PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2007) **365**, 1971–1992
doi:10.1098/rsta.2007.2069
Published online 19 June 2007

Review

Probabilistic climate forecasts and inductive problems

By D. J. Frame[1,*], N. E. Faull[2], M. M. Joshi[3] and M. R. Allen[2]

[1]*Oxford University Centre for the Environment, Dyson Perrins Building, South Parks Road, Oxford OX1 3QY, UK*
[2]*Atmospheric, Oceanic and Planetary Physics, University of Oxford, Parks Road, Oxford OX1 3PU, UK*
[3]*Walker Institute for Climate System Research, Department of Meteorology, University of Reading, Earley Gate, Reading RG6 6BB, UK*

The development of ensemble-based 'probabilistic' climate forecasts is often seen as a promising avenue for climate scientists. Ensemble-based methods allow scientists to produce more informative, nuanced forecasts of climate variables by reflecting uncertainty from various sources, such as similarity to observation and model uncertainty. However, these developments present challenges as well as opportunities, particularly surrounding issues of experimental design and interpretation of forecast results. This paper discusses different approaches and attempts to set out what climate*prediction*.net and other large ensemble, complex model experiments might contribute to this research programme.

Keywords: probabilistic climate forecasts; ensemble-based methods; climate science

1. Introduction

Climate change forecasts are inherently problematic owing to the difficulty in modelling a complex, nonlinear, multiscale system, in which we do not understand how all the relevant interactions evolve over time. Furthermore, this evolution takes place on time scales which are inconveniently long for verification[1] purposes. These factors hamper preferred methods for drawing inferences in the physical sciences, where the normal practice is to attempt to isolate a system for study, control its boundary conditions, attempt to understand all the system interactions, develop a model and then construct and perform a large number of repeatable experiments aimed at both furthering

* Author for correspondence (dframe@atm.ox.ac.uk).

[1] Oreskes *et al.* (1994) point out that the processes scientists usually regard as 'verification' are actually forms of 'confirmation'. Stainforth *et al.* (2007*a*) adopt this terminology; we stick to the phrasing most familiar to climate scientists.

One contribution of 13 to a Theme Issue 'Ensembles and probabilities: a new era in the prediction of climate change'.

understanding of the system and model improvement. Climate modelling is subject to various philosophical problems, including issues surrounding the boundaries of the system under investigation (the processes resolved and how unresolved scales or processes impact upon them), the adequacy of modelled processes, the nature of the relationships between theory and data, issues of chaos and the way these map forward (or back) to represent different future (or past) climates, to name a few. We will not discuss all of these. Interested readers might see Oreskes *et al.* (1994), which discusses the problems inherent in model verification, validation and confirmation in the Earth sciences, pointing out that verification is elusive in open systems, that validation is impossible where the modelled processes are different from those in the real world, and that confirmation, in the Earth sciences, is partial at best.

The present paper is far from comprehensive, focusing on a small subset of these problems: the problem of induction and the problem of the underdetermination of theory by evidence. This paper attempts to explain these and discusses how different methodologies currently used in ensemble climate forecasting all face these problems, making different choices in the way they deal with them.

One further issue the paper touches on is the debate between realism and its adversaries. Science is often conceived as being of special value because the knowledge obtained through scientific inquiry does not depend on the person doing the enquiring. Papineau (1996) characterizes realism (scientific and other) as follows:

> Suppose we take realism… to involve the conjunction of two theses: (1) an independence thesis: our judgements answer for their truth to a world which exists independently of our awareness of it; (2) a knowledge thesis: by and large, we can know which of these judgements are true.

Both theses have been criticized. One can deny that we have access to the independent world, or one can deny that we can know which judgements are true. One can also deny that truth ought to be the appropriate metric for considering theories. All three objections have been much discussed. One quite common view is that scientific knowledge is inherently subjective, and depends on our apprehension of the objects under study. We might call this 'subjectivism'.[2] To some extent this debate runs alongside the interpretive or epistemological division that underpins the ongoing debate between Bayesians and relative frequentists (Gärdenfors & Sahlin 1988). The argument between those who see science as objective and those who see it as subjective continues in the philosophical literature, where it tends to focus on our most elementary physical laws, so as to restrict the focus to those cases which would seem to be the strongest candidates for objective knowledge. Fields such as environmental science, comprising interactions between various multiscale, complex subsystems, are less investigated. This is presumably because, to many philosophers, it would be like shooting fish in a barrel: the case for objectivist conceptions of Earth system science is probably too weak to maintain for very long in the face of experienced philosophical adversaries. However, we argue that in the context of climate research, the objective/subjective debate is probably an unhelpful way to conceive of the issue. Much more relevant is the debate between realism and those of its adversaries that fall (more or less) under the banners of

[2] There is a vast body of philosophical literature on this issue. Interested readers should consult any introductory text to the philosophy of science (e.g. Sklar 1992; Papineau 1996; Chalmers 1999).

'instrumentalism' or 'pragmatism'.[3] We will refer to pragmatism as the view that what matters is not the truth content of a theory, but its utility. It may or may not be the case that science can be true or coherent. What matters more, we argue below, is that science can be useful. On this view, what distinguishes science from other forms of inquiry is not its truth or verisimilitude, but its usefulness and reliability.

This is not intended as anything like the last word on the subject. In fact, the aims of the paper are introductory: we would be happy if this paper raises the awareness of the direct relevance of philosophical issues to ensemble climate forecasting, and, secondarily, we hope that some philosophers might decide that the field of ensemble climate research contains some interesting questions.

2. Induction

Induction is the process through which we generalize about the world from particular instances within it (e.g. Russell 1912). Putative causal relationships are instances of induction in which we infer general laws from the observations we have made up until now. This process is problematic. The basic problem of induction was described by Hume (1739):

> We have no other notion of cause and effect, but that of certain objects, which have been always conjoin'd together, and which in all past instances have been found inseparable. We cannot penetrate into the reason of the conjunction. We only observe the thing itself, and always find that from the constant conjunction the objects acquire an union in the imagination.

In other words, we have to use models, maps or other conceptual schemes to move from observations of regularities to our portrayals of causal connection. In order to develop those conceptual schemes, we need to: (i) imagine potential schemes and (ii) make decisions regarding their applicability or relevance. Developing scientific models is, in this sense, an imaginative exercise: we try to infer causal relationships, and then try to isolate, quantify and test these. Sometimes this is more difficult than other times. It tends to be much simpler when the systems under study are simple or well understood or time independent. It tends to be harder when isolation is elusive and where testing is difficult, as is the case in climate research.

Climate change detection and attribution studies allow us to claim, with more than 95% confidence, that the observed twentieth century warming cannot be explained by reference to natural forcings alone, but can be explained by a combination of natural and anthropogenic forcings (Allen *et al.* 2000). Moreover, we can say that the climate evolved in just the way we would have expected under reasonably simple energy balance considerations (even though we were wrong about certain details (the eruption of Pinatubo; the precise rise in GHG emissions) of the forcings over the period in question). So such evidence, as there is, seems to suggest that we in the mainstream climate community have been right so far.

However, this does not necessarily imply that we will continue to get it right: this prediction could have been accurate even though it neglects climate–carbon cycle feedbacks that will skew predictions on longer time scales; it could have been accurate in the near term but poor in the longer term because it misses some

[3] For example, Papineau (1996) or Worrall (2002).

