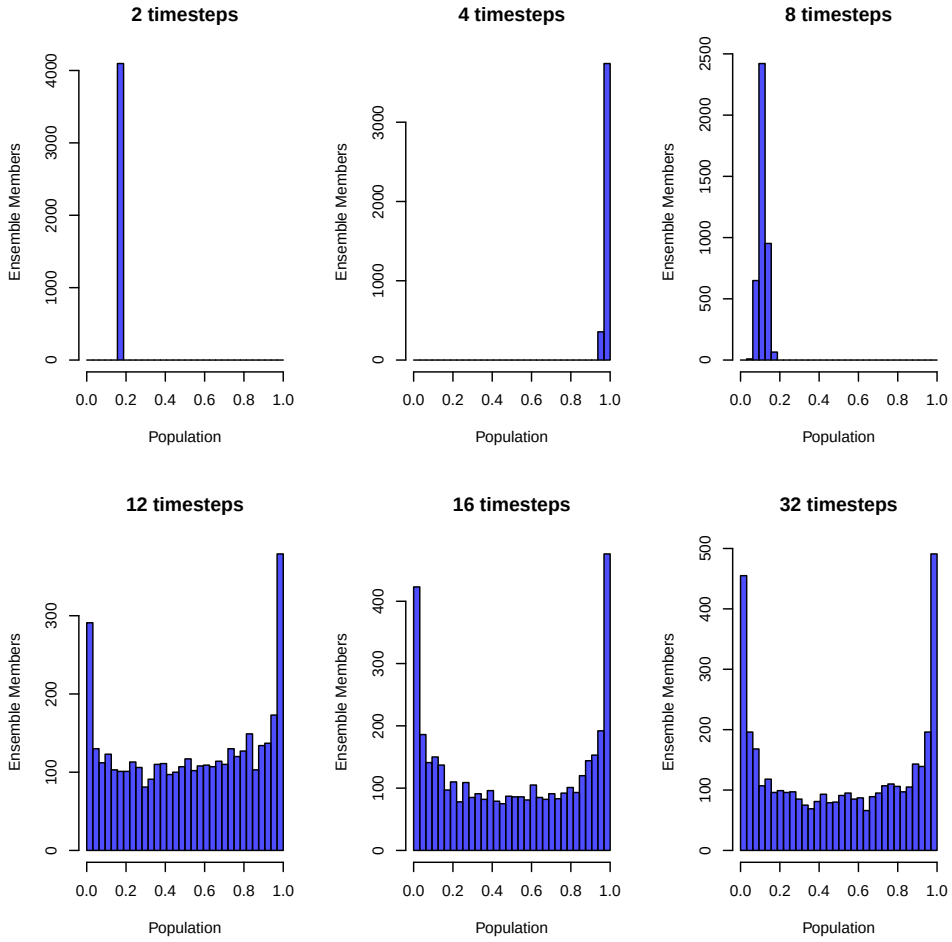

Figure 6: An ensemble of outputs at various lead times

speaking, the weather exhibits some kind of chaotic behaviour: sensitive dependence on initial conditions. So the ensemble method shows you what possibilities there are. Smith (2007) discusses how ensemble forecasting predicted a serious storm that hit England in 1990 which would have been missed by best-guess forecasting (see pp.10–16,139–143).

The "patch prediction" I discussed above also has this feature of highlighting the space of possible evolutions, but ensemble forecasting goes beyond this in "quantifying the uncertainty" by means of the output ensemble. In this way, ensembles can help quantify our knowledge of noise, and so it can deal with inaccuracy (section 2.2), as well as imprecision.

One can make an ensemble of initial conditions, but it is also possible to sample from a distribution on a *parameter value*: this "uncertainty distribution" accounts for possible errors in parameter values in the same way the noise

distribution does for initial conditions. In this way ensemble modelling can help to deal with some worries about model error, as well as initial condition uncertainty. However, we are still working within a parametrised family of functions so similar worries apply as have been already voiced in section 3.3.

Given that ensembles don't deal with all sorts of error, it is a mistake to take ensemble distributions as being probabilities of events in the world. But there is certainly decision-relevant information to be gleaned from these ensembles, and more information that just the possible spread of outcomes; but probably less information than a full probability distribution function over possible outcomes. How much information, and exactly what information there is in ensemble outputs is a question still under discussion (Collins 2007; Parker 2010a,b; Stainforth et al. 2007; Tebaldi and Knutti 2007).

