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

Jonathan Rougier*
Department of Mathematics
University of Bristol, UK

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

David Stainforth
Department of Geography
University of Exeter, UK

February 15, 2008

*Corresponding author: Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW; email j.c.rougier@bristol.ac.uk

Abstract

Projections of future climate are determined by a wide array of detailed and imperfectly understood physical processes. While ensembles of global climate model (GCM) simulations can be run to sample the resulting uncertainties, statistical tools are needed to supplement these simulations, and hence provide a more comprehensive picture of how interactions between different processes influence the response of climate. We illustrate this by combining ensembles from two different experiments to study the response of climate sensitivity in the HadSM3 climate model to 31 model-parameters. We use a Bayesian statistical framework based around linked emulators. Expert judgements are required to quantify the relationship between the two emulators, and these are validated by detailed diagnostics. Using our emulator, we identify the entrainment rate coefficient of the convection scheme as the most important single parameter; find that this interacts strongly with three of the large scale cloud parameters, C_t, C_w (land), and Rh_{crit}; and represent these interactions visually.

KEYWORDS: Bayesian, emulator, diagnostics, convection, large-scale cloud

1. Introduction

HadSM3 comprises 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 the equilibrium response to doubled CO_2. HadSM3 is one of a number of such climate models, developed at different institutions worldwide, and used to investigate global and regional characteristics of the response of climate processes to increases in greenhouse gases. These models contain different choices of horizontal and vertical resolution, different numerical integration schemes, and different parameterisations of sub-grid-scale processes. Therefore, they simulate differently the response of global climate sensitivity, and of the regional and global feedback processes which determine it (Webb et al. 2006). Results from such a multi-model ensemble provide insights into these feedback processes: for example analysis of the latest generation of models suggests that feedbacks associated with low cloud provide the largest contribution to uncertainty in climate sensitivity (Bony and Dufresne 2005; Webb et al. 2006). However, detailed analysis is limited by the small number of ensemble members, and their status as an 'ensemble of opportunity', lacking a systematic approach to the sampling of modelling uncertainties (Tebaldi and Knutti 2007).

An alternative approach is that of the 'perturbed physics' ensemble (PPE), in which simulations are designed to sample variations in parameters controlling the simulation of key climate processes, within a single model. To date, most published PPE studies have focused on HadSM3 and HadCM3, the related configuration in which HadAM3 is coupled to a three-dimensional dynamic ocean component (Murphy et al. 2004; Stainforth et al. 2005; Collins et al. 2006; Harris et al. 2006) (though see also Annan et al. 2005). The advantage of the perturbed physics approach is that it supports a more systematic exploration of modelling uncertainties, in which variations in simulated responses can, in principle, be traced back to particular processes. Their limitation is that they do not explore 'structural' modelling uncertainties, such as the choice of resolution, or of alternative approaches for parameterising sub-grid-scale processes. However, results

indicate that the spread of global and large-scale regional climate responses is similar to that found in multi-model ensembles (Collins et al. 2006; Webb et al. 2006), suggesting that both approaches provide a useful means of exploring the range of simulated climate responses in the current generation of climate models.

In the case of PPEs, the basic approach involves defining a space $\mathcal{X}$ of possible model variants, by asking experts to specify prior distributions for poorly-constrained parameters controlling key climate system processes. Then an ensemble of simulations is run to span or sample that space. The results can be used to understand and quantify simulated responses (Webb et al. 2006), or to construct probabilistic estimates of the response using Bayesian techniques in which locations in $\mathcal{X}$ are weighted according to their relative likelihood, quantified through comparison of simulations of historical climate against a set of observations. Murphy et al. (2004) give an early example of this type of approach. Rougier (2007b) describes a more comprehensive Bayesian framework, including the effects of structural differences between the model used for the PPE and the real world, which cannot be resolved by varying the parameters. Murphy et al. (2007) describe a method for applying this statistical framework in practice, with the aim of providing probabilistic predictions of 21st century climate.

A key requirement of this approach, for the purposes of both understanding the model response and constructing probabilistic predictions, is to be able to estimate the climate model's response at any location in $\mathcal{X}$, in order to explore model behaviour across the entire parameter space, rather than at the subset of locations at which we have run the climate model. We would like to be able to ask the question "What happens when we evaluate the model at x?", where x is any point in $\mathcal{X}$. In this way we could, for example, track the response of the model as we changed one of the parameters, or varied several simultaneously.

In this paper we focus on one particular response, HadSM3's climate sensitivity: the equilibrium change in globally averaged surface temperature following a doubling of the atmospheric concentration of CO_2. This represents a standard benchmark for the response of climate to increases in greenhouse

gases. Thus HadSM3 can be thought of as a function that maps the parameter vector x into a climate sensitivity value $g(x)$. In our ensemble we have a collection of inputs $X = \{x_1, \ldots, x_n\}$ and a corresponding collection of outputs, $y = \{g(x_1), \ldots, g(x_n)\}$. A Bayesian statistical framework termed an 'emulator' allows us to predict $g(x)$ at any x, based on the ensemble and on our judgements about the model. Crucially, this prediction takes the form of a distribution, so that we get not just a point estimate, such as the mean, but also a measure of uncertainty, such as the standard deviation. This uncertainty has two parts. First, there is the irreducible uncertainty from the model's internal variability. Second, there is the uncertainty that arises from not having evaluated the model at or near to x, termed 'code uncertainty' (O'Hagan 2006). Constructing emulators is part of the statistical field of Computer Experiments (see, e.g., Koehler and Owen 1996; Santner et al. 2003). The Bayesian treatment of emulators was initiated by Currin et al. (1991), and continues to develop: current practice is reviewed in O'Hagan (2006) and discussed in Rougier and Sexton (2007).

In this paper we construct an emulator for HadSM3's climate sensitivity as a function of 31 model parameters. This would seem an impossible task given that our ensemble contains only 281 evaluations. But since we quantify our uncertainty, we can show, below, that a large amount of information about HadSM3 can be extracted. Partly, this is because many of the parameters are not important determinants of climate sensitivity (we would not expect this to be true for other types of model output). But also, we use additional information and expert judgements to augment our ensemble. The additional information comes from a second ensemble of HadSM3 simulations, and the expert judgement concerns the relationship between the two ensembles. As our judgements are subjective, we pay close attention to diagnostic information.

Using our emulator we are able to identify the main parameters for determining climate sensitivity, and also to investigate a complex interaction between four parameters controlling some key aspects of the parameterisation of large-scale clouds and convection. In section 2 we describe the two experiments that generate

our two ensembles. Section 3 describes the process of building an emulator for HadSM3's climate sensitivity. Section 4 uses the resulting emulator to investigate the response to the model parameters, both singly and in combination. Section 5 concludes with a summary of our findings and a discussion of our approach.

2. Two experiments on HadSM3

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

2a. *The QUMP experiment*

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 more 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 identified 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 climate 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 climate sensitivity at x was computed in a three-phase experiment. The first phase was a 25-year calibration run in which sea surface temperatures (SSTs) are continu-

ously restored to prescribed values from a historical climatology. The heat fluxes required to achieve this were averaged to provide heat convergence fields intended to represent the effects of ocean heat transport (not simulated explicitly in the mixed layer ocean of HadSM3), and also to offset errors in simulated atmosphere-ocean fluxes. These heat convergences (which vary with position and season, but not from year to year) were then prescribed in phases two and three, consisting of a control simulation with pre-industrial CO_2, and a run with doubled CO_2, both run to equilibrium. The heat convergences should ensure that multi-year averages of SST in the control simulations remain close to observed climatology, subject to the assumption that internal climate variability in SST (suppressed during the first phase, but not in phases two and three) does not give rise to non-linear feedbacks which could cause SSTs to drift.

Climate 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 experiment targeted the areas of model physics thought to be influential for a wide range of global and regional aspects of historical climate, and of the forced response to external changes in radiative forcing. The initial evaluations in the ensemble consisted of single parameter perturbations, augmented by a small number of multi-parameter perturbations. Since that original experiment, we have access to a further 231 evaluations, all multi-parameter perturbations. The first 128 of these are described in Webb et al. (2006), and were chosen to span a wide range of climate sensitivities, subject to the additional constraints of achieving credible simulations of present day climate, and sampling the parameter space as widely as possible. Additional simulations were chosen to populate regions of the parameter space thought likely to be influenced by important interactions. These can be added directly to the original ensemble, to give the 297 evaluations.

