Emulating the sensitivity of the HadSM3 climate model using ensembles from different but related experiments

Jonathan Rougier[*]
Department of Mathematical Sciences
Durham University, UK

David M.H. Sexton and James M. Murphy
Hadley Centre
Met Office, UK

David Stainforth
Department of Physics
University of Oxford

October 2, 2006

Abstract

We combine the ensembles from two different experiments to study the climate sensitivity of the HadSM3 climate model subject to parametric uncertainty. We use a statistical framework based around linked emulators, where expert judgements are required to quantify the relationship between the two ensembles. Detailed diagnostics are presented. As an application, we perform a sensitivity analysis for various choices of distribution describing parametric uncertainty.

KEYWORDS: QUMP, CPNET, Bayesian, emulator, parametric uncertainty

1. Introduction

When we use models as the basis for inference about an underlying system, like climate, there are three sources of uncertainty to account for (Goldstein and Rougier 2004; O'Hagan 2006). First, there is uncertainty about the relationship between our particular implementation of the model, which we term the *simulator*, and the system

[*] Corresponding author: Department of Mathematical Sciences, Durham University, Science Laboratories, Durham DH1 3LE, UK; email J.C.Rougier@durham.ac.uk

itself. Second, there are usually measurement errors in system data used to calibrate the simulator. Third, there are technology and budget constraints that prevent us from evaluating our simulator as much as we would like. In this paper we present a statistical treatment of this third source of uncertainty, termed *code uncertainty* by O'Hagan, which involves using the evaluations we do have to construct an *emulator* for our simulator.

In general, an emulator is a stochastic representation of a function; that is to say, it allows us to predict the output of a function at any point in the input space. If we write our function as

$$x \to g(x)$$

where we will be treating x as a vector of variables and $g(x)$ as a scalar, then the emulator of $g(\cdot)$ consists of the probability distribution function

$$F_{g(x)}(v) \triangleq \Pr\bigl(g(x) \leq v\bigr) , \tag{1}$$

that is, the probability that $g(\cdot)$ when evaluated at x returns a value less than v.

In this paper, we develop and use such an emulator to predict the response of a complex climate model. The response is the equilibrium change in globally averaged surface temperature following a doubling of the atmospheric concentration of CO_2. This quantity is referred to as the *climate sensitivity*, and represents a standard benchmark of the response of climate to increases in greenhouse gases. The climate model consists of the HadAM3 atmospheric general circulation model (Pope et al. 2000) coupled to a simple non-dynamic mixed layer ocean, a standard set-up for the simulation of climate sensitivity. We refer to it hereafter as HadSM3. In common with other climate models, HadSM3 contains many poorly constrained parameters, which represent the effects of sub-grid scale physical processes such as cloud formation, convection, radiative transfer, and turbulent boundary layer mixing. Here we use our emulator to explore the variation of climate sensitivity according to thirty-one HadSM3 parameters controlling key physical processes in the model. Thus our $g(\cdot)$ is climate sensitivity, and our x is a thirty-one dimensional vector of the parameter values; precise definitions are given in section 2.

For inferential purposes we would like to know the value of HadSM3's sensitivity at any value for x. However, HadSM3 is a very expensive model to run, and we have available only a limited number of evaluations, at input values $X \triangleq \{x_1, \dots, x_n\}$, giving rise to the collection of outputs $y \triangleq \{g(x_1), \dots, g(x_n)\}$. Together, we term these the *ensemble* of evaluations, $(y; X)$. The key feature of an emulator is that it quantifies the uncertainty that arises from having only a limited number of evalutions. Thus if x is very close to a point $x_i \in X$ we might be relatively certain about $g(x)$, which we might expect to be close to $g(x_i) \in y$. On the other hand, if x is a long way from the evaluations in X then we might expect to be very uncertain about the value of $g(x)$. Emulator construction is discussed in section 3.

We use an emulator wherever we would like to use the underlying model $g(\cdot)$, but are prevented for reasons of cost: if models were costless to evaluate then emulators would be redundant. Thus the emulator can be used to answer simple questions like "I wonder what would happen if we evaluated $g(\cdot)$ at x?". The answer would be a probability distribution for $g(x)$, which we could use directly, or which we could summarise in terms of the mean and standard deviation, or in terms of an interval. By extension, we can use the emulator to analyse the impact of uncertainty in the 'correct' value for x, which we might term x^*. If we attach a probability distribution to x^*, then $g(x^*)$ is also an uncertain value. What is often not appreciated is that our uncertainty about $g(x^*)$ comes from *two* sources: uncertainty about x^* and uncertainty about $g(\cdot)$. This will be discussed in more detail in section 5, where we treat prediction as an application of emulators. Another important role for emulators, one which will be the focus of this paper, is that they allow us to combine information from different but related experiments, because they provide a natural way for us to quantify our judgement of the degree to which the experiments are similar. Section 4 shows how we use emulators to combine information from two experiments on HadSM3.

One point that must be stressed right at the start is that we are using probability in this paper to quantify our uncertainty. This makes our approach *Bayesian*. Throughout the paper we will be exercising our judgement to create the best emulator that we can, subject to various constraints such as transparency and tractability; we favour these constraints because they allow our approach to be easily replicated. In no sense could our approach be described as 'objective'. Where we make choices we state them clearly and we back them up with diagnostic information. But we do not claim that these choices are uniquely acceptable across the whole spectrum of climate experts, and consequently our results are very much *our* results. There is no single best emulator for HadSM3, and there is no single best probability distribution for HadSM3's climate sensitivity. What we aim to do here is to provide a framework within which it is possible to work out a number of different choices, and illustrate one particular choice, namely our own.

2. Two experiments on HadSM3

Two recent high-profile studies have attempted to quantify our uncertainty about the climate sensitivity in a CO_2 doubling experiment using HadSM3: an atmospheric model coupled to a mixed-layer ocean. This section outlines these two experiments, and the resulting ensembles of evaluations. Details of the two studies can be found in the original papers and their Supplementary Information; here we summarise those aspects that are relevant for the statistical analysis.

2a. *The QUMP study*

In the Quantifying Uncertainty in Model Predictions (QUMP) experiment of Murphy et al. (2004), thirty-one model parameters were identified as being potentially important, out of a possible 100 or so candidates. These thirty-one will be referred to as *variables*, and they are described in Table 1, which also gives the short names by which they will be referred in this paper. Thirteen of the variables are *factors*, i.e. variables that take values in a discrete set. Most of the factors have 2 levels, but two have 3 levels (`GWST` and `NFSL`) and one has 4 levels (`FRF`). Of the eighteen continuous variables, four are contingent on the setting of certain factors; for example, the value of `RHCV` only affects sensitivity when `RHC` is '`off`'; these contingent variables are the reason that Murphy et al. (2004) count twenty-nine rather than thirty-one variables in their description (they did not include `CAPE` and `ANV`).

We denote a particular choice for the values of the variables as x. The sensitivity at x was computed in a three-phase experiment. The first phase was a 25-year calibration run to deduce the appropriate ocean heat flux convergence field to be used in the subsequent two phases. The second and third phases were runs to equilibrium, once with pre-industrial CO_2, and once with doubled CO_2. Sensitivity, or $g(x)$, was defined as the difference in global mean temperature between the second and third phases. The choice of variables in the original study was strongly determined by the initial belief that the $1/$sensitivity was additive in the factors, and additive in simple terms in each of the the continuous variables. Consequently the initial evaluations in the ensemble consisted of single parameter perturbations, augmented by a small number of multi-parameter perturbations. Since that original study, we have access to a further 231 evaluations, all multi-parameter perturbations. The first 128 of these are described in Webb et al. (2006). The additional evaluations were initially chosen to restore balance to the overall ensemble, and then subsequently to populate regions of the parameter space which were thought to have important interactions. These can be added directly to the original ensemble, to give the 297 evaluations that we use in this paper.

2b. *The CPNET study*

Here we focus on the differences between QUMP and the `climate`*prediction*`.net` (CPNET) experiment of Stainforth et al. (2005). The CPNET study varied six of the continuous variables, used in the processes for large scale clouds and convection. Their ensemble comprises a factorial design with five variables at three levels (`VF1`, `CT`, `CW`, `RHCV`, `ENT`; `RHC` was always ‘Off’) and one at two levels (`CFS`). All the other variables in Table 1 are set at their standard values. However, each x was evaluated with a number of different initial conditions, introducing a structured source of uncertainty that is not present in the QUMP study. On analysing the CPNET data, we find that the choice of initial condition does not appear to be predictively important, and so we pool the evaluations across the initial conditions, effectively discarding the extra information that is present in the choice of initial conditions; a similar approach was used in the CPNET study, where different initial conditions for the same x were averaged, to reduce variability.

