Probabilistic Inference for Future Climate Using an Ensemble of Simulator Evaluations

Jonathan Rougier*

Department of Mathematical Sciences

University of Durham, UK

Abstract

Predictions about future climate are typically based on an ensemble of evaluations of a climate model at different parameterisations. If we wish to describe our uncertainty about future climate in terms of probabilities, then this constrains the way in which such an ensemble is generated and used. A key part of the process is a probabilistic description of the way in which evaluations of the climate model are informative about the climate system. Or, to put it more starkly, a probabilistic description of the model's inadequacy as a representation of climate. This paper describes the probabilistic approach, and makes a number of suggestions in order to simplify the challenging task of specifying this description of model inadequacy, to make the calculations more tractable, and to improve the quality of the resulting probability estimates.

Keywords: Model Inadequacy, Simulator Discrepancy, Calibration, Calibrated Prediction, Computer Experiment, Climate Sensitivity

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

1 Introduction

A simple question will help to motivate this paper:

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

This seems to be a well-posed question (subject to technical clarifications that need not concern us here), and certainly a topical one. It is the kind of question that a stakeholder might ask a climate scientist, on the basis that they believe it is the kind of question that climate scientists are funded to answer.

There are two aspects of this question that ought to be highlighted. First, the question asks explicitly for probabilities; second, it asks about the behaviour of the climate itself. So it is important to establish exactly what is meant by 'probability' in this context, and it is also important to understand that answers which focus on the climate sensitivity of this or that climate simulator are not sufficient. In order to satisfy the stakeholder, climate scientists must make a quantitative bridge from their particular climate simulator to the climate system, in order that their statements about climate sensitivity are relevant to the decisions that stakeholders need to take, and are directly comparable to those of other scientists. At the moment we have a form of beauty contest being played out in the journals, where different research groups produce different results depending on their climate simulators and their methods. Ultimately the winner of this contest in the judgement of the stakeholders will be the group that makes authoritative statements about the climate itself, using a transparent method with secure probabilistic foundations.

Any probability that might be quoted in answer to the question posed above is clearly a subjective assessment of uncertainty, as indeed are most probabilities when examined carefully. In this context it is not possible to

make a definitive statement of *the* probability, because a subjective assessment of probability will typically vary from person to person. But it *is* possible to make an authoritative and transparent statement, and this is what we should aim for. In the same way we cannot claim that the science that we do is not subjective, but we do strive to make it transparent, and where we make choices we try to make them wisely. One way to produce an authoritative statement about future climate is within a framework that is informed by the natural laws that are thought to govern climate. Typically the applicability of these laws is widely accepted across the climate community, and so this part of the inference can be considered to be less subjective. Thus the probability becomes a logical deduction that arises from a smaller and more primitive set of uncertainties via the laws of nature and the rules of the probability calculus.

This paper describes the correct way in which probabilities can be deduced in this way: there is no ambiguity since the rules of probability themselves do not admit of any uncertainty. This is addressed in section 2, which indicates the role played by ensembles of evaluations. The reason the paper does not stop there is because the task of specifying a probability distribution over the primitive quantities in all their generality is overwhelming. Section 3 discusses one way to diminish this problem, which is to introduce large amounts of climate data for the purposes of calibration. Section 4 discusses the key assertions that a climate scientist might make in order to simplify the process of specifying the joint distribution. Section 5 focuses on the most difficult quantity to specify: the probabilistic description of a climate simulator's inadequacy. Section 6 introduces further assertions that lead to relatively tractable inferential calculations. Section 7 discusses strategies for choosing the evaluation points in the ensemble. Section 8 concludes, and there are two short Appendices with slightly more technical material.

2 The climate system, and the climate simulator

2.1 The role of the climate model

The collection of quantities that comprise the climate system are denoted as the vector $y \in \mathcal{Y} \subset \mathbb{R}^k$. This collection includes historic climate and 'future' climate: in our case 'future' climate is climate in a future where atmospheric CO_2 doubles according to some prescribed schedule. Denote by Q the subset of $\mathcal{Y}$ for which 'future' global mean temperature increases by at least 2°C following the CO_2 doubling. If we could specify a distribution for y, then we could answer the initial question by adding up the probability assigned to the region Q.

However, it is hard to write down a convincing probability distribution for y. The components of y have complicated dependencies that follow from natural principles such as conservation laws and equations of state. These laws imply that y occupies a complicated manifold $\mathcal{Y}$ rather than the whole of $\mathbb{R}^k$. We can think of this manifold as being constrained by various numerical values which could be operationally defined: primitive physical constants such as gravitational acceleration or molecular viscosity, physical functions (of space and time) such as solar forcing or sources and sinks of atmospheric CO_2, boundary conditions such as the earth's topography, and the initial value of the climate state vector (a function of space). If we knew all of these values and denoted them by x^* then we could write

$$y = g(x^*), \tag{1}$$

i.e. the climate y is a point in $\mathcal{Y}$ that corresponds to the true values x^*; traditionally we might refer to the mapping as 'f' rather than 'g', but statisticians often use F for distribution functions and as this practice is followed below,

so 'g' is used here to avoid any confusion. Now even though x^* could be operationally-defined, its value would not necessarily be known to us with certainty, because the operation may not have been performed. For example, a volcano is a source of atmospheric ash, water vapour and CO_2, but for a given volcano we may have only an intermittent record of its emissions (or perhaps none at all).

Once we have $g(\cdot)$, we can use it map our uncertainty about x^* into uncertainty about y; i.e., we use $g(\cdot)$ to induce a probability distribution for y based on the probability distribution we choose for x^*. This is sometimes referred to as *uncertainty analysis* (O'Hagan et al., 1999). In order to answer the initial question when x^* is uncertain, we add up the probability assigned to those values of x^* for which $g(x^*) \in Q$, giving

$$\mathcal{P} \triangleq \int_x \mathbf{1}_Q\big(g(x)\big)\ dF_{x^*}(x) \tag{2}$$

where '$\triangleq$' denotes 'is defined as', F_{x^*} is our (cumulative) distribution function for x^*, and $\mathbf{1}_Q(y)$ is the *indicator function*, which takes the value 1 when $y \in Q$ and 0 otherwise. The notation $\int \cdots dF_{x^*}(x)$ indicates a *Lebesgue-Stieltjes integral* (see, e.g., Ross, 1988, ch. 7, sec. 9), which generalises the notion of expectation in order that x^* may describe both discrete and continuous quantities. Although this generalisation is necessary for technical reasons, the expression '$dF_{x^*}(x)$' can be thought of simply as a compact way of writing '$f_{x^*}(x)\,dx$' where we treat x^* as absolutely continuous and $f_{x^*}(x)$ is its probability density function evaluated at x.