A small minority of these evaluations produced control simulations of SST significantly cooler than the historical values used to deduce the heat convergence fields. The cooling results from the absence of a dynamical representation of ocean heat transport in HadSM3 (excluded to make the simulations of climate

Table 1: Description of the QUMP variables. Comparable to Murphy et al. (2004), Supplementary Information, Table 2. Each parameter controls a key aspect of one of the schemes for the parameterisation of sub-grid-scale processes in HadSM3 (large scale cloud, convection, sea-ice, etc). 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	Label	Only when
Large-scale cloud			
V_{f1} (ms^{-1})	(0.5, **1**, 2)	VF1†	
C_t ($\times 10^{-4}\,\mathrm{s}^{-1}$)	(0.5, **1**, 4)	CT†	
C_w (land, $\times 10^{-4}\,\mathrm{kg\,m}^{-3}$)	(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 $0\,^{\circ}\mathrm{C}$)	(**0.50**, 0.57, 0.65)	SIA	
Ocean-ice diffusion ($\times 10^{-4}\,\mathrm{m}^2\,\mathrm{s}^{-1}$)	(0.25, 1.00, **3.75**)	OID	
Radiation			

(continued . . .)

Table 1: QUMP variables (continued)

Parameter / Property	Values	Label	Only when
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	

sensitivity computationally feasible, and also because changes in ocean circulation are not likely to be a major determinant of climate sensitivity (e.g., Senior and Mitchell 2000; Boer and Yu 2003). For example, situations can arise in which variability in simulated atmosphere-ocean heat fluxes cause negative anomalies in SST, which then increase boundary layer stability and hence low cloud cover, thus providing a positive feedback leading to further cooling of SST. If the local prescribed heat convergence also happens to be negative in the relevant region (for example because the model simulates too little low cloud on average), then a rapid cooling of local SST can then occur, the effects of which then spread to other regions. In models with a dynamical ocean component the initial positive feedback from local surface exchanges would be offset by a negative feedback from changes in ocean heat transport, but the latter process is missing in atmosphere-mixed layer ocean models. We find 16 model variants in which global mean SST in the control simulation cools in this way. The absence of interactive ocean heat transport in HadSM3 therefore prevents us from being able to obtain credible estimates of climate sensitivity by direct simulation in these 16 experiments, so we exclude them from our analysis. The 281 evaluations that remain provide estimates of sensitivity free from non-physical side-effects of the experimental design. This is demonstrated, for example, by the result that a close relationship is found between the equilibrium surface warming found in 128 of these evaluations, and the transient climate response obtained using corresponding parameter settings in simulations with a dynamical three dimensional ocean component (Collins et al. 2006; Harris et al. 2006). We rely on our emulator, trained on the 281 reliable evaluations, to supply estimates of climate sensitivity for other locations in model parameter space, including those 16 locations for which cooling HadAM3 simulations were excluded.

2b. *The CPNET experiment*

Here we focus on the differences between QUMP and the `climate`*prediction*`.net` (CPNET) experiment of Stainforth et al. (2005). This experiment varied six of the

continuous variables, used in the parameterisation of large-scale clouds and convection. The 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 to the value used in the standard published version of HadAM3. Hereafter these are referred to as the 'standard values', although note that a number of these values are set to an extreme of the expert-specified ranges (Murphy et al. 2004, Supplementary Information). This reflects the practice of tuning climate model parameters to improve the overall simulation of a range of climate variables by adjusting error balances between different physical processes. Each choice for the variables was evaluated with a number of different initial conditions, introducing a structured source of uncertainty that is not present in the QUMP experiment. On analysing the CPNET ensemble, 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; a similar approach was used in the Stainforth et al. (2005) experiment, where different initial conditions for the same x were averaged, to reduce variability.

The CPNET experiment 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 in Stainforth et al. (2005) 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 climate sensitivity.

In our sample from the CPNET experiment we have a total of $3^5 \times 2^1 = 486$ distinguishable 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 (as described in section 2a). We choose to omit these from the CPNET ensemble in the same way as Stainforth et al. (2005).

Comparing these two experiments, we judge there to be sufficient differences that it is not possible to combine the two ensembles directly, or indirectly by reweighting one or the other; in fact they are two different but related experiments. In other words, the relationship between the CPNET climate sensitivity and the six CPNET variables is not simply a noisier version of the QUMP relationship with the same variables, but it is actually a different relationship, affected by the transient behaviour of the HadSM3 model. This informs our statistical modelling choices in section 3c.

2c. *Outline of our approach*

The two experiments outlined in this section have different but complementary strengths. The QUMP experiment has a conventional definition for climate sensitivity, and includes a large number of variables. The CPNET experiment, on the other hand, has a more detailed analysis over six of the most important variables (the CPNET project has subsequently explored many more variables, allowing for a more extensive analysis in the future). Our intention is to combine the ensembles from these two experiments into an emulator for QUMP climate sensitivity defined over the full set of thirty-one variables.

As already described, an emulator is a probability distribution function for $g(x)$. There are many ways of specifying such a function. In a Bayesian statistical approach we probabilistically condition our beliefs about $g(\cdot)$ on the evaluations 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 experiments requires a different emulator, because of the different definitions of climate 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 (281 evalu-

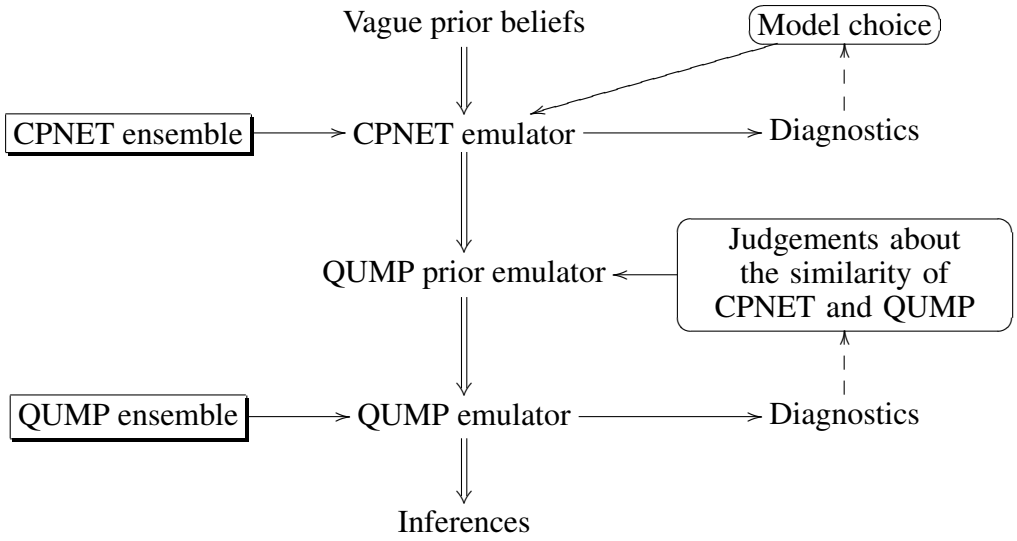

Figure 1: The main stages of our approach for combining information from the CPNET and QUMP ensembles into an emulator for QUMP climate sensitivity. Starting with value prior beliefs, we create the CPNET emulator using the CPNET ensemble. Then we use our judgements about the similarities of CPNET and QUMP to construct a QUMP prior emulator. Finally, we update this emulator with the QUMP ensemble, to construct the QUMP emulator.

ations 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 climate 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. These are discussed in more detail in sections 3d and 3e.

Kennedy and O'Hagan (2000) have proposed a different approach, designed to combine ensembles from the same model solved at different resolutions. However, it is not easily applicable here, due to the complexities of the model parameters, as discussed in section 3a.

3. Emulating HadSM3's climate sensitivity

In this section we describe our approach for emulating HadSM3's climate sensitivity, as outlined in Figure 1. Section 3a outlines a simple emulation framework, based on the Bayesian treatment of the Gaussian Linear Model. Section 3b details the choices we make within this framework, to emulate CPNET's measure of climate sensitivity. Section 3c describes how we quantify our judgements about the relationship between the CPNET and QUMP experiments, in terms of the relationship between the CPNET and QUMP emulators. Section 3d introduces the QUMP ensemble, which is used to generate diagnostic information about the statistical choices we have made, before being assimilated into the QUMP emulator in section 3e. In section 4 we will use the QUMP emulator to investigate the response of HadSM3 to its 31 variables.

