Bayesian Prediction of Climate Using Ensembles of Simulator Evaluations

Jonathan Rougier*

Department of Mathematical Sciences

University of Durham, UK

Abstract

Computer simulators are an important source of information when we make predictive statements about future climate. If these predictive statements are to be probabilistic, it is necessary to construct a joint probability distribution over evaluations of the simulator and the climate itself. In other words, we must quantify the way in which the climate simulator is believed to be informative about the climate. This paper describes a 'minimal' probabilistic framework for linking these two things, which should be sufficiently general to accommodate a wide variety of computer-based climate predictions. It clarifies and extends recent developments in the statistics literature on prediction for physical systems using computer simulators, often known as *computer experiments*.

Keywords: Best Input, Calibrated Prediction, Calibration, Computer Experiment, Data Assimilation, History Matching, Emulator, Tuning

*Science Site, Stockton Road, Durham DH1 3LE, UK. Tel +44(0)191 334 3331, email J.C.Rougier@durham.ac.uk.

1 Introduction

The most recent Assessment Report of the Intergovernmental Panel on Climate Change (IPCC) contained a telling statement concerning uncertainty about future climate, among its proposals for future work:

> Improve methods to quantify uncertainties of climate projections and scenarios, including development and exploration of long-term ensemble simulations using complex models. The climate system is a coupled non-linear chaotic system, and therefore the long-term prediction of future climate states is not possible. Rather the focus must be upon the prediction of the probability distribution of the system's future possible states by the generation of ensembles of model solutions. Addressing adequately the statistical nature of climate is computationally intensive and requires the application of new methods of model diagnosis, but such statistical information is essential. (Moore et al., 2001, p. 771)

Writing as a statistician, I am happy to see that quantifying uncertainty is recognised as an essential aspect of climate projection, and that probability is identified as the appropriate measure of uncertainty. This is timely, as statisticians are actively developing inferential procedures for computer experiments (some references are given in 4.2.1). One point to note, however. The assertion that "prediction of future climate states is not possible" does not accord with statistical usage, since by 'prediction' we would typically mean the determination of a future probability distribution, as clarified in the next sentence of the quotation. In this context one would not "predict the probability distribution", but rather a probability distribution would be the prediction. Clearly statisticians and climate scientists should both be aware of this difference in interpretation of the word 'prediction'.

Most striking in the above quotation is the reason adduced for the impossibility of an exact prediction. This is attributed to the fact that the climate system is a coupled non-linear chaotic system which, in the logic of the sentence, is taken to be a sufficient condition for

predictive uncertainty. I would certainly not disagree with this, but I would question whether the authors are correct in drawing attention to this aspect as the primary cause of predictive uncertainty. If they are correct then our lack of knowledge of the initial value of the climate state-vector is the primary impediment to an exact prediction. I believe that most climate scientists would disagree, citing rather imperfections in the models used, both in terms of our incomplete understanding of climate physics, and the constraints imposed by current computing technology. Certainly the current emphasis seems to be on building better models, rather than performing more evaluations of existing models. The problem of an unknown initial value can be solved, in principle, by brute-force methods. But the question of how it is that evaluations of a very imperfect computer simulator can be informative about the climate is extremely subtle.

The outline of this paper is as follows. Section 2 clarifies the notion of prediction within a subjective probabilistic framework. Section 3 describes climate prediction without a simulator, introducing the likelihood function as the means of packaging climate data. Section 4 introduces the simulator, and describes a probabilistic model for expressing how the simulator is thought to be informative about the climate itself. Section 5 outlines two practical methods of inference about future climate, and describes the ways in which we can learn about climate using the climate data and the simulator. Section 6 presents a simple toy example, showing that inference in practice. Section 7 discusses some practical issues that are often raised by climate scientists, while section 8 concludes.

2 Subjective probability and prediction

When predicting future climate it is necessary to synthesise information from three different sources. First, there are historic and current measurements of the climate, and of quantities affected by the climate. Second, there are evaluations of the climate simulator at different settings. Finally, there are the climate experts. The end-result of the synthesis is primarily a better understanding of future climate, although it can also help to foster a better understanding of the simulator.

2.1 Subjective probability

There are a number of ways in which the above synthesis might be attempted. This paper focuses entirely on probability. To forestall any confusion it must be stressed at the outset that these probabilities *must* be based on the subjective assessment of experts. From the subjective point of view, the rules of the probability calculus are seen as necessary and sufficient conditions to avoid the possibility of being made a sure loser in a series of bets, a treatment associated with de Finetti (1972, 1974). The subjectivist point of view is usually termed *Bayesian*; Lindley (2000) gives a robust and non-technical defence of the Bayesian position, while Bernado and Smith (1994) provides a general reference.

The Bayesian view-point has its champions in the climate community, most prominently perhaps Stephen Schneider (see, e.g., Moss and Schneider, 2000; Schneider, 2001, 2002). It has also been popular in climate change attribution (see, e.g., Berliner et al., 2000). In my experience, though, there remains a large amount of confusion, not to say scepticism. The point which is often not fully grasped is that probability is used as a measure of *uncertainty*, and this entirely subsumes notions of 'randomness'. This can only be a good thing, as the no-

tion of randomness itself is very poorly defined, except perhaps at the quantum level. To this end, it has been advocated that we ought to interpret probabilities on future events within a 'many world' framework, in what appears to be an attempt to make climate prediction conform to the same inferential laws as tossing a coin. But the point is that this kind of sophistry is unnecessary. Anyone who can ask the question "What is the probability that the world will get 6°C hotter by 2100?" without needing to indulge in an ontological debate is already a Bayesian at heart.

The Achilles' Heel of the Bayesian approach is often taken to be our need to make an initial quantification of uncertainty. This occurs in what is usually referred to as the 'prior' Probability Density Function (PDF). Two troublesome situations are envisaged. The first is where our expert cannot choose between two (or more than two) prior PDFs as representations of her beliefs.[1] The second is where two (or more than two) experts disagree about their prior PDFs. These two situations are clearly different from each other, but at a practical level the response is the same: try them both, and give extra weight to those conclusions that appear to be invariant to the choice. In my opinion this is a great *strength* of the Bayesian approach: it allows us to show, formally and quite explicitly, how our conclusions depend on our assumptions. That this is a good idea for climate modelling assumptions is enshrined in the guiding philosophy of the IPCC, and in the many model inter-comparison experiments currently underway. We should demand nothing less of our inferential assumptions, and that is exactly what the Bayesian approach gives us.

In the context of climate prediction, the Bayesian approach implies

[1]Throughout the paper I adopt the convention that there is a single female expert.

that there is no such thing as *the* probability that world will get 6°C hotter by 2100. Instead, there are many probabilities, expressed as "the probability according to expert A (or B, C, ...) and conditional on data D (or D', D'', ...)". A consensus will arise if we can establish that certain probabilities are robust across experts and some kind of union of the datasets. In this case might we talk, with forgivable imprecision, about 'the probability'. One encouraging feature of the Bayesian paradigm is that with sufficient data we can show that even the most radically-opposed experts may be reconciled in their predictions.

2.2 Probabilistic prediction

Although the expert's beliefs are subjective, the procedure for a Bayesian analysis is completely unambiguous. A prediction is a probability distribution conditional on data. The data in this case comprise observations on quantities affected by climate, and evaluations of the simulator. The conditional or *predictive* distribution may be extracted by routine operations from the joint distribution over climate and data. *This joint distribution is the only subjective element.* Therefore whenever we encounter a probabilistic prediction for future climate the only issue open to debate among climate experts is "What joint distribution has been specified?" It is incumbent on anyone making such a prediction to ensure that the answer to this question is absolutely clear.

Now although it is possible to imagine a world in which climate experts deliver a joint distribution and the data to a statistician, and then come back a week later to collect a prediction for the year 2100, it is not the world we live in. I think most climate scientists would react with bemusement were they instructed to write down

$$\mathsf{Pr}\big(y, z, F\big)\,, \tag{1}$$

namely the joint PDF of the climate, observations on quantities affected by climate, and evaluations of a climate simulator. One key role for the statistician is to study the ways in which we can make this very complicated PDF understandable in terms of well-defined primitive quantities. Naturally we want to do this by introducing only the minimum of statistical modelling assumptions, so as to leave the expert with as much flexibility as possible. In a sense this is a meta-role for the statistician, as a sufficiently flexible statistical model should be applicable or generalisable across a wide range of inferential problems.

The approach taken in the following sections is to structure the PDF given above according to two sets of beliefs. First, broadly *qualitative* beliefs, that may be acceptable across a wide range of experts and inferential problems. These qualitative beliefs define the primitive objects about which the expert is required to supply a PDF. If these qualitative beliefs are acceptable to a particular expert in a particular problem then that expert has a free rein to *quantify* her beliefs, in terms of the specification of the PDF, as she sees appropriate. It is the qualitative beliefs that are the subject of this paper, and the primitive objects they give rise to. They are expressed in terms of *conditional independencies*, which are now widely used in statistics to construct inferential models. In this paper I will follow the presentation of Smith (1990).