The CPNET study adopted a Public Resource Distributed Computing (PRDC) approach, performing thousands of evaluations using spare cycles on volunteers’ home and office computers. Within this approach it was not feasible to integrate HadSM3 to equilibrium twice. Instead, three phases of fifteen years each were used. The third phase in particular was too short to establish equilibrium, and so an exponential curve was fitted to global mean temperature in this phase, and then extrapolated to its horizontal asymptote to give a point value for sensitivity. Comparing these simplifications with the QUMP study, assisted by some direct comparisons, we judge that there are sufficient differences that it is not possible to combine the two ensembles directly, or indirectly by reweighting the CPNET ensemble, but that they are in fact two different but related experiments. This informs our statistical modelling choices in section 4.

In our sample from the CPNET study we have a total of $3^5 \times 2^1 = 486$ dis-

Table 1: Description of the QUMP variables. Comparable to Murphy et al. (2004), Supplementary Information, Table 2. Values in parentheses indicate ‘low’, ‘intermediate’ and ‘high’ values of continuous variables. Values not in parentheses indicate levels of discrete variables, or *factors*. Bold values indicate the standard setting. Variables with short names followed by ‘†’ are also used in CPNET.

Parameter / Property	Values	Short name	Only when:
Large-scale cloud			
V_{f1} (ms^{-1})	(0.5, **1**, 2)	`VF1`†	
C_t ($\times 10^{-4}$ s^{-1})	(0.5, **1**, 4)	`CT`†	
C_w (land, $\times 10^{-4}$ kg m^{-3})	(1, **2**, 10)	`CW`†	
Flow-dependent Rh_{crit}	**Off**, On	`RHC`	
Rh_{crit}	(0.6, **0.7**, 0.9)	`RHCV`†	`RHC` ‘Off’
Cloud fraction at saturation (%)	(**0.5**, 0.7, 0.8)	`CFS`†	
Vertical gradient of cloud water	**Off**, On	`VGCW`	
Convection			
Entrainment rate coefficient	(0.6, **3**, 9)	`ENT`†	
CAPE closure	**Off**, On	`CAPE`	
CAPE closure time-scale (hrs)	(1, 2, 4)	`CAPEV`	`CAPE` ‘On’
Convective anvils	**Off**, On	`ANV`	
Convective anvils, shape	(1, 2, 3)	`ANVS`	`ANV` ‘On’
Convective anvils, updraught	(0.1, 0.5, 1)	`ANVU`	`ANV` ‘On’
Sea ice			
Sea ice albedo (at 0 °C)	(**0.50**, 0.57, 0.65)	`SIA`	
Ocean-ice diffusion ($\times 10^{-4}$ m^2 s^{-1})	(0.25, 1.00, **3.75**)	`OID`	
Radiation			
Ice particle size (μm)	(25, **30**, 40)	`IPS`	
Non-spherical ice particles	**Off**, On	`NSIP`	
Shortwave water vapour continuum absorption	**Off**, On	`SWV`	
Sulphur cycle	**Off**, On	`SCYC`	
Dynamics			
Order of diffusion operator	4, **6**	`ODD`	
Diffusion e-folding time (hrs)	(6, **12**, 24)	`DDTS`	
Starting level, gravity wave drag	**3**, 4, 5	`GWST`	
Surface gravity wave wavelength ($\times 10^4$ m)	(1, 1.5, **2**)	`GWWL`	
Land surface			
Surface-canopy energy exchange	**Off**, On	`SCEE`	
Forest-roughness lengths	**1**, 2, 3, 4	`FRF`	
Dependence of stomatal conductance on CO_2	Off, **On**	`STOM`	
Number of forest soil levels for evapotranspiration (grass)	1, 2, **3**	`NFSL`	
Boundary layer			
Charnock constant ($\times 10^{-3}$)	(**12**, 16, 20)	`CHAR`	
Free convective roughness length over sea ($\times 10^{-4}$ m)	(2, **13**, 50)	`FCRL`	
Boundary layer flux profile, G_0	(5, **10**, 20)	`BLFP`	
Asymptotic neutral mixing length, λ ($\times 10^{-2}$)	(5, **15**, 50)	`ANML`	

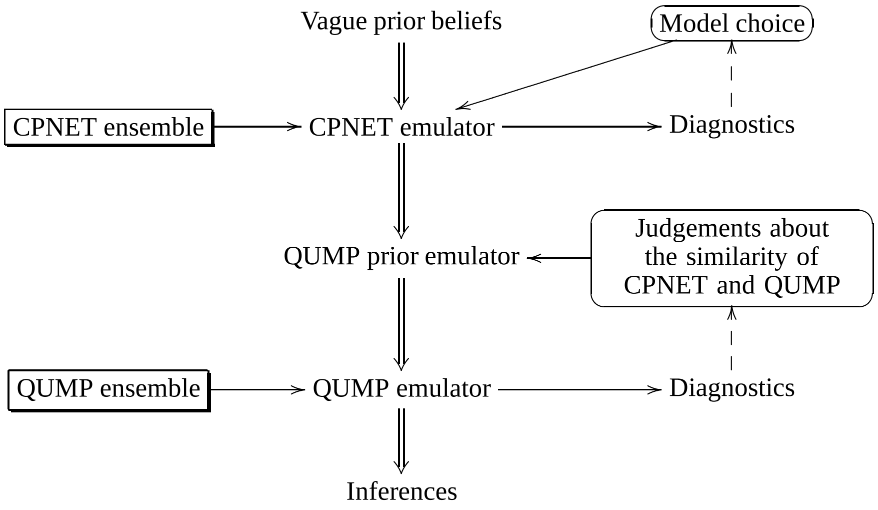

Figure 1: The main stages of our approach for combining information from the CPNET and QUMP ensembles into an emulator for QUMP sensitivity.

tinguishable evaluations (in terms of the x values), and 2377 evaluations overall (accounting for variations in the initial conditions). Many of these produced unstable or non-physical responses, particularly cooling ('drifters'). We choose to omit the drifters in the CPNET study in the same way as Stainforth et al. (2005). Some of the QUMP evaluations also display this type of cooling in the early stages, but so far all of these have equilibriated and the $2 \times CO_2$ run always remains warmer than the $1 \times CO_2$ run, both of which start from the end point of the calibration phase.

2c. *Outline of our approach*

The two studies outlined in this section have different but complementary strengths. The QUMP study has a 'standard' definition for sensitivity, and provides greater flexibility for future inferences through its large number of variables. The CPNET study, on the other hand, has a more detailed analysis over six of the most important variables. Our intention is to combine the ensembles from these two studies into an emulator for QUMP sensitivity defined over the full set of thirty-one variables.

As already described, an emulator is a probability distribution function $F_{g(x)}$, as defined in (1). There are many ways of coming up with such a function; in a *Bayesian emulator* we probabilistically condition our beliefs about $g(\cdot)$ on the observations in the ensemble. Therefore a Bayesian emulator combines two sources of information: prior judgements about $g(\cdot)$, and data from evaluations in the ensemble $(y; X)$. The main stages of our approach are summarised in Figure 1. Each of the two studies requires a different emulator, because of the different definitions of sensitivity. For the CPNET emulator we have plentiful information from the CPNET ensemble, which comprises 421 evaluations in a six-dimensional space. Therefore we start with only vague prior information, because we are content to let the information from the ensemble dominate. For the QUMP emulator, on the other hand, we have only limited information in the ensemble (297 evaluations in a 31-dimensional space). Therefore we combine this with detailed prior information taken from the CPNET emulator, and from our judgement concerning the similarity of the CPNET and QUMP definitions of sensitivity. Figure 1 also shows two diagnostic loops: wherever we have data, we can investigate the propriety of our choices and, to a limited extent, we can modify those choices. This is discussed in more detail in section 4d.

3. Building an emulator from the CPNET ensemble

We illustrate our approach, as outlined in Figure 1, in this section and the next. In this section we develop an emulator for CPNET sensitivity with vague prior information. In section 3a we describe a simple and general framework for specifying an emulator, and in section 3b we make specific choices within that framework to construct an emulator for CPNET sensitivity.

3a. *A general Bayesian emulator*

We describe here a simple Bayesian treatment of the emulator. We impose a certain structure on the general problem of constructing an emulator, as this helps us to define clearly the choices we must make. The emulator is written

$$g(x) = h(x)^T \beta + u(x) \tag{2}$$

where $g(x)$ is the climate sensitivity of our simulator, or some monotonic transformation of the same, termed the *response*; $h(\cdot)$ is a known vector-valued function of the variables, collectively termed the *regressors*; β is an unknown vector of *(regression) coefficients*, and $u(x)$ is a scalar random field, termed the *residual*. Within the regressors we would expect to include non-linear functions of the variables, such as ${x_i}^2$ or $x_i \times x_j$. We must use our judgement, in conjunction with the data where possible, to make choices for the transformation of $g(\cdot)$ and the components of $h(\cdot)$: statistical model choice is a subtle balancing-act between fidelity, efficiency and 'interpretability'—much the same is true of building climate models. The challenge becomes greater as the number of components in x goes up, because the range of possible terms for inclusion among the regressors becomes much larger, and it becomes difficult to contrast alternative choices in terms of standard diagnostics like residual behaviour.