3a. *A general Bayesian emulator*

We describe here a simple Bayesian treatment of the emulator. The emulator is written

$$g(x) = h(x)^T \beta + u(x) \tag{1}$$

where $g(x)$ is the climate sensitivity of HadSM3, 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* (k in total); β is an unknown k-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.

For our given choice for the response and the regressors, we make the following additional choices. First, $u(x)$ has zero mean and a constant unknown variance, σ^2; second, $u(x)$ and $u(x')$ are uncorrelated when $x \neq x'$; third, β, $u(x)$, and σ^2 have a Normal-Inverse-Gamma (NIG) distribution, which may be summarised as

$$\beta \perp\!\!\!\perp u(x) \mid \sigma^2 \tag{2a}$$

$$\beta \mid \sigma^2 \sim \mathrm{N}_k(m, \sigma^2 V) \tag{2b}$$

$$u(x) \mid \sigma^2 \sim \mathrm{N}_1(0, \sigma^2) \tag{2c}$$

$$\sigma^2 \sim \mathrm{IG}(a, d) \tag{2d}$$

where '$\perp\!\!\!\perp$' denotes probabilistically independent, '|' denotes conditional upon, $\mathrm{N}_k(\cdot)$ denotes the k-dimensional Gaussian distribution, and $\mathrm{IG}(\cdot)$ the scalar Inverse Gamma distribution; we must specify the collection $\{a, d, m, V\}$, termed the 'hyperparameters'. With these distributional choices the emulator for $g(x)$ has a Student-t distribution, where both the mean and the scale will depend on x. We have outlined here the standard Bayesian treatment of the Gaussian Linear

Model; full details may be found in O'Hagan and Forster (2004, ch. 11).

At this point our statistical choices have been made for tractability and transparency. The NIG approach is a standard framework for emulation, see, e.g., Rougier (2007a) for a full description, and Rougier et al. (2007) for an example; however, it has some undesirable features (see, e.g., the second half of O'Hagan and Forster 2004, ch. 11). But we have made one unusual choice, which is to treat the residual as having zero correlation length, i.e. to set $\mathrm{Cov}\big(u(x), u(x')\big) = 0$ for $x \neq x'$. The residual accounts for internal variability, for which a zero (or near-zero) correlation length is quite appropriate. However, it also accounts for systematic effects excluded from the regressors, and these have a positive correlation length. Overall, therefore, we have understated the correlation length of the residual: the implications are discussed further in section 4. We have a compelling reason for making this choice, which is that Statisticians have yet to develop flexible covariance structures for $u(x)$ which can be specified over a collection of both continuous variables and factors. This is an active area of research; see, e.g., Han et al. (2007) and Qian et al. (2007). An alternative strategy would be to build a different emulator over the continuous variables for each factor combination; however, our ensembles are not large enough to allow this, because there are 13 factors giving rise to $2^{10} \times 3^2 \times 4^1 = 36864$ factor combinations. As long as the residual does not play a large part in the emulator, our understatement of the residual correlation length is unlikely to be predictively important. In our emulators of QUMP climate sensitivity we find that the regression R^2 is at least 90% and typically more than 95%, depending on the precise choices we make for the transformation of the response and the regressors. 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. In the light of this choice we place strong reliance on diagnostics, discussed in sections 3d and 3e.

To summarise this section, the challenge of building an emulator for $g(\cdot)$ using the ensemble $(y; X)$ has been restructured to (i) choosing a transformation for

climate sensitivity and a collection of regressors $h(\cdot)$, and, conditional on these choices, (ii) specifying the hyperparameters $\{a, d, m, V\}$ in the NIG prior for $\{\beta, u(x), \sigma^2\}$.

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 $\{\beta, u(x), \sigma^2\}$, 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 $h(\cdot)$, $m = \mathbf{0}$, 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 usual Ordinary Least Squares (OLS) form, although the interpretation is a little different, being Bayesian rather than Frequentist. When we refer to, say, a 95% CI we are referring to a 95% 'Credible Interval': an interval defined by the 2.5th and 97.5th percentiles of the distribution (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 their view that this function would be likely to have a simpler additive structure in terms of the variables. This would only be a reasonable transformation if negative values for climate 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 strictly positive 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 four of these are relevant for the CPNET experiment.

We would like our emulator to include interactions among the variables. In the initial QUMP ensemble it was not possible to estimate interactions from the single-parameter perturbations, but they were found to be influential in CPNET. Our general strategy regarding interactions is to treat variables within different parameterisation schemes as non-interacting (these schemes are shown in Table 1), but to include interactions between variables within each scheme. 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 experiment, or, indeed, recommended single-parameter perturbations for the initial stage of the QUMP experiment, 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

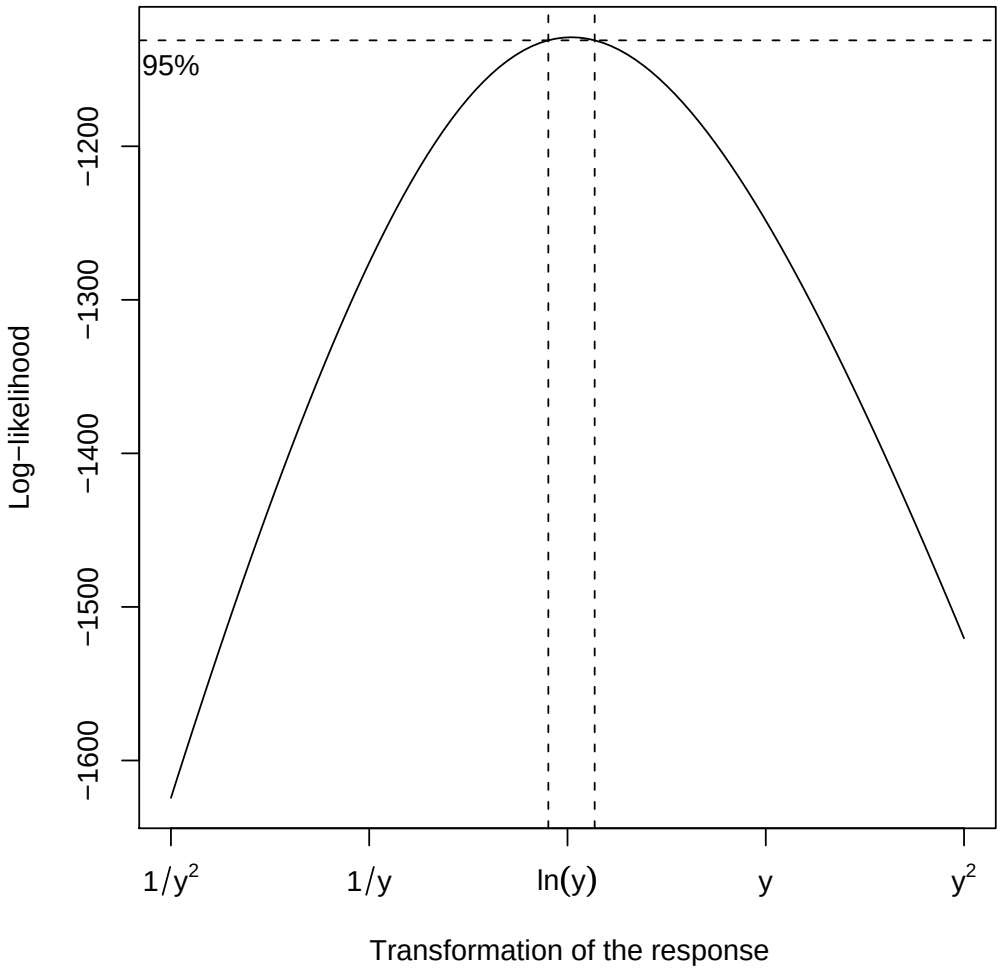

Figure 2: Box-Cox plot to select an appropriate transformation for the response: the high likelihood values are concentrated around the logarithm (the vertical dashed lines indicate an approximate 95% confidence interval).

choice for the transformation of the response; the typical diagnostic for this approach is shown in Figure 2.

We do not want to rule out the possibility of higher-order interactions as well. There 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 *a priori* 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 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 (particularly involving ENT), 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 climate sensitivity, supporting the conclusions of Stainforth et al. (2005).

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 $\log(\text{climate sensitivity})$ and the R^2 is 0.87.

