Inference In Ensemble Experiments

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

1. Department of Mathematics, University of Bristol, 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 Ensemble, Designed Ensemble, 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 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 $g : \boldsymbol{x} \mapsto y$, where $\boldsymbol{x}$ denotes model-inputs, for example initial conditions, forcing functions, and model parameters, $g(\cdot)$ denotes the model, and y denotes a point in the model's output-space. We will focus on one particular type of inference, namely *uncertainty analysis*, which is inference about a model-output given uncertainty in the model-inputs. If we denote by $\boldsymbol{x}^*$ the uncertain model-inputs, then we would like to make inferences about the uncertain scalar quantity $y^* \triangleq g(\boldsymbol{x}^*)$ for some given distribution function F_{x^*}; here '$\triangleq$' denotes 'defined as'. For a climate model we would expect $\boldsymbol{x}$ to comprise both continuous and discrete quantities, and so we cannot assume the existence of a density function for $\boldsymbol{x}^*$. This has both technical and practical consequences. The technical consequences can be minimised by describing our inferences in terms of expectations; the practical consequences will be introduced in §2. We will assume throughout that $g(\cdot)$ is sufficiently well-behaved that $g(\boldsymbol{x}^*)$ is a well-defined uncertain quantity and all the necessary expectations exist.

Our uncertainty analysis is fully-described by the distribution function for y^*,

$$F_{y^*}(v) \triangleq \Pr[y^* \leq v] = \mathrm{E}_{F_{y^*}}[\mathbb{I}(y^* \leq v)] \tag{1.1}$$

where $\mathbb{I}(\cdot) = 1$ if true and 0 otherwise, and $\mathrm{E}_{F_{y^*}}[\cdot]$ is the expectation with respect to the distribution function F_{y^*}. This distribution function is implied by $g(\cdot)$ and

† Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, UK; email J.C.Rougier@bristol.ac.uk

our choice of F_{x^*}, and we can compute $F_{y^*}(v)$ as

$$F_{y^*}(v) = \mathrm{E}_{F_{x^*}}\big[\mathbb{I}(g(\boldsymbol{x}^*) \le v)\big]\,. \tag{1.2}$$

If $g(\cdot)$ is a climate model we do not expect to be able to evaluate this expression directly, but we can approximate it, and so our attention turns to the nature and accuracy of the approximation.

The simplest way to approximate $F_{y^*}(v)$ is to use Monte Carlo (MC) integration

$$F^n_{y^*}(v) \triangleq n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_i \le v) \quad \text{where } y_i \triangleq g\big(\boldsymbol{x}^{(i)}\big) \text{ and } \boldsymbol{x}^{(i)} \overset{\text{iid}}{\sim} F_{x^*}. \tag{1.3}$$

We sample $X \triangleq \big\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\big\}$ independently from F_{x^*}, and we run the climate model at each $\boldsymbol{x}_i$ to compute y_i, which gives us $Y \triangleq \{y_1, \ldots, y_n\}$, an independent and identically-distributed (iid) sample from the density function F_{y^*}. Together, $(Y; X)$ constitute our ensemble of model evaluations. Note, however, that for the inference about y^* only Y is used: generating the $\boldsymbol{x}^{(i)}$ is simply a step in the process of sampling from F_{y^*}. We refer to this as a *Monte Carlo ensemble*.

We can construct an estimate of the entire distribution function for y^* 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 $y_{(1)}, \ldots, y_{(n)}$, where $y_{(i)}$ is the ith order statistic of Y. The empirical distribution function so constructed is only an estimate of F_{y^*}. Sampling effects will tend to shift this empirical distribution function around, and we need to take this into account when determining our uncertainty 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 random lower and upper bounds defining a confidence band with the property

$$\Pr\Big[\ell(v; \boldsymbol{Y}) \le F_{y^*}(v) \le u(v; \boldsymbol{Y}), \text{ for all } v\Big] \ge 1 - \alpha \tag{1.4}$$

where $\boldsymbol{Y}$ denotes an iid sample of size n from F_{y^*}, 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; as this is a vertical band, there is no necessity for the inferred horizontal intervals to be finite, particularly in the tails. A note of caution: a 95% confidence is not the same as a 95% probability that our *observed* interval $\big[\ell(v, Y),\ u(v, Y)\big]$ contains $F_{y^*}(v)$. 'Confidence' is a property of the random interval *before* Y is observed; see, e.g., DeGroot and Schervish (2002, sec. 7.5) for further clarification.