Note that many other types of prediction can be computed within a similar framework, for example the mean and variance of global mean temperature in the year 2100 (or, more generally, the distribution of global mean temperature

in 2100), simply by replacing $\mathbf{1}_Q(\cdot)$ in (2) with a different function of $g(x)$.

2.2 An imperfect climate simulator

The key question is what happens when we replace the true natural principles with one particular approximation of them, based on an incomplete understanding or representation of the physics, and approximations in the solver. This approximation will be referred to as the *climate simulator*, where 'simulator' is used to denote the entity that combines the mathematical model, the simplifications made for tractability, the particular treatment of the model that makes it applicable to a given time and place, and the solver (Goldstein and Rougier, 2005b). The simulator is the computer code, and the vector denoted as x becomes those numbers that must be specified before the computer code will execute. Note the difference between x and x^*: x is a vector of numbers treated as input to a climate simulator, but x^* is a special value among all the possibilities for x that is, in a sense to be explored below, the 'best' value for x to make that particular simulator informative about actual climate.

There are three effects when nature is replaced with a particular simulator. First, the various components of x^* no longer have quite the same meaning as they did before. For example, the quantity labelled as 'viscosity' in one ocean simulator does not *necessarily* take the same value as 'viscosity' in another, and nor does it necessarily take the operational value of viscosity. Second, inadequacies in the simulator imply that there may be no point in the space of possible values for x^* for which (1) holds exactly. Third, the simulator may constrain the information that can be derived about y, for example through

the discretisations in the solver. The general relationship is now

$$y = g(x^*) + \epsilon^* \tag{3}$$

on the understanding that $g(\cdot)$ becomes the climate simulator, y is constrained to be those components of the climate that match the simulator's outputs, the meaning of the vector x^* is no longer completely clear, and $\epsilon^* \in \mathcal{E} \subset \mathbb{R}^k$, referred to as the *simulator discrepancy*, is not necessarily $\mathbf{0}$. But although the concepts are now murkier, the only modification to (2) is to allow for ϵ^*:

$$\mathcal{P} = \iint_{x \times \epsilon} \mathbf{1}_Q\big(g(x) + \epsilon\big)\ dF_{x^*,\epsilon^*}(x, \epsilon), \tag{4}$$

where F is respecified as a joint distribution function for (x^*, ϵ^*) and the integration has an extra k dimensions.

In similar terms to the interpretation of (2), the calculation in (4) adds up the probabilities assigned to those values of (x^*, ϵ^*) for which $g(x^*) + \epsilon^* \in Q$. Note that the effect of including ϵ^* in the calculation of $\mathcal{P}$ is ambiguous: there are some values for (x^*, ϵ^*) for which $g(x^*) \notin Q$ but $g(x^*)+\epsilon^* \in Q$, but likewise there are some values for which $g(x^*) \in Q$ but $g(x^*) + \epsilon^* \notin Q$. Therefore we cannot think of the simpler calculation given in (2) as any kind of bound on the more complicated calculation (4). The relation between the two calculations depends on exactly what beliefs are held about the relationship between the climate simulator and the climate system, as expressed in the distribution function F_{x^*,ϵ^*}.

2.3 Using climate data for calibration

Suppose now that there are observations on some of the components of y, possibly made with error. These observations are used to *calibrate* the simulator,

that is, to learn about (x^*, ϵ^*) and, in so doing, to improve our predictions for the climate behaviour. These observations can be written, in fairly general terms, as

$$z = Hy + e \tag{5}$$

where H is a known *incidence matrix* and e is an unknown measurement error. Sometimes a row of H will pick out an individual component of y, but usually, because y is constrained according to the available outputs from the simulator, the row of H corresponding to any given component of z will interpolate or average across a collection of components of y, in which case e must also account for approximation errors.

In probabilistic terms, the incorporation of information from observations of the climate system into our prediction corresponds to conditioning on the event that the uncertain quantity z takes its observed value $\tilde{z}$, known generally as *calibrated prediction* (Goldstein and Rougier, 2005a). *Bayes's theorem* is used to condition (x^*, ϵ^*) on the event $z = \tilde{z}$, which gives

$$\mathcal{P} = c \iint_{x \times \epsilon} \mathbf{1}_Q\big(g(x) + \epsilon\big) \ \mathrm{Lik}_{\tilde{z}}(x, \epsilon) \ dF_{x^*, \epsilon^*}(x, \epsilon) \tag{6a}$$

where $c \triangleq \Pr\left(z = \tilde{z}\right)^{-1}$, and

$$\begin{aligned} \mathrm{Lik}_{\tilde{z}}(x, \epsilon) &\triangleq \Pr\left(z = \tilde{z} \mid x^* = x, \epsilon^* = \epsilon\right) \\ &= \Pr\big(e = \tilde{z} - H\big(g(x) + \epsilon\big) \mid x^* = x, \epsilon^* = \epsilon\big) \end{aligned} \tag{6b}$$

where '|' denotes 'conditional upon'. Here $\mathrm{Lik}_{\tilde{z}}(\cdot, \cdot)$ is known as the *likelihood function*, and denotes the probability (density) of observing the data $\tilde{z}$ given particular candidate values for x^* and ϵ^*. Using (5), the likelihood can be expressed as the probability of observing the measurement error vector $\tilde{z} - H(g(x) + \epsilon)$. Comparing (4) and (6), the effect of introducing the climate data

is to introduce a weighting function into the integration which is proportional to the likelihood function. This means that candidate values (x, ϵ) which fit the data $\tilde{z}$ better, i.e. for which the measurement error vector is 'smaller', are accorded more weight in the result. A cruder technique is sometimes used, of sampling the space of candidate values for x^* (ϵ^* is usually ignored) and only keeping those for which $Hg(x)$ is sufficiently close to the data $\tilde{z}$. This would not be inconsistent with a probabilistic approach, but a likelihood function that only took the values 0 and 1 would be very unusual, and not easy to defend.

The purpose of ensembles of evaluations of our climate simulator is to approximate integrals such as (4) or (6). There are many ways to go about this, as discussed in books on high-dimensional numerical integration (see, e.g., Evans and Swartz, 2000). Section 7 considers particular strategies that might be useful for climate problems.

This section has shown how the probability calculus provides a transparent route from an initial assessment of our uncertainties and from relevant data, to probabilistic statements about the climate. In order to do this calculation in full generality we need to specify a joint distribution function for the collection (x^*, ϵ^*, e), since this is equivalent to specifying distributions for (x^*, ϵ^*) and for $e \mid (x^*, \epsilon^*)$, where the former distribution appears in the integrand of (6a) and the latter appears in the likelihood function given in (6b). The next sections consider these three uncertain quantities in more detail, and suggest strategies for making the specification of their joint distribution a little easier.

