Jonathan Rougier, David M. H. Sexton, James M. Murphy, and David A. Stainforth

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

Article (Published version)
(Refereed)

Original citation:
Rougier, Jonathan, Sexton, David M. H, Murphy, James M. and Stainforth, David A. (2009) *Analyzing the climate sensitivity of the HadSM3 climate model using ensembles from different but related experiments*. Journal of Climate, 22 (13). pp. 3540-3557. ISSN 0894-8755
DOI: 10.1175/2008JCLI2533.1

Analyzing the Climate Sensitivity of the HadSM3 Climate Model Using Ensembles from Different but Related Experiments

JONATHAN ROUGIER

Department of Mathematics, University of Bristol, Bristol, United Kingdom

DAVID M. H. SEXTON AND JAMES M. MURPHY

Met Office Hadley Centre, Exeter, United Kingdom

DAVID STAINFORTH

Department of Geography, University of Exeter, Exeter, United Kingdom

(Manuscript received 14 March 2008, in final form 15 July 2008)

ABSTRACT

Global climate models (GCMs) contain imprecisely defined parameters that account, approximately, for subgrid-scale physical processes. The response of a GCM to perturbations in its parameters, which is crucial for quantifying uncertainties in simulations of climate change, can—in principle—be assessed by simulating the GCM many times. In practice, however, such ''perturbed physics'' ensembles are small because GCMs are so expensive to simulate. Statistical tools can help in two ways. First, they can be used to combine ensembles from different but related experiments, increasing the effective number of simulations. Second, they can be used to describe the GCM's response in ways that cannot be extracted directly from the ensemble(s). The authors combine two experiments to learn about the response of the Hadley Centre Slab Climate Model version 3 (HadSM3) climate sensitivity to 31 model parameters. A Bayesian statistical framework is used in which expert judgments are required to quantify the relationship between the two experiments; these judgments are validated by detailed diagnostics. The authors identify the entrainment rate coefficient of the convection scheme as the most important single parameter and find that this interacts strongly with three of the large-scale-cloud parameters.

1. Introduction

The Hadley Centre Slab Climate Model version 3 (HadSM3) comprises the Hadley Centre Atmospheric Model version 3 (HadAM3) atmospheric general circulation model (Pope et al. 2000) coupled to a simple nondynamic mixed layer ocean, a standard setup for the simulation of the equilibrium-temperature response to doubled CO_2 (termed climate sensitivity). 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 parameterizations of subgrid-scale processes. Therefore, they simulate global climate sensitivity differently (Webb et al. 2006). Results from such a multimodel 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 modeling 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

Corresponding author address: Jonathan Rougier, Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, United Kingdom.
E-mail: j.c.rougier@bristol.ac.uk

DOI: 10.1175/2008JCLI2533.1

the simulation of key climate processes within a single model. To date, most published PPE studies have focused on HadSM3 and the third climate configuration of the Met Office Unified Model (HadCM3), the related configuration in which HadAM3 is coupled to a three-dimensional dynamic ocean component (see, e.g., Murphy et al. 2004; Stainforth et al. 2005; Collins et al. 2006; Harris et al. 2006; and further references below). The advantage of the perturbed physics approach is that it supports a more systematic exploration of modeling uncertainties, in which variations in simulated responses can be traced back to particular processes. Their limitation is that they do not explore ''structural'' modeling uncertainties, such as the choice of resolution or alternative approaches for parameterizing subgrid-scale processes. However, results indicate that the spread of global- and large-scale regional climate responses is similar to that found in multimodel 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 $\boldsymbol{\chi}$ 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 are 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 $\boldsymbol{\chi}$ 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 (2007) 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 twenty-first-century climate.

This paper focuses 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 $\mathbf{x}$ into a climate sensitivity value $g(\mathbf{x})$. In our PPE, we have a collection of inputs $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and a corresponding collection of outputs, $\mathbf{y} = \{g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n)\}$. A Bayesian statistical framework, termed an emulator, allows us to predict $g(\mathbf{x})$ at any $\mathbf{x}$, based on the ensemble and on our judgments about the model. Crucially, this prediction takes the form of a distribution, comprising 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 $\mathbf{x}$, 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). The use of statistical emulators in climate prediction, taking account of our uncertainty about the model parameters, is discussed in Rougier and Sexton (2007). ''Nonstatistical emulators'' are also possible and have recently appeared in climate science (see, e.g., Knutti et al. 2006; Sanderson et al. 2008a, using neural networks); more widely, these are sometimes known as surrogates.

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 simulations. But because 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 judgments to augment our ensemble. The additional information comes from a second ensemble of HadSM3 simulations, and the expert judgment concerns the relationship between the two ensembles. As our judgments are subjective, we pay close attention to diagnostic information, in which we contrast our statistical predictions with model simulations.

Using our emulator, we are able to identify the main parameters for determining climate sensitivity and to investigate a complex interaction between four parameters controlling some key aspects of the parameterization 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, using two different PPEs. 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 simulations. Details of the two experiments can be found in the original papers and their supplementary information; here, we summarize those aspects that are relevant for our statistical analysis.

a. The QUMP experiment

In the Quantifying Uncertainty in Model Predictions (QUMP) experiment of Murphy et al. (2004), 31 model parameters were identified as being potentially important, out of a possible 100 or more candidates. These 31 model parameters 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 (e.g., switches that are either off or on), but two have 3 levels (GWST and NFSL) and one has 4 levels (FRF). Of the 18 continuous variables, 4 are contingent on the setting of certain factors; for example, the value of RHCV only affects climate sensitivity when RHC = off; these contingent variables are the reason that Murphy et al. (2004) count 29 rather than 31 variables in their description (they did not include CAPE and ANV).

We denote a particular choice for the values of the variables as $\mathbf{x}$. The climate sensitivity at $\mathbf{x}$ was computed in a three-phase experiment. The first phase was a 25-yr calibration simulation in which sea surface temperatures (SSTs) are continuously 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 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 preindustrial CO_2 and a simulation with doubled CO_2, both run to equilibrium. The heat convergences should ensure that multiyear averages of SSTs in the control simulations remain close to observed climatology, subject to the assumption that internal climate variability in SSTs (suppressed during the first phase, but not in phases two and three) does not give rise to nonlinear feedbacks, which could cause SSTs to drift.

Climate sensitivity, or $g(\mathbf{x})$, was defined as the difference in global mean temperature between the second and third phases. The selection of the 31 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 simulations in the ensemble consisted of single-parameter perturbations augmented by a small number of multiparameter perturbations. Since that initial experiment, we have access to a further 231 simulations, all multiparameter 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 for a total of 297 simulations.