3 Prediction without a simulator

The simplest case for prediction is where we do not use a simulator at all. In this case the predictive PDF of y is simply $\mathsf{Pr}\big(y \mid \bar{z}\big)$. Here z denotes the climate data—that is, observations on things that are affected by climate—prior to measurement, i.e. while they are still un-

certain quantities about which the expert has beliefs, while $\bar{z}$ denotes the observed values (which, hopefully, will be not-inconsistent with those beliefs). The predictive distribution for climate is the conditional distribution

$$\mathsf{Pr}\big(y \mid \bar{z}\big) = c\, \mathsf{L}_{\bar{z}}\big(y\big)\, \mathsf{Pr}\big(y\big) \tag{2}$$

where $c \triangleq \mathsf{Pr}\big(\bar{z}\big)^{-1}$, and $\mathsf{L}_{\bar{z}}\big(y\big) \triangleq \mathsf{Pr}\big(\bar{z} \mid y\big)$ is known as the *likelihood* function of the climate. Eq. (2) is often referred to as *Bayes's Theorem*.

The likelihood function describes the way in which different climates give rise to different climate-related observations. If we are interested in climate prediction using climate data we can partition y into (y_h, y_p), where the 'h' and 'p' subscripts corresponding to 'historic' and 'predicted'. 'Historic' is that period up to today, within which the data z have been collected, and 'predicted' is that period in the future for which we want a climate prediction. The expert is almost certain to agree that, in the language of Smith (1990, pp. 90–91), "given the information in y_h, the value of y_p is uninformative about the value of z, whatever the value of y_h might be". This is our first example of a *conditional independence* statement, written

$$z \perp\!\!\!\perp y_p \mid y_h \tag{3}$$

(this relationship is symmetric with respect to z and y_p). Note that this is a statement about the expert's beliefs about the climate data made *before* these data are observed. Where all three components in (3) are taken to be random vectors, this statement is equivalent to a statement about the factorisation of the conditional PDF, namely

$$z \perp\!\!\!\perp y_p \mid y_h \qquad \Longleftrightarrow \qquad \mathsf{Pr}\big(z \mid y_h, y_p\big) = \mathsf{Pr}\big(z \mid y_h\big)\,. \tag{4}$$

Accepting (4), the predictive PDF for y becomes

$$\mathsf{Pr}\big(y \mid \bar{z}\big) = c\,\mathsf{L}_{\bar{z}}\big(y_h\big)\,\mathsf{Pr}\big(y\big)\,. \tag{5}$$

In simple problems it is common to think of the climate data simply as observations made with error on elements of y_h. However, in climate studies a more subtle treatment is required. For example, where z is actually proxy data it makes no sense even to write $\bar{z} - y_h$. A very positive feature of the Bayesian approach is that it is not necessary to map the proxy data into the same space as y_h in order to make a direct comparison; in fact it is usually a mistake. For example, $\mathsf{L}_{\bar{z}}\big(y_h\big)$ might represent the response of plants to climate, as measured by pollen found in lake sediment cores; a probabilistic analysis of this type of data is given by Whiley et al. (2004). The important thing is that the likelihood is treated as a function of y, not of $\bar{z}$. Having said that, if there is a need to map proxy data directly onto climate variables, this task can be achieved quite naturally within the Bayesian approach outlined in this paper. This is discussed in section 5.

Where we have several different sets of proxy data, the expert will often believe that they are each conditionally independent given the climate itself. This would be the natural belief if each set of proxy data was collected by a different expert, using different equipment, from different locations. In this case the likelihood function has a simple product structure

$$\mathsf{L}_{\bar{z}}\big(y_h\big) = \mathsf{L}_{\bar{a}}(y_h)\,\mathsf{L}_{\bar{b}}(y_h)\cdots \tag{6}$$

where $\bar{a}$, $\bar{b}$, ... are different sets of proxy data; perhaps sedimentary pollen, tree rings, ice-cores, and so on. It would be very beneficial if

data experts were able to provide their likelihood functions to an agreed template in y, so that they could simply be bolted onto an inferential problem.

The other quantity the expert must specify is marginal beliefs about the climate itself, in the form $\mathsf{Pr}(y)$. In many cases the expert may be unwilling to specify beliefs about y, in which case she may be tempted to choose $\mathsf{Pr}(y)$ to be a uniform distribution on some finite region, so that it can be neglected in (5). This is almost certain to be an inappropriate reflection on the expert's actual beliefs about y, vague though they may be. The only justification would be that there is enough data in z for the likelihood to be highly concentrated, in which case all vague choices of $\mathsf{Pr}(y)$ will give about the same prediction. In the context of climate this is a very strong assertion that would almost certainly be falsifiable in applications.

There is a second problem with asserting that $\mathsf{Pr}(y)$ is uniform, which is a special case of the more general and clearly incorrect assertion that $y_h \perp\!\!\!\perp y_p$, i.e. beliefs about historic climate on their own are uninformative for predicting future climate. Were this the case it would be impossible to use the data z to predict future climate, because it follows from (4) that $z \perp\!\!\!\perp y_p$. *Therefore the expert who wants to make predictive statements about future climate using climate data is obliged to provide a PDF for y which* does not *have the property that $y_h \perp\!\!\!\perp y_p$.*

4 Introducing a simulator

Following on from the previous discussion, it should be clear that *the role of the climate simulator is to help the expert to express $\mathsf{Pr}(y)$.* Formally this follows immediately from a stronger version of (4): the

assertion that y_h is *sufficient* for z, or

$$z \perp\!\!\!\perp \text{everything else} \mid y_h. \tag{7}$$

In particular, we are going to introduce a climate simulator. In this context (7) states that given the information in y_h, both y_p and evaluations of the simulator are uninformative about the value of z, whatever the value of y_h is. This still seems to be a fairly uncontentious assertion, although it is possible to think of instances where it might not hold. For example, we might have calibrated one of our measuring instruments using evaluations of the simulator, or one like it. But I think it would be rare that the effect would be considered large enough to be worth addressing formally. Where (7) holds, the only place for our simulator in the predictive distribution for y is in the marginal distribution $\mathsf{Pr}\big(y\big)$.

At this stage we do not need to go into too much detail about the climate simulator itself. The simulator is effectively a deterministic function, with an input vector $x \in \mathcal{X}$ that contains quantities that are necessary to the evaluation of f, but about which the expert is uncertain. It can be anything from a compartmental model or an Energy Balance Model (EBM), to an Atmosphere Ocean General Circulation Model (AOGCM). Typical examples of uncertain inputs in the latter would be model parameters such as eddy viscosities, historic forcing such as atmospheric CO_2 concentrations, and the initial value of the state vector.

The output vector $f(x) \in \mathcal{Y}$ is a possible realisation of the climate. In the most general case a point in $\mathcal{Y}$ is a collection of space- and time-indexed state vectors. For an AOGCM each state vector comprises quantities such as velocities, temperature, pressure, and salinity or humidity. The output space $\mathcal{Y}$ also contains the true but unknown

climate value y. Invariably the simulator is believed by the expert to be an *imperfect* model for the climate, in the sense that she does not believe that there exists an input $\hat{x} \in \mathcal{X}$ such that $f(\hat{x}) = y$ exactly. The assumption that there exists such a 'right' input is sometimes referred to as a 'strong constraint' (see, e.g., Annan et al., 2004). Unfortunately this assumption is usually adopted, but there is no need, as I will show below.

If the simulator is to run forward from the present to make predictions about future climate then it must be subject to specified future forcing. Therefore we should think of f as a simulator for past and present climate, and for future climate in, say, the SRES-A1T scenario (Nakićenović, 2000; Schneider, 2002).

4.1 The simulator and the climate

When we introduce a climate simulator, the challenge for the expert is to describe, probabilistically, the way in which she believes that the simulator is informative about the climate itself. That is, she must specify, either directly or implicitly,

$$\mathsf{Pr}\big(F, y \mid X\big) \tag{8}$$

where $X \triangleq \big(x_1, \ldots, x_n\big)$ and $F \triangleq \big(f(x_1), \ldots, f(x_n)\big)$. Here X is any finite collection of simulator inputs, and F the resulting collection of simulator outputs. Some readers may be unfamiliar with the idea that the simulator output is in fact an uncertain quantity, but this means nothing more than the truism that we do not and cannot know $f(x)$ for all values of x. Even after our evaluations we will only know f on a set of measure zero in $\mathcal{X}$. Note that X will always appear to the right of the bar in conditional statements: the expert's beliefs about F are

always for specified inputs.

Now it seems to me to be very unlikely that a climate expert would be able to formulate a PDF for the collection $(F, y \mid X)$ directly. A common practice in this situation is to introduce extra uncertain quantities into the collection which can simplify the joint structure. Many physicists will be familiar with this approach, for example in general relativity. In probability calculations these extra quantities could then be integrated out, but in our case they are meaningful and so are retained. The primary new quantity to be introduced is a 'best input', x_0, the precise meaning of which is discussed in subsection 4.2. Once we admit the notion of a best input, we also have the notion of the *discrepancy*,

$$\epsilon \triangleq y - f(x_0), \tag{9}$$

i.e., the difference between the actual climate and the evaluation of the climate simulator at its best input. This is discussed in subsection 4.3. With these two extra quantities, the joint PDF is $\mathsf{Pr}\big(F, x_0, \epsilon, y \mid X\big)$, but it will be simpler below to use the equivalent form

$$\mathsf{Pr}\big(F, x_0, f_0, \epsilon \mid X\big) \tag{10}$$

where $f_0 \triangleq f(x_0)$, and $y \equiv f_0 + \epsilon$. Hopefully it will be possible to factorise this PDF in a way that makes it easier to compute the joint density as a product of conditional densities. This will depend on the properties that the expert is prepared to ascribe to x_0 and to ϵ.

4.2 How might we describe the 'best input'?

There seem to be two obvious ways in which we might identify a point $x_0 \in \mathcal{X}$ as being 'best'. First, by analogy with the situation in which

the simulator is perfect; and, second, by a simple optimality criterion.