3 The quest for more data

The first question when faced with the challenge of specifying the prior distribution function for (x^*, ϵ^*, e), especially where there is confusion about the

precise meaning of x^* and ϵ^*, is to ask about whether there are circumstances in which it does not have to be carefully-specified. There is an important asymptotic result in Bayesian statistics that states that as the quantity of data grows, the likelihood becomes more concentrated, and so the prior becomes less influential in the resulting conditional distribution (see, e.g., Bernardo and Smith, 1994, sec. 5.3). This is especially the case where we are initially quite uncertain, because in this case it is reasonable to suppose that wherever the likelihood becomes concentrated, the prior at that point is fairly flat, and so may be taken as locally uniform.

Unfortunately this line of reasoning does not extend to our problem in its full generality, because the likelihood function itself requires us to specify the conditional distribution $\Pr(e \mid x^*, \epsilon^*)$, which involves (x^*, ϵ^*). But we can see right away that if the climate scientist was prepared to assert, as a statement of belief, that knowledge of (x^*, ϵ^*) was of no value in predicting e, then the likelihood function could be expressed in terms of the distribution of e alone. The effect in this case would be a large-data result like

$$\mathcal{P} \approx c \iint_{x \times \epsilon} \mathbf{1}_Q\big(g(x) + \epsilon\big)\ \mathrm{Lik}_{\tilde{z}}(x, \epsilon)\ dx\, d\epsilon$$

where now $\mathrm{Lik}_{\tilde{z}}(x, \epsilon) \triangleq \Pr(e = \tilde{z} - H(g(x) + \epsilon))$. A stronger result states that the logarithm of the likelihood function tends to a quadratic form, in which case $c \times \mathrm{Lik}_{\tilde{z}}(x, \epsilon)$ becomes gaussian, with a mean vector and variance matrix that can be inferred from a numerical optimisation over (x, ϵ). Now instead of specifying a probability distribution over the collection (x^*, ϵ^*, e) in order to compute $\mathcal{P}$, the climate scientist would simply have to specify a probability distribution for the measurement error e.

This seems very promising, but the catch is that these asymptotic results

only hold if z is augmented with new *independent* data. Simply adding on more and more observations of recent sea-surface temperature will not have a concentrating effect, because there is almost no information about (x^*, ϵ^*) in the 101^{st} sea-surface temperature that was not present in the first 100 observations. As a general principle, a few sources of well-differentiated data are worth more than many similar data: the worth in this case being measured in terms of reducing the influence of the prior for (x^*, ϵ^*) in the resulting prediction for y, by concentrating the likelihood. What tends to happen is that sea-surface temperatures constrain (x^*, ϵ^*) one way, and pressures constrain it another way. In the log-likelihood function these two constraints look like ridges, and these two ridges combine additively, so that at the point where they cross, the likelihood function becomes more concentrated.

Ideally we want many ridges in the log-likelihood, all crossing in roughly the same place. One of the ways in which this process breaks down even with a lot of different types of data is if the ridges all meet each other at different points. This tends to happen if the discrepancy ϵ^* is left out, or if the amount of measurement error is understated. In this case each set of data creates a very narrow ridge, because the data are wrongly treated as more informative about x^* than they actually are. The result is a very choppy likelihood surface which can be extremely difficult to characterise, and also very difficult to integrate over. Therefore setting $\epsilon^* = \mathbf{0}$ may seem to be very convenient but it is far from harmless: it can jeopardise our inference where the simulator is not a good representation of underlying climate at the scale of the available data.

We can extend this approach further, in our quest for new and independent sources of data. *Proxy data*, namely measurements on processes that are affected by climate but not themselves part of the climate state vector, could constrain (x^*, ϵ^*) in quite different ways to z, not least because some

are available over very long time-periods, from palæo-climate studies. Proxy data can be included into the inference by extending the simulator outputs with a further model mapping (x^*, ϵ^*) into measurements for, say, fossilised tree-ring thickness; recall that (x^*, ϵ^*) allows us to determine (x^*, y) using (3). It is definitely better to do this than to try and extract a set of measurements on the state vector from the proxy and then incorporate these into z. This is because the mapping from (x^*, ϵ^*) into fossilised tree-ring thickness is relatively straightforward if we have a biological model of tree-growth and a physical model of the fossilisation process. But going the other way, tree-ring data cannot be projected onto y alone, because these data also depend on components of x^* such as the cloud model (which affects precipitation) and atmospheric CO_2 concentrations. If we did try such a projection we would have to specify the measurement error distribution conditionally on (x^*, ϵ^*). This would be a very hard distribution to specify, except indirectly using probabilistic inversion to condition (x^*, ϵ^*) on the observed tree-ring data. But this is exactly what we are doing if we extend the simulator to include tree-rings among the outputs. Essentially, we are going to use Bayes's theorem to solve the inverse problem anyhow, so we can simply incorporate the tree ring data as part of the 'forward' problem. Therefore the case for including proxy data in the output of an extended climate simulator is fairly compelling in a probabilistic framework.

4 Simplifying the joint distribution

Generally-speaking, climate scientists are unlikely to have enough data in z to strongly concentrate the likelihood, and render their prior beliefs about (x^*, ϵ^*) immaterial in their inference about climate sensitivity. This is partly due to the poor quality of current climate simulators as representations of weather

(which is where the voluminous data are), but may also reflect chaotic features in the underlying mathematical model (see, e.g., Berliner, 1992; Smith, 2002). Where the simulator is acknowledged to be a poor representation for data on this scale, the ridges in the likelihood function that are induced by the data are wide, and even where they overlap they will not be highly concentrated. This means that a careful assessment of prior beliefs regarding (x^*, ϵ^*, e) will be unavoidable, even though as much data as possible ought to be used for calibration, as a form of insurance.

In thinking about the joint distribution of a collection of quantities the natural starting point is to ask whether they can be treated as mutually independent. If we assert that, say, x^* and e are independent, written as $x^* \perp\!\!\!\perp e$, then we are saying that our predictions for x^* are not affected by the value of e, and *vice versa* (see, e.g., Smith, 1990). Section 3 considered making the assertion

$$e \perp\!\!\!\perp (x^*, \epsilon^*) \tag{7}$$

which in a probabilistic treatment implies that $\Pr(e \mid x^*, \epsilon^*) = \Pr(e)$, allowing likelihood function to be simplified. Even this assertion is not uncontroversial. It implies that $e \perp\!\!\!\perp y$, since y is a deterministic function of (x^*, ϵ^*). This would immediately rule out multiplicative measurement errors, i.e. errors for which the uncertainty is expressed in proportional terms. In fact, these types of errors can often be incorporated by treating some of the simulator outputs in logarithms (transformations of the simulator outputs are discussed in more detail in section 6), or by generalising the measurement equation (5). However, this does not completely resolve the problem. When we compile a list of the ways in which a measurement error can be made we see immediately that many of them are weather-related (e.g. a seasick technician, atmospheric turbulence), and, therefore, climate related. But although (7) may not be a

strict description of beliefs, the climate scientist might well be comfortable with the idea that the impact of y on e is of secondary importance, and that the assertion should be accepted for the time being, in order to move on to more pressing issues.