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

3c. *Linking the two emulators*

Having built an emulator for CPNET climate sensitivity, we turn now to using this emulator as prior information for our emulator for QUMP climate sensitivity. First, 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 25 additional variables. Then we must use our judgement about the relationship between the CPNET and QUMP experiments to map the CPNET hyperparameters, which we here denote $\{a^0, d^0, m^0, V^0\}$, to the QUMP prior hyperparameters, $\{a, d, m, V\}$. In sections 3d and 3e we introduce the QUMP ensemble, to generate diagnostics for our statistical choices, and to update the QUMP ensemble hyperparameters to their final values, $\{a^*, d^*, m^*, V^*\}$.

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. 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 parameterisation scheme, 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.

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.

We can express the extent to which we think that CPNET climate sensitivity and QUMP climate 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

$$\beta_i - c_i = (1 + \omega_i)\,(\beta_i^0 - c_i) + (r_y/r_i)\,\nu_i \tag{3}$$

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 (3). 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 (ranges in both cases). 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 centering 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 2007).

The two Greek terms in (3), ω_i and ν_i, represent independent mean-zero uncertain quantities, for which we must specify standard deviations. We will want to set $\mathrm{Sd}(\nu_i)$ small, so just for the moment we treat ν_i as zero. In this case we have

$$\beta_i \approx (1 + \omega_i)\beta_i^0 \tag{4}$$

and $\mathrm{Sd}(\omega_i)$ controls the probability that β_i has a different sign to β_i^0. Setting $\mathrm{Sd}(\omega_i)$ small relative to 1 would be akin to stating that β_i and β_i^0 were very similar. For example, setting $\mathrm{Sd}(\omega_i) = 1/4$ would state that a change of sign in going from β_i^0 to β_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 is zero or small. We judge that a small value is appropriate here, and we choose $\mathrm{Sd}(\nu_i) = 1/20$ for all matched coefficients. With this value it is very unlikely that regressor i will explain more than one-fifth of the range of the QUMP emulator response in the case where $\beta_i^0 = 0$. It is not easy to choose values for these two standard deviations (or the others below), and to some extent we must be guided by diagnostics.

The unmatched coefficients. The unmatched coefficients are QUMP emulator regression coefficients that do not appear in the CPNET emulator. For these coefficients we use a framework similar to (3), namely

$$\beta_i = (r_y/r_i)\,\nu_i\,. \tag{5}$$

This is just a way of assigning an uncertainty to each unmatched β_i in terms of the scale-free quantity $\mathrm{Sd}(\nu_i)$. We have to decide how much of the response range we believe these additional regressor terms can explain. Our choice is $\mathrm{Sd}(\nu_i) = 1/16$ 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.

The residual. We judge that the residual variance for the QUMP prior emulator will be less than that of the CPNET emulator, because the recorded value of climate 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, which can be inferred from $\{a^0, d^0\}$, and choose a standard deviation equal to the mean, to preserve a large amount of uncertainty. We translate these two values into values for hyperparameters a and d by matching the mean and variance of the Inverse Gamma distribution.

Completing the calculation. Once we have computed $\{a, d\}$, we can use these two values along with the values $\{a^0, d^0, m^0, V^0\}$, the frameworks (3) and (5), and our choices for the standard deviations of the ω_i and ν_i to compute the hyperparameters m and V in the QUMP emulator, by matching the mean and variance of the multivariate Student-t distribution.

3d. *Prior diagnostics*

In constructing our QUMP prior emulator we have used the CPNET ensemble in two ways. We have used it *indirectly*, to select the transformation of the response and to identify important 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. These choices were broadly satisfactory in terms of the diagnostics described below. Now we have decided to screen out more of the drifters (see sections 2a and 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 evalu-

ations in the QUMP ensemble. Each individual prediction, taken marginally, has a Student-t distribution. In Figure 3 we show all 281 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 281 values are well within the 95% CI that we predict. We could 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 QUMP prior emulator, and we prefer to leave things as they are, rather than to invite the suspicion that we have in any way over-tuned our prior.

Note that the cluster of similar evaluations on the lefthand side of the bottom panel of Figure 3 corresponds to the evaluations with single-variable perturbations in the unmatched variables of the QUMP experiment. The CPNET ensemble contains no information about these, and so, according to our statistical choices, they are all predicted the same way. The reason that most of the dots in this cluster are near the median is that most of the unmatched QUMP variables are not important for climate sensitivity (particularly the factors), and so varying them makes little difference. Note, however, that variables which have only a secondary impact on climate sensitivity can still have a primary influence on other aspects of the simulated climate response (see, e.g., Betts et al. 2007).

3e. *Posterior diagnostics*

We also consider a second set of diagnostics, that investigate the posterior predictive properties of the QUMP emulator. One such diagnostic is broadly comparable with the univariate prior prediction given in Figure 3: the leave-one-out diagnostic (see, e.g., Rougier et al. 2007). In this case we update the emulator with all but one evaluation from the QUMP ensemble, and then predict that evaluation. We

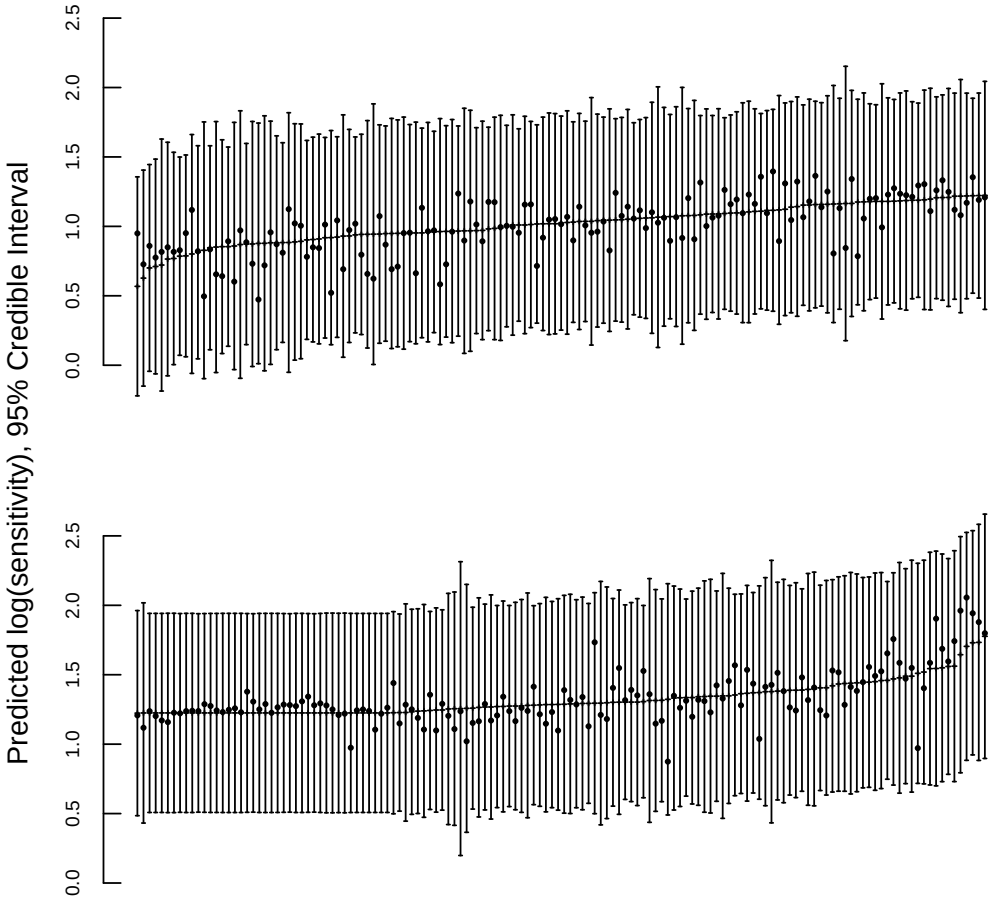

Figure 3: Prior prediction diagnostic showing, for each simulation in the QUMP ensemble, the prior median and 95% CI, along with the actual value of the response (dot). The evaluations are ordered by the median.

can do this with all 281 evaluations; the result is shown in Figure 4. Because 280 is almost the same as 281, the width of the intervals in Figure 4 is a good guide to the amount of uncertainty we will have in our QUMP emulator. By comparing the widths in Figures 3 and 4 we can quantify the contribution of the QUMP emulator in reducing our uncertainty about HadSM3. On the log scale this uncertainty has been reduced by more than 50%.