Questions remain about what interpretation can be given to this ensemble. The world has some particular initial condition, and there is some determinate outcome that will happen, so what sort of probability is the ensemble distribution? This is not a new problem: a similar issue has been extensively discussed in philosophical literature about statistical mechanics (Frigg 2008; Uffink 2007).

There are however differences between interpreting statistical mechanical probabilities and climate projection probabilities. It is standard to interpret ensembles of simulation outputs as probabilistic evidence. That is, we interpret the ensemble of outputs as giving us information about the probability distribution function (PDF) of possible future events in the target system. Recall that the ensemble is discrete, while the PDF in the target system is presumably continuous. So do we treat the output ensemble "as if" it is a representative sample from the system PDF and infer things about the system PDF from it? Parker (2010a) discusses these issues.

4.5 Training and evaluation

As we've already seen (in section 3.7), some processes that happen at scales smaller than the grid scale are important. How are they dealt with? Let's take the example of clouds. Clouds affect the albedo of the Earth. Clouds also have an insulating effect like greenhouse gases do. These two effects act against one another: increasing albedo cools the earth below the cloud layer, insulation warms it. Which of these factors wins out depends on a number of factors related to the make-up of the cloud. The intensity of these effects differs for different kinds of cloud.

Clouds can't be put directly into the model, because they are smaller than the size of the grid square. So we have to add correcting factors to the albedo and insulation parameters for the square. The exact values of these parameters will depend on other parameters that affect clouds (humidity, temperature…) and will vary as these vary.

We aren't really simulating clouds, merely simulating their effects. We aren't calculating what particular values of humidity and so on will generate

what shape clouds, but rather directly what effects these clouds would have on albedo and the like.

How do we work out what values these parameters should have? Derivation from theory is more or less impossible, since these aren't parameters that appear in basic physics theories.[19] We do this by "training" the model on some data. So for example we see what values the parameters should have in order to correctly predict climate data from 1970 to 1990. We then test these parameter values by seeing whether the model can successfully predict climate data from 1990 to 2010. If this is successful, that gives us confidence that the parameter values are successfully compensating for the unsimulated cloud effects.[20] This kind of statistical inference is a fairly widespread practice. Hitchcock and Sober (2004) discuss *why* it is a good practice: in short, it is a good balance between having the model fit the data and avoiding overfitting.

This helps us mitigate against worries about curve fitting (section 3.2) and overfitting (section 3.5). And to some extent, training gives us confidence that we've avoided structure error (section 3.3). That is, if the model predicts "unseen" observations, then that seems to suggest it is getting something right about the structure of the world.

There is something strange about this process, however. We take some parametrised family of models and we use a statistical technique to find a model with good fit. The statistical technique is "blind" to the physical interpretation of the model. That is, we're trying to parametrise clouds, but the statistical technique doesn't "care" about whether the parameter value it spits out is at all physically meaningful. One would have thought that a reasoned process that took account of the physical meaning of the parameters (and their interaction) would easily outperform such blind fitting.[21] So how do we get predictive models out of this process? And what does this mean for our attitude to the models? Can we interpret the parameter value as telling us something about the world?

As Hitchcock and Sober point out, the motivation for the training and evaluation methodology is "purely instrumental in character." They continue: "Even if we know that the true curve is a polynomial of degree r, it may well be that the curve of degree r that best fits the noisy data one has at hand will fare worse in future predictions than a curve of lower degree." (Hitchcock and Sober 2004, p.14).

Petersen (2000) criticises this blind statistical procedure, calling it "bad empiricism", in contrast to "good empiricism" which involves having your parametrisations be informed by basic science. Petersen wants us to derive our "cloudiness" parameter from out understanding of the basic physics of

[19] That said, we can get some information about the possible form they might take from theory

[20] The actual practice is significantly more complicated than this, but the basic idea is the same: fit parameters with some data, validate it on other data.

[21] Katie Steele pointed out how odd this process is.

cloud dynamics. This is an admirable aim, but is it a practicable one?

One might argue that Petersen is right that in the limit, we should try and have all parts of our model be physically motivated. But we need projections of the future evolution of the climate *today* so we had better make them with our current best physics. Our current best knowledge, doesn't include a capacity to have physically meaningful parametrisations for all processes we need to model, so if we want to model those processes, we had better be bad empiricists. Bad empiricism is better than no empiricism. Is failing to model a particular process better than modelling it in an ad hoc way? Covey (2000) thinks not and I agree.