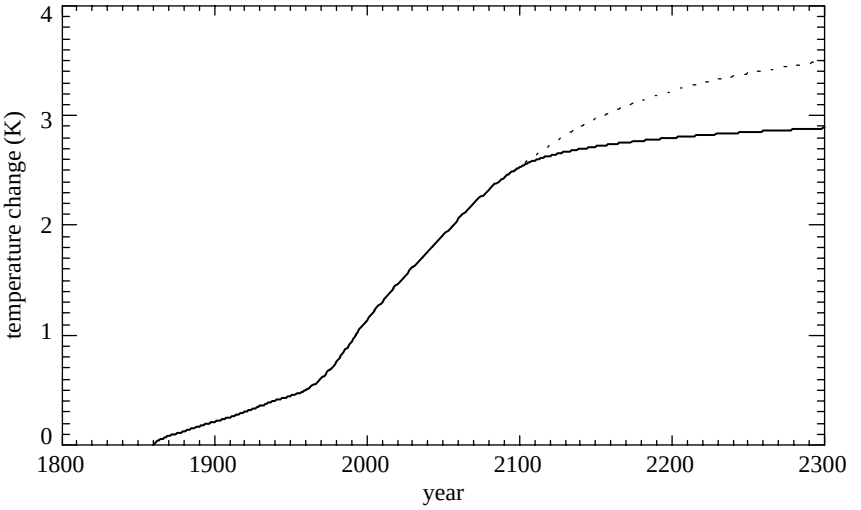

Figure 1. Global mean temperature time series from two models, described in the text, which differ only in the future as a temperature-dependent feedback mechanism is triggered at $\Delta T = 2.50$ K.

as-yet-unquantified temperature-dependent cloud climate or water vapour feedback. Basically, there are any number of ways in which we might have managed to make a good prediction regarding medium-term transient global climate response, while being wrong on longer time scales. The basic Humean point is that we can never exhaust the ways in which we could be wrong. The world can always surprise us. Perhaps the longer we give it, the more likely it is to do so.

The dictum 'make everything as simple as possible, but no simpler' seems to be an appropriate description of how model builders think. But model parsimony is a weak principle: if a process needs to be included for a model to explain a set of observations, then it needs to be included. The problem in environmental modelling is that no one knows exactly when to stop: which processes do not matter? How do we include processes we think matter when we are unsure of their evolution? This is the problem of knowing how to choose between different models of the system which have equal plausibility before the evidence. Imagine two models, both forced with the B1 SRES scenario,

$$c_{\mathrm{eff}} \frac{\mathrm{d}\Delta T}{\mathrm{d}t} = F - \lambda_1 \Delta T, \quad \lambda_1 = \text{const.} = \frac{1}{S}; \quad S = 3.0\,\mathrm{K},$$

$$c_{\mathrm{eff}} \frac{\mathrm{d}\Delta T}{\mathrm{d}t} = F - \lambda_2 \Delta T, \quad \text{if } \Delta T \geq 2.5,\ \lambda_2 = \lambda_1;$$

$$\text{if } \Delta T > 2.5\ \mathrm{K}, \lambda_2 = [3.0 + \ln(1 + \Delta T - 2.5)]^{-1}.$$

In the first model, climate sensitivity S is a constant (3.0 K); in the second, S is the same as in the first model until some threshold temperature, $\Delta T = 2.5$ K, at which point a feedback effect enhances S such that $S = 3.0 + \ln(1 + \Delta T - 2.5)$, increasing the sensitivity and the temperature response of the system, as shown by the dotted line in figure 1. Until the global temperature anomaly rises to 2.5 K, both of these models have equal plausibility before the current evidence: even palaeoclimatic evidence may have little bearing on this problem if, say, the

mechanism that causes the change in sensitivity only occurs with the current configuration of continents and ice sheets. The time-dependent aspects of this problem seem to indicate that it is something like an instance of the 'Grue problem' or the new problem of induction (Goodman 1954).[4] Knowing how to write down the terms that 'matter' for the climate problem is problematic, for at least two reasons: predictions require one to address the problem of induction, usually by assuming either (i) that the past will resemble the future or (ii) that we know the ways in which it will not. Of course, this problem is not new, nor is it peculiar to climate research. The problem of induction has a long history and is much discussed in philosophical literature. The interesting thing in the case of climate research is that some extremely closely related disciplines (numerical weather prediction) are amenable to the usual evaluative loop and other parts of the problem are associated with very long characteristic time scales, over which we lack adequate data constraints. The following discussion focuses on how this sort of problem presents challenges for different methodologies in probabilistic climate research. The primary purpose of the article is to alert climate scientists to the troubles inherent in ensemble climate prediction. It is not intended as a thorough-going investigation into the ways in which the philosophical problem of induction might bedevil climate research. Consequently, we do not reference or discuss the large body of literature on induction specifically or the philosophy of science generally. We simply choose examples from the literature, which seem to highlight the relevant issues for current debates in probabilistic climate research. While philosophers may find this unsatisfactory, our main aim is to provide a practical paper for climate scientists.

(*a*) *Models and their domains*

So how should we conceptualize this problem? Cartwright (1999) suggests that scientific models have domains, and that we ought to be wary of extending the models beyond the domains over which they have been verified. Cartwright (1999) goes a lot further than most physical scientists would be prepared to go in her analysis of the appropriateness of the domains of physical laws, but one need not accept her entire programme to accept that she has a point, and one that is extremely relevant for climate research: we parametrize everything below the scale of a few hundred kilometres (in the horizontal); and though we do resolve many processes at large scales, we make educated guesses about some, too. In meteorology, we frequently make implicit use of the idea that our models work over some domains and not others, especially in the case of dynamical equations: we assume that friction matters near the boundary layer, but not in the free troposphere. The multiscale nature of the system under study, as well as computational limits, leads to judgements which amount to a kind of scientific ellipsis: we drop terms out of our models because we think they are not worth resolving for this particular problem. In the case of our most comprehensive models, we can resolve variables on quite localized scales—Ireland, say—though the information contained in the variable in question is not totally correct. If we are systematically wrong about the whole North Atlantic storm track, then we are probably wrong about plenty of our model predictions about Ireland, though

[4] Note that this is not the problem of different theories offering the same predictions (see Papineau (1996) and essays therein). In this case, the predictions differ; but we cannot choose between them.

perhaps not as predictably wrong about Ireland as we are about the storm track: additional ways of being wrong about Ireland probably hamper even our ability to apply a simple correction for storm track bias.

Our physical laws, for Cartwright, are ceteris paribus laws, at best: their legitimacy holds only over their domains. But it is precisely this ceteris paribus issue that we cannot test in climate research, because: (i) we cannot do repeatable experiments and (ii) the experiments we can conduct have inconveniently long verification time scales. We can read Cartwright's main warning as that we should be wary of extending our claims about the behaviour of the real world beyond that part of the world we have explored in experiments. In other words, we should be careful not to extend our modelling framework beyond its domain. There are times when our models, having been thoroughly tested across the appropriate domains, are adequate to our inferential needs: we can say that the future will probably resemble the past because we have checked out both the relevant model and the relevant domain. When we board an Airbus, we have considerable confidence that the 'model' of 'physical reality' that has been used to construct its jet engines will be adequate for at least the purposes of our flight.[5] When we examine a forecast of today's weather, we take our coat with us because we know that the models have been tested and verified across the relevant temporal and spatial domains. With climate change forecasts, we cannot test our models right across the relevant domain, because we are not able to run the necessary real-world verification experiments.

Exactly what sorts of limit this places on our ability to predict future climates is also unclear, because we cannot close the usual prediction–verification model improvement loop. While 'fast physics' experiments such as that of Rodwell & Palmer (2006) can help us rule out some ways in which the predictions may be wrong, we lack the ability to test the 'slow physics' because none of our models have been verified with the appropriate forcings over the appropriate time periods. Those questions remain open.