In all, 13 of the 281 actual values for $\log(\text{climate sensitivity})$ lie outside the 95% CI of the posterior prediction. In terms of the binomial model, the probability of observing 13 or fewer successes out of 281 *independent* trials with $p = 0.05$ is 0.46, 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 every third evaluation, and update using the others ('leave-93-out'). The joint distribution of all 93 prediction errors after updating should be multivariate Student-t—if our statistical choices are reasonable—so that we can transform the prediction errors to 93 uncorrelated standard Student-t quantities. Figure 5 show the result as a Quantile-Quantile plot (QQ-plot), and a histogram with the standard Student-t density overlaid. 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 choices.

4. Investigating main effects and interactions

As an illustration of the utility of our emulator, now represented in terms of the updated hyperparameters $\{a^*, d^*, m^*, V^*\}$, we investigate the response of HadSM3's climate sensitivity to the 31 variables. Figure 6 shows the effect of each continuous variable in turn, with all of the other variables being set to their standard values. At each specified value on the horizontal axis we show the median, and

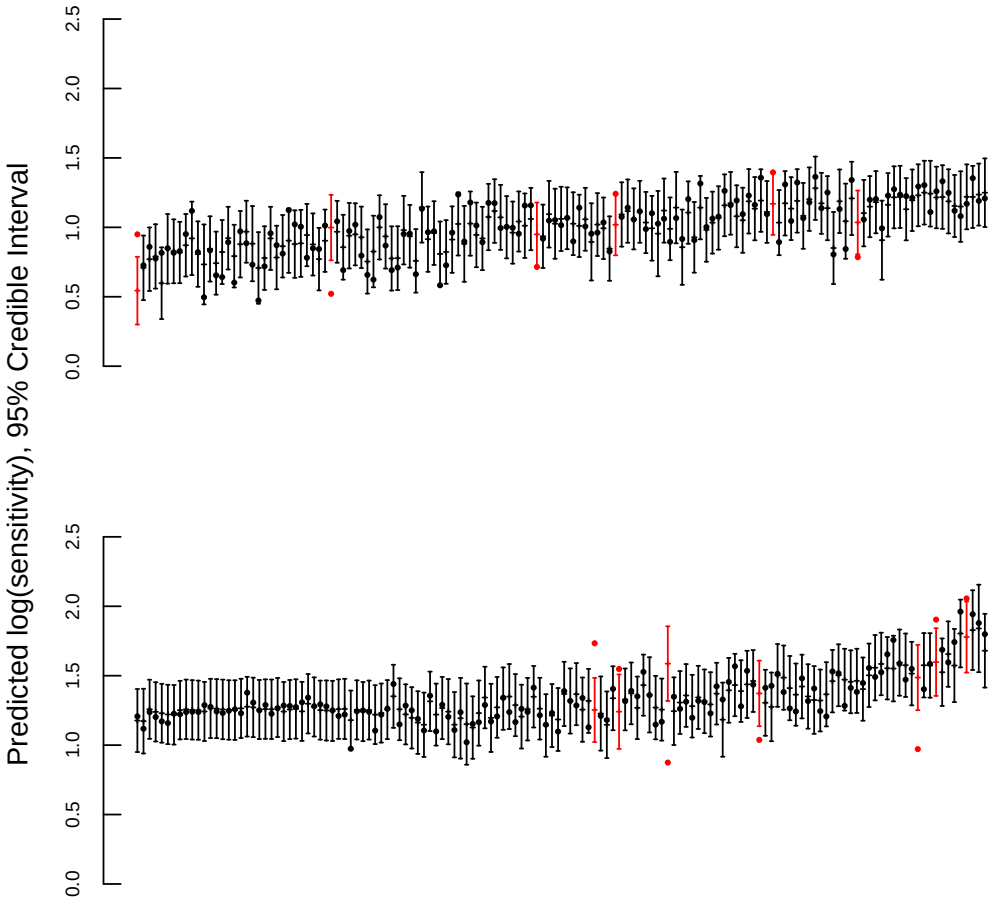

Figure 4: Posterior prediction 'leave-one-out' diagnostic showing, for each simulation in the QUMP ensemble, the posterior median and 95% CI after updating with the other 280 evaluations. The evaluations have the same ordering as in Figure 3.

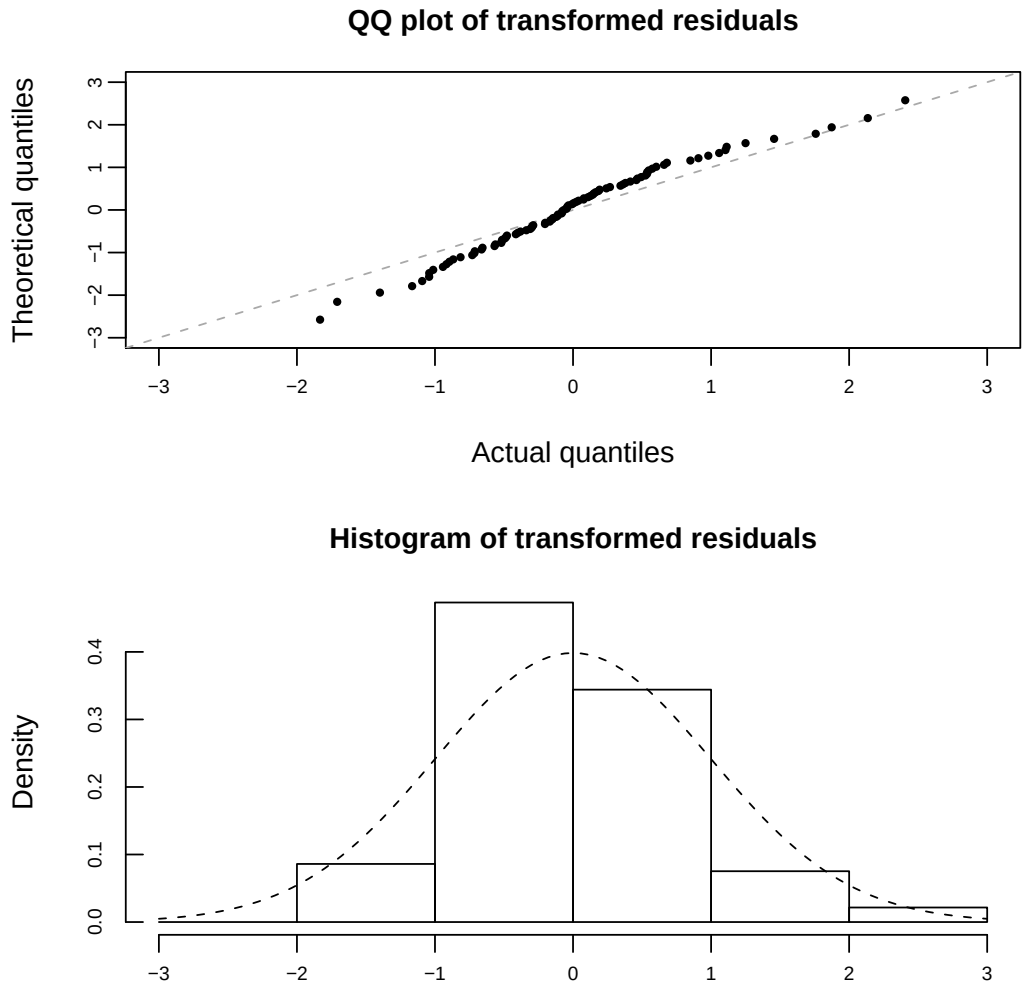

Figure 5: Diagnostics from the joint prediction of every third member of the QUMP ensemble, using the other members (i.e., 'leave-93-out'). After transformation, each prediction error should have a standard Student-t distribution. The first panel shows the QQ plot for the prediction errors, and the second the histogram, with the Student-t density overlaid.