If the choice is between not modelling clouds (implicitly assuming they have no overall effect on the climate) and modelling clouds in an ad hoc "bad empiricist" way, then bad empiricism seems the better choice. But is this a false dichotomy? Would it be better to somehow learn to live with only getting the kinds of predictions we can justifiably secure with good empiricism alone? Going down this line of reasoning might leave us with rather few predictions.

This "theoretical limit" would involve resolving clouds and all other currently parametrised processes, and looks rather like if we were to follow this line to its end, we would be falling into the trap that Borges outlined. So it's not even clear that Bad empiricism is bad in the limit. That said, there is certainly something to the intuition that the more we derive from well confirmed physical theories, the better we will do. So it seems that "avoid ad hoc parametrisations" does serve as some sort of regulative ideal.

What about the worries I raised earlier about simplicity being no guide to the truth of a functional relationship? In fact, Akaike (1973) shows that a measure of predictive accuracy explicitly includes a "simplicity measure" in terms of the number of adjustable parameters. Akaike defines a measure of predictive accuracy that can be broken down into two components: one that measures the functions success so far (how close it is to fitting the training data); and another component that is effectively the number of free parameters (a crude measure of simplicity). Hitchcock and Sober (2004, section 5) discuss this surprising result in the context of model-building. A warning about this result: "Akaike's theorem gives us an estimate of the predictive accuracy of a model, on the assumption that that model is fitted to the data via maximum likelihood estimation." Hitchcock and Sober (2004, p.13)

4.6 Robustness

We don't just have the one model: different groups have different models, built using different data sets and techniques, for different purposes, at different resolutions, incorporating different mechanisms. Agreement between these disparate models is an important source of confidence in their predictions (Lloyd 2009; Parker 2006, 2009). Even a single model can exhibit a kind of robustness if the same predictions come up for different settings of the parameters.

My usage of the word robustness is fairly broad and covers concepts that others might want to distinguish: agreement among models, insensitivity to changes in parameter values... I believe there is enough in common between these various concepts that they can be discussed together, and I choose to use the word "robustness" to refer to them.

Let's first consider parameter robustness. This might also be called "insensitivity". The idea is that if some particular prediction is insensitive to the value of a particular parameter, then this prediction is robust in this sense.

There are some caveats to this characterisation of parameter robustness. If a result is robust across the range $[7,10]$ for some parameter x, and we believe the actual value of x to be near 1, then this robustness is not confidence-warranting. This might seem like a trivial point, but it's not one I've seen discussed anywhere. Or again, even if we have a robust prediction for parameter values $(1,5]$ but pathologically weird behaviour below that, then we still don't have that much confidence in the prediction.

The idea is that robustness is confidence-warranting if we have robustness *for a range of values around the physically expected value of the parameter*. And this important caveat also points to the reason behind the confidence: if we have that robustness, then wherever the actual value is in that expected range, the prediction will hold.

Imagine we knew we had the right model, but we were missing a parameter value. Here, robustness of a result around the possible values of the parameter does warrant confidence. Wherever the parameter is in the possible range, the outcome is the same, so since we've got the model structure right, we know the outcome is assured. To put it the other way, if a result is not robust across different values of an unknown parameter, then we have no reason to be confident in our prediction of that effect based on some estimate of the unknown parameter.

I take this to be the kind of thing people have in mind when thinking about robustness of simple models. We have a simple model motivated by basic physics and an unknown parameter. Say we're investigating heat flow in some liquid for which we don't have a good estimate of thermal conductivity. Some effect that is robust across different values of this unknown parameter is likely to be correct. We think that our model is getting something right about the structure of the world. so robustness in the model suggests that, counterfactually, whatever nearby value Nature could have picked for the parameter, the result would be the same. So this gives us confidence that the result will hold in the actual world. This is similar to the sort of "robustness analysis" that Weisberg (2006) discusses. The details of this line of thought will depend on what position one takes on the theory-world relation.

Introducing model error into this discussion adds another level of complexity. Now consider robustness of a result across different climate models. There are elements of the model that we *know* to be unphysical. There are parts of the simulation that we know aren't mapping onto structure of the

world (or are mapping onto system-structure only very indirectly). So how is robustness of the model evidence that the result is assured in reality?