A small minority of these simulations produced control-period SSTs 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 sensitivity computationally feasible and 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)]. We find 16 model variants in which global mean SST in the control simulation cools in this way (‘‘drifters’’). 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 simulations that remain provide estimates of sensitivity free from nonphysical side effects of the experimental design. This is demonstrated, for example, by the close relationship between the equilibrium surface warming found in 17 of these simulations 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 the 281 reliable simulations to supply estimates of climate sensitivity over the whole model parameter space, including around those 16 locations for which cooling HadAM3 simulations were excluded.

b. 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 parameterization of large-scale clouds and convection. The ensemble

TABLE 1. Description of the QUMP variables. Comparable to Murphy et al. (2004, supplementary information, their Table 2). Each parameter controls a key aspect of one of the schemes for the parameterization of subgrid-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.

Parameter/property	Values	Label	Only when
	Large-scale cloud		
Vf1 (m s^{-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 (h)	(1, 2, 4)	CAPEV	CAPE on
Convective anvils	Off,* on	ANV	
Convective anvils, shape	(1, 2, 3)	ANVS	ANV on
Convective anvils, updraft	(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*)	DID	
	Radiation		
Ice particle size (μm)	(25, 30,* 40)	IPS	
Nonspherical ice particles	Off,* on	NSIP	
Shortwave water vapor continuum absorption	Off,* on	SWV	
Sulfur cycle	Off,* on	SCYC	
	Dynamics		
Order of diffusion operator	4, 6*	ODD	
Diffusion *e*-folding time (h)	(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	

* The standard setting.

** Variables also used in CPNET.

comprises a factorial design with five variables at 3 levels (VF1, CT, CW, RHCV, and ENT; RHC was always off) and one variable at 2 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 in the CPNET experiment was simulated with a number of different initial conditions, introducing a structured source of uncertainty that is not present in the QUMP experiment. On analyzing the CPNET ensemble, we find that the choice of initial condition does not appear to be predictively important for climate sensitivity and so we pool the simulations across the initial conditions; a similar approach was used in the Stainforth et al. (2005) experiment, in which different initial conditions for the same **x** were averaged.

The CPNET experiment adopted a public resource distributed computing (PRDC) approach, performing thousands of simulations 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 15 yr 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 simulations (in terms of the **x** values) and 2377 simulations overall (accounting for variations in the initial conditions). Many of these produced unstable or nonphysical 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).

The following summarizes the differences between the two experiments:

1) Our CPNET ensemble varies 6 parameters, whereas QUMP varies 31.
2) CPNET explores initial condition uncertainty, whereas QUMP does not. (This is not thought to be important for climate sensitivity but may be for other variables.)
3) CPNET uses a 15-yr calibration phase, whereas QUMP uses a 25-yr calibration phase.
4) CPNET does exactly 15 yr each of preindustrial and doubled CO_2 phases, whereas QUMP runs both of these phases to equilibrium.

Both the third and fourth differences will affect the operational definition of climate sensitivity. The fourth difference is the most important because it means that in CPNET the climate sensitivity has to be extrapolated from the simulations, rather than being computed directly. Comparing these two experiments, we judge there to be sufficient differences in 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 behavior of the HadSM3 model. This informs our statistical modeling choices in section 3c.

c. *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, but we restrict ourselves here to the ensemble in Stainforth et al. (2005)]. Our intention is to combine the ensembles from these two experiments into an emulator for QUMP climate sensitivity defined over the full set of 31 variables. It is difficult to draw any firm conclusions about the similarity of the two definitions over the whole of the parameter space, given the limited amount of data we have from the QUMP experiment. But it is our judgment that it would be best (i.e., conservative) to treat them as not only operationally different but also potentially practically different.

As already described, an emulator is a probability distribution function for $g(\mathbf{x})$. There are many ways of specifying such a function. In a Bayesian statistical approach we probabilistically condition our beliefs about $g(\cdot)$ on the simulations in the ensemble. Therefore a Bayesian emulator combines two sources of information: prior judgments about $g(\cdot)$ and data from simulations in the ensemble ($\mathbf{y}$; $\mathbf{X}$), in which $\mathbf{X}$ is the ''design matrix'' of parameter values and $\mathbf{y}$ the resulting vector of climate sensitivities. The main stages of our approach are summarized in Fig. 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 simulations 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 simulations in a 31-dimensional space). Therefore we combine this with detailed prior information taken from the CPNET emulator and our judgment 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.

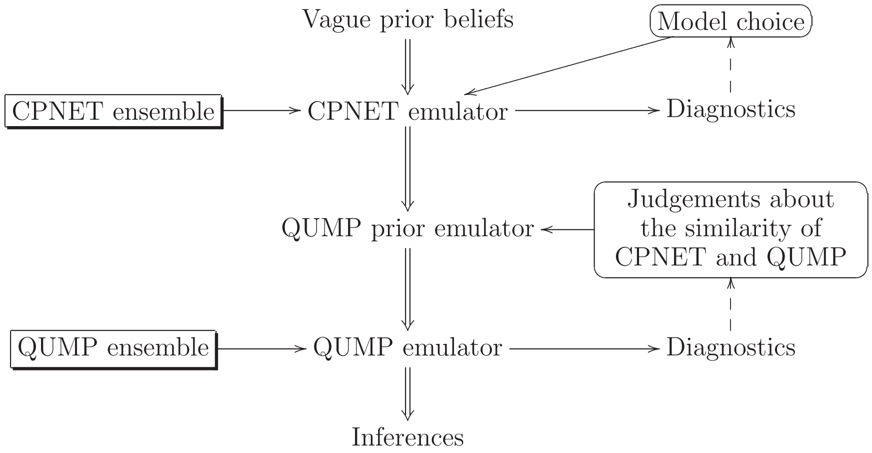

FIG. 1. The main stages of our approach for combining information from the CPNET and QUMP ensembles into an emulator for QUMP climate sensitivity. Starting with vague prior beliefs, we create the CPNET emulator using the CPNET ensemble. Then, we use our judgments about the similarities of the CPNET and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

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, because of the complexities of the model parameters, as discussed in section 3a.