In our ensemble, y is an n-vector of climate sensitivities, possibly transformed, and X is the $n \times p$ design matrix in which row i comprises the values of the variables in the i^{th} evaluation in the ensemble. From this design matrix we can compute the $n \times k$ regression matrix H, where row i of H comprises $h(X_i)^T$. For our given choice for the response and the regressors, we make the following additional choices for the structure of the residual, which serve to simplify the analysis. First, $u(\cdot)$ has zero mean and a constant unknown variance, σ^2; second, the correlation length of $u(x)$ at any x is short compared with the inter-point distances in our ensemble, so that we may treat the observed values of the residual as independent for the purposes of constructing the emulator (in spatial statistics this type of residual is often termed a 'nugget'); third, $u(\cdot)$ is a *gaussian* random field for given σ^2.

These choices allow us to use the standard *conjugate* analysis, i.e. an analysis where the prior and the posterior come from the same family of distributions, so that the update may be described in terms of alterations to distributional parameters. From this we can derive a simple expression for the distribution function $F_{g(x)}$ based on our ensemble $(y; X)$ and our prior assessment of uncertainty concerning the parameters (β, σ^2). The treatment of the residual as a nugget is non-standard. Technically it is inconsistent with the fact that $g(\cdot)$ is a deterministic continuous function, at least on part of its domain, because it prevents $\text{Corr}\big(g(x), g(x')\big) \to 1$ as $x \to x'$. Our attitude is that as long as the residual does not play a large part in the emulator, this type of misspecification is unlikely to be predictively important. In our emulators of QUMP sensitivity we find that the regression R^2 is at least 90% and typically more than 95%. The corresponding R^2 values for CPNET are lower (70–90%), but we are less concerned about the residual behaviour in the CPNET emulator, because the CPNET ensemble is less intensively used. There is an extensive literature on more general types of emulator (see, e.g., Currin et al. 1991; O'Hagan et al. 1999; Kennedy and O'Hagan 2001; Craig et al. 2001; Santner et al. 2003), and these emulators could be deployed in our approach, but only at the expense of much more intricate statistical modelling.

The following outline of the conjugate analysis follows the notation of O'Hagan and Forster (2004, ch. 11). Our prior for $\{\beta, \sigma^2\}$ is Normal-Inverse-Gamma (NIG)

$$(\beta, \sigma^2) \sim \mathit{NIG}\big(a, d, m, V\big) \tag{3a}$$

or, equivalently,

$$\beta \mid \sigma^2 \sim N_k\big(m, \sigma^2 V\big) \quad \text{and} \quad \sigma^2 \sim \mathit{IG}\big(a, d\big) \tag{3b}$$

where '$|$' denotes 'conditional upon', $N_k\big(\cdot\big)$ denotes the k-dimensional Gaussian distribution, and $\mathit{IG}\big(\cdot\big)$ the scalar Inverse Gamma distribution; we must specify the collection $\{a, d, m, V\}$, termed the *hyperparameters*. We have some concerns about the shape of the NIG prior as a representation of our beliefs (e.g., it is not possible to specify that β and σ^2 are probabilistically independent, except in the non-informative case that will be presented below in section 3b), and we adopt it here because in our judgement these concerns are outweighed by its tractability; O'Hagan and Forster (2004, second half of ch. 11) discuss the shape of the NIG distribution in detail, and present various generalisations.

We update using our ensemble by applying Bayes's theorem; the posterior distribution $\{\beta, \sigma^2\} \mid (y; X)$ remains NIG, with updated parameters $\{a^*, d^*, m^*, V^*\}$, where

$$V^* \triangleq (V^{-1} + H^T H)^{-1}, \tag{4a}$$

$$m^* \triangleq V^*(V^{-1}m + H^T y), \tag{4b}$$

$$a^* \triangleq a + m^T V^{-1} m + y^T y - (m^*)^T (V^*)^{-1} m^*, \tag{4c}$$

$$\text{and } d^* \triangleq d + n \tag{4d}$$

(O'Hagan and Forster 2004, sec. 11.10). For our emulator, we use the posterior predictive distribution for $g(x)$ at known x, which is univariate Student-t:

$$g(x) \sim \mathsf{t}_{d^*}\big(h(x)^T m^*,\ (a^*/d^*)w^*(x)\big) \tag{5}$$

providing that $x \notin X$, where $w^*(x) \triangleq h(x)^T V^* h(x) + 1$. For clarity (5) states that

$$\frac{g(x) - h(x)^T m^*}{\sqrt{(a^*/d^*)w^*(x)}}$$

has a standard Student-t distribution with d^* degrees of freedom, and

$$\mathrm{E}\big(g(x)\big) = h(x)^T m^*, \qquad \mathrm{Var}\big(g(x)\big) = \frac{a^*}{d^* - 2} w^*(x).$$

Standard statistical software can compute the distribution function $F_{g(x)}(v)$ for any x and v. All the calculations in this paper were performed using the statistical computing environment R (R Development Core Team 2004).

Therefore the problem of building an emulator for $g(\cdot)$ using the ensemble $(y; X)$ has been restructured to the problem of choosing a transformation for sensitivity, a collection of regressors $h(\cdot)$, and, conditional on these choices, specifying the hyperparameters $\{a, d, m, V\}$ in the NIG prior for $\{\beta, \sigma^2\}$. The two 'big' choices that we have made in this framework are to treat the residual as a nugget, and to adopt a NIG prior. We would be interested in tractable generalisations of either of these choices, but we are satisfied that these are reasonable choices for this application, not just *a priori*, but also in the light of the diagnostic information presented below.

3b. *Building the CPNET emulator*

As explained in section 2c, and illustrated in Figure 1, we are going to simplify the construction of our CPNET emulator by adopting vague prior beliefs, which in terms of the framework from section 3a are vague prior beliefs about $\{\beta, \sigma^2\}$, as summarised in the hyperparameters $\{a, d, m, V\}$. The standard *non-informative prior* has $a = 0$, $d = -k$ where k is the number of regressor functions in $h(\cdot)$, and $V^{-1} = \mathbf{0}$ (O'Hagan and Forster 2004, sec. 11.17–11.19). In this case the posterior distribution for $\beta \mid \sigma^2$ has the classical Ordinary Least Squares (OLS) form—as can be seen from inspection of the updating relations in (4)—although the interpretation is a little different, being Bayesian rather than Frequentist. In particular, the coefficient standard deviations are direct statements of coefficient uncertainty, rather than 'standard errors' arising from a repeated-sampling approach that considers the data themselves to be the source of 'randomisation' (a nonsensical concept in this context). In this paper, when we refer to, say, a 90% CI we are referring to a 90% 'Credible Interval': an interval defined by the 5^{th} and 95^{th} percentiles of the distribution of the coefficient, or $g(x)$, or any other uncertain quantity (O'Hagan and Forster 2004, sec. 2.51).

With this prior, we deploy exactly the same techniques that would be used in a standard analysis to fit an OLS regression (see, e.g., Draper and Smith 1998). In particular, we choose the transformation of y and the regressors together, and we use the residuals for diagnostic information. The QUMP authors, who explicitly construct an emulator for their analysis, choose the transformation $1/y$, based on a general view across the modelling community that this function has a simpler additive structure in terms of the variables. This would only be a reasonable transformation if negative values for sensitivity were judged highly unlikely at any x, because otherwise it would introduce an extreme discontinuity at zero. We subscribe to this view, but we will investigate a wider range of possible power-transformations, including the logarithm, using the Box and Cox (1964) approach (see, e.g., Draper and Smith 1998, sec. 13.2).

For the regressors, the QUMP authors chose linear additive terms for the factors and piecewise linear terms for the continuous variables. We will replace the piecewise linear terms with quadratics—which requires the same number of regression coefficients—as there is no compelling reason to think that HadSM3 has a discontinuous first derivative at the standard setting of its variables. We also choose to take logarithms of some of the continuous variables, namely those for which the intervals in Table 1 have strong positive skewness; this slightly improves the fit of the emulator and reduces the role of the squared terms, making it easier to interpret the emulator coefficients (given below in Table 2). The variables transformed in this way are `VF1`, `CT`, `CW`, `ENT`, `DDTS`, `FCRL`, `BLFP` and `ANML`; only the first three of these are relevant for the CPNET study.