(*b*) *Bayesian induction*

There have been attempts to develop purely Bayesian versions of induction (e.g. Howson & Urbach 1993; Zabell 2005). In these techniques, the problem at hand is usually recast as an urn problem (i.e. recast as an analogous problem in which balls are picked from an urn). Zabell (2005), who gives various examples of urn-based problems, argues that this approach is adequate for a wide range of problems, and that the technique can even deal with 'surprises' in which one encounters some feature one has not seen before. However, as Smith (2007) explains, such techniques rely on the stability of the system being investigated. For the technique to work, one can only be 'surprised' in a small number of ways. One can be surprised by the colour of the balls coming out of the urn, but one cannot have the rules of the urn game change while one is playing it *in ways one does not understand*. It is not obvious exactly how we ought to go about reflecting complex and evolving systems as urn problems, when there are hard-to-predict feedbacks between the balls coming out of the urn and the rules that govern the distribution

[5] We may still be wrong, for Humean reasons. Passengers on various de Havilland Comets in the early 1950s were wrong in their risk assessments because the models employed by de Havilland designers did not include sufficient representation of metal fatigue.

of remaining items inside it. In the climate problem, we are uncertain in a classical sort of way about how the variables interact *at present*, but, more importantly, for the inductive inferences we intend to make regarding the future, we are unsure of how those interactions evolve over time. Thus, the 'modern climate urn' we use for the twentieth century climate may be inappropriate for use in a $2\times CO_2$ world owing to changes to the thermohaline circulation (Manabe & Stouffer 1999) or terrestrial carbon cycle (Cox *et al.* 2004), for instance. Relatedly, in what could be seen as a thinly veiled warning for climate scientists, Smith (2007, pp. 126–130) illustrates the dangers of inferring things about one sort of thing (the real world) from experiments with an imperfect analogue (models).

The problem of induction as it complicates predictions of future climate change is a specific instance of a very general problem regarding assumptions regarding the uniformity of nature. Bertrand Russell (1912) gave the following vivid example:

> And this kind of association is not confined to men; in animals also it is very strong. [...] Domestic animals expect food when they see the person who feeds them. We know that all these rather crude expectations of uniformity are liable to be misleading. The man who has fed the chicken every day throughout its life at last wrings its neck instead, showing that more refined views as to the uniformity of nature would have been useful to the chicken.

Russell's chicken was prone to a grim model error: it missed an important time-dependent term from its model of the world, beholden, as it was, to the 'rosy scenario' (Smith 2002) of the uniformity of nature. However, the omission of this term was not obvious before the fact: even a perfectly rational and/or perfectly Bayesian chicken would not have had any cause to include this term in its best models. To do so would have been alarmist. All the evidence the chicken had suggested that the farmer would keep feeding it, and the evidence for a murderous farmer was no greater than for a comet strike or H5N1 outbreak. The chicken could have stressed itself to death contemplating the possible array of decidedly un-rosy scenarios that it faced, but it had no obvious, unproblematic, uncontestable way of incorporating these into its model of the world.

3. Bayesian approaches in climate research

Most scientists working on Bayesian approaches to climate do not adopt the kind of urn-based approach to induction, preferring to estimate priors across parameters in climate models, and then see what these beliefs, coupled with the evidential data, imply for a posterior distribution (Murphy *et al.* 2004). A thorough-going Bayesian approach to parameter estimation and probabilistic forecasting may yield the best parameters and the best guess distribution *for models with that structure*, but such an approach will, necessarily, suffer the same sort of vulnerability to structural error that we have described above. The only things that can push us towards addressing the model structural error are physically verifiable improvements in the models themselves. For long time-scale processes, these verifiable improvements may take a while.

In the rest of this section, we present a standard Bayesian formulation for relating *data* to a forecast variable F. Methods such as this are becoming

commonplace in the literature (Forest *et al.* 2002; Knutti *et al.* 2002; Murphy *et al.* 2004; Hegerl *et al.* 2006), and represent an avenue for the quantification of uncertainty in climate modelling forecasts, but the Bayesian paradigm is not without its opponents (Stainforth *et al.* 2007*a*). There is an ongoing, often acrimonious, dispute in probability theory between relative frequentists and Bayesians, and the interpretation of any existing probabilistic climate forecasts as providing 'probabilities' is contingent upon accepting certain propositions within that dispute. In particular, where relative frequentists interpret probability as events considered as members of collectives to which the tools of statistical analysis can be applied, Bayesians regard probability as the degree to which a proposition ought to be believed (de Finetti 1964; Gärdenfors & Sahlin 1988). On this view, Bayes' rule is a technique for updating beliefs over the proposition under investigation in the light of new information (Hacking 1967). Although Bayesians and relative frequentists converge in situations where repeated experiments are possible, that is not the case in climate change science, especially for experiments involving either long time scales (where adequate data are a problem) or where the system is stressed in ways we have not observed before. In these situations, the sorts of verification/falsification procedures scientists might like to use are problematic, since we are essentially 'out of sample': if the twenty-first century 'climate urn' is different from the twentieth century climate urn in ways our models do not capture, our predictions will be mistaken in a new way (Smith 2007).

Here, we sketch out a Bayesian approach using an ensemble of models produced by varying parameters in a base model. The model basis for the ensemble could be a single energy conservation equation, an intermediate complexity model, or a full general circulation climate model. In the examples that follow, we will restrict ourselves to a very simple energy balance model for simplicity, but the principles—and the problems—also obtain in more generalized settings. The conditional probability distribution function (PDF) for a forecast variable F, given a set of observations, data, can be expressed in terms of Bayes' theorem

$$P(F|\text{data}) = \frac{P(\text{data}|F)P(F)}{P(\text{data})},$$

where $P(\text{data}|F)$ is proportional to the 'likelihood' that these observations would be simulated by a model which predicts the forecast variable in question to lie within a small distance, $\mathrm{d}F$, of F. In studies where a subset of otherwise equally plausible models all make the same prediction for F, $P(\text{data}|F)$ could simply be the average likelihood of the data taken across this subset (Forest *et al.* 2002; Frame *et al.* 2005) or, perhaps preferably, the maximum likelihood of any model within that subset. Given a 'prior' sampling strategy for models or model parameters, $P(F)$ is proportional to the implied probability that the forecast is within $\mathrm{d}F$ of F before these data are considered. This is the prior distribution of the forecast one needs to assume as one embarks on one's study. $P(\text{data})$ is a constant required to ensure all probabilities sum to 100%.

(*a*) *Forming a prior*

On standard views of subjective Bayesian inference (Gärdenfors & Sahlin 1988; Howson & Urbach 1993) the prior reflects the belief one has in a proposition before (and independently of) the evidence one is using as data is

applied. There is a large literature on how one should do this, but the dominant reading is that this is essentially a choice the experimenter makes. In the growing literature on estimating climate sensitivity (the warming resulting from a doubling of carbon dioxide after the system is allowed to come back into equilibrium), the main choices have been either more-or-less explicit expert priors or uniform priors.

The use of expert prior information in this kind of question suffers from two problems: (i) that one's uncertainty in one's own estimate of one's degree of belief is not captured and (ii) that a joint distribution attempting to capture one's beliefs over the component propositions within a model fails to capture the structural uncertainty which arises from the discrepancy between the model structure and the real system (Smith 2002). There have been attempts to formally capture (i) within an imprecise probability framework (Kriegler & Held 2005). This represents an interesting avenue for research and may help address some of the problems discussed in this essay (though it is unlikely to address them all).

There is also the danger of question begging in the use of expert priors. In the example given earlier, in §2, we had two models of climate change, which give *identical* results until now, but diverge in the future. Following the Bayesian approach of attaching subjective degrees of beliefs to each of these models, *a priori*, simply masks this problem. To give such models different prior weights based on subjective estimates of probability seems to be an exercise in question begging. As Eliot Sober (2002) writes:

> When scientists read research papers, they want information about the phenomenon under study, not autobiographical remarks about the authors of the study. A report of the author's subjective posterior probabilities blends these two inputs together. This is why it would be better to expunge the subjective element and let the objective likelihoods speak for themselves.

