Climatic Change (2007) 81:247–264
DOI 10.1007/s10584-006-9156-9

Probabilistic inference for future climate using an ensemble of climate model evaluations

Jonathan Rougier

Received: 21 May 2004 / Accepted: 20 March 2006 / Published online: 19 January 2007

Abstract This paper describes an approach to computing probabilistic assessments of future climate, using a climate model. It clarifies the nature of probability in this context, and illustrates the kinds of judgements that must be made in order for such a prediction to be consistent with the probability calculus. The climate model is seen as a tool for making probabilistic statements about climate itself, necessarily involving an assessment of the model's imperfections. A climate event, such as a 2 °C increase in global mean temperature, is identified with a region of 'climate-space', and the ensemble of model evaluations is used within a numerical integration designed to estimate the probability assigned to that region.

1 Introduction

A simple question will help to motivate this paper:

> What is the probability that a doubling of atmospheric CO_2 from pre-industrial levels will raise the global mean temperature by at least 2 °C?

This seems to be a well-posed question (subject to technical clarifications which need not concern us here), and certainly a topical one. It is the kind of question a policymaker might ask a climate scientist.

There are two aspects of this question that ought to be highlighted. First, the question asks explicitly for a probability; second, it asks about the behaviour of the climate itself. So it is necessary to establish exactly what is meant by 'probability' in this context, and it is also necessary to understand that answers which focus on the response of this or that climate model are inadequate. In order to satisfy the policymaker, climate scientists must link

J. Rougier
Department of Mathematical Sciences, University of Durham, Science Site, Stockton Road, Durham DH1 3LE, UK
e-mail: J.C.Rougier@durham.ac.uk.

their particular climate model to the climate system, so that their predictive statements about quantities such as future temperature address the needs of policymakers, and are directly comparable to those of other scientists with other models.

This paper clarifies the nature of probabilistic model-based inferences about actual climate, using recent developments in Statistics, notably the field of Computer Experiments. It provides a framework for such inferences, clarifying the role of the climate model, and the purpose of the ensemble of climate model evaluations. In a nutshell, it illustrates 'thinking probabilistically' about climate models and about the climate itself. Therefore even if its detailed prescriptions seem to climate scientists to be inappropriate (although there is no evidence from current practice that they should be), the general approach should serve as a template for all model-based probabilistic inference for climate.

2 Probability and uncertainty

The first thing to understand about probability in this context is that there is no such thing as 'the' probability. To ask for 'the' probability is to make a mistake. A probability is a numerical summary of a person's state of knowledge about a proposition: it is inherently subjective (i.e., it relates to the mind of a subject). Therefore probability takes the possessive article, not the definite one: better to say *your* probability. Inference based on a subjective interpretation of probability is termed *Bayesian Statistics*; see, e.g., O'Hagan and Forster (2004), Bernardo and Smith (1994), or Lad (1996).

Some readers will be concerned about this characterisation of probability. There appears to be a syllogism that runs "Science is objective, your type of probability is subjective, objective and subjective are antonyms, therefore your type of probability has no place in science." It comes up in discussions with scientists often enough to warrant a brief comment. The error is to confuse two meanings of 'subjective'. 'Objective' in this context may be taken to mean disinterested, or uninfluenced by personal prejudice: obviously a hallmark of good science. There is a meaning of 'subjective' which is antonymous to this: emanating from a person's emotions or prejudices. But in dictionaries this is the second meaning. The *first* meaning of 'subjective' is *relating to the mind of the subject*, and this is the appropriate sense when probability is used to describe uncertainty: uncertainty is a property of the mind.

A scientist's prediction will be perforce subjective, but he should aim to be objective as well, by making a disinterested appraisal of the probabilities he attaches to events – this is not paradoxical. Objectivity is not always easy to achieve. For example, if a climate scientist thought too little attention was being given to a certain type of future climate catastrophe, he might be tempted to overstate his probability of the event, in order to attract attention. For the policymaker, though, it is not just what a scientist thinks that is important, but also the extent to which that scientist can justify his assessment. Even though probabilities are subjective statements, not all such statements are demonstrably valid, and of those that are, not all are authoritative. *Valid* statements are those that are consistent with the probability calculus, the axioms of which were clarified by Kolmogorov in the 1930s. The probability calculus allows us to derive potentially complex 'posterior' probabilities from simpler 'prior' ones, and in this sense its practical contribution is to simplify the task of making a prediction. *Authoritative* statements are those for which the scientist is prepared to defend his specification of the prior probabilities as a reasonable summary of his judgements. This does not rule out the judgement "I know very little about this quantity", although too many such judgements might call into question the climate scientist's expertise.

In climate prediction the collection of uncertain quantities for which the climate scientist must specify prior probabilities can be large. Probability distributions over large collections are hard to specify with confidence, in fact the task can often seem overwhelming. They can also be intractable in computations. To make progress it is often necessary to impose additional structure. This is the purpose of constructing a statistical framework (*statistical modelling* is the more usual term, but in this paper 'model' is reserved for the climate model): to find representations of prior probabilities that can be specified in terms of relatively simple numerical summaries, and which are tractable in computations. Note that such a statistical framework is never 'true' – its role is to help structure and summarise a person's judgements.

This paper makes such a *structural* choice, identified below as (S1′). This choice defines the primitive quantities for which a prior distribution must be specified, and does so in such a way that the prior distribution separates into a more manageable set of marginal distributions. This structural choice can easily accommodate current practice in climate science; that is to say, climate scientists should not find it at all restrictive, at least in the short term. This paper also makes two additional *tractability* choices, identified below as (T1) and (T2). These take place within the framework established by the structural choices. The main justification for these tractability choices is computational, and they may easily be generalised.

One of the criticisms levelled at the Bayesian approach is that it is hard to apply in practice. It is hard to quantify judgements about states of knowledge, and structural abstractions, although intended to simplify this process, can also obscure it by their unfamiliarity, or by being too restrictive. Practising Bayesian statisticians know all too well that specifying a prior distribution *per se* is not the difficulty, but specifying a good one is. This is an area full of pitfalls, and often a scientist will find it helpful to work with a statistician to formulate his prior (see, e.g., Garthwaite et al. 2005). The amount of effort and resources devoted to this task ought to reflect the importance of the resulting inference. If it is worth spending thousands of hours constructing a climate model, and millions of dollars collecting climate data, then it does not seem unreasonable to invest a similar amount quantifying our judgements about how these two are related. But this has not yet happened.

3 Predicting future climate

A natural starting-point for predictions about future climate are observations of historical and current climate. In a probabilistic treatment, the predictive distribution for future climate is found by *conditioning* future climate on the observed values for historical and current climate.