Adopting this viewpoint, the prior distribution function factorises as $F_{x^*,\epsilon^*} \times F_e$. The question is, would the climate scientist feel as comfortable with the next step, which would be to assert that

$$x^* \perp\!\!\!\perp \epsilon^*. \tag{8}$$

This assertion has been widely adopted in the statistical literature on computer experiments (see, e.g. Craig et al., 2001; Kennedy and O'Hagan, 2001; Higdon et al., 2005). The meaning of (8) is quite straightforward. It asserts that there exists a simulator input x^* such that, were it to be known, then the climate scientist would be satisfied with the outcome of the single evaluation $g(x^*)$ for the purposes of predicting y, and this would be true no matter what the value of that evaluation turned out to be. This is clearly true in the case where $g(\cdot)$ is perfect and x^* comprises operationally-defined system values, in which case $\epsilon^* = \mathbf{0}$ so that (8) is automatically true. But it is an assertion of belief when $g(\cdot)$ is an imperfect simulator, and it actually serves to define what is meant by x^* in this context, although it does not operationalise it.

There are definitely situations which would cause a climate scientist to reject (8) as a correct statement of his or her beliefs. For example, suppose that x^* was revealed to be an extreme value. Two things might go wrong: the tractability simplifications might break down or the solver might break down, both leading to a simulator output $g(x^*)$ which was trusted less as a representation of the climate system than the output from a central value

for x^*. In this case the climate scientist might believe that the discrepancy could be larger for extreme values of x^*. Note that the problem here is with the simulator, not the underlying model, but the climate scientist should not ignore the fact that the simulator and the model are not the same thing.

Another situation which violates (8) is where the climate scientist believes that a certain value for x is good for predicting one subset of y and a different value, x' say, is good for predicting another subset: these two subsets are often differentiated by type, for example atmospheric pressure and ocean salinity. In this case $\mathrm{Var}\,(\epsilon^*)$ should be specified conditional on x^*, so that, say, $\mathrm{Var}\,(\epsilon_1^* \mid x^* = x) < \mathrm{Var}\,(\epsilon_2^* \mid x^* = x)$ and $\mathrm{Var}\,(\epsilon_1^* \mid x^* = x') > \mathrm{Var}\,(\epsilon_2^* \mid x^* = x')$, where the 1 and 2 subscripts indicate different subsets of y.

Both of these situations involve specifying the variance of the discrepancy conditionally on x^*. In fact this restricted type of dependency can be handled with a simple generalisation of (3), along the lines of

$$y = g(x^*) + \sigma(x^*)\,\epsilon^* \tag{9}$$

where $\sigma(x^*)$ is some specified function. Similarly, the mean of the discrepancy can depend on x^*, for example if the climate scientist believed that the simulator tended to under-predict certain climate properties for certain values of x^*. Providing that beliefs about the discrepancy are restricted to statements about the conditional mean and variance, it could be asserted that

$$x^* \perp\!\!\!\perp \epsilon^* \tag{10}$$

i.e. the joint probability distribution can be can factorised as $F_{x^*,\epsilon^*} = F_{x^*} \times F_{\epsilon^*}$. Then the climate scientist's task of specifying a joint distribution function for (x^*, ϵ^*, e) simplifies to specifying three marginal distribution functions plus,

possibly, additional functions such as $\sigma(\cdot)$ in (9). Another approach to simplifying the specification of the joint distribution function for (x^*, ϵ^*, e) is given at the end of section 5.

Interestingly, assertions such as (8) and (10) have recently been criticised from a foundational point of view, for being inconsistent with the notion that there can be many simulators of the same underlying system (Goldstein and Rougier, 2005a,b). In the second of these papers it is shown that the generalisations required to restore consistency involve analogues of both x^* and ϵ^*. These generalisations turn out to be beneficial from the point of view of the climate scientist, both because they clarify the nature of x^*, which is largely restored to its original operational definition, and because they diminish the role of ϵ^*, which is undoubtedly the most difficult quantity to describe probabilistically. These generalisations will not be discussed further here.

5 The discrepancy

The discrepancy seems to present difficulties to climate scientists who, although aware that their simulators are imperfect, have not, generally, been required to quantify the degree of imperfection. This has lead to a number of predictions for climate sensitivity which are "...conditional on the simulator being correct". Since we know that the simulator is incorrect, and we are given no indication regarding the degree of incorrectness by the experts themselves, this leaves the stakeholders in a difficult position, and certainly not one conducive to taking expensive and irreversible decisions. Climate scientists should be making predictions about the climate's sensitivity, not about their simulator's sensitivity, and it is surely the job of journal editors and stakeholders to see that this happens.

The most important feature of the discrepancy is that if we left it out

of a prediction based on a simulator that was known to be imperfect, then that prediction would be worse than if we included it. If we want to make the best prediction we can, then we ought to include a discrepancy, which means, in the simplest situation, formulating beliefs about ϵ^* in terms of the distribution function F_{ϵ^*}, or in terms of the parameters of that distribution such as the expectation vector and the variance matrix.

In terms of its impact on the inference, the discrepancy plays several roles. First of all, it smooths out the likelihood function, because an appropriate 'gap' between the system data and the simulator output makes the data less informative about the simulator inputs. As mentioned in section 3 this is actually a very important role, because without it we can end up in a situation where there appears to be no value for x^* which is consistent with all of the data. This would appear to suggest that our simulator was not very useful. For example, we would be in a situation where if we used less data we might get a well-fitting choice for x^*, but the value of this choice would move around depending on the data that we left out, and consequently our predictions for climate sensitivity would move around as well. We believe our simulators are useful, but we have to be realistic about how accurate they are, otherwise they will appear to be less useful than they actually are. Therefore effective calibration requires a discrepancy.

Second, the discrepancy variance provides a slot which can quantify the belief that one simulator is better than another with reference to their purpose, which is to learn about actual climate. Without this slot a better simulator can actually give a more uncertain prediction for climate. For example, one way to improve a simulator is to increase the number of inputs that are treated as uncertain. But without a discrepancy, i.e. setting $\epsilon^* = \mathbf{0}$ in (3), the extra uncertainty in x^* can feed through to extra uncertainty in $g(x^*)$, and in y.

The discrepancy can be used to compensate for this by decreasing uncertainty about ϵ^* on the basis that the less constrained are the values for x^*, the better the simulator can represent actual climate. While the actual quantification of relative quality across simulators may be difficult, there is likely to be a broad consensus on the ranking. So somewhere in the inferential calculation about actual climate there has to be a collection of numbers that should be bigger for simulator A than for simulator B, where it is generally agreed that simulator A is not as good as B. The discrepancy variance is a natural place to find these numbers.