We would like our emulator to include interactions among the variables. In the QUMP study it was not possible to estimate interactions from the single-parameter perturbation ensemble, but they were found to be influential in CPNET. Our general strategy regarding interactions is to treat the different physical processes as non-interacting (these processes are shown in Table 1), but to include interactions within each process. Our starting point is to include all two-way interactions in the five CPNET variables in the 'Large Scale Cloud' block, giving a total of

$$1 + \underbrace{6 + (6-1)}_{\text{linear and quad.}} + \underbrace{5 \times 4/2}_{\text{two-way int.}} = 22$$

regression coefficients. The $6 - 1$ is for the quadratic terms: we cannot estimate a quadratic for `CFS` because it only has two levels in the CPNET ensemble. For the same reason we cannot estimate cubic or higher effects in any of the variables. A statistician would not have recommended this type of design for the CPNET study, or, indeed, recommended single-parameter perturbations for the QUMP study, although it must be borne in mind that these types of ensemble study attempt to fulfil a number of different and not necessarily compatible objectives.

Based on this regression, the Box-Cox approach indicates that $\log(y)$ is a good choice for the transformation of the response; the typical diagnostic for this approach is shown in Figure 2. This is a fortuitous outcome, because this particular transformation automatically assigns zero probability to negative sensitivities in the predictive distribution of the emulator. In an earlier analysis the reciprocal had been favoured, which required us to truncate the predictive distribution. This truncation was not a particularly elegant solution, but in practice it made little difference because for most values of x, almost all of the probability mass in the predictive distribution was above zero.

We do not want to rule out the possibility of higher-order interactions as well. There

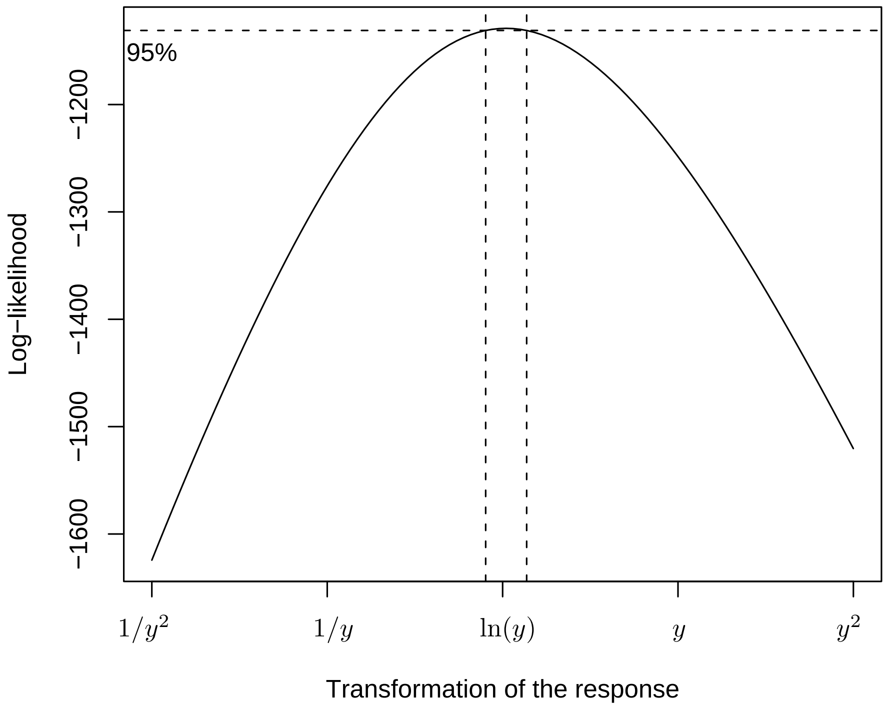

Figure 2: Box-Cox plot to select an appropriate transformation for the response: the logarithm is strongly indicated.

are too many of these to include them all up to a given order, and so we use forward stepwise regression based on the Akaike Information Criterion (AIC) (see, e.g. Draper and Smith 1998, ch. 15) to identify the most important terms among all possible two-, three- and four-way interactions, including interactions between ENT and the 'large Scale Cloud' variables. We do not have strong views about the presence or absence of interactions among these six variables, and so this simple and fairly standard technique seems adequate; had we stronger views we could have adopted a Bayesian hierarchical approach (see, e.g., Chipman et al. 1997). We find fifteen further interactions, namely (in order of acceptance) RHCV:ENT, CT:ENT, CW:ENT, CFS:ENT, CT:CW:ENT, CT:CW:CFS, CT:CW:RHCV, CW:RHCV:ENT, CT:RHCV:ENT, CT:CFS:ENT, VF1:ENT, VF1:RHCV:ENT, VF1:CW:ENT, VF1:CT:CW, and VF1:CT:ENT. We include these higher-order interactions in $h(\cdot)$, but we do not include any others. This gives a total of 37 regressor functions in $h(\cdot)$, including the intercept.

As they may be of some independent interest, the regression coefficients for our CPNET emulator are given in Table 2, along with their standard deviations. The six variables have been re-scaled to lie in the closed interval $[-1, 1]$, according to the minimum and maximum values given in Table 1; this range was chosen rather than, say, the original units or $[0, 1]$, because it makes the linear and quadratic functions orthogonal with respect to a uniform weighting function. There are some influential two-way interactions, and the three-way interactions tend to be the same size as the typical two-way interactions. There is strong evidence here for the importance of interactions in determining HadSM3's sensitivity.

4. An emulator for QUMP sensitivity

Having built an emulator for CPNET sensitivity, we turn now to using this emulator as prior information for our emulator for QUMP sensitivity. We approach this in two stages. First, we construct a prior emulator for QUMP sensitivity. Initially, we must choose a collection of regressors for the QUMP emulator: these will be a superset of the regressors for the CPNET emulator, as QUMP has twenty-five additional variables. Our prior beliefs about QUMP sensitivity are then summarised in terms of hyperparameters $\{a, b, m, V\}$. With the CPNET emulator these hyperparameters took non-informative values, but for the QUMP emulator they will have informative values based on the updated hyperparameters from the CPNET emulator and on our judgement regarding the similarity of the CPNET and QUMP sensitivities. The way we choose to quantify these judgements is discussed in section 4b.

In the second stage we will update these parameters using the QUMP ensemble to give us the posterior values $\{a^*, b^*, m^*, V^*\}$. These form the basis of our QUMP emulator.

4a. *The regressors*

For our QUMP emulator regressors, we start with all those regressors in the CPNET emulator (37 in number) plus the missing quadratic term in CFS. We add all the factors from the QUMP study, and linear and quadratic terms for the new continuous variables.

Table 2: Coefficients from the CPNET emulator ($\times 10^3$). `VF1`, `CT` and `CW` are in logarithms, and all variables are standardised to the interval $[-1, 1]$. Linear terms are shown as `A`, interactions as `A:B` or `A:B:C`, and quadratic terms as `A:A`. The response is $\log(\text{sensitivity})$ and the R^2 is 0.87.

Regressor	Mean	St. dev.
`(Intercept)`	1147.8	30.4
`VF1`	−158.7	11.5
`CT`	283.1	13.0
`CW`	−142.3	12.2
`RHCV`	70.5	12.1
`CFS`	−166.0	12.8
`ENT`	−149.0	13.1
`VF1:VF1`	46.6	15.9
`CT:CT`	−88.6	18.8
`CW:CW`	−66.5	22.8
`RHCV:RHCV`	−4.8	17.8
`ENT:ENT`	239.0	16.3
`VF1:CT`	−21.8	11.8
`VF1:CW`	25.6	11.4
`VF1:RHCV`	−27.6	11.5
`VF1:CFS`	3.0	13.8
`CT:CW`	56.8	13.8
`CT:RHCV`	84.8	12.2
`CT:CFS`	25.4	15.0
`CW:RHCV`	−78.7	12.2
`CW:CFS`	−17.2	13.9
`RHCV:CFS`	21.6	13.9
`RHCV:ENT`	−92.2	13.3
`CT:ENT`	−138.5	15.3
`CW:ENT`	86.3	12.8
`CFS:ENT`	85.0	14.7
`VF1:ENT`	−28.4	12.8
`CT:CW:ENT`	−78.6	13.9
`CT:CW:CFS`	−45.0	16.3
`CT:CW:RHCV`	48.0	13.5
`CW:RHCV:ENT`	42.8	15.2
`CT:RHCV:ENT`	−35.7	14.4
`CT:CFS:ENT`	−52.1	17.6
`VF1:RHCV:ENT`	61.1	14.5
`VF1:CW:ENT`	−35.5	14.2
`VF1:CT:CW`	−24.6	13.2
`VF1:CT:ENT`	−23.9	14.0

We would also like to include some additional two-way interactions. As outlined in section 3b, we choose to include all two-way interactions within each physical process, but we do not include any interactions between processes, bar those between `ENT` and the 'Large Scale Cloud' variables from the CPNET emulator. Taken together this gives

$$37+1+\underbrace{10\times 1+2\times 2+1\times 3}_{\text{QUMP factors}}+\underbrace{12\times 2}_{\text{new cont. vars}}+\underbrace{10+12+1+6+9+17+6}_{\text{new interactions}}=140$$

coefficients. Not all interactions are possible; e.g. `RHC:RHCV` is not possible because `RHCV` is only effective when `RHC` is 'Off'. The physical process 'Dynamics' has 9 interactions because `GWST` is a three-level factor; likewise 'Land Surface' has 17 interactions because `FRF` is a four-level factor and `NFSL` is a three-level factor.

