Clim Dyn (2013) 41:1703–1729
DOI 10.1007/s00382-013-1896-4

History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble

Daniel Williamson · Michael Goldstein · Lesley Allison · Adam Blaker · Peter Challenor · Laura Jackson · Kuniko Yamazaki

Received: 18 June 2012 / Accepted: 23 July 2013 / Published online: 23 August 2013

Abstract We apply an established statistical methodology called history matching to constrain the parameter space of a coupled non-flux-adjusted climate model (the third Hadley Centre Climate Model; HadCM3) by using a 10,000-member perturbed physics ensemble and observational metrics. History matching uses emulators (fast statistical representations of climate models that include a measure of uncertainty in the prediction of climate model output) to rule out regions of the parameter space of the climate model that are inconsistent with physical observations given the relevant uncertainties. Our methods rule out about half of the parameter space of the climate model even though we only use a small number of historical observations. We explore 2 dimensional projections of the remaining space and observe a region whose shape mainly depends on parameters controlling cloud processes and one ocean mixing parameter. We find that global mean surface air temperature (SAT) is the dominant constraint of those used, and that the others provide little further constraint after matching to SAT. The Atlantic meridional overturning circulation (AMOC) has a non linear relationship with SAT and is not a good proxy for the meridional heat transport in the unconstrained parameter space, but these relationships are linear in our reduced space. We find that the transient response of the AMOC to idealised CO_2 forcing at 1 and 2 % per year shows a greater average reduction in strength in the constrained parameter space than in the unconstrained space. We test extended ranges of a number of parameters of HadCM3 and discover that no part of the extended ranges can by ruled out using any of our constraints. Constraining parameter space using easy to emulate observational metrics prior to analysis of more complex processes is an important and powerful tool. It can remove complex and irrelevant behaviour in unrealistic parts of parameter space, allowing the processes in question to be more easily studied or emulated, perhaps as a precursor to the application of further relevant constraints.

Keywords Bayesian uncertainty quantification · History matching · Implausibility · Observations · NROY space

D. Williamson (✉) · M. Goldstein
Department of Mathematical Sciences, Durham University, Durham DH1 3LE, UK
e-mail: d.williamson@exeter.ac.uk

L. Allison
NCAS-Climate, Department of Meteorology, University of Reading, Reading RG6 6BB, UK

A. Blaker · P. Challenor
National Oceanography Centre, Southampton SO14 3ZH, UK

L. Jackson
Met Office Hadley Centre, Exeter EX1 3PU, UK

K. Yamazaki
Atmosphere and Ocean Research Institute, The University of Tokyo, Kashiwa, Chiba 277-8564, Japan

1 Introduction

In order to prepare for the effects of future climate change, decision makers increasingly require quantitative predictions for a variety of climatic variables on a wide range of time scales. Through carefully-designed experiments using coupled atmosphere–ocean general circulation models (AOGCMs), these predictions can be obtained, but are subject to considerable uncertainty; this uncertainty must be quantified if the predictions are to be useful.

One of the principal sources of this uncertainty is parametric uncertainty. This is the extent to which a single model can produce different projections, all consistent with current observations, through changing the values of the parameters in the model. This includes the uncertainty present due to the fact that we cannot observe the model output for every, or indeed for very many, choices of the parameters of the model.

Quantifying the parametric uncertainty in a particular climate model has drawn some attention in the climate community in recent years (Murphy et al. 2004; Collins et al. 2011). This uncertainty must be addressed using perturbed physics ensembles (PPE), where the parameters of a model are all varied in a systematic way and where the model is run at each of these designed settings. We treat the PPE as a sparse sample of behaviours from a vast, high dimensional parameter space. The PPE is used to inform us about the behaviour of the model in this space and, in particular, about regions of the space where model-based projections are not predicted to be inconsistent with current observations.

In this paper we illustrate the application of existing methodology from the statistics with computer experiments literature to rule out regions of the parameter space of the UK Met Office's Third Hadley Centre Ocean–Atmosphere General Circulation Model (HadCM3) (Pope et al. 2000; Gordon et al. 2000) containing model runs that are inconsistent with a handful of physically important observational metrics. In particular, we use emulators, (fast statistical representations of climate models that include a measure of uncertainty in the prediction of climate model output), and four pre-industrial global and hemispheric averages of climatic variables to remove over half of the explored space. We find that of the constraints chosen, global mean surface air temperature (SAT) is the dominant constraint and that the other chosen metrics barely reduce the space further.

We show that the behaviour of more complex climate variables, such as time series of the Atlantic Meridional Overturning Current (AMOC) is easier to capture with emulators in the reduced space than prior to the application of history matching. We also show that perceived relationships between climate variables, such as the Meridional Heat Transport (MHT) and the AMOC, do not necessarily hold in the full parameter space, but do in the constrained parameter space. We show that the response of the AMOC and the MHT to CO_2 forcing is greater in the constrained parameter space. We argue that because of these issues, history matching is an important pre-cursor to any analysis of a complex climate model with a PPE.

We analyse the shape of the constrained parameter space of HadCM3 in the form of 2 dimensional projections. We discover that a parameter controlling the mixing and entrainment between convective clouds and the environment, called entcoef, is the most active, in that varying it causes a wide range of non-physical behaviours. However, we also find that the other parameters are important and can be set so that no single value of entcoef automatically leads to climates that do not match our chosen observational metrics. We explore parts of HadCM3's parameter space previously unstudied by running models with parameter choices outside the ranges established by Murphy et al. (2004). We find many runs in this extended parameter space that are not inconsistent with our observational metrics.

