Inference In Ensemble Experiments

By Jonathan Rougier[1]† and David M.H. Sexton[2]

1. Department of Mathematical Sciences, Durham University, UK
2. Hadley Centre, Met Office, UK

We consider inference based on ensembles of climate model evaluations, and contrast the Monte Carlo approach, in which the evaluations are selected at random from the model-input space, with a more overtly statistical approach using emulators and experimental design.

Keywords: Monte Carlo, Uncertainty, Importance sampling, Emulator, Screening, Climate sensitivity

1. Introduction: Monte Carlo integration

The *raison d'être* of ensemble experiments is uncertainty about the model, usually concerning the relationship between the model and the climate itself. Traditionally the focus has been on varying the initial conditions, to sample internal climate variability. But more recently the focus has broadened to include other uncertain quantities, such as the model parameters. In this paper we describe the lack of precision that results from limits on the number of model evaluations we can perform. This section and §2 consider a simple and robust approach based on random-sampling, while §3 and §4 consider an alternative approach using *emulators*, which leads to a completely different treatment of the model evaluations. Section 5 concludes.

We think of our climate model as the mapping

$$x \to g(x)$$

where x denotes model-inputs, for example initial conditions, forcing functions, and model parameters, and g denotes the model. For simplicity we will focus on one particular type of inference, namely *uncertainty analysis*, which is inference about one or more model-outputs given uncertainty in the model-inputs. However, our analysis generalises directly to other types of inference, for example calibrated prediction, as described in Rougier (2006). If we denote by x^* the uncertain model-inputs, then we would like to make inferences about the uncertain quantity $\delta \triangleq g(x^*)$ for some given probability density function f_{x^*}; we treat x^* as absolutely continuous, for simplicity, but in general x is likely to comprise a mixture of continuous and discrete quantities for a large climate model.

Our uncertainty analysis is fully-described by the distribution function

$$F_\delta(v) \triangleq \Pr(\delta \le v) = \int \mathbb{I}\big(g(x) \le v\big)\ f_{x^*}(x)\,\mathrm{d}x, \tag{1.1}$$

where $\mathbb{I}(\cdot) = 1$ if true and 0 otherwise. In words, the probability that $\delta \le v$ is the probability content of the region of model-inputs which get mapped into values

† Department of Mathematical Sciences, Durham University, Science Site, Durham DH1 3LE, UK; email J.C.Rougier@durham.ac.uk

not more than v. On the righthand side we treat g as though it were a known function: one that may be costlessly evaluated, or admit an analytic solution. In climate models, therefore, this calculation is really just an aspiration: in practice $g(x)$ tends to be very complicated and expensive to evaluate, and the cost of each evaluation is the main impediment to computing the distribution function F_δ. In practice this means that we cannot compute $F_\delta(v)$ exactly, but we can approximate it, and so our attention turns to the accuracy of the approximation.

The simplest approximation to $F_\delta(v)$ is the Monte Carlo (MC) approximation

$$F_\delta^n(v) \triangleq n^{-1} \sum_{i=1}^{n} \mathbb{I}\big(g_i \le v\big) \qquad \text{where } x_i \overset{\text{iid}}{\sim} f_{x^*}. \tag{1.2}$$

We sample $X \triangleq \{x_1, \ldots, x_n\}$ independently from f_{x^*}, and we run the climate model at each x_i to compute $g_i \triangleq g(x_i)$, which gives us $G \triangleq \{g_1, \ldots, g_n\}$. Together, $(G; X)$ constitute our ensemble of model evaluations. In this ensemble we count the number of runs where the model-output is not more than v. We can construct an estimate of the entire distribution function for δ from one sample of size n. Usually this would be plotted as a step-function showing the proportions $(0), 1/n, 2/n, \ldots, 1$ against $g_{(1)}, \ldots, g_{(n)}$, where $g_{(i)}$ is the ith order statistic of G.