4b. *Linking matched coefficients*

When constructing our prior for the QUMP emulator coefficients we distinguish between matched coefficients and new coefficients. The matched coefficients have a direct counterpart in the CPNET emulator. For example, the coefficients on `ENT` and `ENT:ENT` in the QUMP emulator match to corresponding coefficients in the CPNET emulator, but the coefficient on `IPS` in the QUMP emulator is a new coefficient, because `IPS` was not varied in the CPNET study, so that it does not feature in the CPNET emulator, except through its contribution to the constant.

We can express the extent to which we think that CPNET sensitivity and QUMP sensitivity are the same by specifying the degree to which the matched QUMP emulator coefficients are likely to deviate from their counterparts in the CPNET emulator. To quantify the relation between individual pairs of matched coefficients we use the general framework

$$\beta_i - c_i = (1+\omega_i)\,(\beta_i^0 - c_i) + (r_y/r_i)\,\nu_i \tag{6}$$

where β_i^0 and β_i are matched coefficients in the CPNET and QUMP emulators, respectively. Our uncertainty about β_i is induced by our uncertainty about β_i^0, and by the choices we make for the various terms on the righthand side of (6). Two of these terms are straightforward: r_y is the typical scale of the transformed response, and r_i the typical scale of the regressor. These are included so that we can treat both ω_i and ν_i as scale-free, remembering that the units of β_i^0 and β_i are 'response units per regressor units'. This makes it reasonable to use the same choices to link-up all the matched coefficients, if we so choose. The third term, c_i is a centring term for the two coefficients; for this application we will choose $c_i = 0$ for all coefficients, but in other applications a non-zero value might be preferred (see, e.g. Goldstein and Rougier 2006).

The two Greek terms, ω_i and ν_i, are the most important in (6). They represent independent mean-zero uncertain quantities, for which we must specify standard deviations. We will want to set $\text{Sd}(\nu_i)$ small, so just for the moment we treat ν_i as zero. In that case $\text{Sd}(\omega_i)$ controls the probability that $\beta_i - c_i$ has a different sign to $\beta_i^0 - c_i$. Setting $\text{Sd}(\omega_i)$ small relative to 1 would be akin to stating that β_i and β_i^0 were very similar. For example, setting $\text{Sd}(\omega_i) = 1/4$ would state that a change of sign in going from $\beta_i^0 - c_i$ to $\beta_i - c_i$ was judged to be a four-standard-deviation event; crudely, to have a probability of less than 3% if ω_i is unimodal (Pukelsheim 1994), we term this ‘very unlikely’. This is the value that we will choose for all matched coefficients. The second Greek term, ν_i, is included to ensure that β_i can be uncertain even when β_i^0 equals c_i with probability one. A small value is appropriate here, and we choose $\text{Sd}(\nu_i) = 1/20$ for all matched coefficients. With this value is is very unlikely that regressor i will explain more than one-fifth of the range of the QUMP emulator response in the case where $\beta_i^0 = c_i$.

4c. *The rest of the prior emulator*

The choices we make for m_i, $\text{Sd}(\omega_i)$ and $\text{Sd}(\nu_i)$ in section 4b allow us to infer the mean vector and the variance matrix of the matched coefficients in the QUMP prior emulator from the mean vector and variance matrix of the corresponding coefficients in the CPNET emulator. Before we can translate those into values for the hyperparameters m and Σ we must think about the residual process in the QUMP emulator.

We believe that the residual variance for the QUMP prior emulator will be less than that of the CPNET emulator, because the recorded value of sensitivity in the CPNET study includes an extra source of uncertainty, namely the asymptotic approximation to the equilibrium value. Therefore, for σ^2 in the QUMP prior emulator we choose a mean value half of that from the CPNET emulator, and choose a standard deviation equal to the mean, to preserve a large amount of uncertainty. To translate these choices into values for $\{a, d\}$ we need to know that the marginal distribution of σ^2 in the NIG distribution is *IG*$\big(a, d\big)$, and that the mean and variance of this distribution are given by $a/(d-2)$ and $2a^2/\{(d-2)^2\,(d-4)\}$, respectively. Denoting the CPNET hyperparameters with a subscript ‘0’, we can compute the mean value of σ^2 from the CPNET emulator as $s_0^2 \triangleq a_0/(d_0-2)$. Then we can solve $a/(d-2) = s_0^2/2$ and $2a^2/\{(d-2)^2\,(d-4)\} = (s_0^2/2)^2$ simultaneously for $\{a, d\}$. This gives $s_0^2 = 0.023$, $a = 0.035$ and $d = 5$.

Once we have computed $\{a, d\}$, we can use these two values along with the values $\{a_0, d_0, m_0, V_0\}$ to compute the mean and variance of the matched coefficients in the QUMP emulator. We need to know that the marginal distribution of β in the NIG distribution is multivariate Student-t, and that the mean and variance of this distribution are given by m and $\{a/(d-2)\}V$, respectively. Then it is a simple matter to compute the mean and variance of β^0, the CPNET emulator coefficients, use (6) to map these into a mean and variance for β, the QUMP prior emulator matched coefficients, and then infer $\{m, V\}$ from the values of $\{a, d\}$ that we computed above.

The only thing left to specify is a prior mean and variance for the unmatched coefficients in the QUMP emulator. These are the coefficients on regressors that do not appear in the CPNET emulator. For these coefficients we use a framework similar to (6), namely

$$\beta_i = (r_y/r_i)\,\nu_i\,. \tag{7}$$

This is just a way of assigning an uncertainty to each unmatched β_i in terms of the scale-free quantity $\text{Sd}(\nu_i)$. We have to decide how much of the response range we believe these additional regressor terms can explain. Our choice is $\text{Sd}(\nu_i) = 1/16$ for all the new coefficients, so that it is very unlikely that a single regressor can explain more than a quarter of the range of the response.

4d. *Prior diagnostics*

We have used the CPNET ensemble in two ways in constructing our prior emulator for the QUMP experiment. We have used it *indirectly*, to select the transformation of the response and to identify important third-order interactions in the Large-scale cloud parameters and the entrainment rate coefficient. We have also used it *directly*, to choose the prior hyperparameters of the matched coefficients. In the latter we have assigned specific values to quite imprecisely defined quantities. In an ideal world we would arrive at such values through introspection, but in practice it is impossible in a detailed analysis not to incorporate some trial-and-error. For example: originally, we had larger values for $\mathrm{Sd}(\omega_i)$ and $\mathrm{Sd}(\nu_i)$, because at that stage we were screening out fewer of the drifters from the CPNET experiment. These choices were broadly satisfactory in terms of the diagnostics described below. Now we have decided to screen out more of the drifters (see section 2b), we modify our choices, but we cannot escape the knowledge of how our previous choices performed. Statistical purists would regard this as a form of double-counting (the data influencing the prior), but a more pragmatic view is that simple revisions of this kind, taking care to avoid 'over-fitting', tend to approximate an informal type of higher-order learning that we have chosen not to include in the formal analysis.

Our main diagnostic is to use our QUMP prior emulator to predict the evaluations in the QUMP ensemble. Each individual prediction, taken marginally, has a Student-t distribution, as given in (5). In Figure 3 we show all 297 predictions, in terms of their median and 95% CI, and we also show the actual value in each case. The predictions are ordered by the median, which allows us to confirm that our assessment of the hyperparameters has some predictive power; i.e. that our predictions are not insensitive to the values for x. We can also confirm that there is no apparent systematic mis-prediction, with respect to the response. This diagnostic suggests that we have over-stated uncertainty, as all 297 values are well within the 95% CI that we predict. We could if we so chose, impose constraints on $\mathrm{Var}\big(g(x)\big)$, and use these to modify our statistical modelling of NIG hyperparameters such as V. However, we are comfortable with the general principles we have adopted in setting the prior for the QUMP emulator, and we prefer to leave things as they are, rather than to risk the suspicion that we have in any way over-tuned our prior.

Note that the cluster of low-dispersion points on the lefthand side of the bottom panel of Figure 3 correspond to the fifty or so single-parameter perturbations in the QUMP ensemble. We interpret the low dispersion of these points as evidence for the importance of interactions among the variables in determining QUMP sensitivity.

4e. *Updating the QUMP emulator*

Updating the QUMP emulator is a very simple process, following the rules given in (4). This updated emulator will be used in in an application in section 5; it is not detailed here because it has a large number of coefficients. It is also informative to investigate the ways in which the coefficients change following the update: we have done this but do not presented our results here for reasons of space.