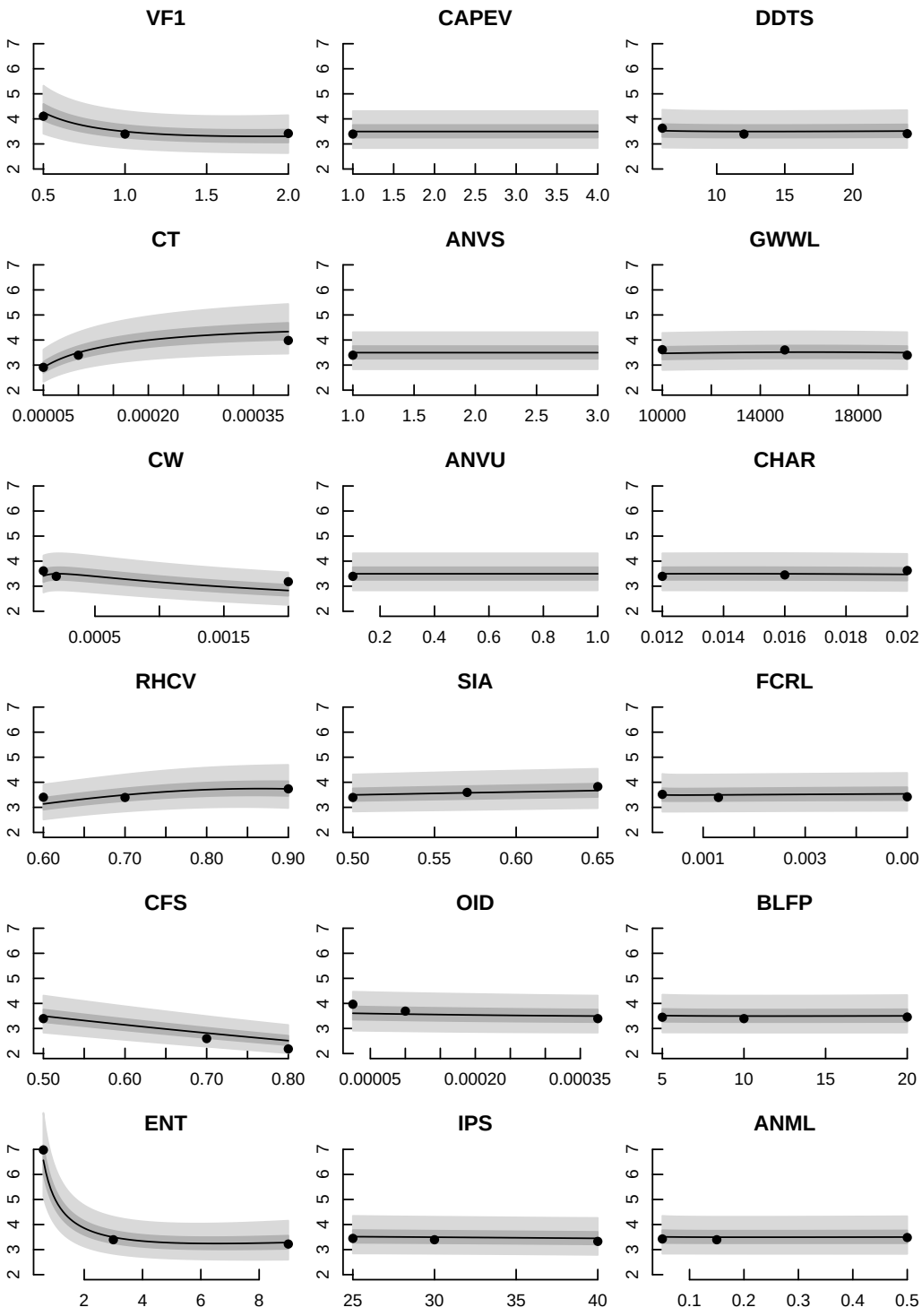

Figure 6: The effect on climate sensitivity of each of the continuous variables, where in each panel all other variables are set to their standard values. The line shows the median, the two envelopes show the pointwise 50% and 95% credible intervals. The dots show actual values from the initial stage of the QUMP experiment.

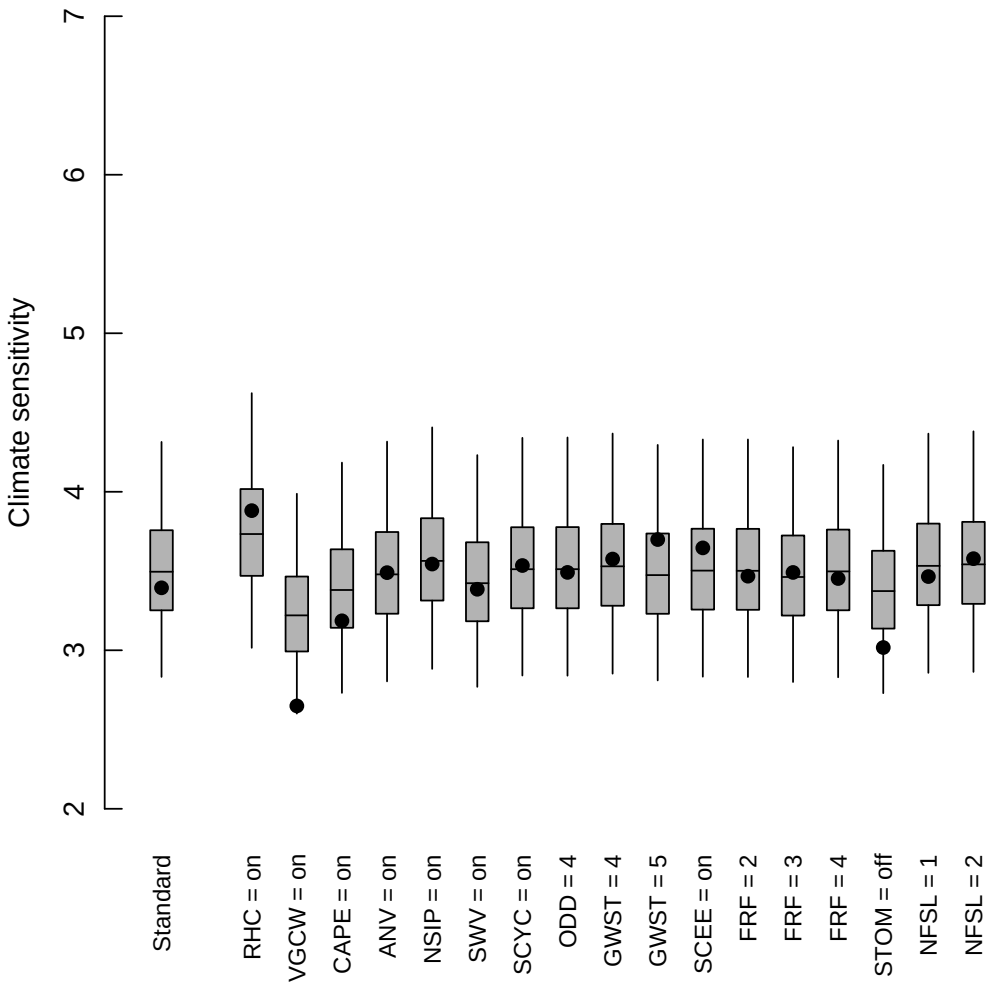

Figure 7: The effect on climate sensitivity of each of the factors. The lefthand column shows climate sensitivity predicted at the standard settings. The other columns show the effect of changing one factor at a time. The box shows the 50% CI, the whiskers the 95% CI, and the central bar the median. The dots show actual values from the initial QUMP ensemble. The vertical scale is the same as in Figure 6.

two envelopes showing the 50% and 95% CIs. Where we have them, we have also shown the values from the corresponding members of the QUMP ensemble, as dots. A similar figure for the factors is shown in Figure 7.

As simple diagnostics, these two figures confirm that our predictions are well-calibrated (although this is not as strict a test as leave-one-out, as the predicted values are included in the emulator). They indicate that the large-scale cloud parameters plus the entrainment coefficient are the important variables (lefthand column of Figure 6). In particular, climate sensitivity is highly sensitive to low values of the entrainment coefficient. Any analysis that accounts for uncertainty in the 'correct' value of entrainment will be sensitive to the choice of distribution: for example, uniform in entrainment and uniform in the reciprocal of entrainment on the full range given in Table 1 will give quite different results (Rougier and Sexton 2007), although our current work suggests that the difference is diminished when `ENT` is calibrated using historical climate, which tends to rule out low values.

At this point we can clarify the practical implication of having a correlation length of zero in the emulator residual, $u(x)$, discussed in section 3a. Ideally, our emulator should interpolate the values in the ensemble to within the uncertainty due to internal variability, roughly $\pm 0.2\,^{\circ}\mathrm{C}$. By inspection the width is typically more like $\pm 0.8\,^{\circ}\mathrm{C}$. We cannot easily reduce this uncertainty by doing further evaluations of HadSM3, as it represents a limitation of the Statistics, not of the data. Note, however, that this noise, while comparable in size to the main effects of each variable, is much less than the combined effect of several variables, as we now illustrate.