In any study where a climate model is combined with observations or used to make inferences about reality, it is important to acknowledge and give a careful treatment of model discrepancy, the extent to which missing processes and approximations in the model affect its ability to inform us about the true climate system (Goldstein and Rougier 2004; Sexton et al. 2011; Kennedy and O'Hagan 2001). As we are uncertain about model discrepancy we treat it as a random quantity within the Bayesian paradigm, where uncertainty for a collection of random quantities is a representation of our subjective beliefs about the collection rather than an actual property of it (see, for example, de Finetti 1974). An estimate for the variance of the model discrepancy is a key requirement for history matching. We discuss this issue and derive an estimate for it under an interpretation of model discrepancy variance as a tolerance to climate model error, using a multi-model ensemble for the World Climate Research Programme's Coupled Model Intercomparison Project phase 3 (CMIP3) (Meehl et al. 2007).

In Sect. 2 we describe current methods for using a PPE to assess parametric uncertainty of a climate model and introduce history matching as an alternative. Section 3 describes our ensemble, the observational constraints used, and the emulators of those constraints in HadCM3 built for this analysis. In Sect. 4 we use CMIP3 to derive model discrepancy for HadCM3 in order to facilitate history matching. In Sects. 5 and 6 we present the results of the history match and explore the "not ruled out yet" region of parameter space in HadCM3. Section 7 contains discussion. Appendices A–C contains details about the parameter space of HadCM3, our emulators and a more technical derivation and discussion regarding the model discrepancy.

2 History matching

In the following section we introduce a number of statistical models that are used to establish a relationship between climate models, reality and observations. To avoid ambiguity with the use of the word model in the climate

literature, the word "model" will be used to refer to a climate model (specifically an AOGCM) and "statistical model" will be used to refer to these mathematical relationships.

2.1 Assessing parametric uncertainty with PPEs

In order to asses the parametric uncertainty of a climate model using a PPE, there have been essentially two approaches used in the climate literature. One has been to treat the ensemble as a representative collection of the possible behaviours of the climate model and to do any analysis on the ensemble itself. The other is the Bayesian approach, which we describe later.

The first step, under the first approach, is to screen out ensemble members that differ markedly from historical observations. This is called ensemble filtering. Some metric of comparison with observed climate is defined by the analyst, and models some tolerance away from the observations are discarded. Rowlands et al. (2012), for example, first discard any ensemble member that requires a global annual mean flux adjustment of absolute magnitude greater than 5 W m^{-2}.

The next step under this approach is to choose a metric with which to weight each ensemble member so that the weighted ensemble represents parametric uncertainty. For example, Rowlands et al. (2012) define a 'likely' set of ensemble members within the sub-ensemble of not-discarded runs based on a selecting those sub-ensemble members below the 66th percentile of a weighted mean squared error that they compute. Murphy et al. (2004) also define a reliability metric which they call the Climate Prediction Index (CPI) and use it to weight each ensemble member according to an estimated relative likelihood that they will correctly predict climate change in the real world.

The issue with this approach is that it does not account for the uncertainty in the unsampled, high dimensional parameter space. Specifically, for any parameter choice at which we have not yet run the climate model, we are uncertain as to what the model climate will look like. By filtering the ensemble, important information about the parameter space, both at the locations of the filtered runs and in a neighbourhood around them, is discarded without being used. That information could be used to update our beliefs about the behaviour of the model climate in those parts of parameter space. By weighting individual members, the fact that there may be very many other projections within the unsampled regions of parameter space that are consistent with the observations and inconsistent with the set of future projections currently in the ensemble, is ignored.

The Bayesian approach uses the PPE to learn about the behaviour of the model throughout its parameter space in order to find regions of the space where model based projections are not inconsistent with current observations. The most accepted way of doing this is Bayesian calibration (Kennedy and O'Hagan 2001), a method that derives a probability distribution over the 'best input' in parameter space that is conditioned on (updated by) observations. We describe this in 2.3 and the process applied to climate models is described by Rougier (2007). Bayesian methods have gained increasing attention in the study of climate models (for example Berliner and Kim 2008; Furrer et al. 2007), and a Bayesian calibration provided the UK Climate Projections (Murphy et al. 2009; Sexton et al. 2011).

Bayesian calibration requires a stochastic representation of the climate model, called an *emulator* (Sacks et al. 1989), to be constructed and to be reliable across the whole of parameter space. This may require highly sophisticated statistical models that accurately capture those behaviours we are interested in whilst simultaneously capturing extremely unphysical behaviours and irrelevant extremes. When the size of a perturbed physics ensemble is restricted by computational resources, this requirement will often mean spending a large proportion of our budget of model runs on exploring parameter space we already know to be unphysical and a significant proportion of our effort in modelling these irrelevant behaviours.

Another potential problem with Bayesian calibration occurs because the calculation implies that there must be some region of parameter space with positive probability of containing the best input. However, there could be no regions of parameter space that contain climates that match observations to the accuracy required considering all relevant uncertainties. If we calibrate without some form of preliminary analysis to check that the parameter space contains settings that lead to climates consistent with current observations, the results of a Bayesian calibration could be meaningless. We discuss this further in Sect. 2.4

History matching (Craig et al. 1996), is an established statistical methodology that uses a PPE and historical observations of a complex system (such as climate) to rule out regions of parameter space of a model. It can, therefore, be a very useful tool as part of an analysis to assess the parametric uncertainty in a climate model. It is a technique that follows the Bayesian approach, however, unlike Bayesian calibration, it can be used to discover whether or not there may be no "physical" regions of parameter space. History matching can be used in tandem with Bayesian calibration, with history matching first locating parts of parameter space that cannot be ruled out given physical observations, and Bayesian calibration then used to construct a probability distribution for the best input within this not-implausible region.