We now have access to a second set of diagnostics, that investigate the posterior predictive properties of the QUMP emulator. One such diagnostic is broadly compara-

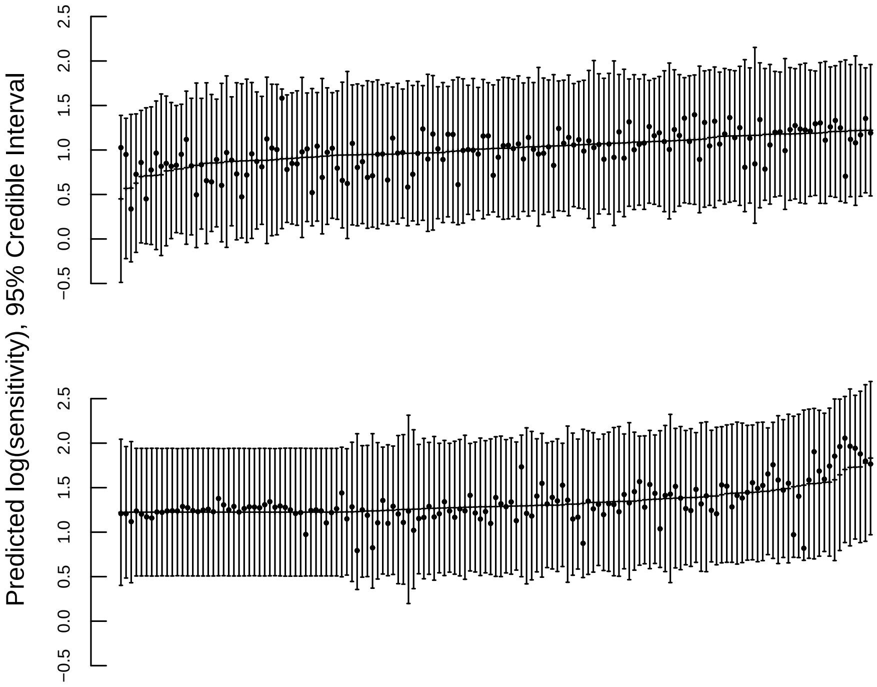

Figure 3: Prior prediction diagnostic showing, for each evaluation in the QUMP ensemble, the prior median and 95% CI, along with the actual value of the response (dot). The evaluations are ordered by the median.

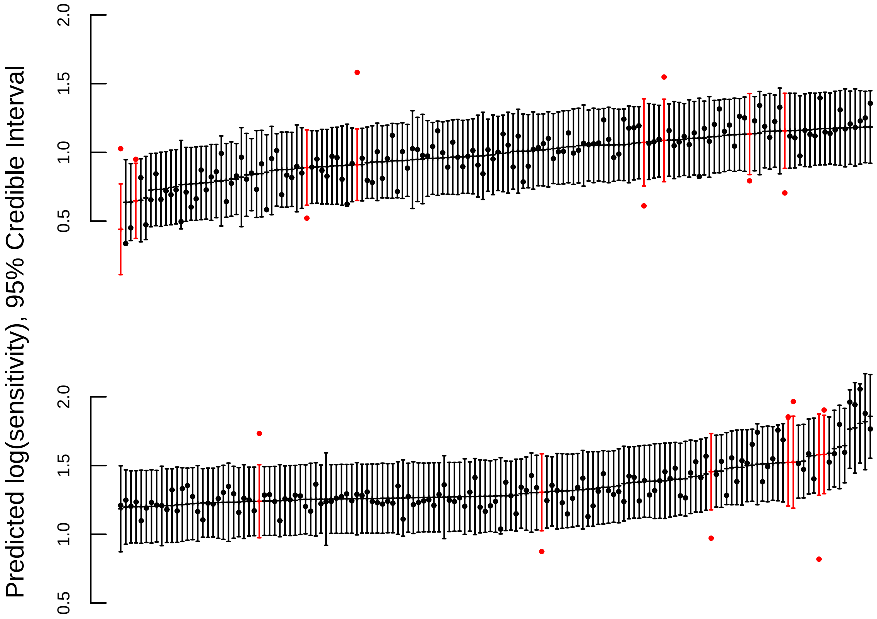

Figure 4: Posterior prediction 'leave-one-out' diagnostic showing, for each evaluation in the QUMP ensemble, the posterior median and 95% CI after updating with the other 296 evaluations. The evaluations are ordered by the median. Note that the vertical scale differs from that in Figure 3.

ble with the univariate prior prediction given in Figure 3: the leave-one-out diagnostic. In this case we update the emulator with all but one evaluation from the QUMP ensemble, and then predict that evaluation. We can do this with all 297 evaluations; the result is shown in Figure 4. In all, 15 of the 297 actual values for $\log(\text{sensitivity})$ lie outside the 95% CI of the posterior prediction. In terms of the binomial model, the probability of observing 15 or fewer successes out of 297 *independent* trials with $p = 0.05$ is 0.58, i.e. not unusual and therefore supportive of our statistical modelling choices; this is only suggestive, however, as our trials are not independent, because the predictions are correlated across the ensemble members.

A sterner diagnostic is to consider the multivariate behaviour of a collection of predictions, taking this correlation into account. For this purpose we select 150 evaluations, and update using the other 147 ('leave-150-out'). The joint distribution of all 150 predictions after updating should be multivariate Student-t, say $y_{150} \sim t_d\big(\mu, \Sigma\big)$. It follows that $y'_{150} \triangleq Q^{-T}(y_{150} - \mu)$ should have distribution $t_d\big(\mathbf{0}, I\big)$, where $Q^T Q \equiv \Sigma$; i.e., should have uncorrelated standardised components. Figure 5 show the result of one such random sample as a Quantile-Quantile plot (QQ-plot), and a histogram with the standard Student-t density overlaid: this is a fairly typical pattern across different possible random samples. Here it is clear from the QQ-plot in particular that there is some mis-fitting, but the differences appear to be relatively minor. These diagnostics appear to be broadly supportive of our statistical modelling choices.

5. The predictive distribution of sensitivity

We present here one application of our QUMP emulator: predicting sensitivity taking account of uncertainty in the correct parametrisation of the HadSM3 simulator. The notion of correct parameter values in this context, although widely used, is not straightforward, and has received some attention in the statistics literature on computer experiments (see, e.g., Kennedy and O'Hagan 2001; Craig et al. 2001; Goldstein and Rougier 2004, 2006); Rougier (2006) discusses the more general approach in the context of ensemble-based climate prediction. For simplicity, we proceed on the basis that such values exist. We define x^* as the vector of correct values, and we use the distribution function F_{x^*} to describe our uncertainty about x^*. Our purpose in this section is not to come up with a 'better' prediction for climate sensitivity than the model-based predictions currently in the literature, but simply to clarify that such a notion would require a consensus about F_{x^*}: something that does not currently exist.

5a. *Estimating the predictive distribution*

Our objective is to compute the (cumulative) distribution function for $\delta \triangleq g(x^*)$, namely

$$F_\delta(v) \triangleq \Pr(\delta \leq v) = \int_x \mathbf{1}\big(g(x) \leq v\big) \; dF_{x^*}(x) \tag{8}$$

where $\mathbf{1}(\cdot)$ is the indicator function. The notation $\int \cdots dF_{x^*}(x)$ denotes a *Lebesgue-Stieltjes* integral, which generalises the idea of expectation to include random quantities such as x^* which include both discrete and continuous components (see, e.g., Ross 1988, sec. 7.9). For any given value for v, (8) simply sums the probability content of the region of the input space for which $\delta \leq v$; i.e. our uncertainty about δ is a consequence of our uncertainty about x^*. This is referred to as the *prior predictive distribution* for δ, the "prior" in this case indicating "prior to the inclusion of actual system data (possibly measured with error) for calibration purposes". In this paper we will not be considering the effect of calibration, and so we may refer without ambiguity to (8) as the *predictive distribution*.