We denote climate by the vector (y_h, y_f), collectively y, where y_h corresponds to historical and current climate, and y_f to future climate. The vector y is a large collection of quantities, where each component is typically indexed by type, and by location and time. The components of y_h will depend on the available data, while the interpretation of y_f will depend on what particular future is to be predicted. In the context of the question posed at the start of the Introduction, our y_f would be a future in which concentrations of atmospheric CO_2 are double their pre-industrial levels; more generally y_f might correspond to one of the SRES scenarios (Nakićenović 2000).

We can express the climate data, z say, in general terms as

$$z \equiv y_h + e \tag{1}$$

where e is an uncertain residual quantity defined as $e \triangleq z - y_h$; '$\triangleq$' denotes 'defined as' and '$\equiv$' 'equivalent by definition'. Here we make the simplifying assumption that it is possible to match climate and climate data one-to-one, but a generalisation would be straightforward. Note that there is no sense in which (1) is the 'true' relationship. It functions as a way for the climate scientist to structure his judgements about about z and y_h, so that a distribution for $\{y, e\}$ induces a distribution for $\{y, z\}$.

At this point we make a structural choice to simplify the specification of the prior distribution $\Pr(y, e)$: that e and y can be treated as probabilistically independent, written

$$e \perp\!\!\!\perp y \tag{S1}$$

If I judge the two quantities to be probabilistically independent then knowing the value of y will not change my predictions about e, and *vice versa*. This can also be expressed in terms of conditional probabilities, as $\Pr(e|y) = \Pr(e)$, where '|' denotes 'conditional upon'. Condition (S1) should be recognised as a choice, and a pragmatic one at that. The quantity e might be thought of as 'measurement error'. A list of the ways in which a measurement error can occur shows that many of them are weather-related (e.g., a seasick technician, atmospheric turbulence), and, therefore, climate-related. By adopting (S1) we are making the judgement that the impact of y on e is of secondary importance in the inference as a whole. Note that (1) and (S1) together rule out multiplicative errors in the original units; if required, these can be incorporated using logarithms.

We will also make a tractability choice: that the marginal distribution of e is Gaussian (also referred to as 'normal') with zero mean and specified variance matrix Σ^e, written

$$e \sim \text{Gau}(\mathbf{0}, \Sigma^e), \quad \text{with } \Sigma^e \text{ specified} \tag{T1}$$

(known measurement biases can be incorporated with a non-zero mean). The variance matrix Σ^e might be taken to be diagonal, implying that the measurement errors are treated as mutually independent. However, measurements and measurement errors typically share characteristics, and consideration should be given to constructing Σ^e hierarchically by type of instrument, by instrument ID, and by proximity in space and time; the shared structure would involve non-zero off-diagonal components.

A prediction for future climate is written as the probability density function $\Pr(y_f|z = \tilde{z})$, thought of as a function of y_f; here $\tilde{z}$ denotes the measured values of z, the observational data on historical and current climate. For this section it is clearer, notationally, to predict the whole of y rather than just y_f; the marginal distribution of $(y_f|z = \tilde{z})$ can be extracted from that of $(y|z = \tilde{z})$ straightforwardly, both in theory and practice. In general terms, the prediction for y is

$$\begin{aligned}
\Pr(y|z = \tilde{z}) &= c \Pr(z = \tilde{z}|y) \Pr(y) \\
&= c \Pr(e = y_h - \tilde{z}|y) \Pr(y) \\
&= c \Pr(e = y_h - \tilde{z}) \Pr(y) \\
&= c\varphi(y_h - \tilde{z}; \mathbf{0}, \Sigma^e) \Pr(y)
\end{aligned} \tag{2}$$

where $c \triangleq \Pr(z = \tilde{z})^{-1}$, and $\varphi(\cdot)$ is the Gaussian density function with specified mean and variance. The first line is sometimes referred to as Bayes's Theorem, although it is simply a consequence of the definition of conditional probability in terms of joint and marginal

probabilities, as laid down by the probability calculus. The second line uses the definition of e in (1), the third line uses (S1), and the final line uses (T1). In computations we can often ignore c, which fulfils the role of a normalising constant, ensuring that the probability density function integrates to 1.

Equation (2) demonstrates that a prediction for climate based on historical and current climate data requires a specification of the prior distribution $\Pr(y)$; that is, a probability distribution over climate itself. There is a partial exception, when prior information about y_h is judged so weak relative to the information in $z = \tilde{z}$ that the marginal distribution $\Pr(y_h)$ is taken to be 'locally uniform' (see, e.g., Box and Tiao 1973, Section 1.2.5). In this case the distribution

$$\Pr(y_h|z = \tilde{z}) = c\varphi(y_h - \tilde{z}; \mathbf{0}, \Sigma^e)\Pr(y_h)$$

is approximately Gaussian, and the predictive distribution for y is

$$\Pr(y|z = \tilde{z}) \approx \Pr(y_f|y_h)\,\varphi(y_h - \tilde{z}; \mathbf{0}, \Sigma^e).$$

This generalises to the case where the marginal distribution for e is something other than Gaussian, although huge amounts of climate data may be required (Bernardo and Smith 1994, Section 5.3). But even in this case it is still necessary to specify the conditional prior distribution $\Pr(y_f|y_h)$ in order to make a prediction about future climate.

This is the basic message of this section: *it is not possible to make a probabilistic prediction about future climate without specifying a probability distribution for climate*. Or, to put it another way, anyone who claims to make such a prediction without such a specification has either made an implicit choice – a lack of transparency which is unfortunate in the context of science and policy – or has violated the probability calculus, and so has made a demonstrable error.

4 The role of the climate model

This section explains that *the role of the climate model is to induce a distribution for climate itself*. This is probably not how climate scientists view their models, but it is the appropriate interpretation of the use of these models within the context of probabilistic *climate* prediction.

As outlined in Section 3, the climate scientist needs to specify his prior probabilities $\Pr(y)$. This is challenging both because y is such a large collection of quantities, and because these quantities are linked by complex interdependencies, such as those arising from laws of nature. For example, conservation laws strongly constrain the evolution of the climate state-vector from one moment to the next; a naïve treatment of $\Pr(y)$ such as local uniformity (invoked in Section 3), would violate conservation by treating the climate state-vector at time $t + dt$ as independent of the value at time t.

In such situations statisticians have found it convenient to specify distributions for quantities such as y by *elaboration* from a simpler collection of uncertain quantities (O'Hagan and Forster 2004, Section 4.28). For example, when asked about the number of heads from 10 spins of a coin, a statistician is unlikely to specify directly a probability for each of the eleven outcomes. Rather, additional structure will be imposed on the distribution, because this is judged to give rise to a better quantification. In the case of the spun coin, the probability for each of the outcomes may be inferred from a binomial model and a single uncertain quantity,

the probability that a single spin comes up heads. In this case the additional structure is that the spins are exchangeable.