History matching got its name from the oil industry, where it was first applied to the simulation of oil wells

(Craig et al. 1997; Cumming and Goldstein 2010). It has also been applied to the simulation of galaxy formation at the beginning of the universe (Vernon et al. 2010) and to an intermediate complexity climate model (Edwards et al. 2011). In this application we apply it to a non-flux-adjusted Atmosphere–Ocean Generalised Circulation Model (AOGCM), HadCM3, as described in Sect. 3.

2.2 Emulators: statistical models for climate models

In order to present the ideas behind history matching, some notation is required. We represent aspects of the climate system of interest by a vector, y, of quantities which are, in principle, observable in the real world. We represent any actual (imperfect) observations of y by z. We write the climate model as the vector function, $f(x)$, where x is the vector of model parameters occupying a defined parameter space $\mathcal{X}$. Each component of y corresponds, in principle, to one of the outputs of $f(x)$. Though y, z and $f(x)$ for any x are defined as objective, non-random quantities, we are uncertain as to their values before they are observed. Hence we make subjective uncertainty statements about these quantities, treating them as "random" within the Bayesian paradigm, and use statistical models to establish relationships between them.

In order to exhaustively evaluate the set of possible simulated climates $\{f(x) : x \in \mathcal{X}\}$, we would have to run the climate model for every possible $x \in \mathcal{X}$. This cannot be done for any computer model with $\mathcal{X}$ continuous. Further, climate models are computationally expensive to run, often taking weeks or months to evaluate for one choice of x. The alternative to this is to run an n member perturbed physics ensemble, $F_{[n]} = (f(x_1)\ldots f(x_n))$, corresponding to parameter choices $X_{[n]} = (x_1, \ldots, x_n)$, called the ensemble design. The ensemble is then used to build a statistical model called an emulator.

An emulator is a stochastic representation of a climate model that gives a prediction of the value of $f(x)$, and associated uncertainty, for any choice of $x \in \mathcal{X}$. Typically an emulator for output i of a climate model $f(\cdot)$ has the form

$$f_i(x) = \sum_j \beta_{ij} g_j(x) \oplus \epsilon_i(x), \tag{1}$$

where the operator $\oplus$ indicates the addition of components that are judged to be uncorrelated and where the $g(x)$ is a vector of specified functions of x, β is a matrix of coefficients and $\epsilon(x)$ is a mean zero residual stochastic process. There is a large literature on emulation and many ways to fit these models and train them using the ensemble $F_{[n]}$. We present one example in Sect. 3 below. Emulators of climate simulators have been built by Challenor et al. (2009), Rougier et al. (2009), Sexton et al. (2011), and Williamson et al. (2012), and there is an extensive general literature (see, for example, Santner et al. 2003).

An emulator is a key tool required for history matching. The required elements for history matching are expectation, $\mathrm{E}[f(x)]$, and variance, $\mathrm{Var}[f(x)]$, for any choice of x. If probability distributions are used to model our uncertainty about $\{\beta, \epsilon(x)\}$, then $\mathrm{E}[f(x)]$ and $\mathrm{Var}[f(x)]$ can be derived directly. However, probability need not be used. We may choose instead to make expectation statements directly, based on the fundamental work of de Finetti (1974, 1975). de Finetti provides a rigorous and operational development in which expectation (termed "prevision" by de Finetti) is the primitive quantity that may be specified directly, and the probability of any event is identified with the expectation for the indicator function for that event (an indicator function is 1 if the event occurs and 0 otherwise). This approach is valuable in problems which are sufficiently complex that we are unable or unwilling first to make a full joint probability specification over the space of possibilities. Such considerations are of particular importance when dealing with high dimensional computer models for complex physical processes, and many of the references in this paper describe analyses carried out within a framework based on expectation (for example Craig et al. 1997, 2001; Vernon et al. 2010; Williamson et al. 2012). For an overview of prevision, see Goldstein(1986b), for a detailed treatment of the development of probability from expectation see Whittle (1992) and for a detailed treatment of the statistical rationale and the practical implications of this approach to uncertainty specification and analysis see Goldstein and Wooff (2007).

We adopt this approach throughout so that our uncertainty specifications in this paper take the form of specifying means, variances, and covariances for quantities of interest without assuming underlying probability distributions. We shall obey the convention and refer to this as the Bayes linear approach. We make this choice to simplify both the specifications required to derive the statistical models and the resulting computation required for our history matching procedures.

2.3 Ruling out regions of parameter space

History matching, like Bayesian calibration, requires a statistical model that relates the climate model to reality. The most popular model for users of both methods is

$$y = f(x^*) + \eta \tag{2}$$

where x^* is the 'best input' of the climate model and η is the model discrepancy, the residual difference between the best representation of climate available from the model and the truth. x^* and $f(x)$ are taken to be independent of η for any x (probability density $\mathrm{P}(f(x)\eta) = \mathrm{P}(f(x))\mathrm{P}(\eta)$ for all x

and $P(x^*\eta) = P(x^*)P(\eta)$) (Goldstein and Rougier 2004; Sexton et al. 2011; Kennedy and O'Hagan 2001). This statistical model implies that learning about $f(x)$ by running the climate model is not informative for η. Though we will use a different statistical model to (2) and introduce it in Sect. 4, the notion of the existence of x^*, a setting of the inputs that is sufficient for the climate model (meaning that if we could evaluate f at x^*, we would know everything the climate model can tell us about y), is important to both methods. Bayesian calibration aims to find the probability distribution $p(x^*|z)$, the distribution of the best input conditioned on historical observations of climate z with

$$z = y \oplus e, \tag{3}$$

where e is mean zero observation error with a probability distribution that is to be specified.