The empirical distribution function so constructed is only an estimate of F_δ. Sampling effects—what happens with different X—will tend to shift this empirical distribution function around, and we need to take this into account when constructing confidence intervals (CIs) for quantiles such as the 90th percentile. A simple way to do this is to invert the Kolmogorov-Smirnov (KS) test, as described in Hollander and Wolfe (1999, §11.5 and Table A.38). This gives lower and upper bounds defining a confidence band with the property

$$\Pr\big(\ell(v; X) \le F_\delta(v) \le u(v; X), \text{ for all } v\big) \ge 1 - \alpha$$

where the probability represents a relative frequency taken over all possible X of size n, and $1 - \alpha$ is the confidence level, typically 95%. Asymptotically, say $n \ge 40$, the 95% confidence band of the underlying distribution function is $\pm 1.36/\sqrt{n}$ vertically about the empirical distribution function.

An important feature of the KS approach is that it gives us a consistent set of horizontal intervals for any collection of percentiles; however, it is conservative for a given percentile, so that the coverage of the horizontal interval defined by the lower and upper bounds with $\alpha = 0.05$ is greater than 95%. For a given percentile we can also compute a point estimate and a horizontal interval directly, for example using the method of Harrell and Davis (1982) (HD). Such an interval will tend to be narrower, but it is less robust to the shape of the underlying distribution for small n, and not necessarily conservative.

We illustrate the results of a MC inference using the climate sensitivity of HadAM3: an atmospheric model coupled to a mixed-layer ocean. Our analysis of two ensembles from this model (Murphy et al., 2004; Stainforth et al., 2005) is described in Rougier et al. (2006). As part of our analysis we construct a statistical emulator of HadAM3. Emulators will be described in more detail in §3. For the time being we note that one outcome of constructing an emulator is a mean function, and this mean function can stand-in for the model itself in applications where the

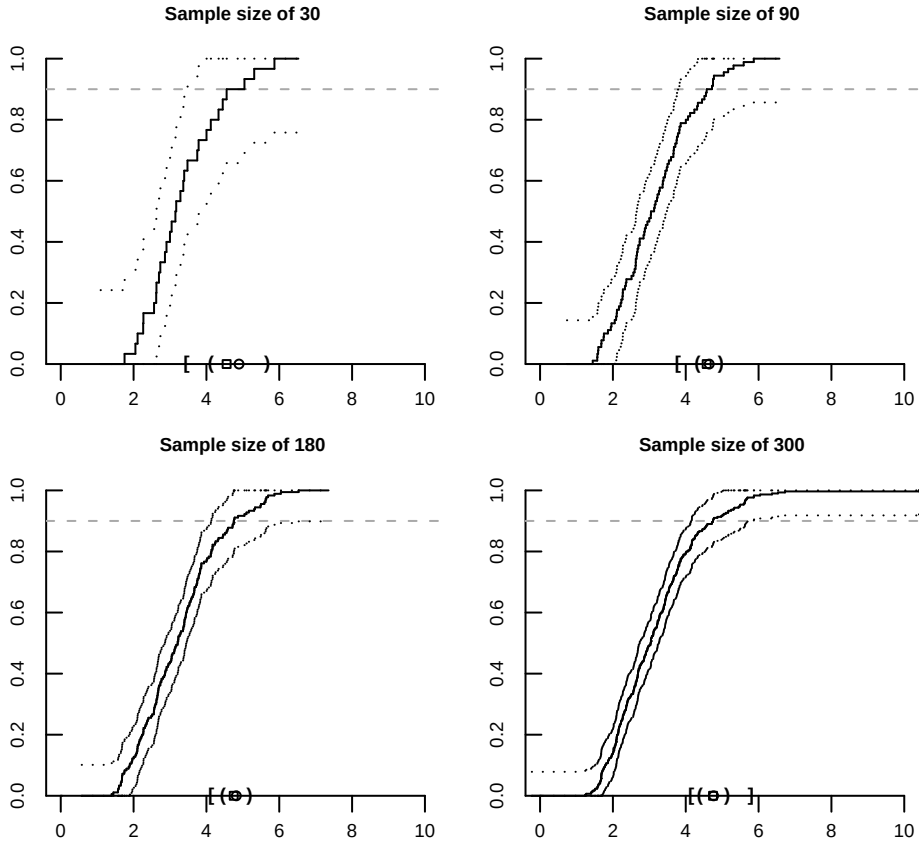

Figure 1. Climate sensitivity (K), uniform prior. Monte Carlo estimated distribution functions from a single ensemble, using four different sizes. The dotted lines indicate the 95% confidence band for the distribution function, using the Kolmogorov-Smirnov approach. On the horizontal axis, the open square and square brackets indicate the point estimate and KS 95% confidence interval (CI) for the 90th percentile; the upper value is undefined in the first three cases. The open circle and round brackets indicate the Harrell-Davis point estimate and asymptotic 95% CI for the 90th percentile. For reference, the true 90th percentile for this distribution is 4.3K.

model would be too expensive to evaluate. Therefore in this section and the next we use the mean function from the emulator in place of HadAM3 itself, to illustrate the effect of different numbers of evaluations in a MC uncertainty analysis.