In the case of climate, the additional structure is supplied by a *climate model*. A climate model is a collection of equations and the means for solving them, at least approximately. These equations embody regularities, both theoretical and empirical, such as conservation laws, equations of state, and coupling equations where the model-domain is described over a number of interacting sub-domains. If we treat some of the coefficients within the equations as uncertain, i.e., we assign them a probability distribution, then the climate model induces a probability distribution over the model-outputs. The same is true if we treat as uncertain the initial value of the state-vector, or the value of various forcing functions. If we can then relate the now-uncertain model-outputs to the climate vector y, we induce a distributionfor y.

Relating an imperfect model to the system it purports to represent is an extremely challenging task: there is no reason to expect a simple resolution. In fact there appear to be foundational difficulties that prevent us from expressing such a relationship in terms of operationally-defined quantities pertinent to the model itself (Goldstein and Rougier 2004, 2006b). These objections notwithstanding, this paper adopts the simplest framework that takes explicit account of the model's imperfections; this is also the *de facto* standard in the statistical treatment of model-based inference for complex systems (see, e.g., Craig et al. 2001; Kennedy and O'Hagan 2001; Higdon et al. 2005; Goldstein and Rougier 2006a).

We treat the climate model as a deterministic function $g(\cdot)$ of a specified collection of model-inputs x (equation coefficients, initial conditions, forcing functions). The model represents the mapping

$$x \to g(x),$$

where the components of $g(x)$ are made to correspond one-to-one with the components of y itself, insofar as this is possible. The input-space $\mathcal{X}$ is the set of all values of x that cannot be ruled out by *a priori* considerations, and the output-space $\mathcal{G}$ is the image of input-space; i.e., $\mathcal{G} \triangleq \{g(x), x \in \mathcal{X}\}$. A point in $\mathcal{G}$ will represent a simplified climate, in the same way that the climate model $g(\cdot)$ is a simplification of the processes that are involved in the real climate.

When a climate scientist considers the ways in which evaluations of the model can be informative about actual climate, it is likely that he will consider, on *a priori* grounds, that some choices of $x \in \mathcal{X}$ are better than others. We posit the existence of a special uniquely-defined model-input value, x^* say, for which

$$y \equiv g(x^*) + \varepsilon^* \tag{3}$$

where $\varepsilon^* \triangleq y - g(x^*)$, is termed the model's *discrepancy*. Together, x^* and ε^* link the model $g(\cdot)$ to the actual climate y, so that assigning a prior distribution to $\{x^*, \varepsilon^*\}$ induces a prior distribution for y. As in (1), Equation (3) is not a 'true' representation of the relationship between $g(\cdot)$ and y, but rather a way for a climate scientist to structure his judgements about such a relationship. Its usefulness depends on our ability to attach meaning to the special model-input value x^*.

The simplest way to provide x^* with a unique definition is to treat it as the 'best' value in the input-space $\mathcal{X}$. In this case we can think of the difference between the climate itself and

any model evaluation in two parts:

$$y - g(x) \equiv \underbrace{g(x^*) - g(x)}_{\text{reducible}} + \varepsilon^* \tag{4}$$

where the first part represents a contribution that may be reduced by a better choice of model-input, and the second part is an irreducible contribution that arises from the model's imperfections. Any analysis of a model based on a specific choice for x, such as the 'standard parameterisation', needs to account for uncertainty about each of these two contributions, unless the climate scientist is prepared to assert that his choice of x is optimal (see the discussion on Tuning in Section 6.1).

The real benefit to defining x^* as the 'best' input is that we can relate it to operationally-defined quantities in the climate itself. Therefore x^* is not just a 'statistical parameter', devoid of meaning: it derives its meaning from the physics in the climate model being approximately the same as the physics in the climate. On this basis, one of the tangible benefits of building higher-resolution climate models is that judgements about x^* become easier to specify, because the physics in the model is closer to that in the climate. Higher-resolution models may also have smaller input-spaces, because of the reduced need for flux-corrections and sub-grid-scale parameterisations. In this paper 'best' is written in scare quotes, because it is hard to give an operational definition that is simultaneously consistent with a particular imperfect model and the climate itself. As a practical consequence the climate scientist can be expected to have quite a vague prior distribution for x^*, particularly for low-resolution models, although this does *not* mean that any old choice will do (see the discussion on $\Pr(x^*)$ in Section 8).

If we impose additional structure on the choice for $\Pr(x^*, \varepsilon^*)$, then we also impose additional structure on $\Pr(y)$. For example, asserting that $\varepsilon^* = \mathbf{0}$ is equivalent to asserting that y *must* lie in $\mathcal{G}$, sometimes referred to as the 'strong constraint'. The grounds for such a strong assertion are unclear. After all, a climate scientist who really judged this to be true of his model would not need to make any further improvements – he would spend the rest of his career exploring his model's input-space. Journal editors and policymakers condone this misjudgement by accepting results that are "conditional on the model being correct". Such results are not *quantitatively* informative about the climate itself, unless the strong constraint has been justified. A justification such as "specifying an appropriate distribution for ε^* is hard, so we set it to $\mathbf{0}$" should be recognised as a calculated decision to make life easier for climate scientists but harder for policymakers.

Therefore we do not want to assert that $\varepsilon^* = \mathbf{0}$. We do, however, impose some lesser structure on $\Pr(x^*, \varepsilon^*)$, in the form of a revised structural choice: that the 'best' input, the discrepancy, and the measurement error are mutually independent, written

$$x^* \perp\!\!\!\perp \varepsilon^* \perp\!\!\!\perp e; \tag{S1$'$}$$

note that (S1$'$) implies (S1), because y is a deterministic function of x^* and ε^*. A simple interpretation is in terms of the two contributions in (4). For any given model evaluation $g(x)$, (S1$'$) implies that our uncertainty about the difference $y - g(x)$ decomposes cleanly into (i) uncertainty about the effect of having a 'non-best' input, and (ii) uncertainty about the discrepancy. We are making the judgement that interactions between these two sources of uncertainty are of secondary importance in the inference as a whole. There are some situations where this may be too strong, and generalisations are possible (Section 7.1).

We also make another tractability choice: that the marginal distribution of ε^* is Gaussian with zero mean and specified variance matrix Σ^ε, written

$$\varepsilon^* \sim \text{Gau}(\mathbf{0}, \Sigma^\varepsilon), \text{ with } \Sigma^\varepsilon \text{ specified} \tag{T2}$$

(known model-biases can be incorporated with a non-zero mean), where the rows and columns of Σ^ε are partitioned in 'h' and 'f', so that, for example, Σ^ε_{hh} and Σ^e both match y_h.