Conversely, history matching aims to discover, for any parameter setting x_0, whether we can rule out x_0 as a candidate for x^*. A single bad $f_i(x_0)$ can rule out a choice x_0 so that we may build emulators for selected constraints only. The analysis leads to the derivation of a membership rule for a subspace of parameter values that are not inconsistent with the observations and our uncertainty specification, defined in this study as the Not Ruled Out Yet (NROY) space.

It is important to clarify that both history matching and Bayesian calibration require a pre-defined parameter space $\mathcal{X}$, and that the results of both methods are only valid within that space. Bayesian calibration begins with a prior probability distribution for x^*, $p(x^*)$, with support (the region over which the distribution is defined) within $\mathcal{X}$ and uses data in the form of a PPE and observations to update it so that the posterior for x^* will still only have support within $\mathcal{X}$. Conversely, history matching begins with parameter space $\mathcal{X}$ and uses a PPE within $\mathcal{X}$ and observations to remove regions of $\mathcal{X}$ that lead to poor representations of climate. NROY space is therefore not defined outside of $\mathcal{X}$.

To determine NROY space, we would like to compare, for any x_0, the value of $f(x_0)$ with the state of the climate y. However, we actually observe z rather than y and the climate model is too expensive to evaluate at every possible choice of the parameters. We, therefore, use the distance between z and our expectation for the value of $f(x_0)$ derived from an emulator for $f(x)$.

The distance we use is called an *implausibility measure*, $\mathcal{I}(x_0)$, and is defined on each scalar component of $f(x_0)$ via

$$\mathcal{I}_i(x_0) = \frac{|z_i - \mathrm{E}[f_i(x_0)]|}{\sqrt{\mathrm{Var}[z_i - \mathrm{E}[f_i(x_0)]]}} \tag{4}$$

with the overall implausibility for $f(x_0)$ given as $\mathcal{I}(x_0) = \max\{\mathcal{I}_i(x_0)\}$. An alternative multivariate definition is available in Vernon et al. (2010).

Large values of the implausibility measure suggest that it is implausible that x_0 is consistent with our uncertainty specification and with the observations, so x_0 can be ruled out as a possible candidate match to x^*. More formally, NROY space is defined to be the subspace of $\mathcal{X}$ such that the absolute value of the implausibility at any x in this subspace is less than or equal to a,

$$\{x \in \mathcal{X} : |\mathcal{I}(x)| \leq a\}$$

for some threshold a that is chosen depending on how large an $\mathcal{I}(x)$ the analyst is prepared to accept.

Calculation of $\mathcal{I}(x_0)$ requires an emulator for $f(x)$ and a specification for $\mathrm{Var}[z - \mathrm{E}[f(x_0)]]$. In order to compute the latter, a statistical model that relates z to the model output at x^* is required (for example, (2) together with (3) constitutes such a statistical model). These requirements are far less burdensome than those made by full Bayesian calibration, which require, in addition to the above, a probability distribution for the output of the model at any setting of the parameters as well as a prior probability distribution on the input space, $p(x^*)$. For example, with (2) the variance required to calculate (4) becomes $\mathrm{Var}[e] + \mathrm{Var}[\eta] + \mathrm{Var}[f(x_0)]$ (Craig et al. 1996), meaning that in order to history match we only require the observational error variance, $\mathrm{Var}[e]$, and the model discrepancy variance $\mathrm{Var}[\eta]$, in addition to an emulator for $f(x)$.

2.4 Benefits of history matching

Unlike methods that focus on filtering a PPE for runs that seem broadly 'physical', history matching uses all of the PPE data to rule out regions of parameter space that are implausible within a well defined statistical framework. It is much simpler to perform than a full Bayesian calibration, requiring a far less complex uncertainty specification both for emulation and when relating the model to reality.

Bayesian calibration requires a probabilistic emulator that is accurate throughout parameter space for every quantity we wish to calibrate on. For many climate model outputs, this can be very difficult as it may mean simultaneous capture of 'physical behaviour' and of extremely unphysical, non linear behaviour and irrelevant extremes. Models such as this require very large ensembles to fit, where most of the effort could be spent modelling the unphysical behaviour.

History matching, on the other hand, allows us to use even the most simple outputs of the climate model in order to begin ruling out regions of parameter space. We can take simple outputs that are relatively easy to model statistically, use them to rule out parameter space via history matching, then focus on emulating more complex model output within NROY space, where they may be easier to model statistically. As such, history matching is a useful

first step to perform before a Bayesian calibration exercise. Indeed Edwards et al. (2011) refer to history matching as 'precalibration' for this reason.

We need not proceed directly to calibration. Having identified an initial NROY space, if time and budget permits we can construct a second emulator using a second ensemble at a new set of points $x_1^{[2]}, \ldots, x_n^{[2]}$, chosen to be within NROY space. We can then repeat history matching, reducing further to a second stage NROY space. This is termed refocussing. We can continue reducing space in this way until the emulator variance is such that there would be no further gain in running another ensemble. Vernon et al. (2010) gives a good overview of this process. It may turn out that we can rule out all of parameter space by history matching, suggesting that the climate model is not informative for climate with respect to our specified model discrepancy.