In this section we will take f_{x^*} to be independent across components, uniform in the continuous inputs, and equally-probable across levels in the discrete components. As a representation of expert judgement about the uncertain model-inputs this is not a particularly appealing choice of distribution (we will investigate another choice in §2), but it is, at the moment, a common choice among climate scientists. Assigning probability distributions to quantities such as model-inputs is discussed in O'Hagan and Oakley (2004) and Garthwaite et al. (2005).

Figure 1 shows the result of an experiment with $n = 30$, 90, 180, and 300 evaluations of the mean function: these evaluations were nested in the sense that

the larger samples are extensions of the smaller ones. So we are addressing the question: what happens if we stop at 30? at 90? and so on. With fewer than 200 evaluations, already a large number for many ensemble experiments using climate models, we cannot get an upper value on the 95% CI of the 90th percentile of climate sensitivity using the KS method, because this is too far into the upper tail of the distribution. The HD asymptotic 95% CIs for the 90th percentile are shown as round brackets in Figure 1; these are only suggestive, particularly for small n.

Information about the uncertainty engendered by the sample size should *always* accompany MC studies with estimated distribution functions, or estimated probabilities.

2. Importance sampling

A major drawback of the MC approach is that it commits us to a particular choice of f_{x^*} through the way in which we select X by random sampling. Often x^* will represent some kind of 'correct' or 'best' model-input: a subtle concept, and perhaps even a flawed one, although certainly useful (Goldstein and Rougier, 2004, 2006b). But it is clear that specifying f_{x^*} involves a choice: there is no obvious 'right' candidate. It is an undoubted weakness of any inferential calculation if we cannot try different choices of f_{x^*}, to examine the sensitivity of our conclusions to choices about which there is no consensus.

With MC inference we can in fact try different distributions for f_{x^*} in (1.1), even after having generated X and evaluated the model at these inputs, using *Importance Sampling* (IS) (see, e.g., Robert and Casella, 1999, §3.3), but this method has its limitations. We will suppose that our evaluations were sampled according to f'_{x^*}, rather than f_{x^*}. In this case we refer to f_{x^*} as the *target* distribution and f'_{x^*} as the *proposal* distribution. IS is based on the simple notion that we can introduce the ratio $f'_{x^*}(x)/f'_{x^*}(x) \equiv 1$ into the integrand in (1.1), with the result that

$$F_\delta(v) \equiv \int \mathbb{I}\big(g(x) \leq v\big)\, \frac{f_{x^*}(x)}{f'_{x^*}(x)}\, f'_{x^*}(x)\, \mathrm{d}x$$

can be estimated by

$$F^{n\prime}_\delta(v) \triangleq n^{-1} \sum_{i=1}^{n} w_i\, \mathbb{I}\big(g_i \leq v\big) \qquad \text{where } x_i \overset{\text{iid}}{\sim} f'_{x^*}$$

and $w_i \triangleq f_{x^*}(x_i)/f'_{x^*}(x_i)$. We can plot our estimate of the distribution function as a step-function showing the cumulative weights $(0), w_{(1)}/n, (w_{(1)} + w_{(2)})/n, \ldots$ against $g_{(1)}, g_{(2)}, \ldots$. This is a generalisation of the original case, where we would have $w_{(i)} = 1$.