4.2.1 Not a perfect simulator

Perfect simulators do not exist, but we may imagine them, if we refrain from being too pedantic. A perfect simulator has an input x^* with the property that

$$f(x^*) = y. \tag{11}$$

In other words we believe that somewhere in $\mathcal{X}$ there is an input x^* with the property that the simulator output when evaluated at x^* will exactly match the climate itself. For perfect simulators of physical systems like climate we may take x^* to be unique, because it is operationally defined with reference to the real world. So, for example, the appropriate value for the gravitational constant G in the perfect simulator is the value of G for which we could imagine an experimental determination quite independently of the simulator.

Obviously if we knew x^* we would only have to perform a single evaluation of the simulator, and it makes sense in this case to speak of the 'right' input value. In practice we will not know x^*, simply because we have not made the appropriate measurements (or not made them sufficiently accurately). For example, we might not have an exact measurement of the initial value for the state vector, or for the historic forcing. In this case the expert must provide a PDF for x^*, and it will be necessary to evaluate the simulator many times.

By extension, where the simulator is not perfect and the notion of the 'right' input is no longer tenable, the expert might still feel that *the simulator is sufficiently good that one evaluation at a unique 'best input' x_0 would be enough to predict y.* In this case, however, the outcome would not be exactly y, but would differ from y by some

unknown amount ϵ with $\mathsf{Var}(\epsilon) > \mathbf{0}$.

The belief that a single evaluation at x_0 would be enough to predict y is represented as

$$F \perp\!\!\!\perp y \mid f_0, x_0, X \tag{12}$$

remembering that $f_0 \triangleq f(x_0)$. In words, (12) asserts that were the best input value x_0 to be revealed, and were we then to evaluate the simulator at x_0 to compute f_0, then further evaluations over any finite collection of inputs X, which give rise to outputs F, would be uninformative about y. Of course in reality we never get to know x_0, and consequently we are obliged to perform many evaluations of the simulator, where each one may be thought of as a candidate for x_0. This is really no different from the case of the perfect simulator where we did not know x^*, with the exception (crucial in practice) that for a perfect simulator $\epsilon \equiv \mathbf{0}$.

The notion of a best input satisfying (12) was developed in a series of papers by Craig et al. (1996, 1997, 1998, 2001). In a sense this has been a formalisation of what Bayesian statisticians were already doing; see, e.g., Haylock and O'Hagan (1996), Kennedy and O'Hagan (2001) and Higdon et al. (2004), and the references therein. A related approach suitable for very quick simulators is Generalised Likelihood Uncertainty Estimation, known as GLUE (Beven and Binley, 1992); more information is available at `http://www.es.lancs.ac.uk/hfdg/glue.html`. But Craig et al. have been particularly concerned by the *meaning* of x_0, in order to facilitate the elicitation process for both x_0 and ϵ. The point is that x_0 is not just a statistical parameter, whose value we may select by, say, least squares. Rather, it is strongly related to measurable quantities, to the best inputs of other existing simulators, and to simulators yet to be built. It is true that at the moment the

expert's beliefs about x_0 for any given simulator tend to be fairly crude (e.g., independent intervals). But we can expect that as more and better climate simulators are built and analysed, the pool of common knowledge about good simulator inputs will grow, and we will want to borrow strength from that pool in our climate predictions, particularly when we want to combine results from two or more simulators. This is easiest if we have a consistent notion of x_0 that transcends, insofar as this is possible, any one particular simulator.

Naturally, there will be situations in which the expert will decide that the simulator is not good enough for there to exist an x_0 satisfying (12). Goldstein and Rougier (2004) discuss strategies for generalising the best input model, albeit expressed in a more restrictive form than is done here; this paper also considers extensions to more than one simulator.

4.2.2 An optimal input value

For any input x we can define the *bias* to be

$$\text{Bias}(x) \triangleq y - f(x) \tag{13}$$

i.e. the difference between the actual climate and the output of the simulator at any given x. The bias at x is an uncertain vector quantity, so it does not make sense simply to state that the input value x is better than x' if $\text{Bias}(x) < \text{Bias}(x')$, since this is a comparison of two uncertain vectors. A natural metric with which to make such a comparison is the expectation of the squared bias, or the *Mean Squared Bias*

$$\text{MSB}(x) \triangleq \mathsf{E}\big(\text{Bias}(x)^2\big) \tag{14}$$

where 'v^2' is used informally to denote the outer product of the vector quantity v. Now we can say that x_0 is the 'best input' if it has the property that, for any given value of x_0, the mean squared bias at x_0 is no greater than the mean squared bias at any other given x, or

$$\mathrm{MSB}(x_0 \mid x_0, x) \leq \mathrm{MSB}(x \mid x_0, x) \tag{15}$$

for all $x_0 \in \mathcal{X}$. Here $\mathrm{MSB}(\cdot)$ is a variance matrix (i.e. square, real-valued, non-negative definite, symmetric), and the statement that $A \leq B$ where both A and B are variance matrices implies that $B - A$ is also a variance matrix.

4.2.3 A sufficient condition

The way to think of (12) and (15) is as constraints on $\mathsf{Pr}\big(F, x_0, f_0, \epsilon \mid X\big)$ that give x_0 a desirable property, and so help the expert to understand how it is that x_0 is special, and, indeed, whether or not x_0 exists for a given problem. In general these constraints are hard to implement, and therefore it is natural to search for a sufficient condition that ensures that one or other, or both, of these two conditions hold automatically. I now show that that

$$f_0, F \perp\!\!\!\perp \epsilon \mid x_0, X \tag{16}$$

is sufficient for both conditions, that is

$$f_0, F \perp\!\!\!\perp \epsilon \mid x_0, X \quad \Longrightarrow \quad \begin{cases} F \perp\!\!\!\perp y \mid f_0, x_0, X \\ \mathrm{MSB}(x_0 \mid x_0, x) \leq \mathrm{MSB}(x \mid x_0, x). \end{cases} \tag{17}$$

The proofs use the simple rules of manipulating conditional independencies, as given, for example, in Smith (1990, p. 91), plus the property

of random vectors that, in the notation and numbering of Smith,

$$X \perp\!\!\!\perp Y \mid Z \Rightarrow g(X) \perp\!\!\!\perp Y \mid Z \tag{P4}$$

for any bounded function $g(\cdot)$.

To prove the first statement in (17), note that, starting with (16),

$$\begin{aligned} f_0, F \perp\!\!\!\perp \epsilon \mid x_0, X &\Rightarrow F \perp\!\!\!\perp \epsilon \mid f_0, x_0, X \\ &\Rightarrow F \perp\!\!\!\perp f_0 + \epsilon \mid f_0, x_0, X \\ &\equiv F \perp\!\!\!\perp y \mid f_0, x_0, X \end{aligned} \tag{18}$$

where the first line is Smith's P3, the second line is (P4) given above, and the final line is by definition.

To prove the second statement in (17), start with (16) and set $X = \{x\}$, from which we find

$$f_0, f(x) \perp\!\!\!\perp \epsilon \mid x_0, x \Rightarrow f_0 - f(x) \perp\!\!\!\perp \epsilon \mid x_0, x \tag{19}$$

by (P4) above. We can decompose the bias as

$$\text{Bias}(x) = y - f_0 + f_0 - f(x) \equiv \epsilon + \big(f_0 - f(x)\big). \tag{20}$$

Then (19) and (20) together imply that

$$\mathsf{Var}\big(\,\text{Bias}(x) \mid x_0, x\big) = \mathsf{Var}\big(\epsilon \mid x_0\big) + \mathsf{Var}\big(f_0 - f(x) \mid x_0, x\big)\,, \tag{21}$$

from which the second statement follows straightforwardly.

As it stands, (16) allows us to factorise the joint distribution (10) as

$$\mathsf{Pr}\big(F, x_0, f_0, \epsilon \mid X\big) = \mathsf{Pr}\big(f_0, F \mid x_0, X\big)\ \mathsf{Pr}\big(x_0, \epsilon\big) \tag{22}$$

where we have been able to remove ϵ from the conditioning statement in the first PDF on the righthand side. Rather than dwell on the meaning of (16) *per se*, I see it as providing an explanation of why the factorisation given in (22) ensures that x_0 is the best input, in both of the senses described above. In practice, the expert should ask herself whether she believes that there exists a best input for her simulator, and then, if satisfied, select (22) as an expedient way to achieve an appropriate probabilistic formulation.

Formulating the first of the two PDFs on the righthand side of (22) is still not an easy task, because of the potential for x_0 to play two roles in (f_0, F). First, x_0 is the input value for which f_0 is the output. But second, x_0 might itself be informative for the expert about F for given X. This may seem arcane, but it undoubtedly happens in practice. Consider, for example, the case of two inputs, x and x', where x is a 'realistic' input value, according to the expert's beliefs about x_0, and x' is not. Familiarity with the climate, and of evaluations of other simulators like f, will lead the expert to ascribe a distribution to $f(x)$ which is different from that of $f(x')$ not just because x is different from x', but because x is realistic while x' is not. The difficult question for the expert to resolve is whether she is prepared to ignore this information for the purposes of simplifying the structure of the inferential calculations.