3 History matching HadCM3

3.1 The RAPIT ensemble

Learning about the parameter space of a fully coupled, non-flux-adjusted climate model like HadCM3 requires a great deal of computing power. Specifically, we need a perturbed physics ensemble of sufficient size that we are able to accurately model HadCM3's response as a function of its input parameters.

RAPIT (Risk Analysis, Probability and Impacts Team) is a National Environment Research Council funded project with the goal of assessing the risk of collapse of the Atlantic Meridional Overturning Circulation (AMOC) using data from climate models and the real world. The RAPIT ensemble was generated using Climate Prediction Dot Net (CPDN, http://climateprediction.net), a distributed computing project through which different climate models are distributed to run on personal computers volunteered by members of the public. A copy of the model, along with a specific prescribed setting of the model parameters, is downloaded by the "client" computer, where it runs in the background, using any spare computing resources available. Data is returned to CPDN where it is stored and made available for access by the general public. At present there are more than 35,000 active hosts, many of which are multi-core machines.

Thanks to the power of CPDN, we were able to create a 10,000-member perturbed physics ensemble of the fully coupled, non-flux-adjusted HadCM3. The price for being able to run such a large PPE using personal computers is that each ensemble member takes a very long time to run. Although run-times differ from machine to machine and from user to user, it takes approximately 28 days for a reasonably fast and dedicated PC to run a 40-year simulation of HadCM3. However, many simulations take a lot longer and some are never returned.

Our ensemble design reflects the requirement for short simulations, with a short spin-up phase and idealised forcing applied straight from the pre-industrial conditions. The model is spun up for 50 years at each of the 10,000 settings of the model parameters, followed by a 70-year transient CO_2 simulation where CO_2 concentrations are increased by $(100 * \Delta)\%$ each year where Δ is treated as another model parameter to be perturbed within the range [0, 0.04]. All members are initialised with preindustrial CO_2 concentrations and with initial conditions from the standard, unperturbed model. Parallel control simulations, where CO_2 remains at preindustrial levels, are also conducted to allow us to monitor continued drift from the spin up.

The parameters to be perturbed are listed in Table 4. These are a subset of perturbations from previous studies with perturbations to atmosphere, sea ice and land models (Murphy et al. 2004; Sanderson et al. 2010) and perturbations to ocean parameters (Collins et al. 2007; Sanderson et al. 2010). Not all of the parameters identified in those studies were perturbed. The nature of the CPDN project also meant that switches between different parameterisation choices and schemes could not be perturbed for this study. We were also given a list of parameter pairs by the Met Office that were to be varied jointly following relationships they specified. For each pair we only include one parameter in the list of independent parameters to be varied in our design and used the given relationships to calculate the values of the others. See Appendix C for further details.

We chose the settings of the input parameters to satisfy two competing experimental design goals. Our first goal was to have a large ensemble of model runs within the area of parameter space used by previous studies. The second was to explore parameter choices outside of this space that may yield interesting (and as yet unseen in HadCM3) behaviour, such as abrupt collapse of the AMOC under increasing CO_2. The boundaries of the space used in previous studies can be seen in Tables 5 and 6 and will be referred to as the "standard" space. We devoted 80 % of our ensemble to a design in this parameter space.

We devoted the remaining 20 % of our ensemble to the expanded parameter space defined in columns 5 and 6 of Table 5 in Appendix C. We restricted any parameters with known dependencies on other parameters to be within the standard ranges on the advice of our Met Office collaborators. Of the remaining parameters, there was no formal rule used to choose the new limits, though generally each range was extended uniformly by between 25 and 50 %, whilst avoiding letting any parameter get too close to or cross 0.

In Murphy et al. (2004), the parameters are set at low, intermediate, and high settings. The size of our ensemble allows us to use a space filling design in the input parameters. We choose the actual values of the inputs by uniform Latin Hypercube sampling (Santner et al. 2003) within the specified ranges of each subspace. This method divides the range of each parameter into equal intervals and ensures that exactly one point is sampled from each interval. Maximin Latin Hypercubes, Latin Hypercubes chosen to maximise the minimum distance between design points (Morris and Mitchell 1995), are often used to select space filling designs because they offer a good coverage of parameter space. Whilst good coverage of parameter space is very important, we must also ensure that our continuous parameters are not highly correlated with one another or with any of the switch variables. If they were this would make it difficult to identify the parameters that drive the variability during emulation.

We ensure that our design in each of the input parameters is as uncorrelated and well spread out as possible by satisfying S-optimality within the class of Latin Hypercubes satisfying a constraint on the maximum correlation (0.1) between any two variables. S-optimality seeks to maximise the harmonic mean of all pairwise inter-point distances within the given class of latin hypercubes (Zaglauer 2012). The resulting design covers the parameter space well and ensures none of the variables are highly correlated. We check that there are no very small eigenvalues of the covariance matrices of each sub-design and the full design to make sure there is no degeneracy.

Finally, we treat the mixed layer parameters slightly differently to the way they are treated in Collins et al. (2007). In that experiment, they are a switch with three settings. Acreman and Jeffery (2007) treat these parameters as continuous within given ranges, and we follow this treatment for all but 2000 of the ensemble members within the standard ranges. Acreman and Jeffery (2007) find that there is significant interaction between the parameters in the Kraus and Turner (1967) bulk mixed layer model. However, to fully explore any interactions with other parameters in the fully coupled model, and to help build emulators, we vary them separately.