For IS to be valid it must be possible to sample all of the values in the target distribution using the proposal distribution. However, this condition is far from sufficient for a reasonable performance with limited n. The problem is that unless the proposal and target distributions are quite similar, the proposal distribution can miss regions of high probability in the target distribution: this is not a problem that re-weighting can fix. Simple diagnostics can be instructive; for example,

$$n^{-1} \sum_{i=1}^{n} w_i \approx \int \frac{f_{x^*}(x)}{f'_{x^*}(x)}\, f'_{x^*}(x)\, \mathrm{d}x = \int f_{x^*}(x)\, \mathrm{d}x = 1$$

and so if the weights do not sum approximately to n, then we should suspect that the proposal distribution has not done a good job. Note that unless $\sum_{i=1}^{n} w_i$ is exactly n, the empirical distribution function will not limit to 1. It is acceptable to normalise the weights to sum to 1, absorbing the factor of n^{-1} (Geweke, 1989, §2), but this would not be appropriate if the sum were not already close to n. Evans and Swartz (2000, §6.2) suggest evaluating the normalised weights, $\tilde{w}_i$ say, in terms of their summed squared values, for which

$$1 \leq n \sum_{i=1}^{n} (\tilde{w}_i)^2 \leq n;$$

values well away from 1 indicate a problem with f'_{x^*} as a proposal for f_{x^*}.

In our illustration, suppose we decided to replace the uniform marginal distribution for each independent continuous model-input with a symmetric triangular distribution over the same interval. This seems like a plausible description of the fact that central values of the parameters are judged more likely to be 'correct' than extreme ones. But because there are 14 such model-inputs in the HadAM3 model, the ratio $f_{x^*}(x)/f'_{x^*}(x)$ involves the fourteenth power of the univariate ratio. This illustrates that there can be a dimensional effect in IS that is not present in MC sampling, because small marginal changes in the distribution of each component of x^* become magnified. In the case of our sample with $n = 300$, the sum of the weights is 128.3 (not close to 300), and n times the sum of the squared normalised weights is 13.4 (a long way from 1). IS cannot be considered reliable in this case.

To show that IS *can* be useful, we also consider a choice of f_{x^*} much closer to our f'_{x^*}, namely a distribution in which just five of the independent continuous variables have triangular distributions (`VF1`, `CT`, `CW`, `CFS`, and `ENT` in Rougier et al., 2006). In this case the sum of the weights ($n = 300$) is 267.7 and the sum of n times the squared normalised weights is 4.3; these values are much better than before, but still suggest caution. Figure 2 shows the result of this different choice for f_{x^*} for $n = 180$ and $n = 300$, with normalised weights. Large weights show up as vertical segments in the empirical distribution function. KS confidence bands are also shown: these are less reliable for IS than for MC, because large weights behave like tied observations, and the KS statistic is less reliable for tied observations. The change of distribution appears to have affected δ, possibly raising the 90th percentile.

Therefore IS is useful if we want to start with a particular choice for f_{x^*} and then look at the effect of small perturbations, but it cannot help us if we are quite uncertain about f_{x^*}, and would like to try out a number of alternatives. Likewise, it cannot be used to calibrate an MC sample by reweighting if the calibration data are highly informative about the model-inputs—which is, of course, what we hope they would be.

3. Emulators

There are three attractive features of the MC approach. First, it is simple to understand and implement. Second, it is sequential, so we can easily add more evaluations if required (other integration methods, like gaussian quadrature, do not have this feature). Third, it is relatively easy to compute a measure of uncertainty about our

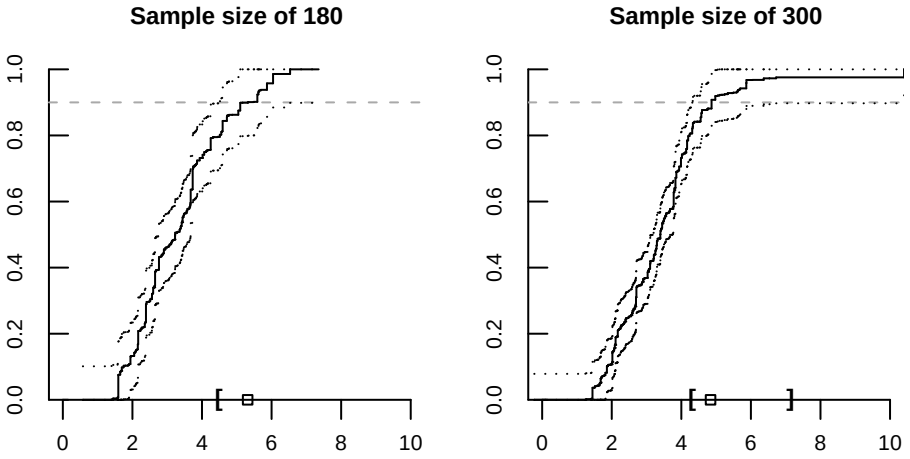

Figure 2. Climate sensitivity (K), triangular distribution for five continuous model-inputs. Computed from the original uniform sample using Importance Sampling. On the horizontal axis, the open square and square brackets indicate the point estimate and Kolmogorov-Smirnov (KS) 95% confidence interval for the 90th percentile; the upper value is undefined in the first case. The KS confidence band can be unreliable with large weights.