An important feature of the KS approach is that it gives us a consistent set of horizontal CIs for any collection of percentiles; however, it is conservative for a given percentile, so that the coverage of the horizontal interval 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 more sensitive to the shape of the underlying distribution for, say, $n \le 30$.

We illustrate the results of a MC inference using the climate sensitivity of HadSM3: an atmospheric model coupled to a mixed-layer ocean, which combines

both continuous and discrete inputs. 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 HadSM3. 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 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 HadSM3 itself, to illustrate the effect of different numbers of evaluations in a MC uncertainty analysis.

In this section we take F_{x^*} to be independent across components (subject to restrictions discussed in §2), 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 et al. (2006).

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 95% CIs for the 90th percentile are shown as round brackets in Figure 1.

Finally, a comment on the sensitivity of the MC approach to the number of model-inputs. It is sometimes averred that MC integration is unaffected by the number of inputs, and the KS result appears to support this. However, our uncertainty about the model output is expressed horizontally, not vertically. When we translate the vertical KS band into a horizontal interval, e.g., for the 90th percentile, two factors are important: the height of the band, *and* the slope of the distribution function around the 90th percentile. The slope of the distribution function often *does* depend on the number of inputs. Suppose $g(\cdot)$ is a climate model with a crude cloud scheme, and $h(\cdot, \cdot)$ is a model with a complicated cloud scheme which requires additional inputs $\boldsymbol{w}$. If the cloud scheme is important, then, typically, $h(\boldsymbol{x}^*, \boldsymbol{w}^*)$ will be more uncertain than $g(\boldsymbol{x}^*)$; the slope of the distribution function around the 90th percentile will be shallower, and uncertainty about the 90th percentile will be larger.

2. Importance sampling

A major drawback of the MC approach is that it commits us to a particular sampling distribution on the model-inputs $\boldsymbol{x}^*$. Often $\boldsymbol{x}^*$ will represent some kind of 'correct' or 'best' input (Goldstein and Rougier, 2004; Rougier, 2007). 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.