Both of these reasons suggest that the variance matrix of the discrepancy should have a non-zero diagonal, i.e. it is not advisable to assert that ϵ^* takes the value $\mathbf{0}$ with probability 1. There is a third reason, though, particularly important in future climate prediction, that concerns the off-diagonal elements in $\mathrm{Var}(\epsilon^*)$. Climate scientists tend to believe 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 climate scientist 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 climate scientist 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 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 $\mathrm{Var}(\epsilon^*)$. They provide a way for the data $\tilde{z}$ to correct systematic errors that are believed to exist in the simulator.

Craig et al. (2001, p. 722) give an example of how these types of beliefs about systematic errors in the simulator 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 of the general form

$$\mathrm{Cov}\left(\epsilon^*_{it}, \epsilon^*_{i't'}\right) = \sigma_1^2 \exp\left\{-\theta_1 \left(t - t'\right)^2\right\} + \sigma_2^2 \, \delta_{ii'} \, \exp\left\{-\theta_2 \left(t - t'\right)^2\right\} \qquad (11)$$

where i represents 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 the *hyperparameters*, have 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 (11) 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.

Goldstein and Rougier (2005b, section 7) provide another route to specifying the off-diagonal structure of the variance of the discrepancy, by considering the extent to which simple relationships among the outputs of the simulator might reflect similar simple relationships in the climate itself. Again, this approach is parameterised by given values for a small collection of hyperparameters. Within a probabilistic framework it is always possible to learn about the values of the hyperparameters using the data $\tilde{z}$, but in practice this tends to be challenging unless the data are carefully chosen to be informative for variance learning (see, e.g., the discussion following Kennedy and O'Hagan,

2001).

Finally in this section we consider another way of thinking about the discrepancy, which provides us with a way to reparameterising (x^*, ϵ^*, e) that can sidestep the assertion that $x^* \perp\!\!\!\perp \epsilon^*$. Climate simulators typically solve the underlying model by stepping through time. We write $x^* \equiv (y_0, \theta)$, where y_0 is the initial value of the state vector and θ comprises uncertain model parameters and historic forcing functions, and $y \equiv (y_1, y_2, \dots)$. Then the simulator applied to a single timestep would be

$$y_t = g_t(y_{t-1}; \theta) + \omega_t \tag{12}$$

where ω_t is the single-step discrepancy that accounts for the fact that the simulator $g_t(\cdot\,; \theta)$ is not a perfect representation of the climate over the interval $(t-1, t]$. But for a simulator that makes a sequence of steps, the starting point of step t is not y_{t-1}, the actual climate at the end of interval $t-1$, but g_{t-1}, where g_t is defined recursively as

$$g_t \triangleq \begin{cases} g_t(g_{t-1}; \theta) & t = \dots, 2, 1 \\ y_0 & t = 0. \end{cases} \tag{13}$$

The additive discrepancy $\epsilon \equiv (\epsilon_1, \epsilon_2, \dots)$ is then

$$\epsilon_t = y_t - g_t(g_{t-1}; \theta). \tag{14}$$

Using linearisation and back-substitution this gives a stochastic approximation to ϵ in terms of θ and $\omega \equiv (\omega_1, \omega_2, \dots)$,

$$\epsilon_t \approx \nabla g_t(g_{t-1}; \theta)\, \epsilon_{t-1} + \omega_t \tag{15}$$

where $\nabla g_t(\cdot\,;\theta)$ is the Jacobian matrix, and $\epsilon_0 \triangleq \mathbf{0}$. In this formulation θ and ϵ^* are not probabilistically independent. The climate scientist could infer the joint distribution of $(y_0, \theta, \epsilon^*)$ from more primitive beliefs about the simulator, stated in terms of the Jacobian matrices (which could be described probabilistically) and the evolution of the stochastic process ω. Alternatively, and more straightforwardly, the inference could be reparameterised to replace ϵ^* with ω, although in this case ω would have to be thought of as an input into the 'all-timesteps' simulator $y = g(y_0, \theta, \omega)$, which has the discrepancy built in. This would involve a fundamental modification of the underlying computer code, but it serves to illustrate a general point. If inference about climate is our goal, then the structure of tools such as climate simulators ought to reflect the inferential calculation, not constrain it. In other words, perhaps climate scientists should consult statisticians when designing climate simulators.

6 Gaussian assertions for tractability

Any calculation designed to approximate an integral such as (6) is going to have to span a $(p+k)$-dimensional space, where p is the number of uncertain quantities in x^* and k is the number of uncertain quantities in ϵ^*. Here k has to be at least as big as the number of components in the observations z. If z is large—in order to reduce the impact of prior beliefs about (x^*, ϵ^*)—then k will be large. So anything that reduces the size of the integral in (6) from $(p+k)$ dimensions to p dimensions is going to make a big difference to the computability of $\mathcal{P}$. One choice that allows us to do exactly that is to treat $(\epsilon^*, e) \mid x^*$ as gaussian. Effectively, in this case ϵ^* may be integrated out of (6) analytically.

For simplicity suppose that x^*, ϵ^* and e can be treated as mutually independent, and (ϵ^*, e) as gaussian; the approach generalises straightforwardly

to the case where the three quantities are not independent. In this case the distribution of $(y, z) \mid x^*$ is gaussian, and

$$\mathcal{P} = c \int_x \mathrm{E}\left(\mathbf{1}_Q(y) \mid z = \tilde{z}, x^* = x\right) \mathrm{Lik}_{\tilde{z}}(x)\ dF_{x^*}(x) \tag{16}$$

where $\mathrm{Lik}_{\tilde{z}}(x) \triangleq \Pr\left(z = \tilde{z} \mid x^* = x\right)$. Here both $y \mid (z, x^*)$ and $z \mid x^*$ are gaussian; the forms of the terms are given in Appendix A.1. Generally, the integrand in (16) takes almost no time to compute beyond that taken to evaluate the simulator to find $g(x)$.

Some climate simulators can be time-consuming to evaluate, and in this case the inference can be generalised to include an *emulator* of the simulator, which is a probabilistic framework for predicting the simulator output at any given value x based on the outcomes of a carefully-chosen set of evaluations. Emulator construction is quite subtle but the principles are well-established: Santner et al. (2003, ch. 3-4) provides a review; Currin et al. (1991) and Kennedy and O'Hagan (2001) describe a relatively simple Bayesian approach using a gaussian process as a prior; Craig et al. (1997, 2001) describe a more general approach using a greater amount of expert knowledge and evaluations of cut-down versions of the full simulator. There is at least one example of a simple emulator in the climate literature (Murphy et al., 2004; Rougier, 2004).