Finally, it may be helpful to compare our approach with a similar treatment by Sanderson et al. (2008a). In this paper, an emulator is constructed for CPNET climate sensitivity from a more recent, much larger ensemble (6096 simulations); 11 parameters were varied independently, of which the original CPNET 6 parameters are a subset. A neural network is used rather than a statistical model, with uncertainty estimates derived from bootstrap resampling. This approach would not be effective for emulating QUMP climate sensitivity because it relies on a large ensemble for both training the neural network and for uncertainty estimation. One interesting feature of the Sanderson et al. (2008a) emulator is the treatment of the 2-level factors as continuous on the interval [0, 1], even though only the values 0 and 1 are attainable in the model. By contrast, we have treated all of the factors as factors (i.e., as discrete quantities without numerically equivalent values); in other words, we do not make the additional assertion that interpolated values of the switches also give rise to physically meaningful outcomes. We also ensure that combinations of factors and continuous variables operate correctly; for example, RHCV only affects climate sensitivity when RHC = off (see Table 1). Sanderson et al. (2008a) do not have to tackle this issue because the parameter space of their experiment is simpler. But there is no reason why it should not be implemented in a neural network with a suitable reorganization of the input layer.

3. Emulating HadSM3's climate sensitivity

In this section we describe our approach for emulating HadSM3's climate sensitivity, as outlined in Fig. 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 judgments 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.

a. A general Bayesian emulator

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

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

in which $g(\mathbf{x})$ is the climate sensitivity of HadSM3 or some monotonic transformation of the same, termed the response; $\mathbf{h}(\cdot)$ is a known vector-valued function of the variables, collectively termed the regressors (k in total); $\boldsymbol{\beta}$ is an unknown k vector of (regression) coefficients; and

$u(\mathbf{x})$ is a scalar stochastic process, termed the residual. Within the regressors, we would expect to include nonlinear functions of the variables, such as x_i^2 or $x_i \times x_j$. We must use our judgment, in conjunction with the data where possible, to make choices for the transformation of $g(\cdot)$ and the components of $\mathbf{h}(\cdot)$: 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 $\mathbf{x}$ 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 behavior.

For our given choice for the response and the regressors, we make the following additional choices: First, $u(\mathbf{x})$ has zero mean and a constant unknown variance σ^2; second, $u(\mathbf{x})$ and $u(\mathbf{x}')$ are uncorrelated when $\mathbf{x} \neq \mathbf{x}'$; and third, $\boldsymbol{\beta}$, $u(\mathbf{x})$, and σ^2 have a normal-inverse-gamma (NIG) distribution, which may be summarized as

$$\boldsymbol{\beta} \perp\!\!\!\perp u(\mathbf{x})|\sigma^2, \tag{2a}$$

$$\boldsymbol{\beta}|\sigma^2 \sim N_k(\mathbf{m}, \sigma^2\mathbf{V}), \tag{2b}$$

$$u(\mathbf{x})|\sigma^2 \sim N_1(0, \sigma^2), \quad \text{and} \tag{2c}$$

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

where $\perp\!\!\!\perp$ denotes probabilistically independent, | denotes conditional upon, $N_k(\cdot)$ denotes the k-dimensional Gaussian distribution, and IG$(\cdot)$ denotes the scalar inverse gamma distribution; we must specify the collection $\{a, d, \mathbf{m}, \mathbf{V}\}$, termed the hyperparameters. With these distributional choices, the emulator for $g(\mathbf{x})$ has a Student's t distribution in which both the mean and the scale will depend on $\mathbf{x}$. We have outlined here the standard Bayesian treatment of the Gaussian linear model; full details may be found in O'Hagan and Forster (2004, chapter 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 (2008b) for a full description and Rougier et al. (2008, manuscript submitted to *Technometrics*) for an example]; however, it has some undesirable features (see, e.g., O'Hagan and Forster 2004, sections 11.43–11.70). But we have made one unusual choice, which is to treat the residual as having zero correlation length {i.e., to set Cov$[u(\mathbf{x}), u(\mathbf{x}')] = 0$ for $\mathbf{x} \neq \mathbf{x}'$}. 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 (Rougier 2008a). 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 $u(\mathbf{x})$ that 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. 2009; Qian et al. 2008). An alternative approach 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 $2^{10} \times 3^2 \times 4^1 = 36\,864$ factor combinations. Another alternative would be to treat the factors as though they were continuous, as done in Sanderson et al. (2008a), but we prefer to leave them as ''switches'' and avoid an additional assertion about the model (as already discussed at the end of section 2c).

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 predicatively 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 behavior 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 summarize this section, the challenge of building an emulator for $g(\cdot)$ using the ensemble $(\mathbf{y}; \mathbf{X})$ has been restructured to (i) choosing a transformation for climate sensitivity and a collection of regressors $\mathbf{h}(\cdot)$, and, conditional on these choices, (ii) specifying the hyperparameters $\{a, d, \mathbf{m}, \mathbf{V}\}$ in the NIG prior for $\{\boldsymbol{\beta}, u(\mathbf{x}), \sigma^2\}$.

b. Building the CPNET emulator

As explained in section 2c and illustrated in Fig. 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 $\{\boldsymbol{\beta}, u(\mathbf{x}), \sigma^2\}$, as summarized in the hyperparameters $\{a, d, \mathbf{m}, \mathbf{V}\}$. The standard noninformative prior has $a = 0$; $d = -k$, where k is the number of regressor functions in $\mathbf{h}(\cdot)$; $\mathbf{m} = \mathbf{0}$; and $\mathbf{V}^{-1} = \mathbf{0}$ (O'Hagan and Forster 2004, sections 11.17–11.19). In this case, the posterior distribution for $\boldsymbol{\beta}|\sigma^2$ has the usual ordinary–least squares (OLS) form, although the interpretation is a little different, being Bayesian rather than Frequentist. When we refer to, for example, 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, section 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 use the residuals for diagnostic information. The QUMP authors, who explicitly constructed an emulator for their analysis, chose 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 (Sanderson et al. 2008b make the same choice). 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 will investigate a wider range of possible power transformations, including the logarithm, using the Box–Cox approach (see, e.g., Draper and Smith 1998, section 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. 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 parameterization schemes as noninteracting (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 2 levels in the CPNET ensemble. For the same reason we cannot estimate cubic or higher effects in any of the variables. A statistician wishing to understand how the model response varies across parameter space 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 studies attempt to fulfill a number of different and not necessarily compatible objectives.