At the time of writing, we have 3,500 control simulations and 5,000 transient simulations completed. If we include all partial simulations from which we have data returned, we have approximately 2 million years of HadCM3 data thanks to the generosity of the members of the general public around the world who have contributed their spare computing resources.

3.2 AMOC in the RAPIT ensemble

The AMOC is a limb of the global ocean conveyor that transports warm, saline surface waters from the tropics into the higher latitude North Atlantic (Broecker 1987; Dickson and Brown 1994; Kuhlbrodt et al. 2007). Paleoclimate records suggest that the AMOC has previously undergone major rapid fluctuations (McManus et al. 2004). The heat transported northward by the AMOC, recently estimated to be in the region of 1.33 ± 0.4 petawatts at 26.5°N (Johns et al. 2011), is gradually released to the atmosphere where it is carried towards Europe by the predominantly westerly winds. This heat transport contributes to the comparatively milder winters experienced in northwestern Europe than at similar latitudes in Western Canada (Rhines and Häkkinen 2003; Broecker 1987).

Experiments with climate models under predicted future emissions scenarios have suggested that there could be around a 30 % reduction in the strength of the AMOC by the year 2100 (Solomon et al. 2007; Gregory et al. 2005), with a corresponding impact on the northward heat transport. However, the risk of a sudden collapse of the AMOC is not confidently known, and improving our understanding of this risk and its uncertainties is a key objective.

Changes in the AMOC are particularly important because of their effect on the Atlantic ocean Meridional Heat Transport (MHT), which in turn has an impact on the climate of the North Atlantic. Although it might be expected that a model with a weaker AMOC has less MHT, there are other factors that play a role. The relative temperature of the water being advected by the AMOC may be different, and the contribution of the gyre transport is also important. We will investigate the how the AMOC and MHT strengths and their sensitivities to increasing CO_2 are affected by our history matching.

In Fig. 1 we plot the AMOC time series at 26°N and 1,000 m depth from the complete runs for both the transient and control runs separately. The transient simulations are coloured by Δ. From this picture, we can see that we have a very wide range of AMOC behaviours in our ensemble. Some runs with technical faults, as a result of issues in writing or transferring model output, are removed from the ensemble and treated as missing data. The two visible common peaks at around 18 years and around 50 years may be the result of a mechanism of climate variability which has been artificially triggered by the parameter perturbations and is the subject of a forthcoming paper.

In order to set up any HadCM3 based inference about the effect of CO_2 forcing on the true AMOC and the corresponding heat transport, we would have to construct an emulator for the AMOC, and one of the longer term goals of our project is to emulate its evolution in time. If attempted within the full parameter space, this already daunting task would be much harder. This is because our statistical models would have to do most of their work attempting to capture the extreme non linear behaviours that we suspect are non-physical and may be removed by

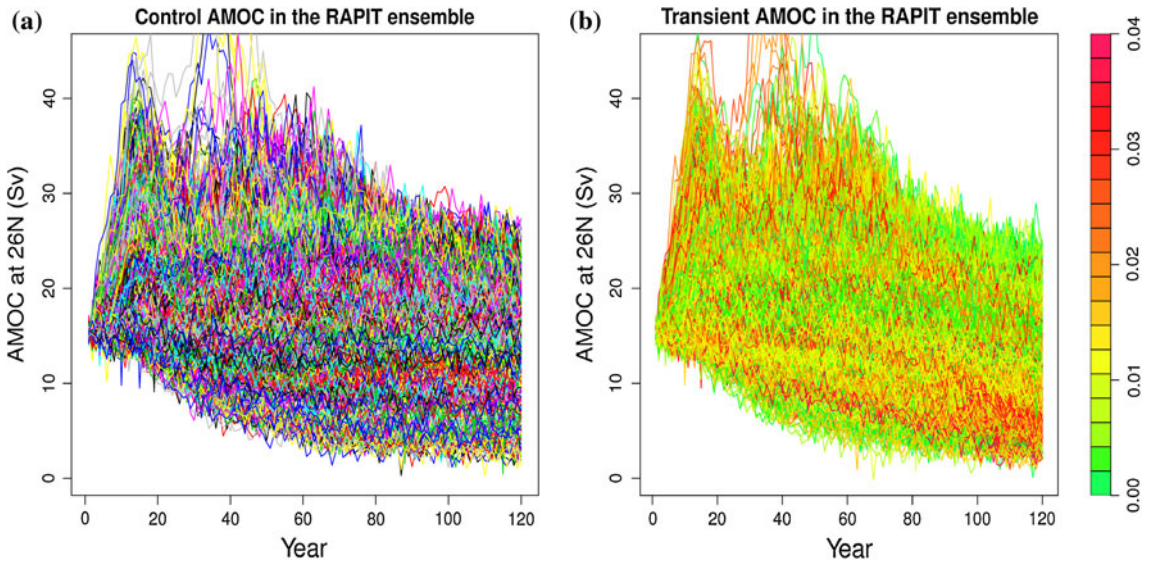


Fig. 1 **a** Control AMOC at 26°N in the RAPIT ensemble. *Colours* are assigned randomly to allow the reader to get an impression of the variability of individual runs. **b** Transient AMOC at 26°N in the RAPIT ensemble coloured by Δ, the annual rate increase of CO_2 forcing applied after year 50

an initial history match. A principal motivation for history matching here, and in any application involving a large ensemble of a climate model, is to make our emulators for key and difficult to model quantities easier to fit and more accurate than they would be if fitted using all data within the unconstrained parameter space.