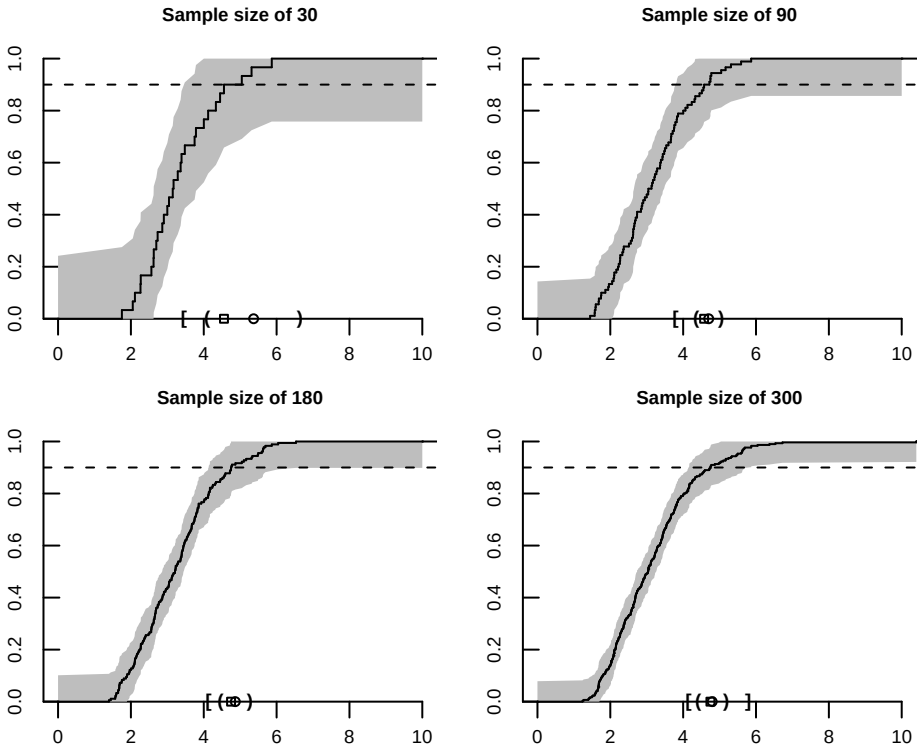

Figure 1. Climate sensitivity (K), uniform prior. Monte Carlo estimated distribution functions from a single ensemble, using four different sizes. The shaded polygons indicate the 95% confidence band for the distribution function, using the Kolmogorov-Smirnov approach. On the horizontal axis, the open 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 is 4.3K.

With MC inference we can in fact try different distributions for F_{x^*}, even after having generated Y, using *Importance Sampling* (IS); see, e.g., Robert and Casella (1999, §3.3). For technical reasons we must introduce the additional requirements that $F_{\boldsymbol{x}^*}$ factorises into a part with only continuous inputs, and the remainder, i.e. $F_{x^*}(\boldsymbol{x}) = F_{c^*}(\boldsymbol{c})F_{r^*}(\boldsymbol{r})$ where $\boldsymbol{x} = (\boldsymbol{c}, \boldsymbol{r})$ and $\boldsymbol{c}$ are all continuous. Suppose we want to investigate the distribution of y after sampling from F'_{x^*}, where we require $F'_{x^*}(\boldsymbol{x}) = F'_{c^*}(\boldsymbol{c})F_{r^*}(\boldsymbol{r})$; i.e., only the distribution for $\boldsymbol{c}$ is different. We refer to F_{x^*} as the *proposal* distribution and F'_{x^*} as the *target* distribution. Providing that f_{c^*} is non-zero wherever f'_{c^*} is non-zero, where a small f denotes a probability density function, we can write (1.1) as

$$F'_{y^*}(v) = \mathrm{E}_{F'_{x^*}}\big[\mathbb{I}(g(\boldsymbol{x}^*) \le v)\big] = \mathrm{E}_{F_{x^*}}\big[\mathbb{I}(g(\boldsymbol{x}^*) \le v)\, w(\boldsymbol{x})\big] \tag{2.1}$$

where $w(\boldsymbol{x}) \triangleq f'_{c^*}(\boldsymbol{c})/f_{c^*}(\boldsymbol{c})$. This relation follows after introducing the value $1 \equiv$

$f_{c^*}(\boldsymbol{c})/f_{c^*}(\boldsymbol{c})$ into (1.2). Eq. (2.1) gives rise to the MC estimator

$$F'_{y^*}(v) \approx n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_i \leq v)\, w_i \tag{2.2}$$

where $w_i \triangleq w\big(\boldsymbol{x}^{(i)}\big)$; this calculation is based on our original sample $(Y; X)$, where $\boldsymbol{x}$ was sampled from F_{x^*}. The sum of the weights should be approximately n, and in this case it is acceptable to normalise them, so that

$$F'_{y^*}(v) \approx \sum_{i=1}^{n} \mathbb{I}(y_i \leq v)\, \tilde{w}_i \tag{2.3}$$

where $\tilde{w}_i \triangleq w_i / \big(\sum_{j=1}^{n} w_j\big)$. We can plot our estimate of the distribution function as a step-function showing the cumulative weights $(0), \tilde{w}_{(1)}, \tilde{w}_{(1)} + \tilde{w}_{(2)}, \ldots$ against $y_{(1)}, y_{(2)}, \ldots$. This is a generalisation of the original case, where we would have $\tilde{w}_{(i)} = 1/n$.