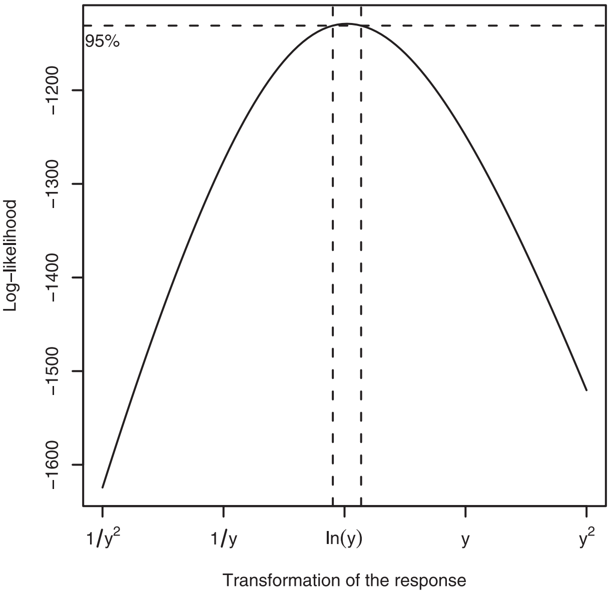

FIG. 2. Box–Cox plot to select an appropriate transformation for the response: the high likelihood values are concentrated around the logarithm rather than the reciprocal (the vertical dashed lines indicate an approximate 95% confidence interval).

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 Fig. 2. Note that log(**y**) is strongly favored over 1/**y**, or climate feedback, which is one way in which our emulator differs from other approaches.

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 (see, e.g., Draper and Smith 1998, chapter 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 15 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 $\mathbf{h}(\cdot)$, but we do not include any others. This gives a total of 37 regressor functions in $\mathbf{h}(\cdot)$, including the intercept.

To summarize the resulting emulator, 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. Thus, there is strong evidence for the importance of interactions in determining HadSM3's climate sensitivity, supporting the conclusions—for CPNET—of Stainforth et al. (2005) and Sanderson et al. (2008b).

c. *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: recollect that the operational definition of climate sensitivity in the CPNET and QUMP experiments is different. First, we must choose a collection of regressors for the QUMP emulator: these will be a superset of the regressors for the CPNET emulator, because QUMP has 25 additional variables. Then, we must use our judgment about the relationship between the CPNET and QUMP experiments to map the CPNET emulator hyperparameters $\{a^0, d^0, \mathbf{m}^0, \mathbf{V}^0\}$ to the QUMP prior hyperparameters $\{a, d, \mathbf{m}, \mathbf{V}\}$. In sections 3d and 3e, we introduce the QUMP ensemble to generate diagnostics for our statistical choices and to update the QUMP hyperparameters to their final values $\{a^*, d^*, \mathbf{m}^*, \mathbf{V}^*\}$.

1) The regressors

For our QUMP-emulator regressors, we start with all the 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 also the 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 parameterization scheme, but we do not include any interactions between processes, barring 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; for example, 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 3-level factor; likewise "land surface" has 17 interactions because FRF is a 4-level factor and NFSL is a 3-level factor.

2) 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 following general framework:

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

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 right-hand side of (3). Two of these terms are straightforward: r_y is the typical scale of the transformed response and r_i is 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 of 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 nonzero value might be preferred (see, e.g., Goldstein and Rougier 2009).

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 the following:

$$\beta_i \approx (1 + \omega_i)\beta_i^0, \qquad (4)$$

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'' (note that 3% is the largest probability consistent with a unimodal distribution: for a Gaussian distribution it would be a small fraction of 1%). 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.

3) 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.

4) 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.

5) Completing the calculation

Once we have computed $\{a, d\}$, we can use these two values along with the values $\{a^0, d^0, \mathbf{m}^0, \mathbf{V}^0\}$, the frameworks (3) and (5), and our choices for the standard deviations of the ω_i and ν_i to compute the hyperparameters $\mathbf{m}$ and $\mathbf{V}$ in the QUMP emulator, by matching the mean and variance of the multivariate Student's t distribution.

d. 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 ''overfitting,'' 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 simulations in the QUMP ensemble. Each individual prediction, taken marginally, has a Student's t distribution. In Fig. 3, we show all 281 predictions in terms of their median and 95% CI and 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; that is, that our predictions are not insensitive to the values for $\mathbf{x}$. We can also confirm that there is no apparent systematic misprediction with respect to the response. This diagnostic suggests that we have overstated uncertainty, as all 281 values are well within the 95% CI that we predict. We could impose constraints on $\mathrm{Var}[g(\mathbf{x})]$ and use these to modify our statistical modeling of NIG hyperparameters such as $\mathbf{V}$. 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 overtuned our prior.

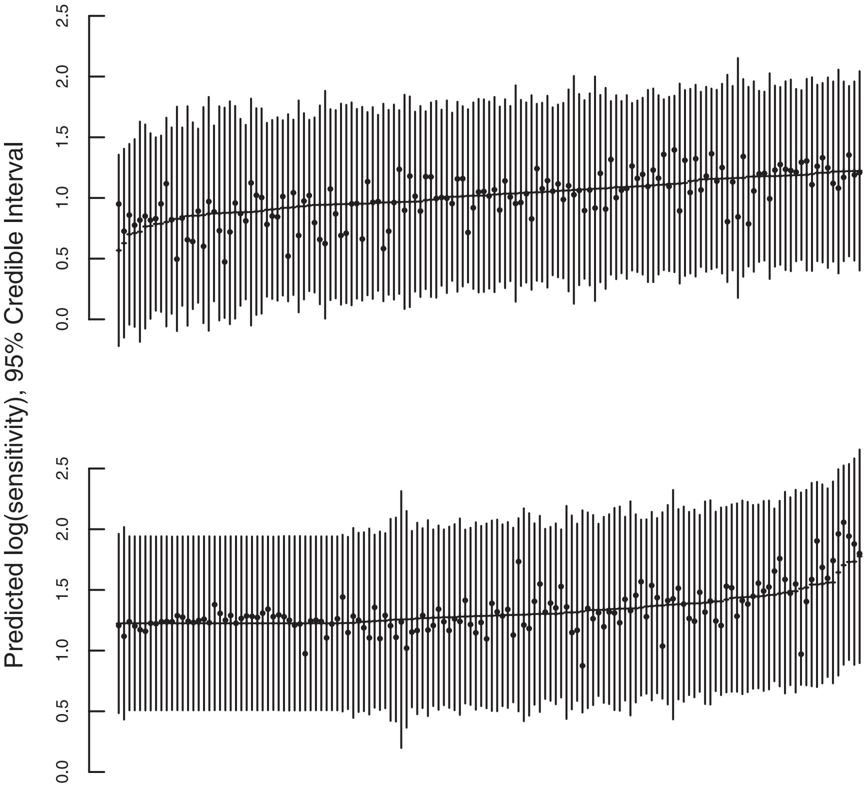

FIG. 3. Prior prediction diagnostic showing, for each simulation in the QUMP ensemble, the prior median and 95% CI, along with the actual value of the response (dot). The evaluations are ordered by the median.