This touches on a deep and difficult issue in inference: how much of our climate data do we treat explicitly, i.e. bundle into z, and how much do we treat implicitly, i.e. allow to influence our choice of $\mathsf{Pr}\big(F, x_0, f_0, \epsilon \mid X\big)$? If the expert wants to simplify the structure of $\mathsf{Pr}\big(f_0, F \mid x_0, X\big)$ by asserting that

$$F \perp\!\!\!\perp x_0 \mid X, \tag{23}$$

i.e. the value of x_0 is uninformative about the simulator outputs at any given collection of inputs, then she will need to put as much information into z as possible, and put all available evaluations of the simulator into (X, F). Currently no one, as far as I know, is contemplating jointly modelling F and x_0 for given X in climate. We may take this, for the time being, as strong circumstantial evidence for the acceptability of (23). In this case x_0 is just another input in the first PDF of (22), which is simply the joint distribution of simulator outputs at any given collection of $n + 1$ inputs. Now we may condition this distribution on the actual evaluations, $\bar{F}$, to give

$$\mathsf{Pr}\big(x_0, f_0, \epsilon \mid \bar{F}, X\big) = \mathsf{Pr}\big(f_0 \mid x_0, \bar{F}, X\big)\ \mathsf{Pr}\big(x_0, \epsilon\big)\,. \tag{24}$$

The first PDF on the righthand side is known as an *emulator*. An emulator is a statistical model for the simulator that is conditioned on the outcome of all available evaluations. In other words, an emulator can be represented as

$$\mathsf{Em}\big(v \mid x\big) \triangleq \mathsf{Pr}\big(f(x){=}v \mid x, \bar{F}, X\big)\,. \tag{25}$$

In the same way that the likelihood function is the container for all of the climate-related data, the emulator is the container for all of the evaluations of the simulator. Emulators are extensively discussed in the papers by Craig et al. and in Kennedy and O'Hagan (2001), which also gives a number of references for the practice of using gaussian random fields as prior distributions for unknown deterministic functions.

From (25) we can rewrite (24) as

$$\mathsf{Pr}\big(x_0, f_0, \epsilon \mid \bar{F}, X\big) = \mathsf{Em}\big(f_0 \mid x_0\big)\ \mathsf{Pr}\big(x_0, \epsilon\big)\,. \tag{26}$$

Eq. (26) is, in effect, a PDF for the climate itself, remembering that $y \equiv f_0 + \epsilon$. Therefore we have achieved the aim set out at the beginning of this section, namely to help the expert to specify $\mathsf{Pr}\big(y\big)$, and we have done it by building an emulator based on evaluations of the simulator, and by specifying the joint distribution $\mathsf{Pr}\big(\epsilon, x_0\big)$. But all of this has been predicated on certain beliefs of the expert, namely those given in (16) and (23), or, more heuristically, that there exists a best input and that little information is lost by treating the best input as uninformative about the behaviour of the simulator.

4.3 The role of the discrepancy

A natural way to consider the joint PDF of (x_0, ϵ) in (26) is in terms of the factorisation $\mathsf{Pr}\big(\epsilon \mid x_0\big)\ \mathsf{Pr}\big(x_0\big)$, i.e. the expert must provide a marginal distribution for the best input, and a conditional distribution for the discrepancy given the best input. The discrepancy PDF may be specified by the expert to be independent of x_0, which is certainly conceptually simpler but also loses some of the richness of the statistical model.

First, note that we may take $\mathsf{E}\big(\epsilon \mid x_0\big) = \mathbf{0}$ without loss of generality, because we can always modify the simulator itself such that any non-zero mean components of ϵ at a given x_0 are incorporated directly:

$$f(x) \leftarrow f(x) + \mathsf{E}\big(\epsilon \mid x_0{=}x\big)\,. \tag{27}$$

This is a formal statement of the informal practice of correcting the simulator for known offsets. Therefore we can consider the role of ϵ primarily in terms of $\mathsf{Var}\big(\epsilon \mid x_0\big)$.

This variance represents the 'distance' between the simulator and the climate itself, insofar as it is a chunk of uncertainty about the

climate that cannot be resolved by the simulator. Thus it allows us to formalise the notion that one simulator is better than another but, as might be expected for such a complicated concept, the answer is not always clear-cut. Each simulator has its own best input value, but these two best input values are related to each other because they are both rooted in the same physics. It is wrong, therefore, simply to assert that simulator A is better than B if

$$\inf_{x} \mathsf{Var}\big(\epsilon_A \mid x_{0A}{=}x\big) < \inf_{x'} \mathsf{Var}\big(\epsilon_B \mid x_{0B}{=}x'\big)\,, \tag{28}$$

either for all components of $\mathcal{Y}$ or for a designated subset, perhaps for the full variance matrix, as shown, or just down the diagonal. Some (x_{0A}, x_{0B}) combinations are far more likely than others, but we do not respect this if we allow ourselves to range over all possible values in search of the smallest variance in each case. One solution, but certainly not the only one, is to compare the variances at the mean of the joint PDF of (x_{0A}, x_{0B}). This discussion illustrates that although the best input model may be rather simple, it does not necessarily impose simplistic structure on issues that we know to be quite subtle.

There are two common situations in which it seems important to allow the variance of the discrepancy to vary with the best input. First, to account for unevenness in the simulator's performance. Typically the simulator is a discretised solver for an underlying set of ordinary or partial differential equations. Often the expert will have the view that the solver works better in some regions of $\mathcal{X}$ than in others. In this case it is natural to set $\mathsf{Var}\big(\epsilon \mid x_0\big)$ to be larger where the solver is less reliable, perhaps in certain 'corners' of $\mathcal{X}$, or perhaps in regions of $\mathcal{X}$ where the simulator is thought to have a highly non-linear response.

Second, it is quite a common belief in physical modelling that for a

given simulator a certain value x' will be a good input for one subset of y, but a different value x'' will be good for another. These subsets might be regions, or times, or components of the state vector. A belief like this may be thought to invalidate the notion of a best input, but this is mistaken. Instead, it should be interpreted as stating that $\mathsf{Var}\big(\epsilon \mid x_0\big)$ must depend on x_0. To give a very elementary example, suppose that f has a scalar input and two outputs, and the expert believes that $x = 3$ is a 'good' value for predicting y_1 and a 'bad' one for predicting y_2, while $x = 7$ is good for y_2 and bad for y_1. This suggests that the expert's beliefs might be represented as

$$\mathsf{Var}\big(\epsilon \mid x_0 = 3\big) = \begin{pmatrix} 1 & 0 \\ 0 & 9 \end{pmatrix} \quad \text{while} \quad \mathsf{Var}\big(\epsilon \mid x_0 = 7\big) = \begin{pmatrix} 9 & 0 \\ 0 & 1 \end{pmatrix}, \quad (29)$$

or something like it. This says precisely that $f(3)$ is thought to give a good prediction for y_1 but a bad one for y_2, while $f(7)$ is thought to give a good prediction for y_2 and a bad one for y_1.

I now turn briefly to the structure of $\mathsf{Var}\big(\epsilon \mid x_0\big)$ and, for simplicity, I will ignore the conditioning on x_0 as the following considerations should hold for all values of x_0. To understand $\mathsf{Var}\big(\epsilon\big)$ the expert needs to perform a thought experiment, and imagine that the value of x_0 has been revealed, and the simulator evaluated, so that she knows f_0. She must also imagine that she knows y, so that she can ask herself about the characteristics that ϵ (that is to say $y - f_0$) might possess. One of these characteristics would be the likely absolute size of ϵ, which gives us the diagonal of $\mathsf{Var}\big(\epsilon\big)$. As explained above, this part corresponds, broadly, to the notion of the simulator's quality.

The second characteristic is that *where the simulator is in error, it is often systematically so.* If, for example, the simulator has under-

represented sea surface temperature off the Azores for the last twenty years, then the expert might believe that there is a more-than-evens chance that this under-representation will continue into the future. Spatially, if the simulator tends to over-represent rainfall in northern France, the expert might believe 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 the expert believes that if the simulator over-represents temperature it contemporaneously (or with a lag) under-represents rainfall. These kinds of effects show up in the *off-diagonal* elements of $\mathsf{Var}\big(\epsilon\big)$.

Craig et al. (2001, p. 722) give an example of how these types of beliefs may be represented in practice. The authors are concerned with the discrepancy between a hydrocarbon reservoir simulator and the measured reservoir well pressures, taken at different wells and at different times. After a discussion with the reservoir engineers, and supported by data analysis on the output of a fast version of the simulator, they selected a discrepancy variance (not depending on x_0) of the general form

$$\mathsf{Cov}\big(\epsilon_{it}, \epsilon_{i't'}\big) = \sigma_1^2 \exp\big\{ -\theta_1 (t-t')^2 \big\} + \sigma_2^2 \, 1_{i=i'} \exp\big\{ -\theta_2 (t-t')^2 \big\} \tag{30}$$

where i represents a well location, and t time, and the parameters $\big\{\sigma_1, \sigma_2, \theta_1, \theta_2\big\}$ had explicit values assigned. In this specification there is a time effect, which says that discrepancies tend to extend through time, and a location effect, which says that discrepancies at the same well tend to be more closely related than discrepancies at different wells. A specification such as (30) 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

Algorithm 5.1 Generating an independent random sample of size k from the predictive distribution, using Rejection Sampling

Require: k, $\mathsf{L}_{\bar{z}}(y_h)$, $\mathsf{Em}(v \mid x)$, $\mathsf{Pr}(\epsilon \mid x_0)$, $\mathsf{Pr}(x_0)$
  $M \leftarrow \sup_{y_h} \mathsf{L}_{\bar{z}}(y_h)$
  $T \leftarrow \emptyset$
  for $i \in \{1, \ldots, k\}$ **do**
    repeat
      Sample $x_0 \sim \mathsf{Pr}(x_0)$
      Sample $f_0 \sim \mathsf{Em}(f_0 \mid x_0)$
      Sample $\epsilon \sim \mathsf{Pr}(\epsilon \mid x_0)$
      $y \leftarrow f_0 + \epsilon$
      Sample $u \sim \mathsf{Uniform}(0, 1)$
    until $uM \leq \mathsf{L}_{\bar{z}}(y_h)$
    $T \leftarrow T \cup \{(f_0, x_0, \epsilon)\}$
  end for
  Return T

adjust the variance parameters if necessary.