The problem arises in the first place because we have no way of choosing between the large number of climate models that are equally plausible (and equally implausible) before the evidence to date.

Various authors (Knutti *et al.* 2002; Frame *et al.* 2005) have argued that either likelihood functions or uniform priors in the relevant forecast variable are a reasonable way of addressing this issue.[6] In this case, one forms a reasonably simple but quite ignorant prior, and then tries to work out what the data can rule out. It has long been argued (e.g. Howson & Urbach 1993; Sklar 1992; Sober 2002) that there is no obvious way to reflect 'no prior knowledge' in the prior. This is undoubtedly correct. Specifying an appropriately ignorant prior is clearly problematic (and has been a significant factor in many rejecting the search for an objective Bayesian paradigm). However, specifying expert priors in the estimation of climate sensitivity is just as problematic: the danger here is that one will double count information using some of the data (over the last 50

[6] A 'uniform prior' essentially amounts to an evenly sampled likelihood weighting in which each model is given equal weight before the data are applied. Its simplicity has proved of enduring appeal, but Rougier (2006), for instance, gives some reasons why it might not be an attractive choice as a belief function.

years, say) in one's prior distribution for sensitivity: how could one ever demonstrate that one has formed one's prior ignorant of the data being used as a constraint?

The problem of sensitive dependence of results on the prior does not matter much where many iterations of the Bayesian loop are possible by 'drawing new balls from the same old urn': eventually, when all the available information has been used, posterior distributions will be so tight the influence of the prior will be minimal. It remains, however, critically important in cases such as the climate problem, where the usual Bayesian loop is hard to close because we cannot be sure how the modern climate urn resembles the '$2\times CO_2$ urn' (or the 'LGM urn'). Assumptions of weak resemblance are open to objections of alarmism or denialism (if they contain 'too much' prior probability at the high or low ends of the range, respectively), while assumptions of strong resemblance are open to objections of question begging, since the prior, rather than the data, provides much of the constraint.

In the case of global mean temperature change, a significant part of the problem, as illustrated by Allen *et al.* (2000) and Knutti *et al.* (2005), is that sensitivity does not scale conveniently with past attributable warming (though the transient response does; Frame *et al.* 2006). In essence, experimenters wishing to rule out high equilibrium sensitivities cannot use recent data alone: further beliefs are required to reduce the probability of high equilibrium warming. The problem, in cases in which data constraints are weak, is that the choice of prior can have a dramatic influence on the posterior distribution. In these cases, specifying the prior can prove deeply problematic. As Chalmers (1999) puts it, the subjective Bayesian's

> probability is not a measure of the degree of belief that a scientist actually has but a measure of a degree of belief they would have if they did not know what they do in fact know. The status of these degrees of belief, and the problem of how to evaluate them, pose serious problems, to put it mildly.

This kind of criticism of the Bayesian approach is standard and is acknowledged as such by most Bayesians (Howson & Urbach 1993; Zabell 2005).

The problems associated with using likelihoods versus expert priors are shown in figure 2. Up to 2000, consistent with practices in the detection and attribution community, ensemble members are given equal weight. Beyond 2000, ensemble members are given expert weights (in this case following Murphy *et al.* (2004), but the basic point also holds for studies that use multiple streams of data (not just the recent past) and/or expert priors). What we see is that the inclusion of the additional prior information tightens the forecast, and, somewhat oddly, we appear to be more confident about the future than we are about the past. Members of the climate prediction community could argue that one might tighten one's estimate of anthropogenic warming by including independent information or expert opinions in D&A studies; members of the D&A community could reply that it seems perverse to change one's opinion of anthropogenic warming in the past 50 years owing to one's ideas about the last glacial maximum. Perhaps the real reason the inconsistency is a problem is that standard IPCC likelihood qualifiers are being used in different ways in chs 9 and 10. Perhaps we ought to use different

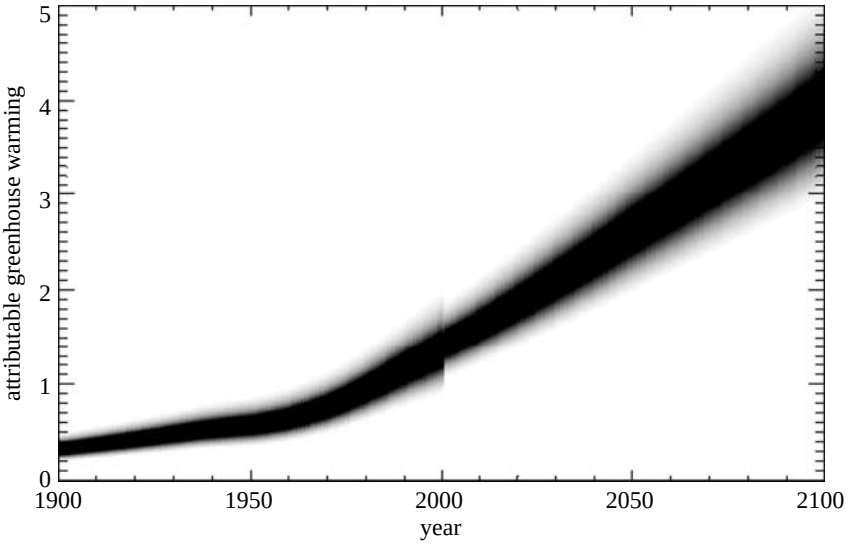

Figure 2. Relative likelihood of past and future greenhouse warming: ensemble estimates of past and future warming, shaded by likelihood conditioned on uniform distribution (until 2000) and expert prior (from 2000).

qualifiers when referring to likelihoods than we do when referring to the subjective distributions that embed expert information.[7]

Focusing on the likelihoods allows scientists to restrict their attention to how the evidence stacks up for each of the models under investigation. As Sober (2002) puts it:

> Likelihoods don't tell you what to believe or how to act or which hypotheses are probably true; they merely tell you how to compare the degree to which the evidence supports the various hypotheses you wish to consider.

We think this is a useful thing for scientists to present, even if they do wish to add further information through the employment of an expert prior. For example, Forest *et al.* (2002) showed the impact of an expert prior on climate sensitivity explicitly by displaying results using both a prior that was uniform over a broad range (which the posterior is close to proportional to the likelihood) and an explicit expert prior. One of us (Allen) argued against the inclusion of the expert prior information at the time, but since both results have been widely and accurately cited in the literature since then, it was clearly the right decision to include it alongside the likelihood information. This seems to us to be an example of 'best practice': by showing the likelihoods, as well as the posterior, they showed the extent to which the prior, rather than the evidence used in that paper, constrained the posterior.

The manifest difficulty of quantifying this kind of uncertainty has led some authors (Stainforth *et al.* 2007*b*) to reject the notion of forecast probabilities as poorly posed, choosing instead to work within explicitly frequentist modelling frameworks for this type of problem. These authors need not deny that Bayes' theorem is a useful way of updating inferences; they merely deny that its use here is appropriate (see also Smith 2007).

[7] Something like 'consensus intersubjective estimate' would seem an accurate, if unwieldy, description.