The problem with IS is that when n is small the proposal distribution can, by chance, easily miss the region of high probability in the target, particularly when the two distributions are not very similar. IS estimates can therefore be very uncertain. Liu (2001, pp. 35–36) shows that the variance of an IS estimator is approximately proportional to one plus the variance of the weights. A useful diagnostic that reflects this is the Effective Sample Size (ESS)

$$\text{ESS} \triangleq \left\{ \sum_{i=1}^{n} (\tilde{w}_i)^2 \right\}^{-1} \tag{2.4}$$

which is 1 when all the weight is concentrated into a single evaluation, and n if it is spread equally across all n evaluations.

Our illustration demonstrates the need for the additional technical requirements for IS. The HadSM3 model has eighteen continuous inputs and thirteen discrete ones. However, four of the continuous inputs are contingent on the discrete inputs; e.g, the two continuous convective anvil parameters (`ANVS` and `ANVU`) will be effectively zero when convective anvils are switched off (`ANV` = Off); see Gregory (1999). These four continuous inputs cannot be taken as probabilistically independent of the discrete ones. Therefore the largest collection of continuous inputs in our factorisation of $\boldsymbol{x}$ is fourteen. Suppose we decided to replace the uniform marginal distribution for each of these inputs 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. If we do this, however, the ratio $w(\boldsymbol{x})$ involves the fourteenth power of the univariate ratio of a triangular to a uniform. This illustrates that there can be an additional dimensional effect in IS, because small marginal changes in the distribution of each component of $\boldsymbol{c}^*$ become magnified. In the case of our sample with $n = 300$, the sum of the weights is 128.3 (not close to 300), and the ESS is only 22. IS cannot be considered reliable in this case.

To show that IS *can* be useful, we also consider a choice for 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 this case the

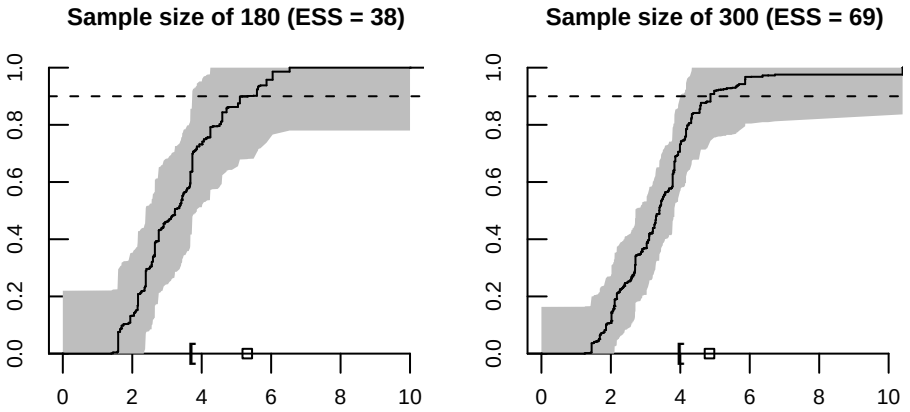

Figure 2. Climate sensitivity (K), triangular distribution for five continuous model-inputs. Computed from the original uniform sample using Importance Sampling. See caption to Figure 1 for details. The KS 95% confidence bands are based on the ESS, see eq. (2.4). For reference, the true 90th percentile is 3.9K.

sum of the weights ($n = 300$) is 267.7 and the ESS is 69; these values are better than before, but still suggest caution. Figure 2 shows the result of this choice for F'_{x^*} for $n = 180$ and $n = 300$. Large individual weights show up as vertical segments in the empirical distribution function. KS confidence bands are also shown, based on the ESS. The HD estimator does not generalise to this case.

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 possibly quite different alternatives.