5 Learning about climate

5.1 Inferential calculations

Our predictive distribution, from (5) and (26) is, finally,

$$\mathsf{Pr}(x_0, f_0, \epsilon \mid \bar{z}, \bar{F}, X) = c\, \mathsf{L}_{\bar{z}}(y_h)\ \mathsf{Em}(f_0 \mid x_0)\ \mathsf{Pr}(\epsilon \mid x_0)\ \mathsf{Pr}(x_0) \quad (31)$$

where $c \triangleq \mathsf{Pr}(\bar{z} \mid \bar{F}, X)^{-1}$, and $y \equiv (y_h, y_p) \equiv f_0 + \epsilon$. A simple way to understand how we might use this specification in practice is to consider a method for independently sampling from the predictive distribution, namely *rejection sampling* (see, e.g., Ripley, 1987; Robert and Casella, 1999), outlined in Algorithm 5.1. This sampling method does not require us to compute the normalisation constant, c.

In this Algorithm, the first three lines inside the **repeat** loop sample $(x_0, f_0, \epsilon \mid \bar{F}, X)$. Each sample is a candidate value for the predictive distribution, but the probability of being selected depends on the value of the likelihood in a stochastic fashion. This is shown in the **until**

line, which uses a randomly generated quantity u which is uniform in the unit interval. With very fast simulators it may be possible to do without an emulator. In this case $f_0 = f(x_0)$ in the Algorithm, i.e. we are able to evaluate the simulator on demand. In probabilistic terms this is equivalent to letting X, the simulator evaluation points, become dense in $\mathcal{X}$, the input space.

If we are more interested in *summarising* the predictive distribution we can use *importance sampling*, which has the attractive feature that, unlike rejection sampling, we never discard a simulator evaluation, although we may down-weight it. The simplest version, illustrated for computing the mean of the predictive distribution of future climate, is

$$\begin{gathered} \hat{I}_k \triangleq \sum_{i=1}^{k} w_i \, y_p^{(i)} \qquad y^{(i)} \overset{\text{iid}}{\sim} \mathsf{Pr}\big(x_0, f_0, \epsilon \mid \bar{F}, X\big) \\ \text{where } w_i \propto \mathsf{L}_{\bar{z}}\Big(y_h^{(i)}\Big) \text{ and } \sum_{i=1}^{k} w_i = 1. \end{gathered} \tag{32}$$

Here $y^{(i)}$ is sampled exactly as in the first three lines of the **repeat** loop in Algorithm 5.1. It is straightforward to show that $\lim_{k\to\infty} \hat{I}_k = \mathsf{E}\big(y_p \mid \bar{z}, \bar{F}, X\big)$, by the Strong Law of Large Numbers (SLLN). This type of weighted average for the mean is superficially similar to some of the inferential calculations proposed in the climate literature, but with at least one crucial difference: (32) contains a discrepancy. As already discussed above, to treat the discrepancy ϵ as identically equal to zero is to assert that there exists a point in the simulator input space for which the simulator output exactly matches the climate. No credible expert could possibly believe this to be true, and consequently an inferential calculation without a discrepancy cannot be treated as well-informed.

Note that more sophisticated sampling and summary methods will

be necessary for anything other than very small problems; Robert and Casella (1999) provides a good general starting point, while Berliner (2001) presents a hybrid scheme motivated by *uncertainty analysis* for climate simulators. Uncertainty analysis is the study of how uncertainty about the simulator input feeds through to the simulator output (see, e.g., O'Hagan et al., 1999).

5.2 How we learn about future climate

In (31) we are simultaneously learning about the climate, the discrepancy and the best input, starting with expert beliefs about the climate, the climate data and the climate simulator, and using climate data and evaluations of the simulator. I like the name *calibrated prediction* (first used, to my knowledge, in the technical report that preceded Kennedy and O'Hagan, 2001) to describe this approach.

I now discuss the components of the predictive distribution in more detail.

5.2.1 The best input

Learning about the best input corresponds to what has been termed 'tuning', 'calibration' or, in the oil industry, 'history matching'. Perhaps the crucial difference is that the approach outlined here makes it clear exactly what the special input that lives in $\mathcal{X}$ actually is, most clearly by stressing that it is not necessary for this quantity to be the 'right' quantity in the sense of satisfying the strong constraint that $f(\hat{x}) = y$.

The predictive probability distribution for the best input is highly informative. First, it can be used diagnostically for the given simulator, and in this role we should favour a large number of inputs. For example, suppose we have an input which has a well-known physical analogue,

say g: gravitational acceleration at the earth's surface. It is tempting to set $g = 9.81$ m/s^2 in the simulator, and admit of no uncertainty. But, as often remarked (see, e.g., Smith, 2002), the two g's are not the same thing, and the expert might reasonably think that some small variation in the simulator g is acceptable within the imperfections of the simulator, if it leads to better prediction. By allowing the simulator g to be slightly uncertain, the expert allows the climate data $\bar{z}$ to modify simulator g if it can improve the correspondence between the simulator and the climate data. Obviously the limits of this modification are entirely controlled by the expert in $\mathsf{Pr}(x_0)$, and for g they might be very tight. From a diagnostic point of view, if the predictive distribution for simulator g is right at the limit of the range of values specified in $\mathsf{Pr}(x_0)$, then this ought to be highly informative about the way in which the simulator is operating. For example, how would the expert feel about the simulator if the simulator g wanted to be 8 m/s^2 rather than 9.81 m/s^2?

Second, learning about x_0 is a good way of transferring information across simulators, because we have specifically acknowledged that x_0 is more than just a statistical parameter for a given simulator. Many components of x will show up as inputs to other simulators: historic forcing for example. It is natural to take the predictive distribution of x_0 from one simulator and use it as the marginal distribution for another. In this way we can partly 'bootstrap' the expert's assessment of $\mathsf{Pr}(x_0)$, by accumulating information across simulators from what might initially be quite vague beliefs about x_0. In practice some care has to be taken in this process not to double-count the climate data.

As a more general point about the size of the input space, it stands to reason that if our statistical model linking the simulator and the climate is predicated on the existence of a point in the input space with

special characteristics, then this model has more chance of being appropriate the larger is the input space. Simply put, an expert who is not sure that the best input model holds for her particular problem should make the collection of inputs over which she expresses uncertainty as large as is feasible.

5.2.2 The discrepancy

The ability to learn about the discrepancy appears to be something entirely new in the field of climate prediction. Many of the same comments that apply to learning about the best input apply here too. Diagnostically the discrepancy allows us to identify subsets of the simulator output that are performing well, or badly, and this feeds straight back into the development of better simulators. In terms of transferring information across simulators, the 'offset' represented by the mean of the predictive distribution for the discrepancy could be used to correct other simulators in the same class, much like a flux correction.

5.2.3 The climate itself

Whether we sample or summarise the predictive distribution we get the whole of y, that is, both historic and future climate. It may not seem necessary to predict historic climate if we already have data, $\bar{z}$, but, first, there is a lot more climate than that for which we have direct measurements, and, second, as explained in section 3, $\bar{z}$ may well comprise observations on proxy data like tree-ring widths, in which case *eq.* (31) *provides a means of mapping proxy data into a probability distribution on the historic components of the climate state vector.* This is just a simple side-effect of the calibrated prediction calculation, but it eliminates the need for separate evaluations of the simulator in order to perform the mapping, and it ensures that the result of the mapping

is consistent with the other aspects of the prediction.

If we turn our attention to future climate, y_p, then (31) performs the task often referred to as 'data assimilation' or 'constraining the model to the data'. We can now see that we learn about future climate in three different ways: through learning about the best input, through learning about the simulator; and through learning about the discrepancy. In this decomposition it is possible to make a clear distinction between weather forecasting and long-range climate prediction. In the latter, we do not expect to be able to reduce the predictive variance of the discrepancy by very much, because the time-point at which we wish to make our prediction is far enough into the future that almost all of the temporal correlation between ϵ_h and ϵ_p will have decayed away. Thus although we can learn a lot about ϵ_h using the climate data $\bar{z}$, little of this reduction in variance will propagate all the way to ϵ_p. Therefore climate prediction is mostly about what can be achieved through learning about the best input and the simulator, and the variance of ϵ_p will typically present an irreducible lower bound on the predictive variance for y_p.

6 A simple example

In this section I present a simple example of calibrated prediction, with a scalar input and a scalar output, which allows a graphical display of the results. The situation in this case is one of 'improving' z, an imprecise measurement on y, by introducing physical information via the simulator and the expert's beliefs about the best input. But there would be no practical difference were f to have a 2-vector output, and the observation z on $f_1(x)$ was used to improve our prediction of $f_2(x)$. For simplicity I only treat the case where the simulator is fast enough

that an emulator is not required.

All the random samples from the predictive distribution are generated using Algorithm 5.1. The calculations used to create the figures in this paper are available from the author as a file to be executed in the R statistical computing environment (R Development Core Team, 2004).