One way to calculate (8)—in principle—is by simple Monte Carlo integration, i.e. to compute

$$F_\delta^{(n)}(v) \triangleq n^{-1} \sum_{i=1}^{n} \mathbf{1}\big(y_i \leq v\big) \tag{9}$$

where $X \triangleq \{x_1, \ldots, x_n\}$ are sampled independently from the distribution F_{x^*}, the climate simulator is evaluated at each x_i, and y_i is the result in each case. By the

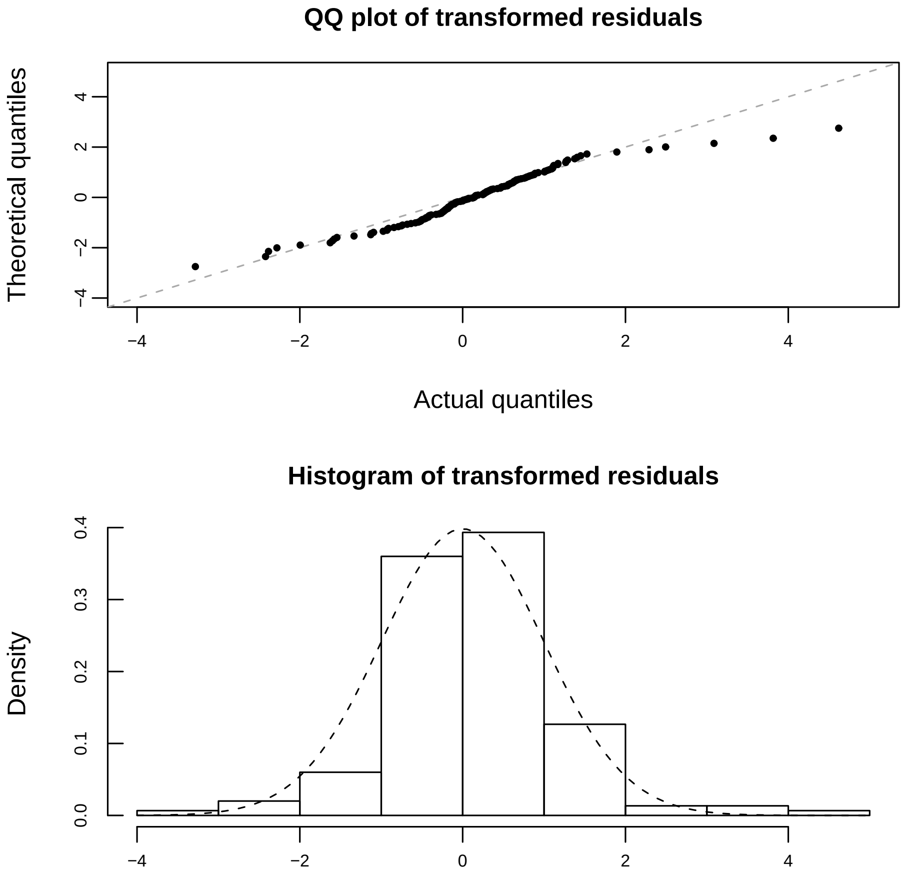

Figure 5: Joint diagnostics from the prediction of a random sample of 150 members of the QUMP ensemble. The theoretical distribution in each case is a standard Student-t with 152 degrees of freedom (effectively, a gaussian distribution).

Strong Law of Large Numbers (see, e.g., Grimmett and Stirzaker 2001, sec. 7.4),

$$\lim_{n\to\infty} F_\delta^{(n)}(v) = F_\delta(v).$$

This is the ‘standard’ approach to ensemble experiments in climate; the ensemble is used directly in the resulting inference. In many problems, however, $F_\delta^{(n)}(\cdot)$ will be a very uncertain approximation because of limitations in the size of n, the number of evaluations in the ensemble. For example, a back-of-the-envelope calculation suggests that with $n = 297$, Monte Carlo probability estimates will be accurate to about $\pm 6\%$. It is also a very restrictive approach, because it requires the ensemble to have been sampled according to a specific distribution. An alternative approach is to use the ensemble to construct an emulator for $g(\cdot)$. Then we replace $g(\cdot)$ in (8) with our emulator, which gives

$$F_\delta(v) = \int_x F_{g(x)}(v)\, dF_{x^*}(x) \tag{10}$$

(without getting into technicalities). This is a simple generalisation of (8) in which the indicator function $\mathbf{1}\big(g(x) \le v\big)$ is replaced with the appropriate probability. In the limit as the number of evaluations in our ensemble becomes large we have $F_{g(x)}(v) \to \mathbf{1}\big(g(x) \le v\big)$ because it becomes more and more likely that somewhere in our ensemble we have actually evaluated that particular choice of x, and so our emulator becomes more and more like a simple look-up table. But the important feature of (10) is that it allows us to incorporate our uncertainty about $g(\cdot)$ in the more usual case where we have only a small collection of evaluations in our ensemble. Now our uncertainty about δ is a consequence of our uncertainty about x^* *and* our uncertainty about $g(\cdot)$; this latter source of uncertainty only tends to zero if the number of evaluations in our ensemble becomes very large.

We can compute (10) using the same Monte Carlo approach given above, giving an estimate

$$F_\delta^{(m)}(v) \triangleq m^{-1} \sum_{i=1}^{m} F_{g(x'_i)}(v), \tag{11}$$

where $x'_1, \dots, x'_m$ are sampled from F_{x^*} as before. The primes indicate that these are *not* the same x values as we have evaluated in our ensemble X. In this case we can have $m \gg n$, where n is the number of evaluations in the ensemble. We can make m as large as necessary to achieve a good estimate of F_δ, because each evaluation of the integrand is the calculation of a distribution function, rather than an evaluation of the simulator. But it is important to bear in mind that this advantage has been purchased with the choices that we must make in order to construct the emulator and derive the distribution function $F_{g(x)}$. Hence the importance of the emulator diagnostics.

By the same token, we do not select our ensemble on the basis of the distribution function F_{x^*}, but rather in order to build as accurate an emulator as possible. This might mean, for example, over-sampling components of x which are thought to be important for determining sensitivity. These types of choice are discussed in the extensive literature on Bayesian experimental design (see, e.g., Chaloner and Verdinelli 1995), and there is also a literature on the particular features of design for computer experiments (see, e.g., McKay et al. 1979; Sacks et al. 1992; Morris and Mitchell 1995; Koehler and Owen 1996).

Note that simple Monte Carlo integration is a very crude approach to determining the value of $F_\delta(v)$. More sophisticated statistical approaches, such as Importance Sampling with variance reduction techniques or Markov chain Monte Carlo (MCMC), will be necessary in more complicated problems; these are discussed in, e.g., Evans and Swartz (2000) or Robert and Casella (1999). We will use simple Monte Carlo integration because our application is straightforward, and we find that $m = 10^3$ is sufficient, and only takes a few seconds.

5b. *Investigating the choice of prior for $\boldsymbol{x}^*$*

Our predictive distribution for δ will depend on our prior for x^*, namely F_{x^*}, and on our emulator for $g(\cdot)$, which in turn will depend on our choice for X, our resulting evaluations y, and the choices we make in the construction of our emulator. In our particular application these latter choices have been described in sections 3 and 4. One of the advantages of using an emulator is that we can investigate different choices for F_{x^*}, to establish to what extent our predictive distribution for δ is affected by aspects of our distribution function for the correct parameters x^*. Here we present a simple experiment to investigate the shape of the distribution. We stress that this is a sensitivity analysis. Our candidates for F_{x^*} below *do not* represent our judgements about the correct parametrisation of HadSM3. O'Hagan and Oakley (2004) discuss the problem of eliciting distributions for model parameters.

For our base-line choice for F_{x^*} we adopt the QUMP prior, namely that all components of x^* are independent; all continuous components are uniform within the limits given in Table 1, and all factors have equal probability on each level (also given in Table 1). We will consider some alternative specifications for the marginal distributions of the continuous components of x^*. Our resulting predictive distributions for δ are shown in Figure 6; the base-line is shown as distribution A.

First, the actual definition of the variables in a simulator is to some extent arbitrary. Often, for example, the choice of whether some variable should be represented as φ or as $\psi \triangleq 1/\varphi$ may come down to what is more efficiently represented in the computer code. Naturally this makes a difference to the choices we make for F_{x^*}, since if φ^* has a uniform distribution then ψ^* does not. In a superficial analysis this can cause some debate, along the lines of "Should our distribution be uniform in φ^* or in ψ^*?", but only because the uniform distribution is presumed to represent some form of 'prior ignorance'. A deeper analysis reveals that it is the concept of 'prior ignorance' which is at fault, not the choice of the uniform distribution. We are not ignorant about the parameters in our simulators; for example it is possible to elicit ranges for them, and also information about symmetry, as shown in Table 1. It must be understood that selecting a uniform distribution is a choice made not of ignorance, but of judgement. Thus if a uniform distribution is chosen for φ^* then the distribution of ψ^*, and indeed of every well-behaved function of φ^*, is also chosen at the same instant. There is no 'default' distribution for F_{x^*}, uniform or otherwise: every choice must be defensible as a description of someone's uncertainty. One thing we can do, however, is a simple experiment to see whether uniform in φ^* or in ψ^* makes any practical difference. The