Note that the cluster of similar simulations on the left side of the bottom panel of Fig. 3 corresponds to the simulations 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 compared to the standard value. Note, however, that variables that 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).

e. *Posterior diagnostics*

We also consider a second set of diagnostics, which investigate the posterior predictive properties of the QUMP emulator. One such diagnostic is broadly comparable with the univariate prior prediction given in Fig. 3: the leave-one-out diagnostic (see, e.g., Rougier et al. 2008, manuscript submitted to *Technometrics*). In this case, we update the emulator with all but one simulation from the QUMP ensemble and then predict that simulation. We can do this with all 281 simulations; the result is shown in Fig. 4. Because 280 is almost the same as 281, the width of the intervals in Fig. 4 is a good guide to the amount of uncertainty we will have in our QUMP emulator. By comparing the widths in Figs. 3 and 4, we can quantify the contribution of the QUMP ensemble in reducing our uncertainty about our chosen definition of climate sensitivity in HadSM3. On the log scale, this uncertainty has been reduced by more than 50%.

In all, 13 of the 281 actual values for log(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 modeling 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 behavior of a collection of predictions, taking this correlation into account. For this purpose, we select every third simulation and update using the others ("leave 93

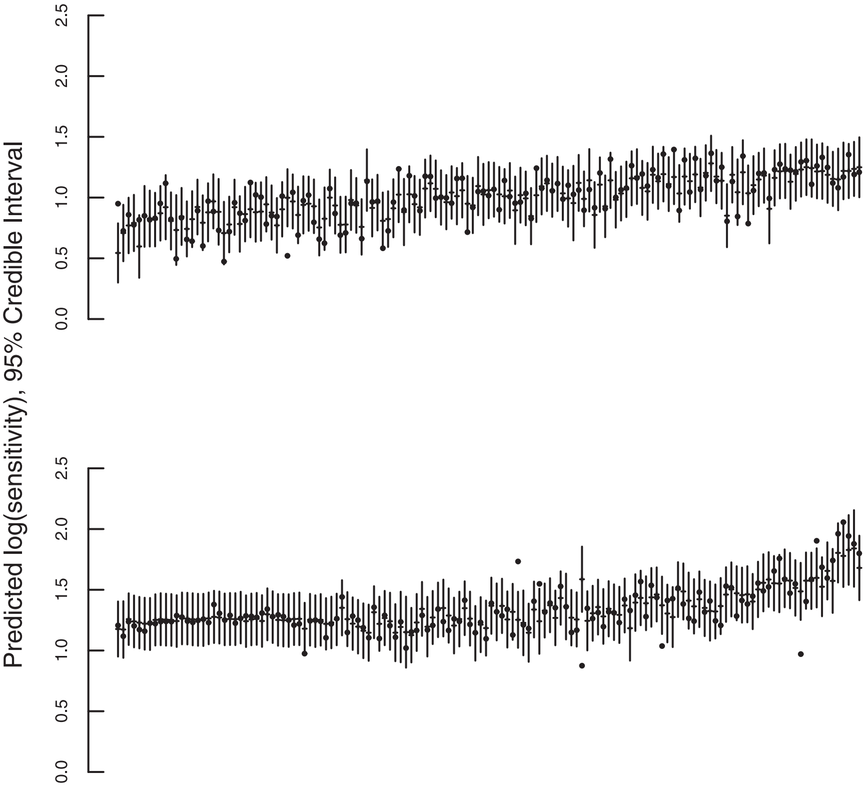

FIG. 4. Posterior prediction leave-one-out diagnostic showing, for each simulation in the QUMP ensemble, the posterior median and 95% CI after updating with the other 280 evaluations. The evaluations have the same ordering as in Fig. 3.

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 shows the result as a quantile–quantile (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 misfitting, 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 $\{a^*, d^*, \mathbf{m}^*, \mathbf{V}^*\}$, we investigate the response of HadSM3's climate sensitivity to the 31 variables.

a. Main effects

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 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 Fig. 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, because the predicted values are included in the emulator). They indicate that the large-scale cloud parameters plus the entrainment coefficient are the important variables (left column of Fig. 6). In particular, climate sensitivity is highly sensitive to low values of the entrainment rate coefficient (ENT). Any analysis that accounts for uncertainty in the ''correct'' value of entrainment will be sensitive to the choice of distribution; for example, uniform in ENT and uniform in the reciprocal of ENT 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.

The main effects shown in these two figures can be compared with the results in Sanderson et al. (2008a, Fig. 6), bearing in mind that they are analyzing CPNET climate sensitivity, not QUMP climate sensitivity, and

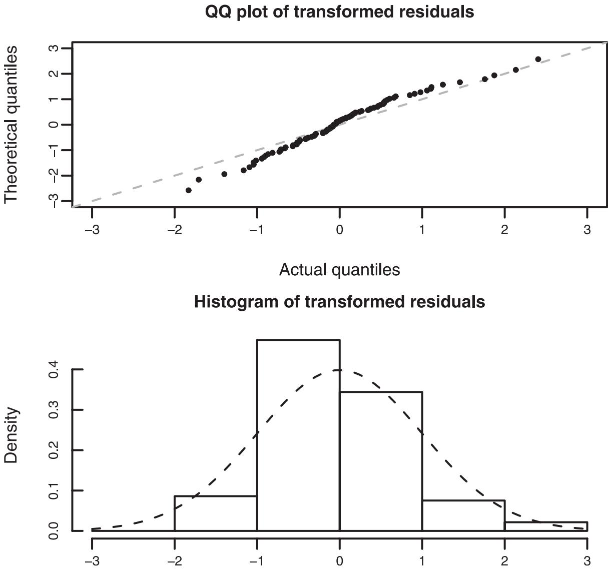

FIG. 5. Diagnostics from the joint prediction of every third member of the QUMP ensemble, using the other members (i.e., leave 93 out). After transformation, each prediction error should have a standard Student's *t* distribution. (left) The QQ plot for the prediction errors and (right) the histogram, with the Student's *t* density overlaid.

use a different approach to emulation and uncertainty estimation.

The main features of the comparison are

1) QUMP climate sensitivity appears to be systematically higher than CPNET climate sensitivity over the full range of parameter values. The amount of the difference varies but appears to be never less than about 1/2°C.
2) Our uncertainty about QUMP climate sensitivity is systematically larger than Sanderson et al.'s (2008a) uncertainty about CPNET climate sensitivity (σ^2 values of about 3/4°C and 1/3°C, respectively). This is attributable in part to the much smaller number of QUMP evaluations but also to our decision to treat switches as factors, rather than as continuous variables (see immediately below).
3) The same main effects dominate, although the effect of varying these dominant parameters seems to be slightly more pronounced in QUMP than in CPNET. The QUMP main effect in CT is concave, whereas it is convex in CPNET.