We examine the effect of interactions between the large-scale cloud variables and entrainment, in determining HadSM3's climate sensitivity. We look at the response of climate sensitivity to `ENT` under different settings for `RHC` and `RHCV`, `CT`, and `CW`. The result is shown in Figure 8: this figure can only be constructed with an emulator. The black line in the lefthand panel is identical to the median line in the `ENT` panel of Figure 6. As a sanity-check we can see that the red lines lie below the blue lines for each line style (`CT`'s main effect is positive, as shown

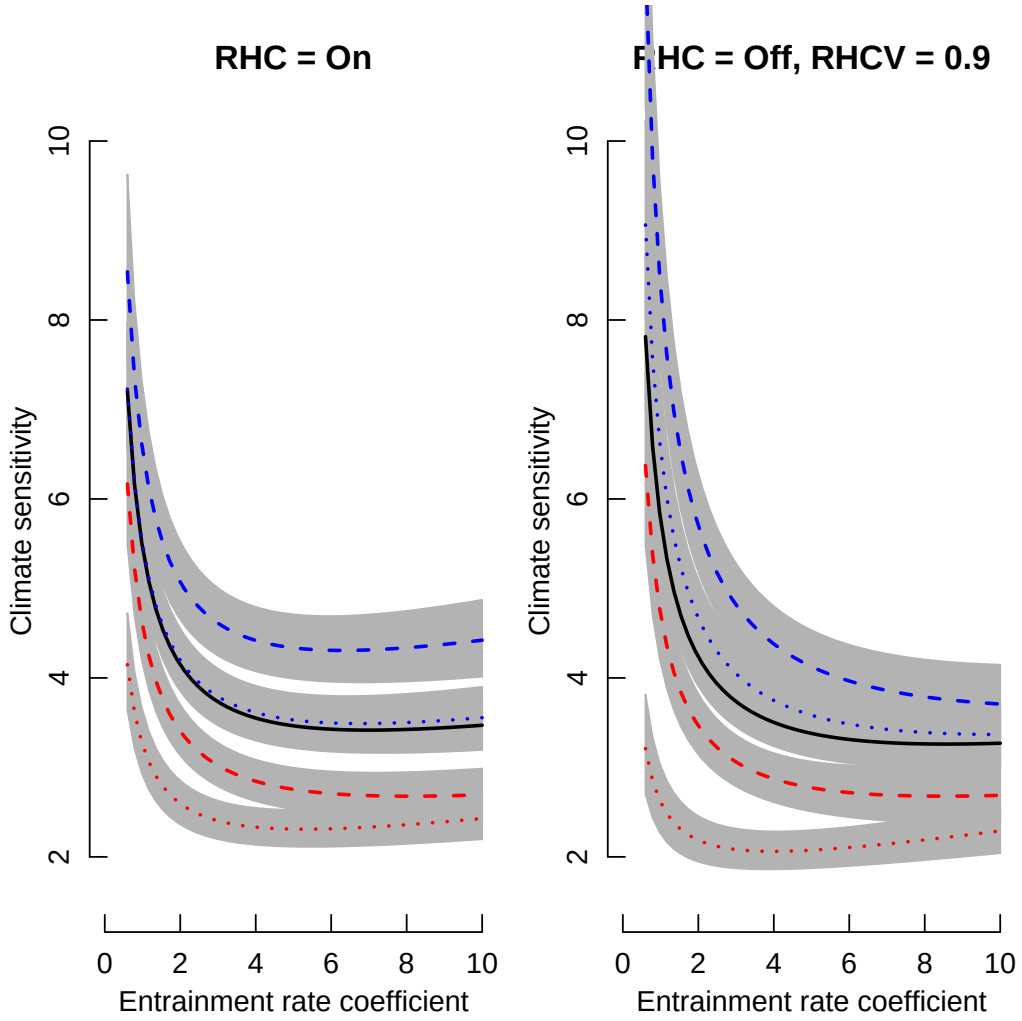

Figure 8: Interaction between entrainment (ENT) and three large-scale cloud variables. Each line shows the median response of climate sensitivity to ENT. For the black line the variables CT and CW are at their standard settings. Four other lines are shown. The colours red and blue indicate low and high values of CT, the line styles dashed and dotted indicate low and high values of CW. The shaded envelope indicates the pointwise 50% CI for each line (n.b. 50% not 95%). In the lefthand panel, RHC is On; in the righthand panel, RHC is Off, and RHCV = 0.9. The black line in the lefthand panel is identical to the line in the ENT panel of Figure 6.

in Figure 6), and the dashed lines lie above the dotted lines for each colour (CW's main effect is negative).

A detailed investigation of these interactions is beyond the scope of this paper, however they appear qualitatively consistent with our understanding of the main physical effects of the relevant variables, which we now summarise.

The effect of reducing ENT is to reduce mixing between air in ascending convective plumes and the surrounding environment, hence increasing the efficiency of convective moisture transport and precipitation. In the control simulation with pre-industrial CO_2, for example, setting $\mathtt{ENT} = 0.6$ (with all other variables kept at their standard values) results in a global balance between precipitation and evaporation being achieved with substantially lower values of cloud and moisture throughout much of the troposphere. In particular, relative humidity values in $\mathtt{ENT} = 0.6$ are much lower in the tropics. The response to doubled CO_2 in $\mathtt{ENT} = 0.6$ shows large increases in tropical relative humidity between 300hPa and 850hPa. This is accompanied by a much weaker negative feedback in the clear-sky component of long wave radiation ($-1.3\,\mathrm{Wm^{-2}K^{-1}}$) than is typically seen in other QUMP simulations, or in simulations with other climate models (values generally range from $-1.7\,\mathrm{Wm^{-2}K^{-1}}$ to $-2.0\,\mathrm{Wm^{-2}K^{-1}}$, see Webb et al. 2006). The difference probably arises mainly from a stronger contribution from water vapour to the clear-sky feedback in $\mathtt{ENT} = 0.6$, compared with typical simulated responses showing much smaller changes in relative humidity (e.g. Soden and Held 2006). If the clear-sky feedback in $\mathtt{ENT} = 0.6$ was altered to a more typical value, the climate sensitivity would be reduced from 7.0°C to ~ 4°C.

In QUMP simulations, a major determinant of variations in climate sensitivity across parameter space (in addition to the impact of ENT on clear-sky fluxes) arises from variations in the contribution of a negative feedback associated with increases in the extent and thickness of low cloud in regions characterised by stable boundary layers (Webb et al. 2006). This feedback tends to be more prevalent in model variants whose control simulations contain relatively high levels of low cloud cover, accompanied by relatively cool and moist boundary layers. The ef-

fect of increasing CW and reducing CT is to inhibit the conversion of cloud water droplets to rain, and therefore favours these characteristics, hence reducing climate sensitivity. We examined a QUMP simulation with low CT, high CW and low ENT, finding that this did not show the large clear-sky feedback discussed above, consistent with the lack of sensitivity to ENT in the dotted red curve of Figure 8 (left panel). This suggests that the negative low cloud feedback in relatively stable regions is able to exert a strong remote influence on surface temperature changes in regions of tropical deep convection, limiting these to a level small enough to avoid triggering the enhanced water vapour feedback seen in model variants with less low cloud in their control simulations (the other curves in Figure 8, left panel). When CT and CW are perturbed to high and low values respectively, the negative low cloud feedback tends to be weaker, hence increasing climate sensitivity. The impact on sensitivity is larger when ENT is smaller (compare blue and black curves in Figure 8, left panel), and is consistent with the standard assumption that individual climate change feedbacks add linearly and independently, and that their effect on the planetary radiation budget scales with the temperature response, implying an inverse relationship with climate sensitivity.

The impact of variables such as CT and CW, which affect the model simulation once cloud is present, is likely to be modulated by variables which affect the ease with which cloud can be formed in the first place. In this regard a key variable is $\mathrm{Rh_{crit}}$, the threshold value of relative humidity for cloud formation (see Table 1). When the switch RHC is Off, $\mathrm{Rh_{crit}}$ takes fixed values prescribed on each model level, and we perturb the value used above the bottom three levels (RHCV). Increasing RHCV reduces the amount of low cloud, and we find that the effect of CT and CW on climate sensitivity (at intermediate and high values of ENT) is smaller for $\mathrm{RHCV} = 0.9$ (shown in Figure 8, right panel) than for lower values of $\mathrm{RHCV} = 0.75$ (not shown). When $\mathrm{RHC} =$ On the model determines $\mathrm{Rh_{crit}}$ dynamically, based on the local variance of cloud water. This has the effect of reducing $\mathrm{Rh_{crit}}$ during episodes of enhanced variability, making it easier to form cloud during the passage of simulated synoptic storms (Cusak et al. 1998). At high values

of ENT, the variation of climate sensitivity with CT and CW when RHC = On is therefore larger than for RHC = Off (*cf* left and right panels of Figure 8), and is in fact very similar to that found with RHC = Off, RHCV = 0.75 (not shown).