estimates. One drawback, as discussed in the previous section, is the inflexibility of being committed to a given distribution f_{x^*}, which is only partially mitigated by IS.

A bigger drawback, though, is that MC is expensive, in terms of the number of evaluations required for a given precision. This will not matter if we have a model with a small number of uncertain inputs that evaluates extremely fast: we might as well use MC and be done with it. But in ensemble experiments with climate models typically the opposite situation prevails: we have a limited number of evaluations of a model with a large input-space. The basis of MC's simplicity is that it assumes nothing about the model: the evaluations in G are simply points in the output-space, and X is discarded. We can do better if we are prepared to exploit the structure in our ensemble, notably the judgement that $g(x')$ is predictable from $g(x)$ when x and x' are not too far apart. In this case we do not discard X, but incorporate it into our inference. We do this by constructing an *emulator*. In many experiments, emulators may be the *only* means of deriving useful probabilistic information, because n is simply too small to be effective in an MC approach.

An emulator is a stochastic representation of a (usually deterministic) complex function. In our case, the emulator is a statistical framework that allows us to compute the distribution function

$$F_{g(x)}(v) \triangleq \Pr\big(g(x) \leq v \mid x, G; X\big)\,,$$

where the model g is now the uncertain quantity on the righthand side, and our information about g is conditional on our observations of the model's behaviour, i.e. on the ensemble. In other words, for any input value x the emulator tells us a probability for the model-output $g(x)$ being no greater than v, based on the information in $(G; X)$. O'Hagan (2006) provides an introduction to emulators, with more detail available in Kennedy and O'Hagan (2001), which includes a summary

of earlier work; the approach of Craig et al. (2001) and Goldstein and Rougier (2006a) may be more appropriate for large models with complex model-inputs, such as climate models.

The role of the emulator is to separate learning about the model from using the model to makes inferences. The purpose of the ensemble is to learn about the model. Once we have distilled the ensemble into the emulator it has no additional value, and the emulator takes the place of the model in our inference. The crucial feature of the emulator is that it does not ignore the uncertainty that arises from having only a limited number of evaluations in the ensemble, what O'Hagan (2006) refers to as *code uncertainty*: this uncertainty feeds through into the inference, and can be substantially reduced by carefully selecting the model evaluations.

In our uncertainty analysis the emulator allows us to focus on what we actually *can* compute, rather than what we aspire to compute:

$$\hat{F}_\delta(v) \triangleq \Pr(\delta \leq v \mid G; X) = \int F_{g(x)}(v)\, f_{x^*}(x)\, \mathrm{d}x \tag{3.1}$$

where, in comparison to (1.1), we see that the emulator distribution function has taken the place of the indicator function. The *quid pro quo* of this realism, though, is the need for a statistical framework that allows us to infer the distribution function $F_{g(x)}$ from the ensemble $(G; X)$. This is both an opportunity and a burden. The statistical framework allows us to incorporate additional information from modellers and from other ensembles; for example, how smooth is the model? and which are the most important model-inputs? But this requires extra work, both in eliciting judgements, and in the painstaking but crucial task of diagnostic assessment.

Staying with MC integration to compute (3.1), we have the approximation

$$\hat{F}^m_\delta(v) \triangleq m^{-1} \sum_{j=1}^{m} F_{g(x_j)}(v) \qquad \text{where } x_j \overset{\text{iid}}{\sim} f_{x^*}. \tag{3.2}$$

The major difference here is that we do not evaluate the model at each x_j, we simply evaluate the emulator distribution function, which is often more-or-less costless. Thus m can be made as large as we need to ensure that there is no sampling uncertainty in the resulting empirical distribution function: it is a precise estimate, although it is important to remember that it is conditional on the n evaluations in the ensemble, and on the statistical framework—hence the importance of diagnostics. This calculation can be repeated for any choice of f_{x^*}, so we can easily compare the effects of, say, a uniform or a triangular distribution.