If different models with different idealisations make the same prediction, that suggests that the prediction is not an artefact of this or that idealisation (Muldoon 2007). What we have here is evidence[22] that the prediction is due to the shared structure of the models which we hope is structure they also share with the world. What we have is evidence that there are no troublesome systematic errors in our models.[23]

However, we can't always be sure of achieving robust predictions. What to do with "discordant evidence" — when our predictions disagree — is an open question (Stegenga 2009).

4.7 Past success

For something like weather forecasting, we have plenty of evidence that our models are doing pretty well at predicting short term weather. This gives us confidence that the model is accurately representing some aspect of the weather system. Parker (2010a) discusses this issue. Past success at predicting a particular weather phenomenon, together with the assumption that the causal structure hasn't changed overmuch does allow a sort of inductive argument to the future success of that model in predicting that phenomenon. Past success suggests that problems with implementation, hardware, and missing physics are not causing the models to go wrong.

Climate modellers do not have access to this source of confidence. First, they don't have the same wealth of past successes: they are predicting things that are still in the future, so they don't know if they've predicted accurately yet! Climate models have only so far managed to retrodict already existing data (see section 4.5).

Climate predictions are conditional on particular CO_2 emission scenarios, as well. So we can't even be sure that the causal structure isn't changing in ways that could undermine our ability to predict. Also, if the world is warming, past data and data from the warmer future will not be similar in all the ways they need to be to underpin the material induction. We can't be sure of the "stable background" we need to ground our induction (Norton 2003).

5 Realism and the "True" model of the world

I want to highlight a couple of points about the above discussion that have a bearing on our attitude towards the model. I mean by this the attitude we ought to take vis-à-vis whether the model is adequately representing the target system.

[22] Defeasible evidence, obviously, but evidence nonetheless

[23] There are, however, almost certainly some systematic errors that have not yet become troublesome…

Throughout this paper I have been talking as if there really is some set of equations that "Nature" uses to evolve the real world system. Our aim is to get our model as close to these equations as possible. Is this true? Does it even make sense? Whatever the answer to those questions, it is certainly the case that it is unreasonable to assume that the "True" equations of the world are in the class of models we are currently exploring. We *know* there are myriad ways our model class *must* fail to contain the True equations if they exist.

So how better ought we talk about the practice of modelling? It seems it can only be a verbal shortcut to talk of our getting at the True model: a metaphorical flourish. Indeed, the model and the system are just *very different kinds of things*. I've pointed to some of these differences above, but let's make them clear again. First, the target system — the physical world — is a physical thing. The model is…what? A computer program? A mathematical construct? Some combination of the above? So it's not even clear what it means to say that the model "gets things right" about the world. What does it mean for a model to be "like" its target system?

One might want to retreat to a kind of instrumentalism here and say that the model is "like" the target system in that the things in the model that represent the observables — temperature, wind speed and the like — are "like" those things in the world. So temperature data in the model is like temperature in the world. But there is still a concern that these are quite different kinds of things. That this particular part of computer memory registers a value interpreted as "13.45 °C" doesn't straightforwardly translate into a fact about the world. Is comparing data measured in the world with data generated by a model comparing apples to oranges?

Perhaps we would like to say that all we are really trying to do when we are modelling is to summarise the actual data as best we can. But then why run simulations into the future?

First, there is a certain sense in which we really *do* take the models as telling us something about the world. These models are supposed to be predictive: they are supposed to tell us how the actual climate will actually vary.[24] So there are certain parts of the model that it seems we really have to be realist about. And this isn't an idle philosophers' realism: decisions costing billions of pounds are made on the basis of these predictions' correctness. How's that for ontological commitment?

And, as Parker (2006) discusses, there is a commitment to model the physical processes in "the right way" as much as possible. Ad hoc fudges that just "save the phenomena" are a last resort: physically meaningful, realistic solutions are much preferred. So it seems there is a sense in which we do take ourselves to be trying to "get at" the equations Nature uses.

However, there are still parts of the model that we know are unphysical:

[24]These projections are conditional on a particular emission scenario, but they are still supposed to be saying something conditional about the actual world.

all the sub-gridscale parametrisations are simply ad-hoc ways of keeping predictive accuracy. There's no attempt to model the phenomena, we are merely trying to accommodate the measurements. So taking this "trying to get at the True Model" line too seriously is misleading. We have a commitment to accurate predictions too and sometimes this leads us to approaching things in unphysical ways. This line might bolster my earlier response to Petersen (2000) on "bad empiricism".