One interesting feature of CPNET sensitivity is that the main effects appear to be monotonic and have quite simple shapes. This emerges as an inference from Sanderson et al.'s (2008a) neural net emulator (which has the flexibility to fit more complicated relationships) but is a choice we impose on our QUMP emulator; although, in a sense, it is an inference for us too, because our emulator satisfies diagnostic checking.

At this point, we can clarify the practical implication of having a correlation length of zero in the emulator residual, $u(\mathbf{x})$, discussed in section 3a. Ideally, our emulator should interpolate the values in the ensemble to within the uncertainty due to internal variability, roughly ±1/5°C, but the uncertainty is typically ±3/4°C. We cannot easily reduce this uncertainty by doing further simulations of HadSM3, because it represents a limitation of the statistical model, not of the data. Note, however, that this uncertainty, although comparable in size to the main effects of each variable, is much less than the combined effect of several variables, as we now illustrate.

b. *Interactions*

We examine the effect of interactions between the large-scale-cloud variables and the entrainment rate 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 Fig. 8. This figure, in which we display the response to a large set of carefully chosen combinations of parameter values, can only be constructed with an emulator, although broadly similar conclusions can be drawn in other ways (Sanderson et al. 2008b).

Figure 8 shows the interaction between ENT, along the horizontal axis, and CT and CW, shown in a four-way layout of low and high values. The two panels vary the setting of the switch RHC. The solid line in the left panel is identical to the median line in the ENT panel of Fig. 6. For clarity, observe that the dotted lines lie above the dashed lines for each symbol style (CT's main effect is positive, as shown in Fig. 6), and the circles lie above the triangles for each line style (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 summarize. This summary illustrates that the availability of a skillful emulator, within the framework of a perturbed physics ensemble in which particular climate feedbacks can be traced back to specific variables, provides the potential to improve our understanding of how detailed physical processes can combine to give rise to different values of climate sensitivity.

First, consider the left panel, with RHC = on. 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

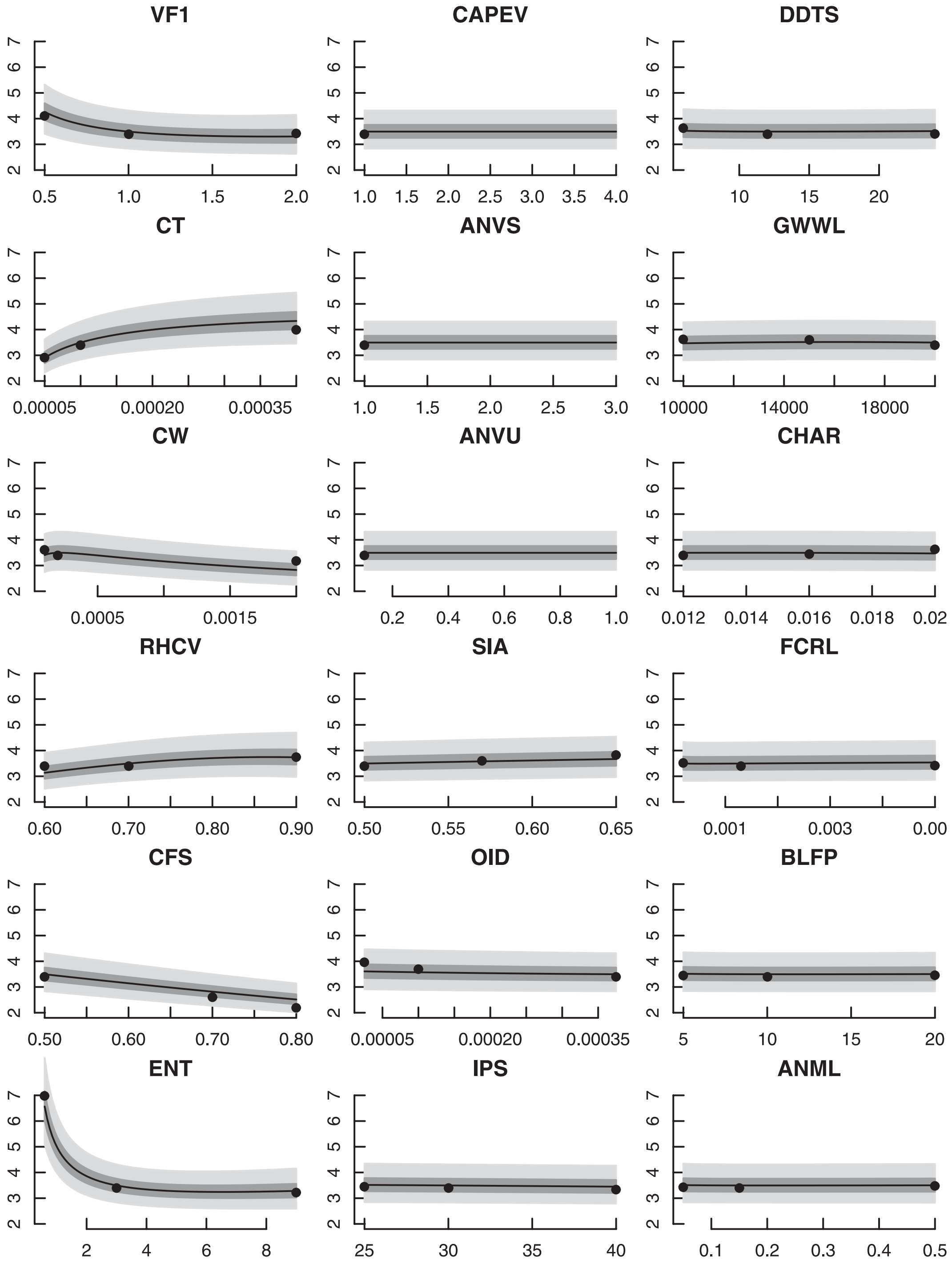

FIG. 6. The effect on climate sensitivity of each of the continuous variables. All other variables are set to their standard values. The line shows the median, the two envelopes show the pointwise 50% and 95% credible intervals. The dots show actual values from the initial stage (single-parameter perturbations) of the QUMP experiment.

moisture transport and precipitation. In the control simulation with preindustrial CO_2, for example, setting 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 ENT = 0.6 are much lower in the tropics (Sanderson