(*b*) *Probabilistic coherence*

The assumption of logical omniscience is the implication that an agent is aware of *all* the logical implications of their beliefs (Hacking 1967). This is necessary if one is to satisfy one of the probability axioms. Strong readings of logical omniscience are counterintuitive; they imply that Einstein 'knew' the general theory of relativity before he sat down to write it. Even more perversely, strong coherence requirements insist one is fully coherent across all one's beliefs, even those that are in error (a very real problem in a field as uncertain as climate research). This problem, among others, has led to controversy regarding the value of Dutch book arguments (arguments that a set of beliefs violate the coherence requirement) even in less uncertain cases than ours (e.g. Kennedy & Chihara 1979; Maher 1997; Vineberg 2001). Ian Hacking, directly addressing problems associated with personal probability, argues that any conception of incoherence needs to be 'tied to a definition of knowledge' (Hacking 1967). In other words, the issue of coherence is to be determined against a broader epistemological background. In the case Hacking discusses, he claims that 'a man is incoherent if a person *knowing no more than that man does* is assured of winning' on a Dutch book. If all that man has is a model, and the person betting against him has only the same model, then maintaining strict coherence of belief would seem to be required for coherence.

In the climate change case, the man has a model (ensemble) but the person has not only the model but also, say, NCEP and ERA-40 reanalyses. He can see where the model shows systematic biases compared to the present-day real world. The proper objects for coherence would seem to be beliefs about the climate system as a whole, rather than the more restricted set of beliefs about model parameters. Yet our beliefs about the models and the real world are already in conflict, since the real world obeys regularities not represented in models. Therefore, since the models are some way from perfect, allowing ensemble weights to vary according to the question under consideration—so as to be less exposed to model systematic error—seems to be a fairly reasonable thing to do.[8]

In effect, all we are doing is saying 'since the models do not represent the system (and do not exhaust the range of relevant scientific beliefs), we are not going to enforce coherence of our model-related beliefs when we are trying to predict the actual system. In this case, we regard coherence across model beliefs as an irrelevance, since the only coherence that would matter applies to beliefs about the system. To try to account for the fact that we know our model-beliefs are wrong, we are going to use an algorithm that perhaps exaggerates our scepticism but that which ought to err on the conservative side

[8] In fact, perhaps the surest way for the man to lose money on his bets would be to treat the model as structurally perfect, subject only to parametric error, and fail to even try to account for the fact that the real world climatology is not replicated in his ensemble. Say the model has a systematic bias that means its Asian monsoon is too weak. This is true across the entire ensemble. If he perturbs parameters in his model, applies his preferred prior and then attempts to place bets on the strength of the Asian monsoon, knowing that the model is in error there but not accounting for it, then he is going to lose in the long run against someone who makes some reasonable attempt to account for systematic bias.

in its real-world predictions.' While everyone accepts that weak coherence requirements are reasonable for beliefs—it seems a bad practice to believe logical contradictions, for instance—strong coherence requirements seem to be extremely dubious in a field as subject to such chronic uncertainty and model inadequacy as climate research. If one wants to make claims about the real world, then to the extent that Dutch book arguments might apply at all, they ought to apply to the full set of climate-related beliefs, rather than the restricted subset embodied in GCMs. The issue turns not on arguments regarding the axioms of probability, but on epistemology. To insist on coherence of belief across the parameters in a model ensemble is to take the models very seriously as representations of reality. Opponents of this view—whom we might call pragmatists—treat the model ensemble as convenient fictions that map physical quantities to other physical quantities. We do not really care that the parameter set corresponding to our best model of Australian climate (α_{Aust}) is different from the parameter set for simulating global climate (α_{global}), or that our method implies that the implied parameter distributions vary somewhat from problem to problem. In the absence of a perfect model of the climate system, we are under no obligations to consider a parameter which governs the rate at which ice falls through cloud in our best model of Australian climate ($\mathrm{VF1}_{\mathrm{Aust}}$) as being especially closely related to the 'same' parameter in our best global model ($\mathrm{VF1}_{\mathrm{Global}}$), since the rate at which ice falls changes from model to model. If we are attempting to illustrate what our best guess of global climate in 2080 under the A1B scenario is, a GCM might be an appropriate mapping tool. If we want to conduct a probabilistic investigation of whether Greenland will melt in the next 1000 years, we might choose an Earth system model of intermediate complexity. Our choice of base model depends on what the problem is. Intuitively, we all know this. Practically, we all do it. Ideally, pragmatists like us would write down our best model for each of the problems we wish to consider, and then develop forecast distributions from those, without thinking terribly hard about whether or not our model parameter-related beliefs are fully coherent.

4. Likelihood-based approaches in climate*prediction*.net

Different modelling groups have taken different approaches to the problem of ensemble experiment design. These use a range of models, of varying complexity and sampling designs that range from exhaustive Monte Carlo methods to various forms of efficient parameter sampling. Given computing constraints, projects generally face trade offs regarding model complexity versus ensemble size, so it is usually not practical for groups to design large ensemble experiments for GCMs. One exception to this is climate*prediction*.net, which was conceived just to circumvent the usual computational limits by employing distributed computing techniques. Climate*prediction*.net does face constraints: by most climate science standards it is not a flexible experiment. It can be costly to set up new experiments, and those analysing the data need to be mindful of the bandwidth and reliability constraints associated with distributed computing: whatever the downsides of queueing for supercomputer

time, HPC staff seldom randomly turn nodes on and off or try to run GCMs on ageing laptops.

The obvious way to evaluate the assumptions that the effects of parameters combine linearly, or that the parameter space is smoothly varying, or other such assumptions regarding model phase space, is to test them but running more exhaustive parameter sampling exercises. This is what climate*prediction*.net can do. It is not necessarily the case that climate*prediction*.net can provide a better forecast than other approaches, but it is probably true, as Smith (2002) notes, that the experiment will give us a better idea of attempting to place limits on what the climate will not do (given the underlying model structure). It has been claimed (I. Held 2005, personal communication) that it would be possible to pick out the standard GFDL model from the entire family of HadCM3L models that form the climate*prediction*.net ensemble. This is undoubtedly true. Because we are limited by underlying the model structure (and the parameters we have chosen to investigate), we cannot explore the space of every possible plausible climate model. Because our ensemble explores only the subspace available to models with the HadCM3 structure, we would expect all of our models to share certain biases (compared to other plausible climate models). The GFDL model, and any detuned version of it, will have biases in different places; there may be ways in which it is very similar to some member of the climate*prediction*.net ensemble, but there will be ways in which it is identifiably dissimilar. So if we reflect all our beliefs regarding parameter uncertainty in HadCM3 parameters, our forecasts will fail to account for the biases that inhere in *all* models with the HadCM3 structure. If we fail to account for the omission of structural errors in our probabilistic forecast, we run the risk of producing quite tight distributions, especially for regional variables, that would completely fail to agree with a sister experiment that uses a different underlying model structure. One can imagine a situation where one group of scientists (the Light Greys) estimate the uncertainties in their model's parameters and then another group (the Dark Greys) repeat the experiment, using a different base model. If both ignore structural uncertainty, they are likely, for at least some variables, to obtain distributions a bit like those in figure 3: the experiments indicated by both the light grey and dark grey curves will claim quite tight distributions (hence predictive skill) while failing to agree with each other. The approach we have argued for above—finding likelihoods of the data over uniformly sampled forecast variables—is likely to give a much more conservative forecast (the black curve in figure 3) that claims less precision regarding that variable, but is less dependent on model structure (as argued in Allen *et al.* 2006). Having mutually inconsistent distributions that depend so much on model structure seems very undesirable: a decision maker, on being handed the light grey 'subjective pdf' and the dark grey subjective pdf might understandably ask the question 'which one is right?' She may not find the answer, 'they are subjective pdfs' particularly useful. It is basically in order to avoid this sort of problem—which we think is related to the common problem of overconfidence—that much of the research in climate*prediction*.net has been using likelihood-based approaches. As probabilistic climate research starts to try to provide highly regionalized products for adaptation purposes, these sorts of problems are likely to become more apparent.