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 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 are simply points in the output-space, and $\boldsymbol{x}$ is discarded. We can do better if we are prepared to exploit the structure in our ensemble, notably the judgement that $g(\boldsymbol{x}')$ is predictable from $g(\boldsymbol{x})$ when $\boldsymbol{x}$ and $\boldsymbol{x}'$ are not too far apart. In this case we do not discard $\boldsymbol{x}$, 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

$$F_{g(x)}(v) \triangleq \Pr\left[g(\boldsymbol{x}) \leq v \mid Y; X\right], \tag{3.1}$$

where the model $g(\cdot)$ is now the uncertain quantity on the righthand side, and our information about $g(\cdot)$ is conditional on our observations of the model's behaviour, i.e. on the ensemble. In other words, for any input value $\boldsymbol{x}$ the emulator tells us a probability for the model-output $g(\boldsymbol{x})$ being no greater than v, based on the information in $(Y; X)$. O'Hagan (2006) provides an introduction to emulators. One simple approach to constructing an emulator is to use a Bayesian treatment of regression, where the regressors are linear and non-linear functions of the model-inputs. This is effectively the approach used in our illustration.

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

$$F_{y^*}(v) \triangleq \Pr\left[y^* \leq v \mid g(\cdot)\right] \tag{3.2}$$

where the conditioning on $g(\cdot)$ makes explicit what was previously implicit, namely that on the righthand side of (1.1) we were treating the model as though it were known. What we can *actually* compute, though, with our n evaluations, is

$$\hat{F}_{y^*}(v) \triangleq \Pr[y^* \leq v \mid Y; X] = \mathrm{E}_{F_{x^*}}\left[F_{g(x^*)}(v)\right], \tag{3.3}$$

where we choose to treat $\boldsymbol{x}^*$ and $g(\cdot)$ as probabilistically independent. Comparing (3.3) with (1.2), the emulator distribution function has taken the place of the indicator function $\mathbb{I}(\cdot)$, because with our finite ensemble it is no longer clear-cut that $g(\boldsymbol{x}) \leq v$, for arbitrarily-chosen $\boldsymbol{x}$. 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 $(Y; X)$. 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 (3.3), we approximate $F_{y^*}(v)$ as

$$\hat{F}^m_{y^*}(v) \triangleq m^{-1} \sum_{j=1}^{m} F_{g(x^{(j)})}(v) \quad \text{where } \boldsymbol{x}^{(j)} \overset{\text{iid}}{\sim} F_{x^*}. \tag{3.4}$$

The major difference here is that we do not evaluate the model at each $\boldsymbol{x}^{(j)}$, 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 of $\hat{F}_{y^*}$. From a practical point of view, the emulator separates 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. Thus the calculation of $\hat{F}_{y^*}(v)$ can be repeated for any choice of F_{x^*}, so we can easily compare the effects of, say, a uniform or a triangular distribution.

The MC approach and the emulator approach have two quite different sources of uncertainty about the distribution for y^*, but they both arise as a consequence of us only having n evaluations in the ensemble. In the MC approach our uncertainty about F_{y^*} 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 $F^n_{y^*}$. In the emulator approach we do not approximate F_{y^*}, instead we compute $\hat{F}_{y^*}$ exactly. By using expert judgements and carefully-chosen evaluations we expect that $\hat{F}_{y^*}$ will be a better approximation than $F^n_{y^*}$, but this will depend on the model. If $g(\cdot)$ 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, and in this way to justify the extra (human) costs involved. For example, if $\boldsymbol{x}$ represents the initial value of the state vector in a large climate model, then it is a common judgement that $g(\boldsymbol{x}')$ may not be predictable from $g(\boldsymbol{x})$ even when $\boldsymbol{x}$ and $\boldsymbol{x}'$ are quite close; in the language of spatial statistics, the *correlation length* for initial conditions is short. This lack of predictability will undermine the efficacy of an emulator, and in this case the MC approach for initial conditions has much to recommend it. By way of contrast, the correlation length for model-parameters is likely to be much greater, and so a perturbed-physics experiment is a natural candidate for an emulator.