It may be helpful to describe the behaviour of the emulator for finite n and as $n \to \infty$. For any ensemble, the emulator can be used to compute a mean function $\mu(x) \triangleq \mathrm{E}\big(g(x) \mid x, G; X\big)$ and a variance function $\Sigma(x) \triangleq \mathrm{Var}\big(g(x) \mid x, G; X\big)$. These are simply summaries of the distribution function $F_{g(x)}$. Using standard relationships in probability, we can find the conditional mean and variance of δ as

$$\mathrm{E}(\delta \mid G; X) = \mathrm{E}\big(\mu(x^*)\big) \quad \text{and} \quad \mathrm{Var}(\delta \mid G; X) = \mathrm{E}\big(\Sigma(x^*)\big) + \mathrm{Var}\big(\mu(x^*)\big)\,.$$

In the limit $n \to \infty$ a well-constructed emulator satisfies $\mu(x) \to g(x)$ and $\Sigma(x) \to 0$, for all x, so that $\mathrm{E}(\delta \mid G; X) \to \mathrm{E}(\delta)$ and $\mathrm{Var}(\delta \mid G; X) \to \mathrm{Var}(\delta)$. So in this limit the emulator gives the same answer as we would get were we to treat the model

as precisely known, i.e. as we did in (1.1). For finite n, however, the emulator also includes the uncertainty that comes from *not* knowing the model precisely.

The MC approach and the emulator approach have two quite different sources of uncertainty about the distribution for δ, but they both arise as a consequence of us only having n evaluations in the ensemble. In the MC approach our uncertainty about F_δ comes from our failure to compute the integral exactly due to limited n, and is summarised in terms of the sampling properties of the empirical distribution function F_δ^n. Our uncertainty about F_δ goes down at a rate proportional to $1/\sqrt{n}$. In the emulator approach we do not approximate F_δ, instead we compute $\hat{F}_\delta$ exactly. By using expert judgements and carefully-chosen evaluations we expect that $\hat{F}_\delta$ will be a better approximation than F_δ^n, but this will depend on g. If g has no structure then an emulator will not improve on the MC approach. But if g has structure that we can exploit, for example being smooth, or having only a limited number of important model-inputs, then we expect the emulator to do better than $1/\sqrt{n}$, and in this way to justify the extra (human) costs involved.

To illustrate one of the benefits of an emulator, we present some results based on an ensemble of evaluations of HadAM3, as described in Rougier et al. (2006). For simplicity, here we will use an emulator constructed from our QUMP ensemble alone (297 evaluations), with a fairly uninformative prior. The way in which the evaluations in our X were chosen is outlined in §4. Crucially, however, there is no way we could interpret X as the outcome of some sampling exercise, so MC was never an option. As a general point pertinent to many ensemble experiments, if the evaluations in X are not sampled from some specific distribution, or do not conform to an explicit integration scheme, then using them to construct an emulator is the only option for probabilistic inference with uncertain model-inputs.

Figure 3 shows two quite different choices for the distribution of x^*: (A) uniform distribution in all of the continuous model-inputs; (B) triangular. It also shows two other choices: (C), like (A) but with the reciprocal of the entrainment rate being uniform; and (D), like (B) but with the reciprocal being triangular; these are included in response to the ongoing debate about whether the entrainment rate or its reciprocal is the more natural parametrisation. A value of $m = 10^4$ in (3.2) was sufficient to make these estimated distribution functions precise. The message from this Figure is unambiguous, if unsurprising: if it makes a difference to scientists and policymakers whether the 90th percentile of climate sensitivity in HadAM3 is 4K or 6K, then climate scientists will have to think hard about the type of distribution they specify for the model-inputs.

4. Experimental design

Once we have liberated the choice of X from any particular sampling scheme, we can choose our evaluations to learn about g in an informative way. The general approach is termed *Bayesian Experimental Design* (Chaloner and Verdinelli, 1995); more specialist information and further references can be found in Koehler and Owen (1996) and Santner et al. (2003). We suggest the following three stages.