In computations we will not need the prior distribution $\Pr(y)$ explicitly, but for reference it is found as

$$\begin{aligned}
\Pr(y) &= \int \Pr(y|x^*)\Pr(x^*)\,dx^* \\
&= \int \Pr(\varepsilon^* = y - g(x^*)|x^*)\Pr(x^*)\,dx^* \\
&= \int \Pr(\varepsilon^* = y - g(x^*))\Pr(x^*)\,dx^* \\
&= \int \varphi(y - g(x^*); \mathbf{0}, \Sigma^\varepsilon)\Pr(x^*)\,dx^*.
\end{aligned} \tag{5}$$

The first line is a standard result from the probability calculus, sometimes referred to as the 'law of total probability'. The second line uses the definition of ε^* from (3), the third line uses (S1′), notably $\varepsilon^* \perp\!\!\!\perp x^*$, and the final line uses (T2). This makes it clear that the task of specifying a prior distribution $\Pr(y)$ given the model $g(\cdot)$ has been repackaged into specifying a prior distribution $\Pr(x^*)$ and a variance matrix Σ^ε, both of which are model-specific. It may not be easy for the climate scientist to quantify his judgements in these terms, but given that he needs a distribution for y in order to proceed, the real question is whether it is better to specify $\Pr(x^*)$ and Σ^ε, or to specify $\Pr(y)$ directly.

Technical aside. To simplify the exposition, in this paper the climate model $g(\cdot)$ is treated as known, rather than known only at a finite set of evaluations. Judgements about $\{x^*, \varepsilon^*\}$ should be thought of as conditional on $g(\cdot)$. This may be reasonable in many applications, but see Goldstein and Rougier (2006a) and the references therein for the more general case, where $g(\cdot)$ is itself treated as uncertain.

5 Model validation

Model validation is about assessing whether the model is indeed about as good as the climate scientist judges it is; in our case, that is according to his specification of $\Pr(x^*)$, Σ^ε, and Σ^e. This can be done by examining the climate scientist's ability to predict the historical and current climate data. This has been termed 'retrodiction' or 'postdiction'. In Bayesian Statistics the term 'prediction' suffices because uncertainty is not exclusively a property of future quantities: anything uncertain can be predicted, and, prior to seeing the climate data $\tilde{z}$, z is simply another uncertain quantity, just like y_h or y_f.

The predictive distribution for z is found using the same steps as in (5), which gives

$$\Pr(z) = \int \varphi\big(z - g_h(x^*); \mathbf{0}, \Sigma^\varepsilon_{hh} + \Sigma^e\big)\Pr(x^*)\,dx^*. \tag{6}$$

Here we have used the fact that $z \equiv g_h(x^*) + \varepsilon_h^* + e$ from (1) and (3), and, using (S1′), (T1), and (T2), the collection $\{\varepsilon^*, e\}$ is jointly Gaussian, so that $\varepsilon_h^* + e$ is Gaussian.

Note that in (6), and some of the expressions below, the two variance matrices for ε_h^* and e are combined additively. It would be possible to specify a single variance matrix in their place, but this would obscure the fact that *model discrepancy and measurement error are two completely different things.* Consequently their variance matrices are quite different. The only situation in which a single matrix might be specified is where the climate scientist judges that one uncertainty dominates the other. For example, if $(\Sigma_{hh}^{\varepsilon})^{-1}\Sigma^{e} \ll I$, where I is the identity matrix, then the discrepancy variance dominates. If this was judged to be the case, the climate scientist would be justified in approximating Σ^e with $\mathbf{0}$, and concentrating his efforts on specifying Σ^{ε}.

The simplest way to construct a diagnostic from $\Pr(z)$ and the observed climate data $\tilde{z}$ is to compute marginal probabilities such as $\Pr(z_i \leq \tilde{z}_i)$, where z_i is the ith component of z; general methods for computing this type of probability are discussed in Section 8. A value close to 0 or to 1 indicates that the observed value $\tilde{z}_i$ is in the tail of the climate scientist's distribution for z_i, which may give cause for concern. Rather than use single components of z, we might also use subsets; for example, all quantities of a certain type, or from a certain location or time inconsistencies. In this way it is possible to get an insight into where inconsistencies arise. Perhaps the climate scientist has overstated the model's ability to replicate precipitation, by choosing too-small values in the appropriate part of the diagonal of Σ^{ε}. In this case, removing precipitation from a subset would improve the value of the diagnostic.

Responding to evidence of inconsistency can be tricky. To a limited extent the climate scientist can modify his choice for, say, Σ^{ε}. This would usually be appropriate if he were initially quite uncertain about his specification of Σ^{ε} and if he implemented changes that were highly aggregated, to avoid the danger of over-fitting to $\tilde{z}$. Statistical purists would regard this as a form of double-counting, but it may also be regarded as an informal implementation of a hierarchical statistical framework. It may be difficult to improve the diagnostic in this way, which would suggest that the problem is more fundamental; for example, the use of Gaussian distributions in (T1) or (T2). Sometimes this can be addressed by transforming some of the climate components (e.g., by using logarithms for small but strictly positive quantities, or the logistic transformation for proportions).

6 Climate inference with a climate model

By this stage the climate scientist should already have validated his model and his specification of $\Pr(x^*)$, $\Sigma_{hh}^{\varepsilon}$ and Σ^{e}, as described in Section 5. There is no direct way to validate his specification of $\Sigma_{hh}^{\varepsilon}$ or $\Sigma_{hf}^{\varepsilon}$, although these ought to be consistent with his choice for $\Sigma_{hh}^{\varepsilon}$. Where possible, the climate scientist should arrange that there are components of y_h that are like the components in y_f.

6.1 Learning about the 'best' model-input

In Statistics learning about x^* is often referred to as *model calibration.* Climate scientists sometimes speak of 'constraining' x^* with the climate data, but this conjures up the wrong image. It is not simply a case of ring-fencing an area of 'good' candidates for x^* within the model's input-space $\mathcal{X}$. Even were the climate data measured without error (i.e., $z = y_h$), it would be difficult to partition the model's input-space $\mathcal{X}$ cleanly into 'good' and 'not-good' regions, because of uncertainties about the relationship between the model and climate.

Instead, calibration consists of computing the distribution $\Pr(x^*|z = \tilde{z})$. We can think of this distribution as 'scoring' the points in $\mathcal{X}$ as candidates for x^*. A key feature of calibration is that it should take account of the model's imperfections. It would make little sense to calibrate an imperfect model on the basis that the model was, in fact, perfect. The model validation step should prevent this type of misjudgement from occurring.