6 Summary

Table 1 summarises how each of the coping strategies discussed fares in relation to each of the sources of error. Imprecision and inaccuracy can be mitigated against by making better measurements and running ensemble forecasts rather than "best-guess" forecasts. Deeper worries about measurement might be offset by agreement among models which suggest that the quantities we are measuring are the right ones. Basic physics might also give us confidence we are on the right track as far as getting the relevant quantities right goes.

Some physical parameters can be measured independently of the climate modelling so better measurements can sometimes help with worries about parameter error. For example, the thermal conductivity of air, or the viscosity of water can be determined independently of the role they play in climate models. Robustness of a prediction across an ensemble of parameter values mitigates against worries about parameter values, even in the absence of such direct techniques. We can also train models on past data to determine unknown parameters.

Error in the structure is somewhat harder to deal with. Finding parameter values through training on past data involves starting with some parametrised family of functions, so we can't train our models to account for error in picking the model class. If basic physics suggests that the structure ought to be roughly *this way*, then that warrants some confidence that we aren't subject to structure error there. If a prediction is robust across different models with different structures, this gives us confidence that the differences in structure do not contribute to making this prediction go wrong. Likewise, past success suggests we have got the model structure roughly right. For similar reasons, past success and robustness give us some confidence that the physics missed out of our models is not leading our models to be seriously wrong. Missing physics can also be accommodated by "training" unphysical parameters that correct for the processes we're missing.

That a model trained on some past data can still retrodict unseen data partially mitigates worries about overfitting. Past success can be seen as warranting extra confidence of the same kind. If the model isn't too sensitive to changes in its parameters, this also suggests that the model is not "over fitted".

	Better measurements	Theory	Intervals	Ensembles	Training	Robustness	Past Success
Imprecision	M		PM	M			
Inaccuracy	PM		PM	M			
Deeper		PM?				WC?	
Parameter	PH	PH		M	M	PH	PH
Missing physics					PH	WC	WC
Structure		WC				WC	WC
Overfitting					PM	WC	PH
Discretisation						PH	WC
Resolution						PH	PH
Implementation						WC	WC

Table 1: Summary of errors and coping strategies

M: Mitigated — PM: Partially Mitigated — WC: Warrants Confidence — PH: Possibly Helpful

Worries about discretisation and resolution effects can be somewhat allayed by results not being sensitive to changes in resolution, and by past success at modelling the system at that level of resolution. Deep worries about implementation can be dealt with in the same way.

I have outlined a variety of sources of error in scientific modelling. I have also explained some ways we have of maintaining confidence in our predictions despite these errors. I tried to diagnose how these methods allowed us to remain confident in the face of error.

Thanks

Thanks to Dean Peters, Katie Steele, Susanne Burri and Charlotte Werndl for discussions that shaped my thinking on these issues. Thanks especially to Dave Stainforth, Lenny Smith and Roman Frigg for detailed comments on drafts of this paper. Any flaws remaining in the reasoning are mine alone.

References

Most of the figures are from Solomon et al. (2007). Figure 1 is FAQ1.2 Figure 1; Figure 3 is Figure 1.2; Figure 4 is part of Figure 1.4; Figure 5 is FAQ 1.1 Figure 1. Figures downloaded from http://www.ipcc.ch/. Figure 2 and Figure 6 drawn in R by the author, details available on request.

Akaike, Hirotsugu (1973). "Information Theory as an Extension of the Maximum Likelihood Principle." In: *Second International Symposium on Information Theory*. Ed. by B. Petrov and F. Csaki. Akademiai Kiado, pp. 267–281.

Borges, Jorge Luis (1999). "On Exactitude in Science." In: *Borges Collected Fictions*. Translated by Andrew Hurley. Penguin, p. 325.

Chang, Hasok (2004). *Inventing Temperature*. Oxford University Press.

Collins, M. (2007). "Ensembles and probabilities: a new era in the prediction of climate change." *Philsophical Transactions of the Royal Society* 365, pp. 1957–1970.

Covey, C. (2000). "Beware the Elegance of the Number Zero." *Climactic Change* 44, pp. 409–411.

Earman, John (1986). *A primer on determinism*. D. Reidel Publishing Company.

Fefferman, Charles L. (n.d.). *The Navier-Stokes equation*. http://www.claymath.org/millennium/Navier-Stokes_Equations/navierstokes.pdf.

Frigg, Roman (2008). "A field guide to recent work on the foundations of statistical mechanics." In: *The Ashgate companion to contemporary philosophy of physics*. Ed. by D. Rickles. Ashgate, pp. 99–196.