The assertion that both ϵ^* and e are gaussian may seem unrealistic in the case where the components of y are constrained by their interpretation to respect certain limits, e.g. to be strictly positive. In these cases the outputs of the simulator can sometimes be transformed so that they are unbounded. As a general point, numerical approximations to integrals such as (6) or (16) work better where the integrands are low-order in x, and it is often beneficial to use, say, logarithmic transformations of strictly positive components of $g(\cdot)$

that might otherwise be squashed up against the origin for large parts of the input space.

7 Design issues

Finally we turn to the practical issues of computing an approximate value for $\mathcal{P}$. We consider the tractable special case of the previous section, i.e. x^*, ϵ^* and e are mutually independent, and (ϵ^*, e) is gaussian. However, the approach outlined below is perfectly general.

Our objective is to evaluate (16). At this point we have to confront the size of x, denoted p. For anything other than a trivial climate simulator, p is almost certainly larger than can be managed with a simple product integration rule such as gaussian quadrature. This is because x includes not only the relatively small collection of uncertain coefficients in the underlying mathematical model, but also the much larger collection of other uncertain numerical values in the code, most notably forcing values with spatial and temporal indices, and the initial value of the state vector.

Denote by x_1 the uncertain model coefficients and by x_2 all of the remaining uncertain values, so that $x \equiv (x_1, x_2)$. Typically, the value of x_1 is much more important than x_2 in determining the general behaviour of $g(x_1, x_2)$. For example, one component of x_2 might be the quantity of particulate matter ejected into the atmosphere in the region containing Mt. Pinatubo in 1991, when the volcano erupted. No-one knows exactly how much matter was ejected. Even if someone did, it is not clear that this would be the best value to use in the simulator, taking account of deficiencies in the modelling of the impact of atmospheric particulate matter in terms of scattering solar radiation and seeding clouds. Therefore the best value for this quantity in the simulator should be treated as uncertain. The impact of this quantity propagates forwards in time

in the simulator, but it cannot propagate backwards so there are simulator outputs which are completely unaffected. The coefficients in x_1 together affect every single output of the simulator, but this does not mean we should concentrate on x_1 and ignore the contribution of uncertainty in x_2. Climate simulators are well-known to exhibit strong non-linearities arising from positive feedback and hysteresis. It would be very misleading to treat some or all of the components of x_2 as fixed, if quite minor changes in the value of x_2 might lead to large changes in $g(x_1, x_2)$. However, this is what is happening at present.

A general rule for numerical integration is that the more knowledge that can be incorporated about the integrand, the more accurate will be the result for a fixed number of evaluations. In our case the climate scientist should attempt to incorporate the knowledge that x_1 has a bigger effect on the simulator output than x_2. Although there are several methods that could be used, for climate simulators a hybrid of deterministic and stochastic methods could be useful, which might be termed *quadrature with stochastic forcing*. The basic idea is to evaluate the simulator over a carefully-chosen collection of candidate values for x_1^*, where each evaluation is made with one or more randomly-sampled candidates for x_2^*.

For simplicity, suppose that the climate scientist is comfortable with the idea that x_1^* and x_2^* are probabilistically independent, so that $F_{x^*} = F_{x_1^*} \times F_{x_2^*}$. Then we can write, starting from (16),

$$\begin{aligned} \mathcal{P} &\equiv c \int_{x_1} J(x_1)\, dF_{x_1^*}(x_1) \\ &\approx \hat{c} \sum_{i=1}^{m} w^{(i)} \hat{J}\big(x_1^{(i)}\big)\, dF_{x_1^*}\big(x_1^{(i)}\big) \end{aligned} \tag{17a}$$

where

$$J(x_1) \triangleq \int_{x_2} \mathrm{E}\left(\mathbf{1}_Q(y) \mid z = \tilde{z}, x^* = (x_1, x_2)\right) \mathrm{Lik}_{\tilde{z}}(x_1, x_2)\, dF_{x_2^*}(x_2) \qquad (17b)$$

and $x_1^{(1)}, \ldots, x_1^{(m)}$ and $w^{(1)}, \ldots, w^{(m)}$ are chosen according to some integration rule over x_1; the hats over c and $J(\cdot)$ denote numerical approximations. To compute $\hat{c}$ and $\hat{J}(\cdot)$ we have to evaluate the simulator, for which we require an x_1 and an x_2 value. The simplest way to proceed (not the best) is to generate a value for x_2 randomly from $F_{x_2^*}$ for each $x_1^{(i)}$. In other words, each simulator input comprises a carefully-chosen value for the uncertain model parameters and a randomly-chosen value for the uncertain other (less important) simulator inputs. Appendix A.2 gives a more general algorithm.

One thing to note is that there is no reason for the integration rule values $x_1^{(1)}, \ldots, x_1^{(m)}$ to be an equally-spaced grid across the individual components of x_1. In fact, gaussian quadrature is superior to Simpson's rule for the same m precisely because the abscissae are *not* equally spaced. This contrasts with current practice (see, e.g., Stainforth et al., 2005). Neither is there any reason for the number of different levels to be the same for each component of x_1 (another feature of Stainforth et al.). If the response of the simulator to the first component was thought to be much more important than the response to the second component, then it would make sense, in terms of deriving a better approximation for $\mathcal{P}$, not to have the same number of levels in both components, but to have one more level in the first and one less in the second (reducing m by 1 in the process). This is another simple example of the way in which the climate scientist's knowledge can be used to improve the calculation of $\mathcal{P}$.

An interesting question arises about whether it is useful to build *stochastic*

climate simulators, i.e. climate simulators that are a function of x_1 alone, with x_2 being chosen at random from some specified distribution (which we would take to be $F_{x_2^*}$). In the context of probabilistic inference centred on $\mathcal{P}$ the answer is clearly negative. This is because the quality of our approximation for $\mathcal{P}$ for a given budget of simulator evaluations can be improved, sometimes dramatically, by variance reduction techniques (references are given in Appendix A.2). To give one example, we might be able to afford two realisations of x_2 for each $x_1^{(i)}$. We could choose to generate two independent random realisations, but it would be better to generate a pair of realisations which were *antithetic* (crudely, negatively correlated). The use of antithetic random variables will improve the quality of our approximation $\hat{J}\big(x_1^{(i)}\big)$. A stochastic simulator in which the random value for x_2 is buried inside the computer code would prevent us from doing this.

One further comment derived from a common misconception in the literature. Choosing a uniform distribution for x_1^*, i.e. setting $dF_{x_1^*}(x_1) \propto \mathbf{1}_{X_1}(x_1)$ for some finite region X_1, is seldom appropriate for parameters in a physical model. For example, do Murphy et al. (2004, see Supplementary Table 2) really believe that, say, all values of the entrainment rate coefficient between 0.6 and 9 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 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. There is absolutely no sense in which choosing a uniform distribution is 'objective', since it requires us to specify limits (except in the case of an improper uniform prior on the whole of the real line, sometimes used in a Bayesian *reference analysis*). Furthermore, there is no sense in which a uniform distribution is especially parsimonious,

since there are other two-parameter distributions, both symmetric (triangular) and asymmetric (gamma), that might have been chosen instead. And there is no compelling reason to select the distribution for $F_{x_1^*}$ on the basis of parsimony anyway.