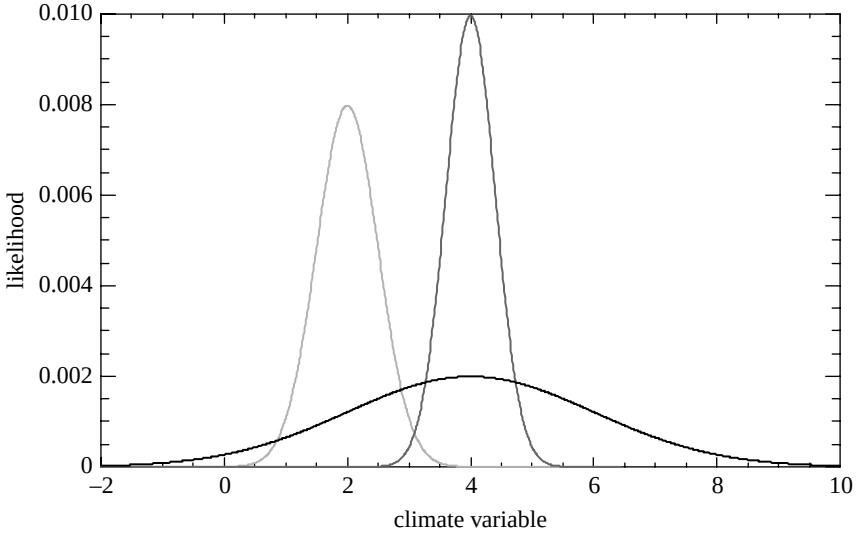

Figure 3. Three (fictitious) predictions of a climate variable. The dark grey and light grey curves are posterior probabilities based on expert priors and the black curve represents the likelihoods, evenly sampled in the forecast variable, that underpin the dark grey posterior.

5. Frequentist bounding box approaches

The approach favoured by Stainforth *et al.* (2007*b*) seeks to avoid many of the problems associated with inductive inferences regarding the future state of the actual system by dealing instead with the modelled world. On this view, all we can say is 'here is what the models do'. This is attractively conservative in terms of the inferences it makes, avoids potential disputes regarding priors and, best of all, is testable: if you want to test whether the models do something, then you run more models and see what happens.

However, even this sort of approach relies on contestable inferences: Stainforth *et al.* (2007*a*,*b*) argue that we ought to take a 'bounding box' approach to the ensemble, in which we plot the model runs on some sort of scatter plot, and then draw a bounding box around the outliers to determine 'what the models do'. But drawing this boundary, and interpreting it, relies on certain decisions or judgements, too. An earlier version of fig. 3 of Stainforth *et al.* (2007*b*) presents a good example of this issue: towards the top right of the figure, there is a region in which no models fall. Further up, and to the right, we see an island of more models. How should we treat this sort of lacuna? Is this lacuna inside or outside the bounding box? As argued above, one of the most attractive features of this methodology is that we can add models and test whether the lacuna persists as we add models. If we run a million (say) models and none falls in the gap, we might think it is a stable feature of the model's phase space. But what exactly does this mean? Are we justified in treating this sort of feature as a bifurcation? If so, how big does the lacuna have to be before it is deemed 'justified'?[9] Should I redo the experiment with other, only obliquely related models and look for the same feature? What would that experiment tell me about the world?

[9] It is tempting, but ambiguous, to refer to the 'reality' of the lacuna.

More practically, though this approach does a nice job of explaining to potential forecast users that climate models are prone to various sorts of chronic error, it is not obvious how it explains to users how the models *are* informative. Because it refrains from weighting models by comparison to the 'real world', it seems to leave to users the difficult job of knowing what the bounding box ought to imply for their real-world decisions. It is true that comparing models to certain (gridded, processed and often reanalysed) real-world observables is contentious. But so is not making that comparison, because it abrogates to decision makers the challenging task of deciding how to infer things about the real world from imperfect models.

6. Inference and second-order uncertainty

The earlier discussion about 'priors' and beliefs invites the question: what attitude should scientists take towards risk? We often express errors in terms of confidence intervals such that we may claim that a variable will have a value within the stated range of 19 times out of 20. The claim represents our best guess of our understanding of the errors, but we can be wrong about this in two sorts of ways: we can overestimate the stated errors, claiming less accuracy than may be justifiable; or we may claim more accuracy than is justifiable. We can call these underconfidence and overconfidence, respectively. When looking at a system as complex as the Earth, scientists will not only be wrong about how relevant processes work, they will also be wrong with regard to their estimates of uncertainty. They are almost certain to commit both types of error. Should they make them in equal measure, or should they be more averse to one sort of error? We suggest that most of us think that science should prefer to be conservative, i.e. to prefer underconfidence to overconfidence. Early estimates of the speed of light contained some that contained the true value within surprisingly large error bars, and some that excluded the true value from within a spuriously tight range of uncertainties (Morgan & Henrion 1990). The latter is usually regarded as the more embarrassing error. Why is overconfidence a sin, while underconfidence mere fussiness? It probably has to do with the scientific practice of building chains of inference: if we are conservative, even sceptical, in our claims and can still reject hypotheses, then the inferential chains we build will be robust. If, however, we admit more speculative hypotheses into our conceptual schemes, we run the risk of building inferential edifices on unstable foundations. The inferential models we thus construct are more likely to be wrong when we set lower burdens of proof. Since robustness is one of the cardinal scientific virtues, this would seem to be an error worth avoiding. Information poor prior beliefs may well lead us to reject some reasonable hypotheses, while ensuring we reject a much larger group of dubious relationships. This leads us towards underconfidence rather than overconfidence, and we argue that this is a desirable scientific practice. While more sanguine priors may well reflect an optimistic scientist's beliefs, we would argue that they are often likely to understate the optimistic scientist's ignorance. In fact, scientists generally make the mistake of overconfidence (Morgan & Henrion 1990):

> The one consistent finding across all elicitation techniques that have been examined is a strong and consistent tendency towards overconfidence.

So we would argue that deliberately and systematically manipulating the prior to give un-confident forecast distributions is not only a good epistemic practice (in that it is more likely to exclude spurious propositions), but also helps us avoid the more common real-world error.

Scientists face incentives to narrow the bounds of their distributions: it makes them look as though they know more. If someone could make a compelling argument that they have constrained climate sensitivity to within a degree, they would receive significant scientific kudos. On the other hand, there is no such pressure back the other way, enticing one to err on the side of caution. The problem is worse than this, in that once probabilistic forecasts start being sold and investment decisions start being made on the basis of them, those responsible for these forecasts will face strong pressure to 'ratchet down' estimates of uncertainty. While few present-day scientists are likely to be around to face the consequences if their forecasts for 2050 climate turn out to be wrong, they are likely to face strong criticism if they revise up their estimates of uncertainty in the relatively near future, effectively admitting that their earlier forecasts were over-optimistic.

7. Decisions in probabilistic climate research

In order to make inferences about the real world, based on the model (or ensemble) at hand, we need to compare its simulations against some aspect of the real world. In the enormously multivariate system that comprises climate, determining what constitutes a relevant test can be problematic. As has been pointed out (C. Piani 2005, personal communication) the issue of which data to compare model ensembles against is problematic: if one includes more and more variables, one rules out more and more models.

Deciding which data matters for which problem is clearly a choice. Reasonable people find contestable decisions in all the available distributions for climate sensitivity, and, in fact, in forecast distributions for any climate variable. To pragmatists, science becomes 'settled' not when it achieves truth or objectivity, but when reasonable and informed people no longer care to dispute decisions to which the claim under debate is sensitive. People either decide that they do not really disagree, or that their disagreements do not actually matter for the final outcome. The way in which scientists usually persuade others that disagreements do not matter is to show that one's results are not sensitive to the contested decisions. Sensitivity analyses help in this regard by demonstrating that, even if one does not agree with the decisions used in a study, they have little bearing on the outcome.