Frigg, Roman and Julian Reiss (2009). "Philosophy of Simulation: Hot New Issues or Same Old Stew?" *Synthese* 169, pp. 593–613.

Hitchcock, Christopher and Elliott Sober (2004). "Prediction Versus Accommodation and the Risk of Overfitting." *British Journal for the Philosophy of Science*, pp. 1–34.

Knight, Frank (1921). *Risk, Uncertainty and Profit*. Houghton and Mifflin.

Laplace, Pierre Simon de (1951 [1816]). *A Philosophical Essay on Probabilities*. Translated from the French by Frederick Wilson Truscott and Frederick Lincoln Emery. Dover.

Lloyd, Elisabeth A. (2009). "Varieties of Support and Confirmation of Climate Models." *Proceedings of the Aristotelian Society Supplementary Volume* LXXXIII, pp. 213–232.

May, Robert (1976). "Simple mathematical models with very complex dynamics." *Nature* 261, pp. 459–467.

McGuffie, Kendal and Ann Henderson-Sellers (2005). *A Climate Modelling Primer*. Third Edition. Wiley.

Muldoon, Ryan (2007). "Robust Simulation." *Philosophy of Science* 74, pp. 873–883.

Norton, John (2003). "A Material Theory of Induction." *Philosophy of Science* 70, pp. 647–670.

Parker, Wendy (2006). "Understanding Pluralism in Climate Modeling." *Foundations of Science* 11, pp. 349–368.

— (2009). "Confirmation and Adequacy-for-Purpose in Climate Modeling." *Proceedings of the Aristotelian Society Supplementary Volume* LXXXIII, pp. 233–249.

— (2010a). "Predicting weather and climate: Uncertainty, ensembles and probability." *Studies in History and Philosophy of Modern Physics* 41, pp. 263–272.

— (2010b). "Whose Probabilities? Predicting Climate Change with Ensembles of Models." *Philosophy of Science* 77, pp. 985–997.

Petersen, Arthur (2000). "Philosophy of Climate Science." *Bulletin of the American Meteorological Society*, pp. 265–271.
Popper, Karl (1982). *The Open Universe*. Routledge.
Regan, Helen, Mark Colyvan, and Mark. A. Burgman (2002). "A Taxonomy and Treatment of Uncertainty for Ecology and Conservation Biology." *Ecological Applications* 12, pp. 618–628.
Schiermeier, Quirin (2010). "The Real Holes in Climate Science." *Nature* 463, pp. 284–287.
Sen, Armartya (1993). "Positional Objectivity." *Philosophy and Public Affairs* 22, pp. 126–145.
Smith, Leonard (1996). "Accountability in ensemble prediction." In: *Proceedings of the ECMWF workshop on predictability*, pp. 351–368.
— (2000). "Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems." In: *Nonlinear Dynamics and Statistics*. Ed. by A. Mees. Birkhauser.
— (2007). *Chaos: A very short introduction*. Oxford University Press.
Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Trignor, and H.L. Miller, eds. (2007). *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press.
Stainforth, David A., Miles R. Allen, E.R. Tredger, and Leonard A. Smith (2007). "Confidence uncertainty and decision-support relevance in climate models." *Philosophical Transactions of the Royal Society* 365, pp. 2145–2161.
Stegenga, Jacob (2009). "Robustness, Discordance and Relevance." *Philosophy of Science* 76, pp. 650–661.
Stirling, Andy (2010). "Keep it complex." *Nature* 468, pp. 1029–1031.
Talagrand, Olivier (1997). "Assimilation of Observations, an Introduction." *Journal of the Meteorological Society of Japan* 75, pp. 191–209.
Tebaldi, Claudia and Reto Knutti (2007). "The use of the multi-model ensemble in probabilistic climate projections." *Philosophical Transactions of the Royal Society* 365, pp. 2053–2075.
Uffink, Jos (2007). "Compendium of the foundations of statistical mechanics." In: *Philosophy of Physics*. Ed. by Jeremy Butterfield and John Earman. North-Holland, pp. 923–1047.
Walker, W., P. Harremoës, J. Rotmans, P. J. van der Sluijs, M. van Asselt, P. Janssen, and M. P. K. von Krauss (2003). "Defining uncertainty: A conceptual basis for uncertainty management in model-based decision support." *Integrated Assessment* 4, pp. 5–17.
Weisberg, Michael (2006). "Robustness Analysis." *Philosophy of Science* 73, pp. 730–742.