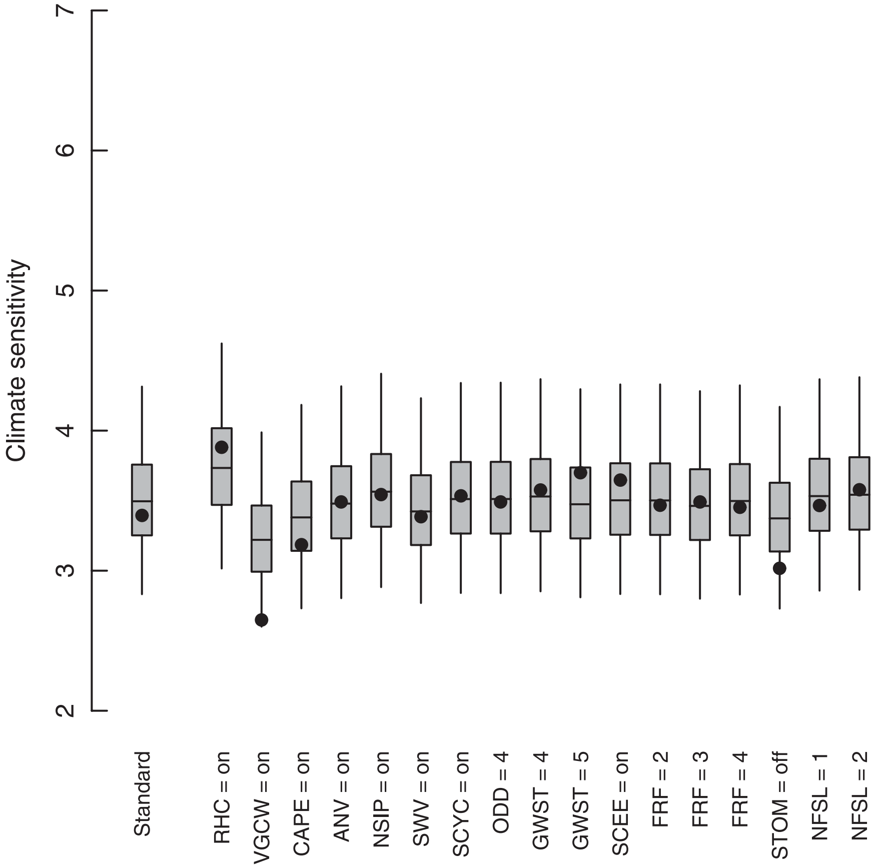

FIG. 7. The effect on climate sensitivity of each of the factors. (left) Climate sensitivity predicted at the standard settings. The other columns show the effect of changing one factor at a time. The box shows the 50% CI, the whiskers show the 95% CI, and the central bar shows the median. The dots show actual values from the initial stage of the QUMP experiment. The vertical scale is the same as in Fig. 6.

et al. 2008b). The response to doubled CO_2 in ENT = 0.6 shows large increases in tropical relative humidity between 300 and 850 hPa. This is accompanied by a much weaker negative feedback in the clear-sky component of longwave radiation (-1.3 W m^{-2} K^{-1}) than is typically seen in other QUMP simulations or in simulations with other climate models (values generally range from -1.7 to -2.0 W m^{-2} K^{-1}; see Webb et al. 2006). The difference probably arises mainly from a stronger contribution from water vapor to the clear-sky feedback (Sanderson et al. 2008b) in 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 ENT = 0.6 was altered to value more typical of other models, the climate sensitivity would be reduced from 7.0° 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 characterized by stable boundary layers (Webb et al. 2006). This feedback tends to be more prevalent in model variants whose control simulations contain relatively large amounts of low cloud cover accompanied by relatively cool and moist boundary layers. The effect of increasing CW and reducing CT is to inhibit the conversion of cloud water droplets to rain, and therefore favors 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 dashed-triangle curve of Fig. 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

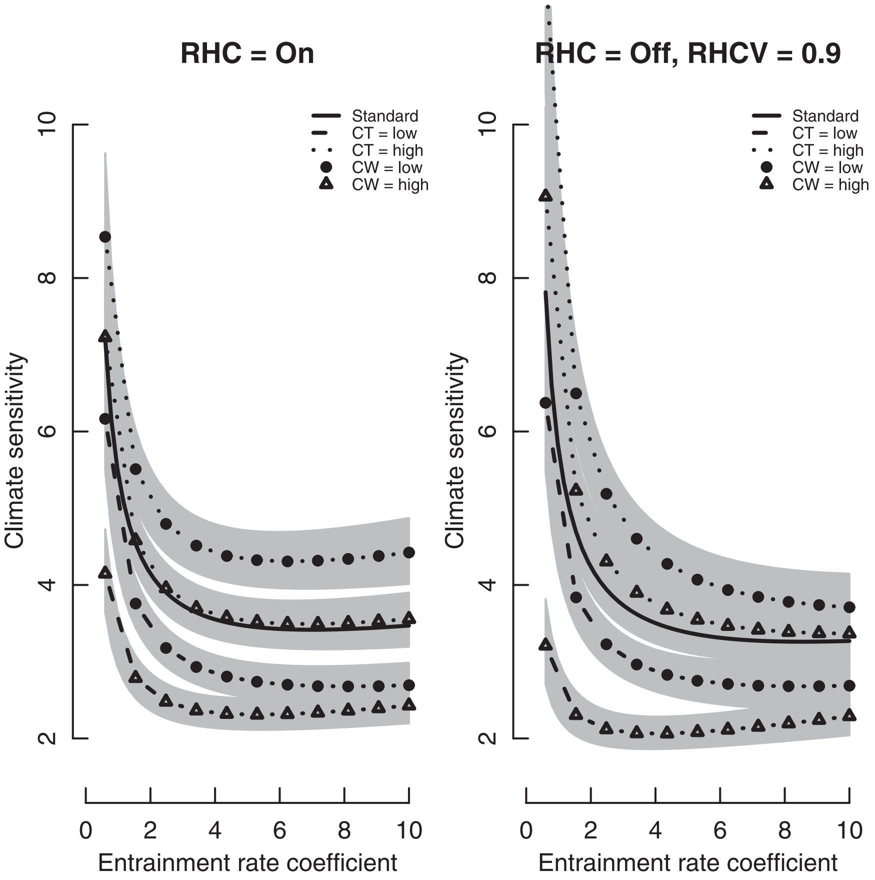

FIG. 8. Interaction between the entrainment rate (ENT) and three large-scale-cloud variables. Each line shows the median response of climate sensitivity to ENT. For the black line, the variables CT and CW are at their standard settings; (left) the black line is identical to the ENT line in Fig. 6. Four other lines are shown: line styles indicate values of CT and symbols indicate values of CW. The shaded envelope indicates the pointwise 50% CI for each line (note that it is 50%, not 95%). (left) RHC is on; (right) RHC is off and RHCV = 0.9.

water vapor feedback seen in model variants with less low cloud in their control simulations (the other curves in Fig. 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.