The calibration distribution is

$$\begin{aligned}\Pr(x^*|z = \tilde{z}) &= c\Pr(z = \tilde{z}|x^*)\Pr(x^*) \\ &= c\varphi\big(\tilde{z} - g_h(x^*); \mathbf{0}, \Sigma^{\varepsilon}_{hh} + \Sigma^{e}\big)\Pr(x^*) \end{aligned} \tag{7}$$

where $c \triangleq \Pr(z = \tilde{z})^{-1}$, as before. The first line is Bayes's Theorem, and the second line follows from (S1′), (T1), and (T2), using the same reasoning as in (5) and (6). One way to summarise this calibration distribution is in terms of its modes, i.e. its local maxima. In our approach we have posited a unique 'best' model-input, but this does not mean that we will find a clear-cut candidate using the climate data. It is quite possible – in fact, it is highly likely – that there will be multiple modes in the calibration distribution.

For example, with the given data $z = \tilde{z}$ and a particular model it might not be possible to distinguish between a candidate value for x^* with high climate sensitivity offset by a strong aerosol forcing, and one with low climate sensitivity offset by a weak aerosol forcing. These might be two modes in the calibration distribution or, if the data and the model permit, they might be two points on a ridge. In situations like this we can use the probabilistic framework to determine what kind of additional observations would best reduce our uncertainty about x^*. The simplest approach is to generate *pseudo-data* using the climate model and the statistical framework, include these data in y_h and z, and examine their impact on the calibration distribution (e.g., in terms of reducing the variance or selecting one mode over the other).

6.2 'Tuning'

This process of computing the calibration distribution should be contrasted with the alternative practice of 'tuning' the model using the data $z = \tilde{z}$. Tuning the model involves searching over $x \in \mathcal{X}$ for an optimal value that minimises some metric defined on the vector of differences $\tilde{z} - g_h(x)$; here $\tilde{z}$ is standing in for y_h, which we cannot observe directly. All big climate models are tuned in this way, and the result is sometimes termed the 'standard parameterisation'. However, because the model evaluation times are so long, typically only a small number of components of x are varied in the tuning procedure, and it is seldom possible to verify that the standard parameterisation is optimal.

There are two issues with this practice, even assuming that it can be effectively implemented. First, what metric should be minimised? Second, what about the situation where many choices of x seem to give the same optimal value? The probabilistic approach gives cogent answers to both of these questions. First – taking the simple case where $\Pr(x^*)$ is locally uniform – the implicit metric is the generalised sum of squares

$$(\tilde{z} - g_h(x))^T \big(\Sigma^{\varepsilon}_{hh} + \Sigma^{e}\big)^{-1}(\tilde{z} - g_h(x)),$$

from (7). This shows explicitly the role of the model discrepancy and the measurement error in quantifying the difference between $\tilde{z}$ and $g_h(x)$. Second, the probabilistic approach outlined in this paper never commits us to using just a single model-input value, but allows us to incorporate, with appropriate weights, all candidates for x^* that are not ruled out by

the climate data (Section 8). Although our approach commits us to accepting a 'best' input, it does not commit us to acting as though we have found it. Nor does it make the concept of the 'best' input contingent on the data we happen to have collected, although we do use that data to learn about the 'best' input's value.

6.3 Learning about future climate

There are two distributions for future climate: the prior prediction $\Pr(y_f)$ and the posterior prediction $\Pr(y_f|z=\tilde{z})$. The prior prediction may be computed from (5). The posterior prediction is the conditional prediction, also referred to as the *calibrated prediction* because it is the prediction made once x^* has been calibrated using $z=\tilde{z}$. The calibrated prediction is computed as

$$\begin{aligned}\Pr(y_f|z=\tilde{z}) &= \int \Pr(y_f|x^*, z=\tilde{z}) \Pr(x^*|z=\tilde{z})\, dx^* \\ &= \int \varphi(y_f; \mu_{f|z}(x^*), \Sigma_{f|z}) \Pr(x^*|z=\tilde{z})\, dx^*, \end{aligned} \tag{8a}$$

where $\Pr(x^*|z=\tilde{z})$ was given in (7), and

$$\mu_{f|z}(x) \triangleq g_f(x) + \Sigma^{\varepsilon}_{fh}\left(\Sigma^{\varepsilon}_{hh} + \Sigma^{e}\right)^{-1}(\tilde{z} - g_h(x)), \tag{8b}$$

$$\Sigma_{f|z} \triangleq \Sigma^{\varepsilon}_{ff} - \Sigma^{\varepsilon}_{fh}\left(\Sigma^{\varepsilon}_{hh} + \Sigma^{e}\right)^{-1}\Sigma^{\varepsilon}_{hf}. \tag{8c}$$

The first line in (8a) is the law of total probability. The second line, and the definitions in (8b) and (8c), follow from the fact that the conditional distribution $\{y, z\}|x^*$ is Gaussian, by (S1′), (T1), and (T2). The expressions in (8b) and (8c) are standard matrix expressions for the conditional mean and variance of a Gaussian random vector (see, e.g., Mardia et al. 1979, Section 3.2).

Equation (8) makes it clear that there are two routes via which the climate data impact on predictions of future climate. First, these data typically have the effect of concentrating the distribution $\Pr(x^*|z=\tilde{z})$ in (8a), relative to its prior form $\Pr(x^*)$. If $g_f(x)$ varies a lot over $x \in \mathcal{X}$, then this concentration reduces our uncertainty about y_f, through reducing our uncertainty about $g_f(x^*)$. The degree of concentration depends not just on the quantity of climate data, but also on its variety. A given type of data will tend to concentrate the calibration distribution in a particular way. Ideally we want these different concentrations to be 'orthogonal', so that the size of their intersection – the joint concentration – is small. A simple approach is to ensure that there are several different types of quantities in y_h (physical insights are helpful here). A strong version of this is to include observations on palaeo-climate proxies, involving an additional 'forward' model mapping the climate state-vector into the proxies.

The second route operates through the discrepancy ε^*, and can be seen in the mean function, (8b). As long as $\Sigma^{\varepsilon}_{fh} \neq \mathbf{0}$, the difference $\tilde{z} - g_h(x)$ is used to adjust the mean of y_f away from $g_f(x)$. This route will not operate for palaeo-climate proxies if – as seems likely – the covariance between the proxies and the post-industrial climate state-vector in both the measurement error and the model discrepancy is judged to be zero.

The posterior predictive distribution can also be used to quantify the value of additional climate data, e.g., from a proposed array of buoys. Pseudo-data representing measurements from the buoys can be included in z, and the impact of these additional data can be quantified in terms of their ability to reduce uncertainty about key aspects of future climate. Ideally the arrangement (i.e., number, location, and telemetry) of the buoys can be optimised in this way, but at the very least it is possible to rule out costly experiments that appear to have little impact on our predictions for future climate.

7 Specifying the discrepancy

The climate scientist can leave the discrepancy ε^* out of the analysis by setting $\Sigma^\varepsilon = \mathbf{0}$, which has the effect of constraining ε^* to be $\mathbf{0}$, according to (T2). Section 4 has already discussed why this would be a misjudgement. This section examines Σ^ε in more detail.