My simulator is the function

$$f(x) = 3 + 8x - 5x^4 \qquad x \in \mathcal{X} = [0, 1]. \tag{33}$$

Being my own expert, I am happy to assert that this simulator is a sufficiently good model of the underlying physical system that condition (26) holds. Moreover, since I would like to keep things as simple as possible, I am also happy to assert that $\epsilon \perp\!\!\!\perp x_0$. My prior beliefs about the simulator and the system are then taken to be

$$x_0 \sim \mathsf{Beta}\big(2, 3\big) \tag{34a}$$

$$\epsilon \sim \mathsf{Gaussian}\big(0, 0.5^2\big) \tag{34b}$$

$$z \mid y \sim \mathsf{Gaussian}\big(y, 1^2\big) \tag{34c}$$

where (34c) is the likelihood function. There is nothing special about the choice of gaussian, as opposed to any other distribution, for ϵ and $z \mid y$. From (33), (34a) and (34b) and the simulator we can infer the joint distribution of (x_0, y). This is shown in Figure 1. At this point, i.e., before introducing the datum z, I have the opportunity to inspect the marginal distribution of y, in order to audit my choices for x_0 and ϵ. We may suppose that I am satisfied with the distribution shown in Figure 1, insofar as it concurs with my beliefs that, say, $y \approx 6 \pm 3$. If my beliefs had been more specific, for example if I had had a view

about the shape of $\mathsf{Pr}(y)$, then I would be able to scrutinise my choices in more detail, and alter them if necessary.

Now we introduce the datum, z. Figure 2 shows the impact on the predictive distribution of four different values of z. A simple maximum likelihood approach to inferring y from z would give a gaussian density for y centred on z with a standard deviation of 1. Compared to this, we observe from the vertical 'histograms' in Figure 2 that the inclusion of information from the simulator affects the mean, the standard deviation, and the shape of the predictive distribution for y. In particular, we can see how my beliefs about the simulator and the system have 'dragged up' a low value for z, and 'dragged down' a high one. The extra information available from the simulator can be quantified in terms of how much smaller the standard deviation of y is than the maximum likelihood value. For the four cases the predictive standard deviations of y are 0.71, 0.86, 0.73, and 0.55, so inclusion of the simulator has lead to sizable reductions in uncertainty over what could be achieved with the datum alone.

We also have revised beliefs about the best input x_0 and the discrepancy ϵ (I'll ignore the latter here). Marginal summaries of x_0 are shown on the horizontal 'histograms' in Figure 2, while the joint distributions for (x_0, y) are indicated by the contours. Once again all of the characteristics of the predictive distribution of the best input appear to be affected by the value of z. Interestingly, though not by design, the predictive distribution for (x_0, y) appears to be bi-modal. This is an example of how much information we miss if we only consider marginal distributions.

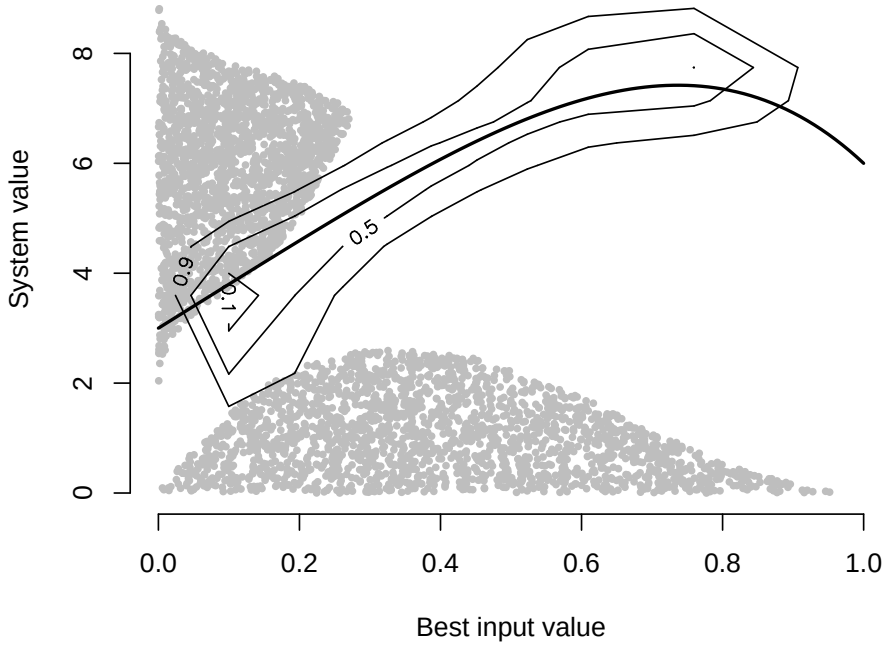

Figure 1: The joint and marginal distributions of x_0 and y. The joint distribution is shown as iso-probability contours delineating high-probability regions (0.9, 0.5 and 0.1), while the shapes of the marginal distributions are shown on the two axes. The thicker line shows the function.

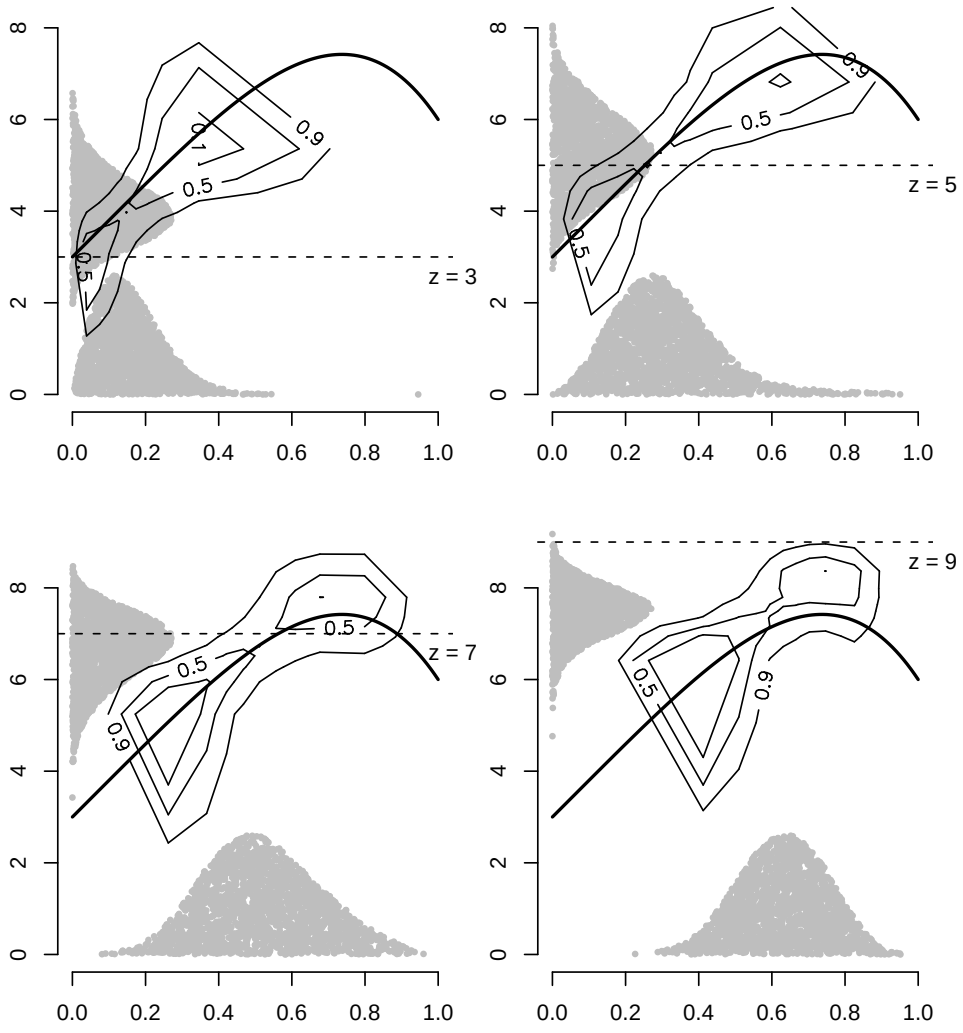

Figure 2: The joint and marginal distributions of x_0 and y conditional on four possible values for the datum z. See the caption of Figure 1 for details.

7 Discussion

This paper has not contained very much practical advice on how to do probabilistic inference for climate prediction, and, frankly, section 7 is not a good place to start. Practical issues, which tend to be more case-specific, are probably best dealt with in explicit studies, and we are developing these at the moment. I would, however, like to make a few brief comments that reflect the kinds of concerns that are raised by climate scientists, as seen from the statistical point of view.

7.1 Calibration via optimisation

There seems to be an accepted wisdom among physical modellers that the right way to calibrate a model, i.e. to choose the appropriate value for the simulator input, is to minimise the sum of squared deviations between the simulator outputs and the system data. The value so chosen can then be thought of as the 'right' input, and a point prediction of, say, future climate can be derived by running the simulator forward at this input value. This optimisation approach is attractive because it appears to be objective, in the sense that the least squares metric does not require any parameters, so that two research groups with the same model and the same data would reach the same conclusion without communicating. It is possible to generalise the least squares metric to introduce weighting, usually in the form of an inverted variance matrix. Often this variance matrix is taken to be diagonal, where the individual variances reflect 'natural variability' plus measurement error. Again, this is reassuringly objective once there is widespread agreement about how to compute the weights.

Of course the whole tenor of this paper is subjective, so it is hardly likely that I will find favour with such an approach. It strikes me as

strange that, after investing so much in the physics during the process of building a model, we should want to abandon the physics at the point where we calibrate the model. There is nothing 'physical' about the least squares criterion. In statistics the advantageous properties of estimators derived from the least squares criterion are restricted to a very small subset of inferential problems, certainly not one that contains the problem of assimilating system data into an imperfect and non-linear physical model. And then there is the problem of how well these optimisation methods work in practice: how do we know that we have found a global rather than a local maximum, and how do we account for the possibility that we have not? This has been referred to as the problem of *equifinality* (see, e.g., Aronica et al., 1998).