To illustrate, we present some results using an emulator for climate sensitivity based on an ensemble of 297 evaluations of HadSM3. 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 the abscissae of an 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 $\boldsymbol{x}^*$: (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 parameterisation. A value of $m = 10^4$ in (3.4) was sufficient to make these estimated distribution functions precise. Treated as a simple sensitivity analysis, this illustration shows that the choice of prior for $\boldsymbol{x}^*$ has an impact of about 2K on the 90th percentile. In a more sophisticated analysis, in which the prior for $\boldsymbol{x}^*$ is calibrated with observations, the choice of prior is not likely to be as influential.

4. Experimental design

Once we have liberated the choice of X from any particular sampling scheme, we can choose our evaluations to learn about $g(\cdot)$ in an informative way. We refer

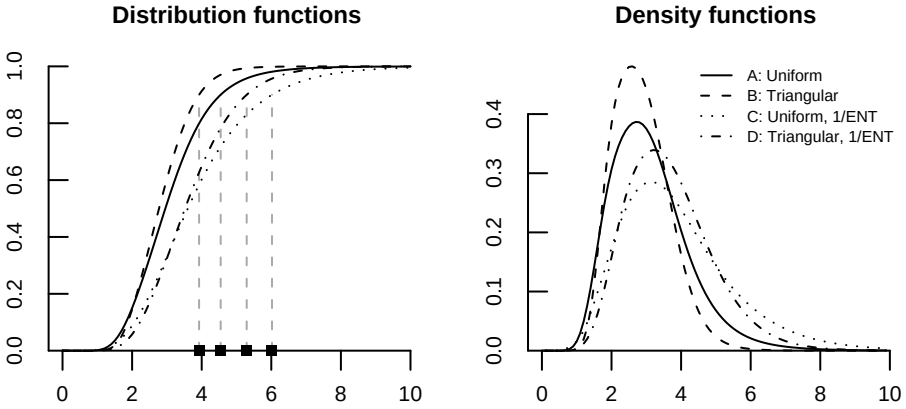

Figure 3. HadSM3 climate sensitivity (K), conditional on an ensemble of 297 evaluations, for four different choices of distribution for F_{x^*} (see text in §3). On the horizontal axis of the lefthand panel the filled squares indicate the four 90th percentiles.

to this as a *Designed ensemble,* as the general approach is informed by *Bayesian Experimental Design* (Chaloner and Verdinelli, 1995); more detailed 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 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). Another option which combines stages (1) and (2) is to generate a screening design, and then assign the likely active inputs to the best subset in the design, e.g., the D-optimal subset.

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 of the model-input space for which the predictive uncertainty, i.e. $\mathrm{Sd}[g(\boldsymbol{x}) \mid Y; X]$, 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.

Our HadSM3 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 HadSM3 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.

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). 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, which is valuable where there is no consensus about what an appropriate choice might be.

J.C. Rougier has been partly funded by NERC, under the RAPID Directed Programme. We would like to thank Michael Goldstein, Peter Craig, Jeremy Oakley, James Annan, and the referees for very helpful observations.

References

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

DeGroot, M. H., Schervish, M., 2002. Probability and Statistics, 3rd Edition. Reading, Mass.: Addison-Wesley Publishing Co.

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.

Gregory, J., 1999. Representation of the radiative effect of convective anvils. Tech. Rep. 7, Met. Office.

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.

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.

Liu, J., 2001. Monte Carlo Strategies in Scientific Computing. New York: Springer.

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., Buck, C. E., Daneshkhah, A., Eiser, J., Garthwaite, P., Jenkinson, D., Oakley, J., Rakow, T., 2006. Uncertain Judgements: Eliciting Expert Probabilities. Chichester: Wiley.

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

Rougier, J., 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations, forthcoming in Climatic Change.

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.