7.1 Specifying the diagonal components

The diagonal components of Σ^ε, denoted diag(Σ^ε), summarise the climate scientist's judgement about the model's ability to replicate climate. For any given model-output component $g_i(\cdot)$ and matching climate component y_i, the climate scientist must specify how accurate he judges the model would be at its 'best' model-input x^*. According to (S1′), this can be done without reference to the actual value of x^*.

Take Sea-Surface Temperature (SST), for example, at a specific spatial location such as the Azores, and time, such as today. How accurately can we expect the model to replicate this climate quantity at its 'best' input? Probably better than $\pm 10^\circ$C, but probably not as well as $\pm 1\,^\circ$C. For his particular model the climate scientist might judge that a value such as $\sqrt{\Sigma^\varepsilon_{ii}} = 2\,^\circ$C would be appropriate, perhaps representing his view that

$$\Pr(|\varepsilon^*_i| \leq 6^\circ\text{C}) \geq 95\%,$$

by a statistical rule of thumb for unimodal distributions known as the 3-Sigma Rule (Pukelsheim 1994). Possibly the same value will do for all SSTs, or maybe it will need to be modified by latitude. This type of reasoning about diag(Σ^ε) may be crude, and subject to much refinement in the future, but it is less crude than choosing diag(Σ^ε) = $\mathbf{0}$.

Another climate scientist might differ, and suggest a standard deviation of 1.5°C for the same quantity. It is natural to ask: who is right? and does it matter? It is not necessary to compare the propriety of the two values directly, although in some cases that may be possible. It may be easier to investigate the impact of the different choices on the resulting climate inference: perhaps the inference is insensitive to the difference. If the difference *does* matter, then diagnostics (Section 5) may be able to indicate that one value better reflects the data than the other. But there can be no certainty of a clear-cut answer to the subtle question of which scientist to favour. One possible response is to use the probabilistic framework to determine which additional climate observations would best resolve the issue, perhaps using pseudo-data (Section 6).

An interesting situation arises when the climate scientist judges that our structural choice (S1′) is not sufficiently general. This makes it difficult to specify the diagonal and non-diagonal components of Σ^ε because the structural framework within ε^* is defined is inadequate. For example, suppose it was revealed that x^* was an extreme value in $\mathcal{X}$. At an extreme model-input value, simplifications in the model might break down or the model's

solver might break down, both leading to a model-output $g(x^*)$ which was less trusted as a representation of the climate system than the output from a central value for x^*. In this case the climate scientist might judge that the discrepancy could be larger for extreme values of x^*. Another situation which violates (S1′) is where the climate scientist judges (e.g., from theoretical considerations) that a certain model-input value is good for predicting one subset of the components of y and a different value is good for predicting another subset: these two subsets might be differentiated by type, for example atmospheric pressure and ocean salinity.

In cases like these, the dependence of the climate scientist's uncertainty about ε^* on the value of x^* can be introduced into the inferential calculations with only minor additional cost, by treating the relationship between y and $g(\cdot)$ given in (3) in the more general form

$$y \equiv g(x^*) + Q(x^*)^T \varepsilon^*$$

where $Q(x^*)$ is the square-root of a specified variance matrix defined as a function of x^*, and then $\varepsilon^* \perp\!\!\!\perp x^*$ as before. In a statistically more sophisticated treatment, Goldstein and Rougier (2004, 2006b) consider a generalisation of (3) and (S1′) which may be thought of as a way to choose $Q(\cdot)$ in the above expression, but which is really a more flexible framework for specifying judgements about the relationship between the climate model and the climate system, within which (3) and (S1′) is a special case. Goldstein and Rougier (2006b) contains the most complete statement in the Statistics literature on linking one or more models and the underlying system.

7.2 Specifying the off-diagonal components

The off-diagonal components of Σ^ε summarise how the discrepancies are related across different model-outputs. If $\Sigma^\varepsilon_{fh} = \mathbf{0}$ then inspection of (8b) shows that the predictive mean of $y_f|(x^* = x, z = \tilde{z})$ is simply $g_f(x)$. But if $\Sigma^\varepsilon_{fh} \neq \mathbf{0}$ then non-zero components in $\tilde{z} - g_h(x)$ modify the mean of y_f relative to $g_f(x)$. A simple way of expressing this is that the prediction in this case can be 'bias-correcting'.

Climate scientists tend to believe that where a climate model is in error, it is often systematically so. If, for example, the model is revealed to have under-represented SST in the Azores for the last twenty years, then the climate scientist might judge that there is a more-than-evens chance that this under-representation will continue into the future. Spatially, if the model is revealed to over-represent rainfall in northern France, the climate scientist might judge that there is a more-than-evens chance that it over-represents rainfall in southern France as well. There may also be other more complicated types of effect: perhaps if the model over-represents temperature it contemporaneously (or with a lag) under-represents rainfall. It is these types of systematic error that the off-diagonal components Σ^ε represent.

Craig et al. (2001, p. 722) give an example of how these types of judgements about systematic errors in the model may be represented in practice. The authors consider a model of a hydrocarbon reservoir, and are concerned with the discrepancy on well pressures, at different wells and at different times. After a discussion with the reservoir engineers, and supported by data analysis on the output of a simplified (low-resolution) version of the model, they selected a discrepancy variance of the general form

$$\mathrm{Cov}\left(\varepsilon^*_{it}, \varepsilon^*_{i't'}\right) = \sigma_1^2 \exp\{-\theta_1(t-t')^2\} + \sigma_2^2 \delta_{ii'} \exp\{-\theta_2(t-t')^2\}$$

where i resents a well location and t represents time, $\delta_{ii'}$ is the Kronecker delta function, and $\{\sigma_1, \sigma_2, \theta_1, \theta_2\}$, termed the *hyperparameters*, have explicit values assigned. In this

specification there is a time effect, which says that discrepancies tend to extend through time, which interacts with a location effect, which says that discrepancies at the same well tend to be more closely related than discrepancies at different wells. This type of parameterisation can be fed back to the reservoir engineers as (random) realisations of the discrepancy vector, plotted by well and by time, so that they can get a feeling for typical behaviour, and then adjust the hyperparameters if necessary. Craig et al. (1998) describe computer-based tools for this purpose.

The restriction that Σ^{ε} be a variance matrix imposes constraints on the type of functional forms that can be used to parameterise the covariance in terms of properties such as type, location, and time. Choosing and quantifying such parameterisations is part of *Spatial Statistics* (see, e.g., Cressie 1991).

8 Computation: The role of the ensemble

This section describes how to go from a climate scientist's posterior prediction $\Pr(y_f|z=\tilde{z})$, given in (8), to his probability for any particular climate event.