In the probabilistic approach to calibration outlined here we acknowledge the existence of a best input without requiring it to be the right input. Therefore the probabilistic approach is clearly applicable in all situations where it was previously thought that optimisation by least squares was a good idea, and many more besides. But rather than focus on a single evaluation at a single input, the predictions we make are averaged over candidate values for the best input, allowing for the possibility that there may be many good matches in different parts of the input space. The physics enters into the expert's choice for the marginal distribution of the best input, the variance of the discrepancy, and the likelihood function of the climate data. The second and third of these should also have played a part in choosing the weights in the least squares criterion, and so are not new although they may have been downplayed. But incorporating beliefs about the best input is a definite departure. With optimisation we have to impose hard boundaries on the input space, but with probabilistic calibration those boundaries can be 'softer', allowing us to down-weight but not exclude

regions of the input space in which we get a good match to the climate data but for which the physics is considered to be less likely.

7.2 Choice of simulator outputs

A climate simulator, especially an AOGCM, produces millions of outputs, and it is inconceivable that we could use them all in an inferential calculation. At the moment, I would suggest a realistic upper limit for the number of outputs that could be treated inferentially is a few hundred, maybe rising one or two orders of magnitude as we become more skilled and computers become faster. I have partitioned these outputs into two types: those that enter into the likelihood function, y_h, and those that we want to predict, y_p. As already discussed in subsection 5.2 , if we want to predict the near future then there is an advantage to having recent-past outputs in y_h, because using these we can learn about near-future values for the discrepancy. But if we want to predict the far future then we should choose the outputs in y_h to be informative about the best input, x_0, since our best hope of reducing the variance of our predictions is reducing our uncertainty about the best input to the climate simulator. Very often this will include outputs from the far-past, because these tend to present a clearer signal, being untainted by the effects of anthropogenic forcing. Usually it is advantageous to include several different sources of information, on the grounds that the joint likelihood constrains y_h, and thus x_0, rather better. Therefore in choosing y_h we might start by identifying a few well-separated sources of high-quality proxy data, plus some direct measurements on the state vector, and then identifying the smallest collection of outputs y_h that is sufficient for the resulting union.

For prediction we have more options, as we can scale the size of the inferential calculation according to the amount of detail we want

about y_p. For example, in the simplest case we would want just the marginal predictive distributions, with no information about the joint structure. In this case we can restrict y_p to a scalar value, and then we can repeat the inferential calculations once for each scalar about which we wish to learn. If we are satisfied with simple summary statistics of the joint predictive distribution like the predictive mean vector and the variance matrix, then we can restrict y_p to a pair of values, and repeat the inferential calculation for each pair of elements about which we wish to learn, and so on.

Our decision about how we treat y_p should be driven by the needs of the stakeholders, since we should envisage that predictive information about y_p goes forward into a decision analysis. In a full decision analysis we need the full joint predictive distribution, but for many purposes, including the more general task of raising awareness of future climate issues, more marginal predictive information will suffice, and y_p can be treated in subsets, making the inferential calculations far easier.

7.3 Physical parameters in the likelihood

We can envisage sources of proxy data for which historic values for the climate state vector are not sufficient on their own. For example, if we consider tree-ring data, these will be affected not just by the climate state vector, namely temperature, humidity and so on, but also by atmospheric information that enters the simulator as an input, namely CO_2 concentrations which may be forcing functions. From an inferential point of view, there is no problem at all in including this information in the likelihood function, which becomes

$$\mathsf{L}_{\bar{z}}\big(x_0, y_h\big) \triangleq \mathsf{Pr}\big(\bar{z} \mid x_0, y_h\big)\,, \tag{35}$$

but this raises a tricky foundational question. The quantity x_0 is defined to be the best simulator input. Crucially, it is not defined to be the 'correct' system value, even in those situations where there is a straightforward analogue between simulator input and system property. This has already been stressed in subsection 5.2 in the discussion about the gravitational constant g. Therefore it is not entirely clear that we want to use x_0 in the likelihood. Rather, we probably need to think more deeply about the extent to which simulator inputs and system properties are similar. I intend to treat this problem in more detail elsewhere, but I raise it here as something worth thinking about.

7.4 What about chaos?

'Chaos' is a recurrent theme in climate prediction, although from a practical point of view its impact seems to be felt primarily in terms of high sensitivity of the simulator outputs to the inputs in certain regions of the input space; Berliner (1992) provides an interesting discussion. It is natural to ask whether our best input approach can accommodate simulators that might display this kind of sensitivity, and the answer is, of course, affirmative.

The best input approach asserts the existence of a simulator input x_0 with special characteristics. It does not say anything about the behaviour of the simulator at inputs away from x_0, and therefore it contains no structure that would allow us to reject it explicitly for simulators for which $\|f(x + dx) - f(x)\|$ was large. That is not to say, however, that knowing that the simulator is very sensitive in certain regions of $\mathcal{X}$ is not important. On the contrary, it is a key part of how we formulate the emulator, defined in (25).

The emulator can be though of as a technology that allows us to predict the value of $f(x')$ from an observation on $f(x)$. Simulators that

are very smooth have $\|f(x) - f(x')\|$ small for x' in the neighbourhood of x, and our prediction of $f(x')$ using $f(x)$ is likely to have a small variance. But if f is thought to be very sensitive to the inputs in the region of x we will want to ensure that even when x' is close to x, our prediction of $f(x')$ using $f(x)$ has a large variance. In the extreme case we can model part of the variance of our emulator as a 'nugget', so that our beliefs about the simulator have the form

$$\mathsf{Cov}\big(f(x), f(x')\big) = \kappa(x, x') + \delta(x - x')\,\Sigma \tag{36}$$

where κ is some differentiable variance function, $\delta(\cdot)$ is the delta function, and Σ is the nugget variance; see, e.g., Bartlett (1978, ch. 5) for an introduction to the theory of random fields, while Cressie (1991) discusses statistical models with nuggets. Along the diagonal of Σ we may well have what is often termed the 'natural variability' of the simulator: this is the variability of $f(x)$ that can *only* be reduced by an evaluation at x itself.

In the inferential calculations, for example as given in Algorithm 5.1, the impact of this sensitivity of the simulator output to the input is found in sampling f_0: even where the sampled x_0 is close to a value that appears in X, say x_1, the sampled value for f_0 will not necessarily be close to that of $f(x_1)$. This seems perfectly right and proper. But of course we can only implement this behaviour in a controlled fashion if we have an emulator. Therefore inferential calculations that do not use emulators are at an immediate disadvantage where the underlying simulator is sensitive in this way. Although a common practice, it seems wrong, for example, to attempt to reflect this sensitivity in the likelihood function or the weights w_i in (32) because, as already explained in section 3, the likelihood function exists entirely indepen-

dently of the simulator, and so can in no way be affected by properties that the simulator is thought to have.

7.5 Advantages of an emulator

It is important to understand the benefits of having an emulator, because while emulators are easy to construct, good emulators are not, and it is important to understand the benefits that follow from investing in a good emulator (usually, by hiring a statistician to build one). A full discussion of inference with emulators would take up a whole paper in itself, and so I am not going to attempt it here, but I would just like to pick out two other advantages, following on from the previous subsection.

First, we seldom know the exact predictive distribution: we can only sample it, or estimate summary statistics like moments. Consider, for example, importance sampling for the predictive mean. As written in (32) we have n evaluations of the simulator, which go into the emulator, and k samples from the marginal distribution of $(x_0, f_0, \epsilon \mid \bar{F}, X)$. Our ability to make k as large as necessary, i.e. to have $k \gg n$, is crucial if we want our estimate $\hat{I}_k$ to have good properties, namely to be reasonably close to the true value $\mathsf{E}\big(y_p \mid \bar{z}, \bar{F}, X\big)$. Typically we can keep increasing k and monitoring the result until we are confident in the estimate. But if we choose to do without an emulator, then k and n are constrained to be the same value, and the cost of increasing k by 1, which was previously more-or-less zero, is now the cost of another simulator evaluation. So if our simulator is expensive to evaluate it is much less likely that we can have confidence in our estimate, because we do not have the freedom to take it toward its asymptotic limit.

Second, with an emulator comes the freedom to separate the location of the n evaluation points in $\mathcal{X}$ from that of the k sampled points.

Only the second collection of k points has to follow a random design, or, in the case of the more elaborate sampling procedures, a prescribed design. The first collection of n points can be anything we like, which affords us the opportunity to select the simulator evaluations that we perform according to optimality criteria. Obviously this is important where the number of evaluations that we can afford to perform is limited, and this will always be the case with AOGCMs, where we can expect to have only 10–100 evaluations in total (remembering that this is for a fixed future forcing scenario). Without an emulator we have only limited freedom to select the n evaluation points, as they must also be consistent with an integration rule over $\mathcal{X}$.

In a simple sequential design framework with an emulator, we can score a candidate evaluation point for x_{n+1} directly in terms of our prediction for y_p. We can use our emulator to predict the simulator output at x, our candidate value for x_{n+1}, and then perform the inference as though this were the actual outcome of the $(n+1)^{\text{th}}$ evaluation, to examine the impact of the choice x on measures such as the predictive variance of y_p. As we vary x we can identify good candidates for x_{n+1}, and quantify their impact. Knowing, at least approximately, this impact in advance allows us to terminate an experiment early on the basis that little further improvement in our prediction can be expected, or to make a strong case for further resources for more evaluations. The issue of optimal design for inference is discussed in Craig et al. (2001, section 8), and, for large problems, in Goldstein and Rougier (2003).