Generally, climate change scientists are unlikely to want to argue about the axioms of logical epistemology, for instance, but might want to argue strenuously about the relevance of a given variable or dataset as a constraint on a given observable or forecast variable, or about the general applicability of a model with a certain functional form. We might, as a first approximation, separate the decisions we make into the following categories:

(i) choice of model,
(ii) choice of dataset,
(iii) choice of model sampling approach,
(iv) choice of weighting data,

Table 1. Approximate taxonomy of some of the decisions made in recent probabilistic climate research.

model	dataset	prior	forecast	study
EBM	recent	uniform	PDF	Andronova & Schlesinger (2000)
—	recent + LGM + volcanic	gamma	PDF	Annan & Hargreaves (2006)
EBM	recent	uniform and expert	PDF	Forest *et al.* (2002)
EBM	recent	uniform	PDF	Frame *et al.* (2005)
EBM	recent + last millennium	uniform	PDF	Hegerl *et al.* (2006)
EMIC	recent	uniform	PDF	Knutti *et al.* (2002)
EMIC	recent	range	imp. prob.	Kriegler & Held (2005)
GCM	recent	expert	PDF	Murphy *et al.* (2004)
EMIC	recent + LGM		PDF	Schneider von Deimling *et al.* (2006)
GCM	recent	—	range	Stainforth *et al.* (2005)

(v) choice of how to combine the information, and
(vi) choice of how to interpret.

Within each of these categories, we may face additional choices. For instance, the choice of model may be further broken down into additional choices:

(i) physical variables to resolve and
(ii) functional form.

In the case of probabilistic climate forecasting, all studies will have to make these sorts of decisions, and we can group them together as in table 1.

Just as there is no algorithm for generating models (it remains imaginative), there is no general theory of relevance. Even if we could deal adequately with issues regarding data relevance, ensemble weighting and so forth, we still become stuck when we consider the issue of model adequacy. As we have seen, any model is prone at least to the theoretical possibility of becoming unhinged through the omission of a small but crucial term. In cases in which we can perform highly repeatable experiments—weather forecasting or aircraft engineering, for instance—this possibility is largely theoretical. In the case of climate change modelling—in which many processes have long time scales and are poorly understood—it is likely to be of first-order importance.

So our 'subjective best guesses' tend to represent tighter estimates than our likelihood functions, which, in turn, are tighter than simple ensemble spread. But the costs of gaining this tightness of distribution come in the form of making increasing numbers of contestable claims. We can stand back a long way from model versus real-world problems and adopt a bounding box approach, or we can choose some sampling strategy and dataset and try to compare them, or we can go for an explicitly subjective best guess distribution. One way of doing the latter

is via expert elicitations (Morgan & Henrion 1990; Morgan & Keith 1995; Morgan *et al.* 2006). At the very least, these make sure that everyone is answering the same questions. They also have the advantage of showing the diversity of expert opinion, and, one might hope, of ensuring that a range of expert opinions are seen, in context, together, rather than in isolation. Again, there is no unproblematic way of combining these distributions but the reflections of diversity represented in their answers to similar questions ought at least to allow decision makers to see how informed scientists might agree or disagree on a particular question. The practical value of this information will differ across decision makers, but for some the concise, structured, qualitative representations of diversity of expert opinion ought to be a useful practical tool.

Among the decisions that different groups make are decisions of how to deal with structural error or the problem of induction posed in §2. Three main approaches have appeared in the literature and they are as follows:

— frequentist approaches,
— likelihood-based methods, and
— subjective Bayesian approaches.

All three are contestable. As we have seen in this article, they are all subject to different sorts of problems, either through the decisions they make (explicitly or implicitly) or through the things they choose to refrain from doing. Frequentists refrain from the problematic step of comparing their ensemble to the real world, but this leaves an interpretive problem for decision makers who want to use that ensemble to make real-world decisions. Likelihood enthusiasts are beholden to their sampling strategy, and can be accused of making partial, though not full, use of available information, and of knowing just how to interpret likelihoods. Subjective Bayesians face considerable problems regarding the choice of prior, and regarding the value of subjective probability.

So how should we think about the information that these models (and ensembles of models) *do* contain? Pragmatists treat physical laws as tools for the manipulation of some system. Realists argue that the laws themselves are true descriptions of those systems. In spite of its attractions for other sorts of physicists, realism seems to be an inappropriate way to conceive of climate physics: arguments regarding the truth content of our best models seem beside the point when one considers that the horizontal extent of every cloud in HadCM3 is some integer multiplied by its grid resolution $2.5 \times 3.75°$. Rather than seeing models as describing literal truth, we ought to see them as convenient fictions which try to provide something useful.

8. Conclusions

There is plenty of scope for disagreement regarding the value of the choices made by various research groups. There is probably less scope for arguing about what we should do about it. Some will favour a standardized methodology, preferably theirs, under which we would all address the problem in basically the same way. The benefits of this approach would seem to be that we are all doing basically similar studies, which ought to, in some sense, be more comparable than the shotgun scatter approach we sometimes have today. This means that the

interpretation of results would be simplified because we could basically ignore voices outside the mainstream methodology, while all agreeing roughly on what the interpretation of the information coming out of the studies is. We would have the incentives in place to use the same language, all agreeing that either the outputs of experiments following the approved methodology are probabilities or they are not. Basically, we would be agreeing on the conditions that underpin the conditionality.

The costs of such an approach are that we ignore legitimate differences of opinion regarding methodology and interpretation. The alternative would seem to let the field of ensemble climate research Balkanise. If we are to follow this approach, we need to behave more like the human sciences, in terms of making it plain that we are working within a given methodological framework, without needing to argue in each paper for the superiority of that framework.

In summary, there is no universally agreed way of approaching the problem of quantifying uncertainty in climate forecasts. Owing to the problems outlined in this paper, especially that of induction, there will never be a single 'right way' of dealing with uncertainty in climate research. The decisions we make regarding (i) the comparison of data to ensembles, (ii) the appropriateness of different 'prior' distributions or families of distributions, and (iii) the adequacy of the model families themselves will always materially affect the results we obtain. Given this situation, the best we can do is to make our methodologies and assumptions as open and transparent as possible. This may sound like a homily to some and dangerously social constructivist to others, but at least this allows simple and direct comparisons between different studies: if everyone followed Forest *et al.* (2002, 2006) in showing the effects of a uniform and expert prior (or the likelihoods as well as the posterior), then we can at least compare the effects of the prior.

D.J.F. thanks the James Martin 21st Century School for its support. N.E.F. thanks the EU Ensembles programme for its support. We would like to thank David Sexton, Peter Taylor and the reviewers for their helpful comments in the preparation of this manuscript. Correspondence and requests for materials should be addressed to D.J.F. (e-mail: dframe@atm.ox.ac.uk).

References

Allen, M. R., Stott, P. A., Mitchell, J., Schnur, R. & Delworth, T. 2000 Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **407**, 617–620. (doi:10.1038/35036559)

Allen, M., Frame, D., Kettleborough, J. & Stainforth, D. 2006 Model error in weather and climate forecasting. In *Predictability in weather and climate* (eds T. Palmer & R. Hagedorn), pp. 391–428. Cambridge, UK: Cambridge University Press.