3.3 Observational constraints

When choosing observational constraints for HadCM3, we select metrics we would expect any reasonable climate simulation to be able to represent. We have chosen to use simple, univariate, global and hemispheric averages as constraints rather than more complex quantities such as spatial fields (as in Sexton et al. 2011), for two main reasons. Firstly, one of the principal goals of this history match is to use quantities that are physically important and easy to emulate across parameter space as an initial way of ruling out implausible regions. Secondly, the lack of twentieth century forcing in our ensemble leaves us limited in terms of the relevant observational data available. We do not have preindustrial observations over sufficiently large spatial fields to warrant their use in a history match. There are likely to be many parameter combinations that produce 'unphysical' climate states, deemed not implausible by this process. These may be ruled out by future history matching or Bayesian calibration, perhaps using further constraints.

In an analysis that goes on to provide probabilistic climate predictions such as that in Murphy et al. (2009), we would run a second ensemble within NROY space and with twentieth century forcing, and either use history matching with more complex constraints to further reduce NROY space or use Bayesian calibration with more complex constraints to generate probabilistic predictions. Note that our calibration may be even more powerful if this second history match is performed beforehand.

The first constraint we use is global mean surface air temperature (SAT). This can vary substantially from observations (with values across the ensemble ranging between -5 and 33 °C) as the parameter perturbations alter the radiative balance at the top of the atmosphere. SAT has also been shown to have a large correlation with the AMOC strength in a previous perturbed physics ensemble based on HadCM3 (Jackson et al. 2012). Observational constraints were calculated using the present day HadCRUT3 climatology (Brohan et al. 2006) to give 14 °C, which is also consistent with the value of 14.0 ± 0.5 °C from Jones et al. (1999).

A preindustrial value is estimated by using HadCRUT3 anomalies to calculate a change since preindustrial times. Since this observation set only started in 1850 and observations were scarce during the early part of the period, values during 1850-1880 were used to represent the preindustrial period with the assumption that temperatures had changed little since preindustrial times.

For each 30 year period (1850–1880 and 1960–1990 for present day) a time average of SAT was calculated at each data grid point for each month. In this way a seasonal climatology was created for, say January, using all data from any grid point containing data from at least one January in that 30 year period. These 12 monthly climatologies are then averaged to make a 30 year climatology, this time only using grid points that have existing data for all months. This method attempts to minimise regions with

missing data whilst avoiding creating a seasonal bias. Finally a comparison of preindustrial and present day climatologies suggest temperature changes of around 0.3 °C that are reasonably spatially consistent, although there are few observations at high latitudes. Combining these studies and accounting for the rounding we have done to arrive at these numbers we obtain the constraint of 13.6 ± 0.5 °C.

The second constraint used is the northern hemisphere meridional SAT gradient (SGRAD, measured as the difference between averages over 0−20°N and 50−70°N to avoid errors associated with sparse observations in the Arctic). The gradient is related to the net (ocean and atmosphere) meridional heat transport, used for ruling out regions of parameter space where an unreasonably low or high transport of heat from lower to higher latitudes is predicted. The third constraint is the average northern hemisphere seasonal cycle of SAT (SCYC, June–August average minus December–February average). The strength of the seasonal cycle is related to the sensitivity of the climate to changing external forcing, in this case to the seasonally varying solar radiation, suggesting that this might be an important observational constraint.

These two constraints are calculated from the present day HadCRUT3 climatologies, with observational errors calculated as the difference between this value and that calculated from an alternative climatology from Legates and Willmott (1990). For SGRAD a difference from preindustrial conditions is assessed by a comparison of the 1960–1990 and 1850–1880 climatologies described above. The meridional SAT gradient only differs by 0.03 °C between the two climatologies, and a comparison of spatial differences suggests little sensitivity to latitude, apart from high latitudes north of 70°N which were excluded from SGRAD because of the sparsity of data. This change is judged to be negligible.

Using the seasonal climatologies from the two time periods to construct a change since preindustrial times for SCYC gives a value of 0.17 °C. This is less than, though of the same order as, the defined observational error. Additionally, the temperature difference is spatially inhomogeneous and concentrated over Europe where a timeseries can be created because of good spatial and temporal observational coverage. This timeseries suggests that this difference is part of the multi-decadal variability, rather than a systematic trend. Hence the difference in SCYC since preindustrial times is taken to be zero.

The final constraint is that of global mean precipitation (PRECIP, kg/m^2/s). Although this is expected to vary with global mean SAT, it provides an important constraint on the hydrological cycle. Present-day estimates of global mean precipitation rate (from a variety of products based on different satellite and gauge measurements) generally lie within the 2.2–2.9 mm/day range (for example, Huffman et al. 2009; Trenberth et al. 2009; Liu and Allan 2012). However, based on recent measurements of surface downward longwave radiation, it has been suggested that these values may be underestimates (Stephens et al. 2012), with the implied global mean precipitation rate being 3.1 ± 0.6 mm/day. This present-day range is used to obtain an estimate for the corresponding pre-industrial value by using an estimated percentage change in global mean precipitation per degree of global warming (2.2 ± 0.52 % K^{-1} Frieler et al. 2011). This reduces the present-day precipitation rate by 0.74 % (a very small reduction when compared with the observational uncertainty).

The constraints, z, above, have each been given observational error in the form of a range. We require a variance for the observation error, e, and so we interpret these ranges as 2 standard deviations. This interpretation is broadly consistent with treating each range as a 95 % confidence interval for the constraint. Though it may be more natural to view the given ranges as bounds on the observational error, the variances we would derive from this view would be far smaller than under our interpretation. Treating these as 2 standard deviations represents a cautious approach that will mean Var[e] is larger than it might have been under an interpretation of the given ranges as bounds and will therefore result in less parameter space removed by history matching.