Any inferential approach which weights members of an ensemble equally, or only with reference to the likelihood, is making the same implicit assertion, and ignoring the widely-held view that central values of the model parameters are probably better candidates than extreme ones. Since the value for $\mathcal{P}$ can only be interpreted as a subjective assessment based on our knowledge, climate scientists should make the best possible use of that knowledge, and definitely not make simplistic assertions that they do not believe and that cannot be defended. Exactly the same situation prevails with regard to the probabilities attached to various future climate scenarios (see, e.g., Moss and Schneider, 2000; Schneider, 2001).

8 Conclusion

The fundamental message of this paper is that making inferences about future climate using an imperfect climate simulator is a very challenging business. If those inferences are required to be probabilistic, then the challenge is to specify a joint distribution function for the collection of uncertain quantities (x^*, ϵ^*, e), where x^* is the 'best' setting for the simulator inputs, ϵ^* is the discrepancy between the 'best' simulator output and the climate system, and e is measurement error on climate data used for calibration (see section 2). The easiest way for the climate scientist to proceed with the inference is to treat the three uncertain quantities as mutually independent, so that the joint distribution function factorises into the product $F_{x^*} \times F_{\epsilon^*} \times F_e$. This has to be an assertion of belief on the part of the climate scientist, and it must involve

a certain amount of pragmatism because, as has been discussed, none of these independencies is completely defensible. That having been said, the climate literature has yet to produce *any* analysis that provides a transparent and defensible quantification of simulator inadequacy which is then incorporated into a climate prediction. So although climate scientists may baulk at the assertion that, for example, x^* and ϵ^* are independent, this is surely a lesser concern than the 'default' assertion that $\epsilon^* = \mathbf{0}$. To give an analogy, we may not understand the precise mechanism by which the ocean and the atmosphere ought to be coupled in a climate simulator, but this does not mean that we set the coupling to zero. It is the same with the discrepancy distribution: we may not know it but we ought to assess it as best we can.

This paper has made a number of other observations about the use of an ensemble of evaluations of a climate simulator. In particular, the purpose of the ensemble is to approximate a high-dimensional integral and, consequently, much guidance regarding the design of the ensemble can be derived from the very large body of literature concerning numerical integration. One suggestion for handling the high dimension using a combination of deterministic and stochastic integration approaches was given in section 7. This section also suggested that a uniform distribution for x^*, although a popular choice, was seldom defensible in the case where components of x^* represented uncertain physical parameters without very strict bounds. Section 3 recommended extending climate simulators to include proxy data among their outputs, in an attempt to lessen the contribution of the climate scientist's probabilistic assessment of (x^*, ϵ^*) in the inference.

In the immediate future, however, during a time when we really need to address questions such as the one posed at the start of the paper, we have to acknowledge that any answer will depend critically on the probabilistic

assessment that the climate scientist makes, and that there can be no 'gold-standard' by which those assessments can be judged. But the sooner that the community of climate scientists confronts this issue, the sooner a consensus might emerge about crucial parameters such as $\mathrm{Var}\,(\epsilon^*)$.

Statisticians can help in this process in facilitating the direct elicitation of a distribution function for (x^*, ϵ^*), by deriving efficient frameworks for structuring the distributional parameters and estimating the hyperparameters, and also by thinking about more general ways in which inference about a physical system can proceed from an ensemble of simulator evaluations, as discussed and illustrated in Goldstein and Rougier (2005b). They also have access to a large literature and a great deal of practical experience in diagnosing conflicts between the prior distribution and the climate data, performing a sensitivity analysis with respect to key parameters in the prior distribution, software testing to validate the performance of the climate simulator, and efficiently implementing large-scale inferential calculations to get the best possible estimator for $\mathcal{P}$ for a given budget of simulator evaluations. In all of these tasks a much greater benefit is derived where statisticians are able to work closely with climate scientists, because a large part of Statistics is about turning expert knowledge to one's advantage.

A Appendix

A.1 Tractable gaussian calculations

The uncertain quantities x^*, ϵ^* and e are taken to be mutually independent; (ϵ^*, e) are gaussian with zero means and marginal variances Σ^ϵ and Σ^e. The value for $\mathcal{P}$ is given in (16). Based on (3) and (5), the likelihood function in (16) has the form

$$\mathrm{Lik}_{\tilde{z}}(x) = \phi\big(\tilde{z}; Hg(x), \Sigma^z\big) \tag{A1a}$$

where $\phi(\cdot;\cdot,\cdot)$ is the gaussian density function with given mean vector and variance matrix, and

$$\Sigma^z \triangleq H\Sigma^\epsilon H^T + \Sigma^e. \tag{A1b}$$

The expectation in (16) has the form

$$\mathrm{E}\left(\mathbf{1}_Q(y) \mid z = \tilde{z}, x^* = x\right) = \int_{y\in Q} \phi\big(y; \mu_{y|z}(x), \Sigma_{y|z}\big)\, dy \tag{A2a}$$

where

$$\mu_{y|z}(x) \triangleq g(x) + \Sigma^\epsilon H^T \big(\Sigma^z\big)^{-1}\big(\tilde{z} - Hg(x)\big) \tag{A2b}$$

$$\Sigma_{y|z} \triangleq \Sigma^\epsilon - \Sigma^\epsilon H^T \big(\Sigma^z\big)^{-1} H\Sigma^\epsilon \tag{A2c}$$

(see, e.g., Mardia et al., 1979, ch. 3). If the simulator has an output component which corresponds directly to post-CO_2-doubling global mean temperature then the integral over the gaussian density function in (A2a) simplifies to an evaluation of a tail probability for a scalar gaussian quantity with a given mean and variance.