Andronova, N. G. & Schlesinger, M. E. 2000 Causes of global temperature changes during the 19th and 20th centuries. *Geophys. Res. Lett.* **27**, 2137–3140. (doi:10.1029/2000GL006109)

Annan, J. & Hargreaves, J. 2006 Using multiple observationally-based constraints to estimate climate sensitivity. *Geophys. Res. Lett.* **33**, L06 704. (doi:10.1029/2005GL025259)

Cartwright, N. 1999 *The dappled world: a study of the boundaries of science.* Cambridge, UK: Cambridge University Press.

Chalmers, A. F. 1999 *What is this thing called science?* 3rd edn. New York, NY: Open University Press.

Cox, P. M., Betts, R. A., Collins, M., Harris, P. P., Huntingford, C. & Jones, C. D. 2004 Amazonian forest dieback under climate-carbon cycle projections for the 21st century. *Theor. Appl. Climatol.* **78**, 137–156. (doi:10.1007/s00704-004-0049-4)

de Finetti, B. 1964 Foresight: its logical laws, its subjective sources. In *Studies in subjective probability* (eds H. E. Kyburg & H. E. Smokler), pp. 93–159. New York, NY; London, UK; Sydney, Australia: Wiley.

Forest, C. E., Stone, P. H., Sokolov, A. P., Allen, M. R. & Webster, D. 2002 Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**, 113–117. (doi:10.1126/science.1064419)

Forest, C. E., Stone, P. H. & Sokolov, A. P. 2006 Estimated PDFs of climate system properties including natural and anthropogenic forcings. *Geophys. Res. Lett.* **33**, L01 705. (doi:10.1029/2005GL023977)

Frame, D. J., Booth, B. B. B., Kettleborough, J. A., Stainforth, D. A., Gregory, J. M., Collins, M. & Allen, M. R. 2005 Constraining climate forecasts: the role of prior assumptions. *Geophys. Res. Lett.* **32**, L09 702. (doi:10.1029/2004GL022241)

Frame, D. J., Stone, D. A., Stott, P. A. & Allen, M. R. 2006 Alternatives to stabilization scenarios. *Geophys. Res. Lett.* **33**, L14 707. (doi:10.1029/2006GL025801)

Gärdenfors, P. & Sahlin, N.-E. (eds) 1988 Introduction: Bayesian decision theory—foundations and problems. In *Decision, probability and utility*, pp. 1–18. Cambridge, UK: Cambridge University Press.

Goodman, N. 1954 *Fact, fiction, and forecast*. Cambridge, MA: Harvard University Press.

Hacking, I. 1967 Slightly more realistic personal probability. *Philos. Sci.* **34**, 311–325. (doi:10.1086/288169)

Hegerl, G. C., Crowley, T. J., Hyde, W. T. & Frame, D. J. 2006 Climate sensitivity constrained by temperature reconstructions over the past seven centuries. *Nature* **440**, 1029–1032. (doi:10.1038/nature04679)

Howson, C. & Urbach, P. 1993 *Scientific reasoning: the Bayesian approach*. La Salle, IL: Open Court.

Hume, D. 1739 *A treatise of human nature*. Oxford, UK: Oxford University Press.

Kennedy, R. & Chihara, C. 1979 The dutch book argument: its logical flaws, its subjective sources. *Philos. Studies* **36**, 19–33. (doi:10.1007/BF00354378)

Knutti, R. T., Stocker, T., Joos, F. & Plattner, G.-K. 2002 Constraints on radiative forcing and future climate change from observations and climate model ensmbles. *Nature* **416**, 719–723. (doi:10.1038/416719a)

Knutti, R., Joos, F., Mueller, S. A., Plattner, G. K. & Stocker, T. F. 2005 Probabilistic climate change projections for CO_2 stabilization profiles. *Geophys. Res. Lett.* **32**, L20 707. (doi:10.1029/2005GL023294)

Kriegler, E. & Held, H. 2005 Utilizing belief functions for the estimation of future climate change. *Int. J. Approx. Reason.* **39**, 185–209. (doi:10.1016/j.ijar.2004.10.005)

Manabe, S. & Stouffer, R. J. 1999 Are two modes of thermohaline circulation stable? *Tellus A* **51**, 400–411. (doi:10.1034/j.1600-0870.1999.t01-3-00005.x)

Maher, P. 1997 Depragmatized dutch book arguments. *Philos. Sci.* **64**, 291–305. (doi:10.1086/392552)

Morgan, M. G. & Henrion, M. 1990 *Uncertainty*. Cambridge, UK: Cambridge University Press.

Morgan, M. G. & Keith, D. 1995 Subjective judgments by climate experts. *Environ. Sci. Technol.* **29**, 468–476. (doi:10.1021/es00010a003)

Morgan, M. G., Adams, P. J. & Keith, D. W. 2006 Elicitation of expert judgments of aerosol forcing. *Clim. Change* **75**, 195–214. (doi:10.1007/s10584-005-9025-y)

Murphy, J., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M. & Stainforth, D. A. 2004 Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**, 768–772. (doi:10.1038/nature02771)

Oreskes, N., Shrader-Frechette, K. & Belitz, K. 1994 Verification, validation, and confirmation of numerical models in the earth sciences. *Science* **263**, 641–646. (doi:10.1126/science.263.5147.641)

Papineau, D. (ed.) 1996 *The philosophy of science*. Oxford, UK: Oxford University Press.

Rodwell, M. J. & Palmer, T. N. 2006 Using numerical weather prediction to assess climate models. *Q. J. R. Meteorol. Soc.* **133**, 129–146. (doi:10.1002/qj.23)

Rougier, J. C. 2006 Probabilistic inference for future climate using an ensemble of climate model evaluations. *Clim. Change* **81**, 247–264. (doi:10.1007/s10584-006-9156-9)

Russell, B. 1912 *The problems of philosophy*. London, UK; New York, NY: Williams and Norgate; Henry Holt and Company.

Schneider von Deimling, T., Held, H., Ganopolski, A. & Rahmstorf, S. 2006 Climate sensitivity estimated from ensemble simulations of glacial climate. *Clim. Dyn.* **27**, 149–163. (doi:10.1007/s00382-006-0126-8)

Sklar, L. 1992 *Philosophy of physics*. Oxford, UK: Oxford University Press.

Smith, L. A. 2002 What might we learn from climate forecasts. *Proc. Natl Acad. Sci. USA* **99**, 2487–2492. (doi:10.1073/pnas.012580599)

Smith, L. A. 2007 *A very brief introduction to chaos*. Oxford, UK: Oxford University Press.

Sober, E. 2002 Bayesianism—its scope and limits. In *Bayes's theorem* (ed. R. Swinburne), pp. 21–38. Oxford, UK: Oxford University Press.

Stainforth, D. *et al.* 2005 Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* **433**, 403–406. (doi:10.1038/nature03301)

Stainforth, D. A., Downing, T. E., Washington, R., Lopez, A. & New, M. 2007*a* Issues in the interpretation of climate model ensembles to inform decisions. *Phil. Trans. R. Soc. A* **365**, 2163–2177. (doi:10.1098/rsta.2007.2073)

Stainforth, D. A., Allen, M. R., Tredger, E. R. & Smith, L. A. 2007*b* Confidence, uncertainty and decision-support relevance in climate predictions. *Phil. Trans. R. Soc. A* **365**, 2145–2161. (doi:10.1098/rsta.2007.2074)

Vineberg, S. 2001 The notion of consistency for partial belief. *Philos. Studies* **102**, 281–296. (doi:10.1023/A:1010309526393)

Worrall, J. 2002 Philosophy of science: classic debates, standard problems, future prospects. In *Philosophy of science* (eds P. Machamer & M. Silberstein), pp. 1–17. Edinburgh, UK: Blackwell.

Zabell, S. L. 2005 *Symmetry and its discontents: essays on the history of inductive probability*. Cambridge, UK: Cambridge University Press.