Now consider the effect of RHC; a simple comparison of the two panels in Fig. 8 indicates that this effect is small but not insubstantial. 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 that affect the ease with which cloud can be formed in the first place. In this regard, a key variable is Rh_{crit}, the threshold value of relative humidity for cloud formation (see Table 1). When the switch RHC is off, Rh_{crit} takes fixed values prescribed on each model level and we perturb the value used above the bottom 3 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 RHCV = 0.9 (shown in Fig. 8, right panel) than for lower values of RHCV = 0.75 (not shown). When RHC = on, the model determines Rh_{crit} dynamically, based on the local variance of cloud water. This has the effect of reducing 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 and is in fact very similar to that found with RHC = off and RHCV = 0.75 (not shown).

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, in conjunction with judgments about the ''best'' value of the parameters and the model discrepancy (Rougier 2007; Rougier and Sexton 2007). Because of 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 about ±3/4°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 simulations over just six of the model parameters. The second ensemble, from the QUMP experiment, comprised a much smaller number of more time-consuming simulations, over 31 model parameters. Simulating 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 judgments about how similar the configurations are. We have chosen to make our judgments explicit, adopting a Bayesian statistical approach that 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 tractable parametric relationship [Eq. (3)] that reduces the quantification to specifying a handful of values. This relationship reduces the burden on the expert but is undoubtedly simplistic. It could easily be generalized, for example, by applying a different relationship within each parameterization scheme.

Throughout the paper we have exercised our judgment to create the best emulator that we can, subject to various constraints such as transparency and tractability; we favor these constraints because they allow our approach to be more easily replicated. Where we make choices, we have stated them clearly and 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 we have illustrated one particular choice, namely, our own.

Acknowledgments. This work was supported by the Joint Defra and MoD Programme, (De-fra) GA01101 (MoD) CBC/2B/0417 Annex C5. Jonathan Rougier has been partly funded by the U.K. Natural Environment Research Council (NERC) RAPID Directed Programme. David Stainforth has been partly funded by NERC under its research fellowship scheme and by the Tyndall Centre for Climate Change. We thank Professor Michael Goldstein for helping to develop the approach of linked emulators and for his comments on the statistical modeling. We would also like to thank a total of five reviewers for their very perceptive comments on two earlier versions of this paper.

REFERENCES

Betts, R. A., and Coauthors, 2007: Projected increase in continental runoff due to plant responses to increasing carbon dioxide. *Nature,* **448,** 1037–1041.

Boer, G. J., and B. Yu, 2003: Dynamical aspects of climate sensitivity. *Geophys. Res. Lett.,* **30,** 1135, doi:10.1029/2002GL016549.

Bony, S., and J.-L. Dufresne, 2005: Marine boundary layer clouds at the heart of cloud feedback uncertainties in climate models. *Geophys. Res. Lett.,* **32,** L20806, doi:10.1029/2005GL023851.

Chipman, H., M. Hamada, and C. F. J. Wu, 1997: A Bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics,* **39,** 372–381.

Collins, M., B. Booth, G. Harris, J. Murphy, D. Sexton, and M. Webb, 2006: Towards quantifying uncertainty in transient climate change. *Climate Dyn.,* **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, Volume III,* C. Gatsonis et al., Eds., Lecture Notes in Statistics, Vol. 121, Springer-Verlag, 37–87.

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. *J. Amer. Stat. Assoc.,* **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. *Quart. J. Roy. Meteor. Soc.,* **125,** 3057–3076.

Draper, N. R., and H. Smith, 1998: *Applied Regression Analysis*. 3rd ed. John Wiley & Sons, 706 pp.

Goldstein, M., and J. Rougier, 2009: Reified Bayesian modelling and inference for physical systems. *J. Stat. Plann. Inference,* **139,** 1121–1239.

Han, G., T. Santner, W. Notz, and D. Bartel, 2009: Prediction for computer experiments having quantitative and qualitative input variables. *Technometrics,* in press.

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 Dyn.,* **27,** 357–375.

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

Knutti, R., G. Meehl, M. Allen, and D. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate,* **19,** 4224–4233.

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

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.

——, 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. *Philos. Trans. Roy. Soc. London,* **365A,** 1993–2028.

O'Hagan, A., 2006: Bayesian analysis of computer code outputs: A tutorial. *Reliab. Eng. Syst. Saf.,* **91,** 1290–1300.

——, and J. Forster, 2004: *Bayesian Inference*. Vol. 2B, *Kendall's Advanced Theory of Statistics,* 2nd ed. Edward Arnold, 480 pp.

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

Pukelsheim, F., 1994: The three sigma rule. *Amer. Stat.,* **48,** 88–91.

Qian, P. Z. G., H. Wu, and C. F. J. Wu, 2008: Gaussian process models for computer experiments with qualitative and quantitative factors. *Technometrics,* **50,** 383–396.

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

——, 2008a: Comment on article by Sansó et al. *Bayesian Anal.,* **3,** 45–56.

——, 2008b: Efficient emulators for multivariate deterministic functions. *J. Comput. Graphical Sci.,* **17,** 827–843, doi:10.1198/106186008X384032.

——, and D. Sexton, 2007: Inference in ensemble experiments. *Philos. Trans. Roy. Soc. London,* **365A,** 2133–2143.

Sanderson, B. M., and Coauthors, 2008a: Constraints on model response to greenhouse gas forcing and the role of subgrid-scale processes. *J. Climate,* **21,** 2384–2400.

——, C. Piani, W. J. Ingram, D. A. Stone, and M. R. Allen, 2008b: Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. *Climate Dyn.,* **30,** 175–190.

Santner, T. J., B. J. Williams, and W. I. Notz, 2003: *The Design and Analysis of Computer Experiments*. Springer, 283 pp.

Senior, C., and J. Mitchell, 2000: The time dependence of climate sensitivity. *Geophys. Res. Lett.,* **27,** 2685–2688.

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

Stainforth, D., and Coauthors, 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. *Philos. Trans. Roy. Soc. London,* **365A,** 2053–2075.

Webb, M., and Coauthors, 2006: On the contribution of local feedback mechanisms to the range of climate sensitivity in two GCM ensembles. *Climate Dyn.,* **27,** 17–38.