1. **Screening runs.** The initial set of evaluations is designed to pick-out basic structure in the model, such as identifying the important or *active* model-inputs, plus some indication of the nature of the model-response to these

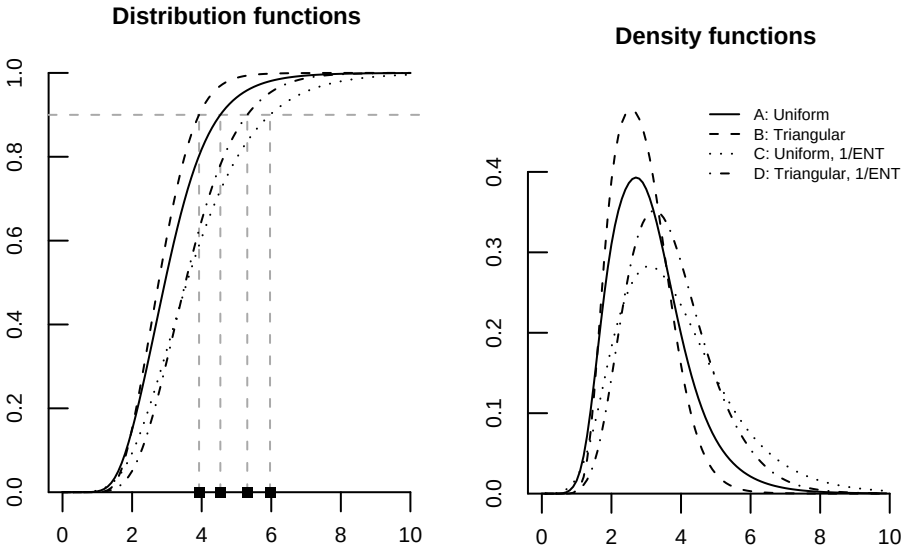

Figure 3. HadAM3 climate sensitivity (K), conditional on an ensemble of 297 evaluations, for four different choices of distribution for f_{x^*} (see text in §3). On the horizontal axis of the lefthand panel the filled squares indicate the four 90th percentiles.

inputs (e.g., linear, quadratic, linear in the log). A *maximin latin hypercube* can be an effective choice. Where we have strong prior information about which inputs are important (often the case with climate models), we may use such a design on the less important model-inputs, and a more structured design in the subspace of active inputs, as described next.

2. **Interactions.** In climate models we expect interactions between model-inputs to be important in determining the model-outputs. With a large number of model-inputs we cannot expect to explore all possible interactions, even if we limit ourselves to two-way effects. Therefore we explore interactions initially in the active inputs. This second set of evaluations could follow a standard experimental design such as a *fractionated factorial*, which allows us to identify low-order interactions (two- and three-way, for example).

3. **Sequential.** After the first two stages we should have enough evaluations to build a useful emulator. In the third stage we can use this emulator to select further evaluations. The simplest approach is to put additional evaluations into regions x of the model-input space for which the predictive uncertainty, i.e. $\mathrm{Sd}(g(x) \mid x, G; X)$, is currently high. Such evaluations will tend quite naturally to avoid the previous evaluations in X.

Where we have calibration data we would expect to iterate these stages, refocusing our approach as these data rule out regions of the model-input space.

These three stages are designed to produce evaluations that are *generally* informative, rather than informative for a particular inference. This reflects the fact that an ensemble from a large climate model is likely to be used a number of different ways. Craig et al. (2001) describe a more targeted sequential approach when it is

possible to perform dedicated evaluations, which generalises to a *batch sequential* approach that makes the best use of a number of additional evaluations (e.g., choosing the next five evaluations jointly). A useful way of evaluating a candidate input x is to use the emulator to generate *pseudo-data* at x: a realisation of what $g(x)$ might be. This can be added to to the ensemble as though it were real, in order to judge the impact on the resulting inference of actually evaluating the model at x.

Our HadAM3 ensemble comprises several different sets of evaluations. Initially, there were single-parameter perturbations in each model-input, and a very limited number of multiple-parameter perturbations, as used in Murphy et al. (2004). Since that time we have augmented the ensemble with batches of evaluations designed to allow us to learn about the HadAM3 model (see Webb et al., 2006, for details). We have adjusted the balance of the ensemble as a whole so that no model-input values were particularly over-represented. We have also filled-in regions identified with the major sub-processes (using fractional factorials and carefully-selected latin hypercubes) to make sure that we have information on low-order interactions between model-inputs within each sub-process. In the future we plan to use sequential design to select additional evaluations.