Each climate event can be identified with a region of climate-space, and the probability of the event is then the probability that actual climate y falls into that region. To address the question posed at the start of the Introduction, denote by Q the region containing all the values for y in which global mean temperature is at least 2°C higher in 2100. To assess the climate scientist's probability of the event, we integrate his posterior prediction for y over the region Q. This is a perfectly general operation, but our Q only involves future climate, so we can concentrate on y_f rather than the whole of y. The result is

$$\begin{aligned}\Pr(y_f \in Q|z=\tilde{z}) &= \int 1_Q(y_f)\Pr(y_f|z=\tilde{z})\,dy_f \\ &= \int 1_Q(y_f)\int \varphi(y_f;x^*)\Pr(x^*|z=\tilde{z})\,dx^*dy_f \\ &= \int \left\{\int 1_Q(y_f)\,\varphi(y_f;x^*)\,dy_f\right\}\Pr(x^*|z=\tilde{z})\,dx^* \qquad (9)\end{aligned}$$

where $1_Q(y_f)$ is the indicator function of the event $y_f \in Q$. The first line integrates over Q, the second line introduces the posterior prediction from (8), writing $\varphi(y_f;x^*)$ for $\varphi(y_f;\mu_{f|z}(x^*),\Sigma_{f|z})$ to simplify the notation, and the final line re-arranges this integral to express it as an expectation of a function of x^* with respect to the calibration distribution, (7). An interesting feature of (9) is how the integration over climate-space is re-formulated as an integration over model-input space, through the use of a climate model to induce a distribution for climate itself.

Typically all model-based calculations of probabilities for climate events will involve the type of double integral given in (9): one integral over y_f, and one over x^*. An advantage of the two tractability choices (T1) and (T2) is that the internal integral with respect to y_f can often be computed directly. For example, if the climate scientist can arrange for one of the model's outputs to correspond to global mean temperature in 2100, then the integral of $1_Q(y_f)\varphi(y_f;x^*)$ is over the righthand tail of a Gaussian distribution with known mean and variance, and this can be computed at effectively no cost. To emphasise this we can write (9)

as

$$\Pr(y_f \in Q|z=\tilde{z}) \equiv \int f(x^*)\Pr(x^*|z=\tilde{z})\,dx^* \tag{9'}$$

where $f(x) \triangleq \int 1_Q(y_f)\,\varphi(y_f;x)\,dy_f$.

The remaining integral, over x^*, is much more challenging. Resource constraints will often prevent us from evaluating the integrand as many times as would be required to determine $\Pr(y \in Q|z=\tilde{z})$ to high accuracy. *The role of the ensemble of model evaluations is to estimate the integral over x^* as well as possible*. It follows that the choice of evaluations in the ensemble should be guided by the principles of numerical integration, for which there is a large literature (see, e.g., Robert and Casella 1999; Evans and Swartz 2000).

One possible stochastic method is Monte Carlo integration; this is not the best method but it is one of the simplest. Treating (9′) as a statistical expectation, we have the approximation

$$I^{(n)} \triangleq n^{-1}\sum_{i=1}^{n} f(X_i) \quad X_i \overset{\text{iid}}{\sim} \Pr(x^*|z=\tilde{z}), \tag{10}$$

where the model-inputs $X_1,\ldots,X_n$ are sampled independently from the calibration distribution, and the model is evaluated at each set of inputs in order to compute $f(X_i)$. This approximation is asymptotically exact, i.e.,

$$\lim_{n\to\infty} I^{(n)} = \Pr(y_f \in Q|z=\tilde{z}),$$

by the Strong Law of Large Numbers (see, e.g., Grimmett and Stirzaker 2001, p. 329). Therefore the value of $I^{(n)}$ can be thought of as an estimate of the required probability. The problem with (10) is that we do not have a simple way to sample from $\Pr(x^*|z=\tilde{z})$. We can solve this problem by sampling from the prior distribution $\Pr(x^*)$ and then re-weighting (see, e.g., Smith and Gelfand 1992). This gives a different approximation

$$J^{(n)} \triangleq \sum_{i=1}^{n} w_i f(X_i) \quad X_i \overset{\text{iid}}{\sim} \Pr(x^*) \tag{11a}$$

where

$$w_i \propto \Pr(z=\tilde{z}|x^*=X_i) = \varphi\big(\tilde{z}-g_h(X_i);\mathbf{0},\,\Sigma^{\varepsilon}_{hh}+\Sigma^{e}\big), \tag{11b}$$

from (7), and $\sum_{i=1}^{n} w_i = 1$. The distribution $\Pr(z=\tilde{z}|x^*)$ is termed the *likelihood function* of x^*, and (11) can be simply expressed as "sample from the prior, weight by the likelihood". The role of the climate data $z=\tilde{z}$ is to up-weight model-inputs that give good matches, and down-weight those that give bad matches. This approximation is also asymptotically exact. Crucially, there are standard assessments of the accuracy of estimates such as $I^{(n)}$ and $J^{(n)}$ which will depend on n.

Equation (11) illustrates the balancing act at the heart of every model-based inference, between the quality of the model and the accuracy of the inferential approximation. A high-resolution model confers two advantages. First, the prior distribution $\Pr(x^*)$ may be easier to specify, as discussed in Section 4. Second, the discrepancy variance Σ^{ε} ought to be smaller, so that the induced probability distribution for y has more physical structure; this also diminishes the sensitivity of the inference to the climate scientist's specification of Σ^{ε},

which can be challenging. On the other hand, a long evaluation time means that n in (11) will be small, compromising the accuracy of the probability estimate. There is a danger that climate scientists' natural inclinations will take them too far in the direction of model quality. A high-resolution model that can only be evaluated once, or a handful of times, is of little use in climate inference because the probability estimates are not reliable. One simple expedient to illustrate this would be to require all climate scientists who use models to publish a measure of the accuracy of their probability estimates, such as the 95% confidence interval. This would show in a relatively short time where the balance ought to lie.

There are better 'generic' methods than simple Monte Carlo integration, for example importance sampling with variance reduction techniques like antithetic variables, or latin hypercube sampling. More sophisticated still, we can use methods that allow us to exploit the particular features of our integrand, based on our knowledge of the climate model from the evaluations as they become available (sequential methods), from other sets of evaluations in related experiments, or from evaluations of other similar climate models. It can be quite hard work designing an effective ensemble to perform a given inference, but if a climate model takes hundreds of hours to evaluate for a given x, then it does not seem unreasonable to spend a few hours choosing the ensemble before the experiment starts, or choosing the next point in the ensemble while the experiment is running. It seems almost reckless to select each point at random using a simple Monte Carlo approach.