A.2 Quadrature with stochastic forcing

The idea is to use a carefully-chosen integration rule over the important inputs x_1 and a stochastic rule over the less important ones, x_2, where $x \equiv (x_1, x_2)$. Algorithm A.1 provides a basic implementation, resulting in an approximation of $\mathcal{P}$, denoted $\hat{\mathcal{P}}_{m,n}$, where the total number of simulator evaluations is $m \times n$. This approximation is a consistent estimate of $\mathcal{P}$ in the sense that

$$\lim_{\min\{m,n\}\to\infty} \hat{\mathcal{P}}_{m,n} = \mathcal{P}.$$

The j loop in Algorithm A.1 is a simple monte carlo approximation to the

Algorithm A.1 Quadrature with stochastic forcing algorithm for approximating $\mathcal{P}$ in $m \times n$ simulator evaluations

Require: $x_1^{(1)}, \ldots, x_1^{(m)}$ and $w^{(1)}, \ldots, w^{(m)}$
  for $i \in \{1, \ldots, m\}$ **do**
    for $j \in \{1, \ldots, n\}$ **do**
      Randomly sample $x_2^{(j)} \sim F_{x_2^*}$
      Evaluate $g\big(x_1^{(i)}, x_2^{(j)}\big)$
      Compute $L_j \triangleq \mathrm{Lik}_{\tilde{z}}\big(x_1^{(i)}, x_2^{(j)}\big)$ using eq. (A1)
      Compute $E_j \triangleq \mathrm{E}\left(\mathbf{1}_Q(y) \mid z = \tilde{z}, x^* = \big(x_1^{(i)}, x_2^{(j)}\big)\right)$ using eq. (A2)
    end for
    Compute $\hat{I}_n\big(x_1^{(i)}\big) \triangleq \frac{1}{n} \sum_{j=1}^{n} L_j$
    Compute $\hat{J}_n\big(x_1^{(i)}\big) \triangleq \frac{1}{n} \sum_{j=1}^{n} E_j \, L_j$
  end for
  Compute $\hat{c}_{m,n} \triangleq \left[\sum_{i=1}^{m} w^{(i)} \, \hat{I}_n\big(x_1^{(i)}\big) \, dF_{x_1^*}\big(x_1^{(i)}\big) \right]^{-1}$
  Compute $\hat{\mathcal{P}}_{m,n} \triangleq \hat{c}_{m,n} \sum_{i=1}^{m} w^{(i)} \, \hat{J}_n\big(x_1^{(i)}\big) \, dF_{x_1^*}\big(x_1^{(i)}\big)$

integral $J\big(x_1^{(i)}\big)$, defined in (17b). It could be improved in a number of ways, for example using importance sampling or using variance reduction methods such as antithetic variables (see, e.g., Ripley, 1987; Robert and Casella, 1999; Evans and Swartz, 2000). The i loop integrates out x_1 according to some specified integration rule, summarised in the abscissae $x_1^{(1)}, \ldots, x_1^{(m)}$ and the weights $w^{(1)}, \ldots, w^{(m)}$. One natural choice would be an integration rule formed from the product of deterministic one-dimensional rules, such as gaussian quadrature. Another would be a space-filling design such as a latin hypercube (McKay et al., 1979), or a quasi-random rule (Niederreiter, 1992); these types of design are more appropriate in the case where the climate scientist has little knowledge about which components of x_1 are particularly influential in $g(x_1, x_2)$.

The approximation $\hat{c}_{m,n}$ for $c \triangleq \Pr(z = \tilde{z})^{-1}$ follows from

$$\begin{aligned}\Pr(z = \tilde{z}) &= \iint_{x_1 \times x_2} \mathrm{Lik}_{\tilde{z}}(x_1, x_2)\; dF_{x_1^*}(x_1)\, dF_{x_2^*}(x_2) \\ &\equiv \int_{x_1} I(x_1)\, dF_{x_1^*}(x_1) \\ &\approx \sum_{i=1}^{m} w^{(i)}\, \hat{I}_n\big(x_1^{(i)}\big)\, dF_{x_1^*}\big(x_1^{(i)}\big) \qquad \text{(A3a)}\end{aligned}$$

where

$$I(x_1) \triangleq \int_{x_2} \mathrm{Lik}_{\tilde{z}}(x_1, x_2)\; dF_{x_2^*}(x_2) \qquad \text{(A3b)}$$

and $\hat{I}_n(x_1)$ is a simple monte carlo approximation to $I(x_1)$.

Acknowledgements

This work is funded by the U.K. Natural Environment Research Council (NERC), grant ref. NER/T/S/2002/00987 (RAPID Thematic Program) and the Tyndall Centre, grant ref. T2/13.

References

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

Bernardo, J. and A. Smith: 1994, *Bayesian Theory.* Chichester, UK: John Wiley & Sons.

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.

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.

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.

Currin, C., T. Mitchell, M. Morris, and D. Ylvisaker: 1991, 'Bayesian Prediction of Deterministic Functions, with Application to the Design and Analysis of Computer Experiments'. *Journal of the American Statistical Association* **86**, 953–963.

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

Goldstein, M. and J. Rougier: 2005a, 'Probabilistic Formulations for Transferring Inferences from Mathematical Models to Physical Systems'. *SIAM Journal on Scientific Computing* **26**(2), 467–487.

Goldstein, M. and J. Rougier: 2005b, 'Reified Bayesian Modelling and Inference for Physical Systems'. Under review, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf`.

Higdon, D., M. Kennedy, J. Cavendish, J. Cafeo, and R. D. Ryne: 2005, 'Combining Field Data and Computer Simulations for Calibration and Prediction'. *SIAM Journal on Scientific Computing* **26**(2), 448–466.

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.

Mardia, K., J. Kent, and J. Bibby: 1979, *Multivariate Analysis*. London: Harcourt Brace & Co.

McKay, M., W. J. Conover, and R. J. Beckham: 1979, 'A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output of Computer Code'. *Technometrics* **21**, 239–245.

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.

Murphy, J. M., D. M. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth: 2004, 'Quantification of Modelling Uncertainties in a Large Ensemble of Climate Change Simulations'. *Nature* **430**, 768–772.

Niederreiter, H.: 1992, *Random Number Generation and Quasi-Monte Carlo Methods*. SIAM: Philadelphia.

O'Hagan, A., M. Kennedy, and J. Oakley: 1999, 'Uncertainty Analysis and Other Inferential Tools for Complex Computer Codes'. In: J. Bernardo, J. Berger, A. Dawid, and A. Smith (eds.): *Bayesian Statistics 6*. pp. 503–519, Oxford University Press. With discussion, pp. 520–524.

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

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

Ross, S.: 1988, *A First Course in Probability*. Macmillan: New York, 3rd edition.

Rougier, J.: 2004, 'Brief Comment Arising re: "Quantification of Modelling Uncertainties in a Large Ensemble of Climate Change Simulations" by Murphy et al (Nature, 2004)'. Unpublished, available at `http://www.maths.dur.ac.uk/stats/people/jcr/newMurph.pdf`.

Santner, T., B. Williams, and W. Notz: 2003, *The Design and Analysis of Computer Experiments*. New York: Springer.

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

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.

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

Stainforth, D., T. Aina, C. Christensen, M. Collins, N. Faull, D. Frame, J. Kettleborough, S. Knight, A. Martin, J. M. Murphy, C. Piani, D. Sexton,

L. A. Smith, R. Spicer, A. Thorpe, and M. Allen: 2005, 'Uncertainty in Predictions of the Climate Response to Rising Levels of Greenhouse Gases'. *Nature* **433**, 403–406.