A more thorough analysis is needed to confirm the physical mechanisms suggested above. However, this discussion illustrates that the availability of a skilful emulator, within the framework of a perturbed physics ensemble in which particular climate feedbacks can be traced back to specific variables, provides potential to improve our understanding of how detailed physical processes can combine to give rise to different values of climate sensitivity. Given such understanding, it may be possible to develop metrics capable of narrowing the uncertainty in climate sensitivity, and perhaps also regional aspects of climate change, by confronting these processes with relevant observations.

5. Summary

We have constructed an emulator that allows us to predict HadSM3's climate sensitivity at any choice of values for the 31 model parameters varied in the QUMP experiment. This emulator is a statistical framework that allows us to quantify the uncertainty in our predictions. Due to the complexity of the model, and in particular the combination of both continuous and discrete parameters, we are obliged to compromise in our statistical framework, which leaves us with an irreducible uncertainty of a little under 2 °C in our 95% CIs. This 'noise', however, is smaller than the 'signal' coming from varying the parameters, and we are able to identify important sources of variation in the climate sensitivity of HadSM3, which are the large-scale cloud parameters and the entrainment rate coefficient, and investigate the interaction between these parameters, which is complex.

We constructed our emulator from two ensembles. These came from the same underlying model, but in different treatments. The first ensemble, from the CPNET experiment, comprised a large number of relatively quick evaluations over just six of the model parameters. The second ensemble, from the QUMP

experiment, comprised a much smaller number of more time-consuming evaluations, over thirty-one model parameters. Evaluating a model in different configurations is a natural way to increase the efficiency of an experiment, although more typically the difference in configurations is in the resolution of the solver (Craig et al. 1997; Kennedy and O'Hagan 2000). Ideally the two versions would be run interactively, and statistical tools would be used to choose, sequentially, which version to run and at what value of the model parameters to run it. In our case, were we to run both experiments again, we might have used the emulator from the CPNET ensemble to identify the presence of important high-order interactions, and then designed the QUMP ensemble to learn more about these; this type of sequential approach is discussed further in Rougier and Sexton (2007).

Any such approach, that uses the same model (or similar models) in multiple configurations requires a method for assimilating both ensembles into an inference. This will inevitably require judgements about how similar the configurations are. We have chosen to make our judgements explicit, adopting a Bayesian statistical approach which obliges us to quantify that similarity, in terms of the relationship between the emulators for each configuration. Our statistical framework links common coefficients in the two emulators, using a simple parametric relationship (eq. 3) that reduces the quantification to specifying a handful of values. This relationship reduces the burden on the expert, but it is undoubtedly simplistic. It could easily be generalised, for example by applying a different relationship within each parameterisation scheme.

Throughout the paper we have exercised 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 more easily replicated. Where we make choices we have stated them clearly and we have backed 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. We have provided a framework within which it is possible

to work out a number of different choices, and illustrated one particular choice, namely our own.

Acknowledgement Jonathan Rougier has been partly-funded by the U.K. Natural Environment Research Council (NERC), RAPID Directed Programme. David Sexton and James Murphy were supported by the Joint Defra and MoD Programme, (Defra) GA01101 (MoD) CBC/2B/0417 Annex C5. David Stainforth has been partly-funded by NERC under its research fellowship scheme and by the Tyndall Centre for Climate Change. We would like to thank Prof. Michael Goldstein for helping to develop the approach of linked emulators, and for his comments on the statistical modelling. We would also like to thank the three referees for their very perceptive comments on an earlier version of this paper, and the Editor, Sandrine Bony, for her patience.

References

Annan, J., J. Hargreaves, R. Ohgaito, A. Abe-Ouchi, and S. Emori, 2005: Efficiently constraining climate sensitivity with ensembles of paleoclimate simulations. *SOLA*, **1**, 181–184.

Betts, R., O. Boucher, M. Collins, P. Cox, P. Falloon, N. Gedney, D. Hemming, C. Huntingford, C. Jones, D. Sexton, and M. Webb, 2007: Projected increase in continental runoff due to plant responses to increasing carbon dioxide. *Nature*, **448**, 1037–41.

Boer, G. and B. Yu, 2003: Dynamical aspects of climate sensitivity. *Geophysical Research Letters*, **30**, 1135.

Bony, S. and J.-L. Dufresne, 2005: Marine boundary layer clouds at the heart of cloud feedback uncertainties in climate models. *Geophysical Research Letters*, **32**, L20 806.

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.

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

Collins, M., B. Booth, G. Harris, J. Murphy, D. Sexton, and M. Webb, 2006: Towards quantifying uncertainty in transient climate change. *Climate Dynamics*, **27**, 127–147.

Craig, P., M. Goldstein, A. Seheult, and J. Smith, 1997: Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes Linear strategies for large computer experiments. *Case Studies in Bayesian Statistics III*, C. Gatsonis, J. Hodges, R. Kass, R. McCulloch, P. Rossi, and N. Singpurwalla, eds., New York: Springer-Verlag, 37–87, with discussion.

Currin, C., T. Mitchell, M. Morris, and D. Ylvisaker, 1991: Bayesian prediction of deterministic functions, with application to the design and analysis of computer experiments. *Journal of the American Statistical Association*, **86**, 953–963.

Cusak, S., J. Edwards, and R. Kershaw, 1998: Estimating the subgrid variance of saturation, and its parametrization for use in a GCM cloud scheme. *Quarterly Journal of the Royal Meteorological Society*, **125**, 3057–3076.

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

Goldstein, M. and J. Rougier, 2007: Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, forthcoming as

a discussion paper, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf`.

Han, G., T. Santner, W. Notz, and D. Bartel, 2007: Prediction for computer experiments having quantitative and qualitative input variables, in submission.

Harris, G., D. Sexton, B. Booth, M. Collins, J. Murphy, and M. Webb, 2006: Frequency distributions of transient regional climate change from perturbed physics ensembles of general circulation model simulations. *Climate Dynamics*, **27**, 357–375.

Kennedy, M. and A. O'Hagan, 2000: Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, **87**, 1–13.

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.

Murphy, J., B. Booth, M. Collins, G. Harris, D. Sexton, and M. Webb, 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philosophical Transactions of the Royal Society, Series A*, **365**, 1993–2028.

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. *Reliability Engineering and System Safety*, **91**, 1290–1300.

O'Hagan, A. and J. Forster, 2004: *Bayesian Inference*, volume 2b of *Kendall's Advanced Theory of Statistics*. London: Edward Arnold, 2nd edition.

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.

Qian, P., H. Wu, and C. Wu, 2007: Gaussian process models for computer experiments with qualitative and quantitative factors, in submission.

Rougier, J., 2007a: Efficient emulators for multivariate deterministic functions, in submission, currently available at `http://www.maths.bris.ac.uk/~mazjcr/OPemulator.pdf`.

— 2007b: Probabilistic inference for future climate using an ensemble of climate model evaluations. *Climatic Change*, **81**, 247–264.

Rougier, J., S. Guillas, A. Maute, and A. Richmond, 2007: Emulating the Thermosphere-Ionosphere Electrodynamics General Circulation Model (TIE-GCM), in submission, currently available at `http://www.maths.bris.ac.uk/~mazjcr/EmulateTIEGCM.pdf`.

Rougier, J. and D. Sexton, 2007: Inference in ensemble experiments. *Philosophical Transactions of the Royal Society, Series A*, **365**, 2133–2143.

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

Senior, C. and J. Mitchell, 2000: The time dependence of climate sensitivity. *Geophysical Research Letters*, **27**, 2685–2688.

Soden, B. and I. Held, 2006: An assessment of climate feedbacks in coupled ocean-atmosphere models. *Journal of Climate*, **19**, 3354–3360.

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.

Tebaldi, C. and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society, Series A*, **365**, 2053–2075.

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.