5. Conclusion

Simple MC inference, for which the ensemble represents a random sample from some specified distribution over model-inputs, is a very robust approach, making no assumptions about the form of the underlying climate model. This is both its strength (generality) and its weakness (inefficiency, inflexibility), and a clear demonstration of the 'no free lunch' principle. The alternative approach is to tune our inference and calculations to our particular climate model. Emulators provide one means for doing this, most clearly seen in the way in which they permit us to do n carefully-chosen evaluations of the model rather than n random evaluations of the model. Emulators also allow us to incorporate expert judgement into their prior specification, although this is less important if we have a reasonable number of evaluations from the Screening and Interaction stages outlined in §4. By separating the ensemble from the inference, emulators also allow us to perform a wide range of inferential calculations over any number of different probabilistic choices: an important feature where there is no consensus about what an appropriate choice might be. The emulator used for the illustrations in this paper is described in Rougier et al. (2006).

J.C. Rougier is currently funded by NERC, under the RAPID Directed Programme. We would like to thank Michael Goldstein, Peter Craig, and James Annan for very helpful discussions.

References

Chaloner, K., Verdinelli, I., 1995. Bayesian experimental design: A review. Statistical Science 10 (3), 273–304.

Craig, P., Goldstein, M., Rougier, J., Seheult, A., 2001. Bayesian forecasting for complex systems using computer simulators. Journal of the American Statistical Association 96, 717–729.

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

Garthwaite, P., Kadane, J., O'Hagan, A., 2005. Statistical methods for eliciting probability distributions. Journal of the American Statistical Association 100, 680–701.

Geweke, J., 1989. Bayesian inference in econometric models using Monte Carlo integration. Econometrica 57 (6), 1317–1339.

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

Goldstein, M., Rougier, J., 2006a. Bayes linear calibrated prediction for complex systems, forthcoming in the Journal of the American Statistical Association, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/BLCP.pdf`.

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

Harrell, F., Davis, C., 1982. A new distribution-free quantile estimator. Biometrika 69, 635–640.

Hollander, M., Wolfe, D., 1999. Nonparametric Statistical Methods, 2nd Edition. New York: John Wiley & Sons, Inc.

Kennedy, M., O'Hagan, A., 2001. Bayesian calibration of computer models. Journal of the Royal Statistical Society, Series B 63, 425–464, with discussion.

Koehler, J., Owen, A., 1996. Computer experiments. In: Ghosh, S., Rao, C. (Eds.), Handbook of Statistics, 13: Design and Analysis of Experiments. North-Holland: Amsterdam, pp. 261–308.

Murphy, J., Sexton, D., Barnett, D., Jones, G., Webb, M., Collins, M., Stainforth, D., 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430, 768–772.

O'Hagan, A., 2006. Bayesian analysis of computer code outputs: A tutorial, forthcoming in Reliability Engineering and System Safety, currently available at `http://www.tonyohagan.co.uk/academic/pdf/BACCO-tutorial.pdf`.

O'Hagan, A., Oakley, J., 2004. Probability is perfect, but we can't elicit it perfectly. Reliability Engineering and System Safety 85, 239–248.

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

Rougier, J., 2006. Probabilistic inference for future climate using an ensemble of climate model evaluations, forthcoming in Climatic Change, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/CCfinal.pdf`.

Rougier, J., Sexton, D., Murphy, J., Stainforth, D., 2006. Emulating the sensitiviy of the HadAM3 climate model using ensembles from different but related experiments, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/hadSens.pdf`.

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

Stainforth, D., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D., Kettleborough, J., Knight, S., Martin, A., Murphy, J. M., Piani, C., Sexton, D., Smith, L. A., Spicer, R., Thorpe, A., Allen, M., 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. Nature 433, 403–406.

Webb, M., Senior, C., Sexton, D., Ingram, W., Williams, K., Ringer, M., McAveney, B., Colman, R., Soden, B., Gudgel, R., Knutson, T., Emori, S., Ogura, T., Tsushima, Y., Andronova, N., Li, B., Musat, I., Bony, S., Taylor, K., 2006. On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles. Climate Dynamics 27, 17–38.