8.1 Choices for $\Pr(x^*)$

This paper has not made any explicit suggestions for specifying the probabilistic framework, beyond the structural choice (S1′) and the two tractability choices (T1) and (T2). It is pitched at a very general level, and actual specifications will depend on both the application and the climate scientist. However, some bad choices have been highlighted: both $\Sigma^e = \mathbf{0}$ and $\Sigma^\varepsilon = \mathbf{0}$ are bad choices, in the sense that they would seldom represent the climate scientist's own judgement about his data and his model.

One more bad choice should be highlighted, which is to assign x^* a uniform distribution on $\mathcal{X}$. This asserts that the climate scientist judges that every value in $\mathcal{X}$ is an equally-good candidate for x^*, no matter whether it is in the centre of $\mathcal{X}$, or tucked into a corner. To take one example, how many of the authors in Murphy et al. (2004) really judge that, for their model, all values for the 'best' entrainment rate coefficient between 0.60 and 9.00 are equally-probable even though the standard setting is 3? Or that values of 0.59 or 9.01 are simply impossible? If a value of 9.01 is impossible, then common sense suggests that a value of 9.00 ought to be highly improbable, and certainly less probable than a value of 3. So at the very least a triangular distribution would have been more defensible.

Referring back to the discussion in Section 2, there is no sense in which any distributional choice is 'objective', unless it is a disinterested assessment of uncertainty. So a simpler shape like a rectangle does not confer more objectivity than a more complicated shape like a triangle. The only thing that confers objectivity is a climate scientist's ability to justify his choice without compromising his reputation.

9 Summary: Pertinent questions

Instead of a standard summary, I offer here a selection of questions for policymakers to ask climate scientists, and for climate scientists to ask each other. The purpose of the questions is to

illuminate the degree to which climate scientists have objectively quantified their uncertainty, and so to assess the value of their probabilistic predictions as a guide to future climate.

0. *Probability*. What do your probability statements about future climate represent? Why should we believe that your probability is a better guide to the future than someone else's?

A good guide to the second part of this question should be found in the answers to the following.

1. *Measurements.* Do you have exact observations on historical and current climate data? If not, how have you quantified the measurement errors? (Σ^e in our treatment.)
2. *The 'best' model-input*. How have you related your climate model to the climate itself: are you adopting the 'best' model-input approach? (If not, see below.) If so, do you judge that extreme values of the 'best' input are as likely as central values? ($\Pr(x^*)$ in our treatment.)
3. *Model imperfections*. Do you believe that if you knew the 'best' modelinput x^*, then the model-output $g(x^*)$ would exactly replicate climate itself? If not, how have you quantified the model's imperfections? (Σ^ε in our treatment.)
4. *Model validation*. What diagnostics did you compute to check that your climate data and your model are about as good as your specification of Σ^e, $\Pr(x^*)$, and Σ^ε imply?
5. *Computation*. How did you estimate your probability? What uncertainty do you have about your estimate? Possible follow-up: Isn't it rather reckless to use a random design (like Monte Carlo integration) if the model evaluations are very expensive?

Any attempt to derive model-based predictions for future climate needs to address these same issues. The particular contribution of this paper has been to focus on *probabilistic* prediction, and to channel these questions into particular quantities, according to the probabilistic framework summarised by the choices given in (S1′), (T1), and (T2). A climate scientist is welcome to reject these choices as being too simplistic to represent his judgements about the model, real climate, and the climate data. What he cannot do is reject the probability calculus if he wants to make probabilistic predictions. Consequently he must show, mathematically, how his more general framework can be used to perform the operations of model validation, calibration and calibrated prediction, and provide a precise statement about the role of the ensemble of model evaluations, exactly as has been done here.

Acknowledgements This work is funded by the U.K. Natural Environment Research Council (RAPID Directed Program) and the Tyndall Centre for Climate Change Research. I would like to thank James Annan, Peter Challenor, Neil Edwards, Michael Goldstein, David Sexton, and the Editor and Referees for their very helpful comments on two earlier versions of this paper.

References

Bernardo JM, Smith AFM (1994) Bayesian Theory. Chichester, UK: John Wiley & Sons

Box GEP, Tiao GC (1973) Bayesian Inference in Statistical Analysis. Reading, Massachusetts: Addison-Wesley

Craig PS, Goldstein M, Rougier JC, Seheult AH (2001) Bayesian forecasting for complex systems using computer simulators. Journal of the American Statistical Association 96:717–729

Craig PS, Goldstein M, Seheult AH, Smith JA (1998) Constructing partial prior specifications for models of complex physical systems. The Statistician 47:37–53. With discussion

Cressie NAC (1991) Statistics for Spatial Data. New York: John Wiley & Sons

Evans M, Swartz T (2000) Approximating Integrals via Monte Carlo and Deterministic Methods. Oxford: Oxford University Press

Garthwaite PH, Kadane JB, O'Hagan A (2005) Statistical methods for eliciting probability distributions. Journal of the American Statistical Association 100:680–701

Goldstein M, Rougier JC (2004) Probabilistic formulations for transferring inferences from mathematical models to physical systems. SIAM Journal on Scientific Computing 26(2):467–487

Goldstein M, Rougier JC (2006a) Bayes Linear calibrated prediction for complex systems. Journal of the American Statistical Association. Forthcoming, available at http://www.maths.dur.ac.uk/stats/people/jcr/BLCP.pdf

Goldstein M, Rougier JC (2006b) Reified Bayesian modelling and inference for physical systems. Journal of Statistical Planning and Inference. Accepted as a discussion paper subject to revisions, available at http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf

Grimmett GR, Stirzaker DR (2001) Probability and random processes. Oxford: Oxford University Press, 3rd edition

Higdon D, Kennedy MC, Cavendish J, Cafeo J, Ryne RD (2005) Combining field data and computer simulations for calibration and prediction. SIAM Journal on Scientific Computing 26(2):448–466

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. Journal of the Royal Statistical Society, Series B 63:425–464. With discussion

Lad F (1996) Operational Subjective Statistical Methods. New York: John Wiley & Sons

Mardia KV, Kent JT, Bibby JM (1979) Multivariate Analysis. London: Harcourt Brace & Co

Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M. Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430:768–772

Nakićenović NJ, editor (2000) IPCC Special Report on Emmissions Scenarios. Cambridge UK: Cambridge University Press

O'Hagan A, Forster J (2004) Bayesian Inference. Volume 2b of Kendall's Advanced Theory of Statistics. London: Edward Arnold, 2nd edition

Pukelsheim F (1994) The three sigma rule. The American Statistician 48:88–91

Robert CP, Casella G (1999) Monte Carlo Statistical Methods. New York: Springer

Smith AFM, Gelfand AE (1992) Bayesian statistics without tears: A sampling-resampling perspective. The American Statistician 46:84–88