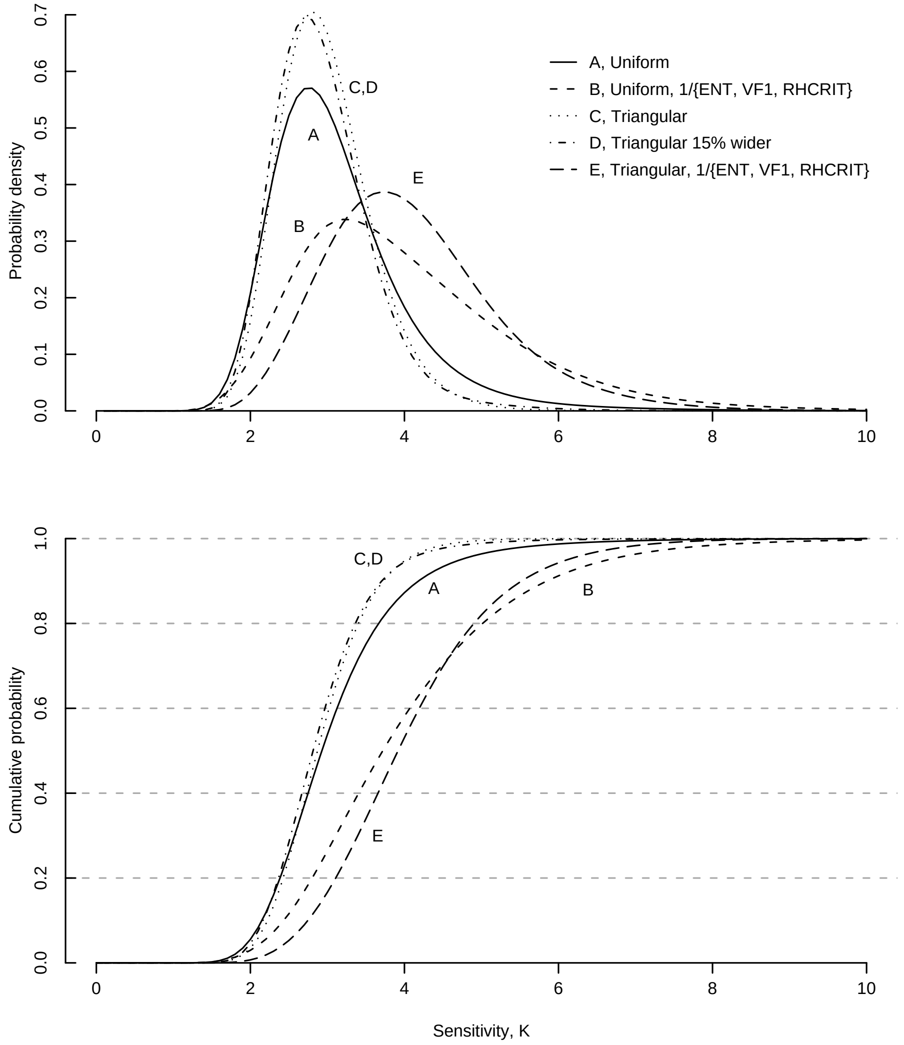

Figure 6: Various predictions for $\delta \triangleq g(x^*)$ for different choices of F_{x^*}. A: Uniform prior on all continuous components of x^*; B: same as A except with a uniform prior on the reciprocal for components `ENT`, `VF1`, `RHCRIT`; C: Symmetric triangular prior on all continuous components of x^*, same ranges as A and B; D: Same as C, except with ranges increased by 15%; E: triangular prior, reciprocal in `ENT`, `VF1`, `RHCRIT`.

marginal priors for the components of x^* corresponding to `ENT`, `VF1` and `RHCRIT` can be treated as uniform in the reciprocal; shown as distribution B in Figure 6.

Second, as has frequently been noted, the uniform is actually a very poor choice of shape for a distribution on the correct value of a continuous variable; see, e.g., Garthwaite et al. (2005). The uniform asserts, for the correct value of a continuous variable, that all values within some range are equally likely, and all values outside that range are impossible. Thus for `CW`, which we believe is an important determinant of HadSM3's sensitivity, all values between 1 and 10 are deemed equally likely, even though the standard value is 2, and a value such as 0.9 is deemed impossible. This does not seem a very defensible position. A more natural choice in this situation would be to favour a prior which had a continuous density function, rather than a step at either end, so that, for example, an impossible value like 0.9 is only a little less probable than the nearby possible value of 1.1. The simplest such distribution is symmetric triangular, which has the same number of parameters as the uniform distribution. We investigate using this distribution for the correct values of all of the continuous variables; shown as distribution C in Figure 6.

Third, we investigated increasing the width of the triangular prior for each of the continuous variables by 15% (subject to a non-negativity constraint, and `CFS` > 0.5 and `ANVS` > 1), to account for the possibility that the experts who set the widths might have underestimated their parametric uncertainty (see, e.g., Soll and Klayman 2004). This is shown as distribution D in Figure 6.

Finally, we combined the triangular prior with the reciprocal expression of `ENT`, `VF1` and `RHCRIT`, shown as distribution E in Figure 6.

Comparing the predictive distribution for δ derived under our different specification for F_{x^*} we can see immediately that the shape of the prior *does* matter, particularly in determining the length of the upper tail, where the most risks lie from a decision-making point of view. The biggest difference comes from switching to the reciprocal for the components of x^* corresponding to `ENT`, `VF1` and `RHCRIT`. In both cases the triangular distribution gives a smaller uncertainty for δ than the uniform, with a materially lower probability of extremely high values. This reflects the fact that extreme values for HadSM3's sensitivity are found in the corners of the parameter-space, and the triangular distribution downweights these relative to the uniform. Similarly, the uniform and reciprocal-uniform are more different than the triangular and reciprocal-triangular. It is also interesting to note that the predictive distribution for δ is relatively insensitive to the width of the marginal distributions, at least for our experiment of increasing that width by 15%.

All of these findings are conditional upon our emulator for HadSM3, and consequently on the CPNET and QUMP ensembles. They are also prior to calibrating HadSM3 with climate data: the necessary steps for this are described in Rougier (2006).

6. Summary

An emulator allows us to separate the process of choosing the evaluations in the ensemble from the inference that we intend to do. Consequently we can choose our evaluations wisely, rather than 'randomly', and we can perform a variety of inferences, including a detailed sensitivity analysis, such that that suggested for F_{x^*}. This is important because there can be no 'right' choice for F_{x^*}, although we might hope that a broad consensus might emerge.

But the main part of this paper has addressed another feature of emulators; they provide us with an opportunity to exercise our judgement when using a simulator such as HadSM3 in a CO_2-doubling experiment. We can, if we so choose, *and* if we have sufficient evaluations, delegate all such judgements to standard statistical tools. To a large extent, this is what we did in section 3b, when we built an emulator for CPNET sensitivity, using the CPNET ensemble. We used the Box-Cox approach to select an appropriate transformation of the response; we used stepwise selection to help choose the regressors; and we did this within an emulator framework consistent with a non-informative prior. Our judgements were exercised much more in section 4, where we had to make explicit quantitative choices that described the extent to which we believed that the CPNET and QUMP sensitivities were related. We did not make these judgements in isolation, however, but with the support of detailed diagnostic information.

We see this feature as the crucial advantage of emulators, either in providing the opportunity to augment a small ensemble with prior information, or in allowing us to combine information from different but related studies.

Acknowledgement Jonathan Rougier is funded by the U.K. Natural Environment Research Council (NERC), RAPID Directed Programme. We would like to thank Prof. Michael Goldstein for helping to develop the approach of linked emulators, and for his perceptive comments on the statistical modelling.

References

Box, G. and D. Cox, 1964: An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, **26**, 211–243, with discussion, pp. 244–252.

Chaloner, K. and I. Verdinelli, 1995: Bayesian experimental design: A review. *Statistical Science*, **10**, 273–304.

Chipman, H., M. Hamada, and C. Wu, 1997: A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, **39**, 372–281.

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.

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.

Draper, N. and H. Smith, 1998: *Applied Regression Analysis*. New York: John Wiley & Sons, 3rd edition.

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

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

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

— 2006: 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`.

Grimmett, G. and D. Stirzaker, 2001: *Probability and Random Processes*. Oxford University Press, 3rd edition.

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.

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

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.

Morris, M. and T. Mitchell, 1995: Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, **43**, 381–402.

Murphy, J., D. Sexton, D. Barnett, G. Jones, M. Webb, M. Collins, and D. Stainforth, 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. and J. Forster, 2004: *Bayesian Inference*, volume 2b of *Kendall's Advanced Theory of Statistics*. London: Edward Arnold, 2nd edition.

O'Hagan, A., M. Kennedy, and J. Oakley, 1999: Uncertainty analysis and other inferential tools for complex computer codes. *Bayesian Statistics 6*, J. Bernardo, J. Berger, A. Dawid, and A. Smith, eds., Oxford University Press, 503–519, with discussion, pp. 520–524.

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

Pope, V., M. Gallani, P. Rowntree, and R. Stratton, 2000: The impact of new physical parameterizations in the Hadley Centre climate model, HadAM3. *Climate Dynamics*, **16**, 123–146.

Pukelsheim, F., 1994: The three sigma rule. *The American Statistician*, **48**, 88–91.

R Development Core Team, 2004: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3, `http://www.R-project.org`.

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

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

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`.

Sacks, J., S. Schiller, and W. Welch, 1992: Design for computer experiments. *Technometrics*, **31**, 41–47.

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

Soll, J. and J. Klayman, 2004: Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **30**, 299–314.

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.

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