8 Conclusion

In this paper I have attempted to present a coherent view of how we might use climate simulators to make predictive statements about cli-

mate. The approach I have suggested here is not that which is currently adopted in climate prediction, where most of the effort appears to be directed at data collection and model building. My personal view is that creating the simulator is only half the battle. Understanding and quantifying how that simulator is informative about the actual climate is the other half, and is a prerequisite for calibrating the simulator using climate data, and predicting climate using the simulator. My natural inclination is to quantify in terms of probabilities, because I feel that probability is the right metric for uncertainty, and I think this is a view shared by many stakeholders.

The Bayesian approach outlined in this paper focuses expert knowledge on the specification of well-defined primitive quantities, namely the best input and the discrepancy, and relieves the expert of the broadly mechanical task of assimilating knowledge and data into climate predictions, which proceeds using standard tools such as those outlined in section 5. In this respect the Bayesian approach is exactly in harmony with the approach taken by climate scientists themselves. We would naturally be cautious of a statement about climate in 2100 that did not seem to accommodate the mechanical features of, say, the various conservation laws. The computer-based climate simulator is the means of imposing these conservation laws on scientists' beliefs about other more subjective aspects, such as subgrid-scale parameterisations or equations of state. In exactly the same way we should be cautious of probabilistic statements about the climate in 2100 that are not demonstrably conditional probabilities. We cannot simply compute some weighted average from the output of our ensemble of evaluations and call it a mean: it is only a mean if we have followed the rules of the probability calculus.

The statistical approach I have outlined in this paper, based on the

notion of a best input value, appears to me to be somewhat restrictive (as discussed in Goldstein and Rougier, 2004). But it is far more general than that implicit in current approaches. Therefore I think it makes a natural staging post between where we are now, and where we might want to be in, say, ten years time.

Acknowledgements

This research has been supported by the Tyndall Centre for Climate Change Research, grant ref. T2/13, and by the Natural Environment Research Council (NERC), grant ref. NER/T/S/2002/00987. I must acknowledge the crucial contribution of my colleagues at Durham, in particular Peter Craig, Michael Goldstein and Allan Seheult, while absolving them of all responsibility. I would also like to thank Myles Allen, Peter Challenor, Jim Hall, Marcel van Oijen, and David Sexton for their perceptive comments on earlier drafts.

References

Annan, J., J. Hargreaves, N. Edwards, and R. Marsh: 2004, 'Parameter Estimation in an Intermediate Complexity Earth System Model Using an Ensemble Kalman Filter'. *Ocean Modelling*. Forthcoming, preprint available from http://www.climate.unibe.ch/~edwards/jandj.pdf. 12

Aronica, G., B. Hankin, and K. Beven: 1998, 'Uncertainty and Equifinality in Calibrating Distributed Roughness Coefficients in a Flood Propagation Model with Limited Data'. *Advances in Water Resources* **22**, 349–365. 36

Bartlett, M.: 1978, *An Introduction to Stochastic Processes*. Cambridge UK: Cambridge University Press, 3rd edition. 40

Berliner, L.: 2001, 'Monte Carlo Based Ensemble Forecasting'. *Statistics and Computing* **11**, 269–275. 27

Berliner, L., R. Levine, and D. Shea: 2000, 'Bayesian climate change assessment'. *Journal of Climate* **13**, 3805–3820. 4

Berliner, L. M.: 1992, 'Statistics, Probability and Chaos'. *Statistical Science* **7**, 69–122. 39

Bernado, J. and A. Smith: 1994, *Bayesian Theory*. John Wiley & Sons. 4

Beven, K. and A. Binley: 1992, 'The Future of Distributed Models: Model Calibration and Uncertainty Prediction'. *Hydrological Processes* **6**, 279–298. 15

Craig, P., M. Goldstein, J. Rougier, and A. Seheult: 2001, 'Bayesian Forecasting for Complex Systems Using Computer Simulators'. *Journal of the American Statistical Association* **96**, 717–729. 15, 20, 24, 42

Craig, P., M. Goldstein, A. Seheult, and J. Smith: 1996, 'Bayes Linear Strategies for Matching Hydrocarbon Reservoir History'. In: J. Bernado, J. Berger, A. Dawid, and A. Smith (eds.): *Bayesian Statistics 5*. Oxford: Clarendon Press, pp. 69–95. 15

Craig, P., M. Goldstein, A. Seheult, and J. Smith: 1997, 'Pressure Matching for Hydrocarbon Reservoirs: A Case Study in the Use of Bayes Linear Strategies for Large Computer Experiments'. In: C. Gatsonis, J. Hodges, R. Kass, R. McCulloch, P. Rossi, and N. Singpurwalla (eds.): *Case Studies in Bayesian Statistics III*. New York: Springer-Verlag, pp. 37–87. With discussion. 15

Craig, P., M. Goldstein, A. Seheult, and J. Smith: 1998, 'Constructing Partial Prior Specifications for Models of Complex Physical Systems'. *The Statistician* **47**, 37–53. With discussion. 15

Cressie, N. A. C.: 1991, *Statistics for Spatial Data*. New York: John Wiley & Sons. 40

de Finetti, B.: 1972, *Probability, Induction and Statistics*. London: John Wiley & Sons. 4

de Finetti, B.: 1974, *Theory of Probability*. London: John Wiley & Sons. Two volumes (2nd vol. 1975); A.F.M. Smith and A. Machi (trs.). 4

Goldstein, M. and J. Rougier: 2003, 'Calibrated Bayesian Forecasting Using Large Computer Simulators'. Under submission, currently available at http://www.maths.dur.ac.uk/stats/physpred/papers/CalibratedBayesian.ps. 42

Goldstein, M. and J. Rougier: 2004, 'Probabilistic formulations for transferring inferences from mathematical models to physical systems'. *SIAM Journal on Scientific Computing*. Forthcoming. 16, 44

Haylock, R. and A. O'Hagan: 1996, 'On inference for outputs of computationally expensive algorithms with uncertainty on the inputs'. In: J. Bernado, J. Berger, A. Dawid, and A. Smith (eds.): *Bayesian Statistics 5*. Oxford, UK: Oxford University Press, pp. 629–637. 15

Higdon, D., M. Kennedy, J. Cavendish, J. Cafeo, and R. D. Ryne: 2004, 'Combining Field Data and Computer Simulations for Calibration and Prediction'. *SIAM Journal on Scientific Computing*. Forthcoming. 15

Kennedy, M. and A. O'Hagan: 2001, 'Bayesian calibration of computer models'. *Journal of the Royal Statistical Society, Series B* **63**, 425–464. With discussion. 15, 20, 27

Lindley, D.: 2000, 'The Philosophy of Statistics'. *The Statistician* **49**, 293–337. With discussion. 4

Moore, B., W. Gates, L. Mata, and A. Underdal: 2001, 'Advancing Our Understanding'. In: J. Houghton, Y. Ding, D. Griggs, M. Noguer, P. van de Linden, X. Dai, K. Maskell, and C. Johnson (eds.): *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge: Cambridge University Press, pp. 769–785. 2

Moss, R. and S. Schneider: 2000, 'Uncertainties in the IPCC TAR: Recommendations to Lead Authors for More Consistent Assessment and Reporting'. In: R. Pachauri, T. Taniguchi, and K. Tanaka (eds.): *Guidance Papers on the Cross Cutting Issues of the Third Assessment Report*. Geneva: World Meteorological Organisation, pp. 33–57. 4

Nakićenović, N. (ed.): 2000, *IPCC Special Report on Emmissions Scenarios*. Cambridge UK: Cambridge University Press. 12

O'Hagan, A., M. Kennedy, and J. Oakley: 1999, 'Uncertainty Analysis and Other Inference Tools for Complex Computer Codes'. In: J. Bernado, J. B. A. Dawid, and A. Smith (eds.): *Bayesian Statistics 6*. pp. 503–524, Oxford, UK: Clarendon Press. With discussion. 27

R Development Core Team: 2004, 'R: A language and environment for statistical computing'. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-00-3, http://www.R-project.org. 31

Ripley, B.: 1987, *Stochastic Simulation*. New York: John Wiley & Sons. 25

Robert, C. and G. Casella: 1999, *Monte Carlo Statistical Methods*. New York: Springer. 25, 27

Schneider, S.: 2001, 'What is 'Dangerous' Climate Change?'. *Nature* **411**, 17–19. 4

Schneider, S.: 2002, 'Can We Estimate the Likelihood of Climatic Changes at 2100?'. *Climate Change* **52**, 441–451. 4, 12

Smith, J.: 1990, 'Statistical Principles on Graphs'. In: R. Oliver and J. Smith (eds.): *Influence Diagrams, Belief Nets and Decision Analysis*. John Wiley & Sons, Ltd., Chapt. 5, pp. 89–120. With discussion. 7, 8, 17, 18

Smith, L.: 2002, 'What Might We Learn From Climate Forecasts?'. *Proceedings of the National Academy of Sciences* **99**, 2487–2492. 28

Whiley, M., J. Haslett, S. Bhattacharya, J. Allen, B. Huntley, and F. Mitchell: 2004, 'Bayesian Palæoclimate Reconstruction'. Currently under submission. Available from http://www.tcd.ie/Statistics/staff/personalfolders/johnhaslett/jrss.pdf. 9