The values of z and Var[e], are presented in Table 1. Note that if our interpretation of the ranges were actually correct, the given variances may be underestimates due to the fact that we do not have real pre-industrial observations, only estimates for them. We discuss this further in the context of our implausibility measure and explore the sensitivity to variance underestimation in Sects. 4 and 5.

3.4 Emulating the constraints in HadCM3

We build 4 separate emulators for the constraints in HadCM3, each capturing the mean and variance of the constraint as a function of the parameters. We calculate the constraints by computing their final decadal mean (years 71–80) for ensemble members that have completed 80 years. At the time this work was completed there were

Table 1 The observational constraints, errors, and emulator variances for each constraint

Constraint	Units	z	Var[e]	Var[$f_i(x)$]
SAT	°C	13.6	0.0625	0.437
SGRAD	°C	27.7	0.25	0.271
SCYC	°C	12.0	0.01	0.0734
PRECIP	kg/m^2/s	3.56×10^{-5}	1.19×10^{-11}	3.06×10^{-13}

The emulators are described in Sect. 3.4. The numbers in the table are rounded

insufficient ensemble members complete to year 120 to enable us to build the emulators, whilst there were >1,350 ensemble members which had completed 80 years of integration. Hence we use results from after 80 years of spin up with the caveat on our results that many runs will drift further from initial conditions later. For SAT, in particular, it is unlikely that a model deemed to be implausible after 80 years would later become not implausible since SAT drifts from initial conditions (close to observational values) towards a new state as the model adjusts to its altered radiative fluxes. A decadal mean is chosen to remove interannual variability, but to include as little of the early spin up as possible.

We began our emulation by fitting a mean function, $\beta g(x)$ in Eq. (1), via ordinary least squares regression using a stepwise selection method explained in Appendix B. Although it is known that entcoef, a parameter controlling the mixing and entrainment between convective clouds and the environment, drives a lot of the variability in HadCM3 (Joshi et al. 2010), it has also been shown that other parameters affect the response of the model and interact with entcoef (Rougier et al. 2009). We therefore permit our emulators to include functions of all of the parameters. This enables us to capture explicitly many non-linear behaviours present in the climate model output in our mean function. The type of functions we include are discussed in Appendix B. We discuss the fitting of $\epsilon(x)$ from Eq. (1) in Appendix B, where we also illustrate the performance of our statistical models.

4 History matching with multi-model ensembles

In order to perform a history match of HadCM3 using the 4 identified observational metrics, we require an emulator for the relevant model outputs as described in Sect 3.4, and a statistical model relating true climate y to $f(x)$ so that $\mathrm{Var}[z - \mathrm{E}[f(x)]]$ in Eq. (4), can be computed. Taking this statistical model to be (2) requires us to specify a meaningful assessment of model discrepancy variance. This is a challenging task that necessarily requires a number of subjective judgements from the climate modellers themselves regarding model deficiency and the potential for model improvement. Such judgements, ideally, would be elicited carefully from the model developers as they were for a computer model simulating the evolution of galaxies since the big bang in Vernon et al. (2010).

Expert judgement for model discrepancy is important here as any information we have about the discrepancy, before using our PPE, must be external to the PPE. Equation (2) specifies that the set $\{f(x), x^*\}$ and η are independent for all x (Kennedy and O'Hagan 2001; Sexton et al. 2011). This statistical model implies that the PPE is not informative for model discrepancy. The richest source of information about discrepancy therefore comes directly from the modellers themselves, who are aware of their model's deficiencies and the ways in which they might be improved in order to make them 'closer' to the true climate.

Previous climate studies, such as that undertaken to provide the UK Climate Projections (Murphy et al. 2009), have avoided an elicitation of model discrepancy variance from model developers. Instead, model discrepancy variance has been specified using a multi-model ensemble (MME) from the World Climate Research Programme's Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset (Meehl et al. 2007). Sexton et al. (2011) derive a discrepancy variance from a subset of the CMIP3 models they judged to be "state-of-the-art" using relationship (2) and their HadSM3 emulators for the different observations. The emulators are used to find the closest match of HadSM3 to each CMIP3 model and a discrepancy variance is calculated by computing the variance of all differences between the CMIP3 model output and the emulator prediction.

There are a number of uncertainties ignored under this approach, the principal of which being that each of the CMIP3 members itself has its own model discrepancy so that the difference between each state-of-the-art model and its closest analogue in the model used to generate the PPE might be a tiny fraction of the difference between true climate and the model at its best input.

Without an elicitation involving the modellers of HadCM3, we view model discrepancy variance as our 'tolerance to error'. This means that in specifying uncertainty for model discrepancy, we are implicitly commenting on how far we are happy for HadCM3, run at its best input, to be from the true climate. Note that this interpretation of model discrepancy has implications for the results of a history match, particularly if all of the parameter space is deemed not implausible. With an expert-judgement based model discrepancy variance, ruling out all of parameter space would be seen as a sign that the discrepancy variance had been mis-specified as the degree to which the climate model output differs from the true climate had been under estimated. However, under a tolerance to error interpretation, ruling out all of space is a sign that, given the level of climate model deficiency we are prepared to accept for the purposes of any analysis, the model under study is uninformative for the aspects of climate in question. Under this interpretation of the model discrepancy variance it is natural to explore the effect of different tolerances to error on our inferences.

However, we do prefer to have a physical basis for selecting those tolerances that we do. We therefore follow previous studies and base our specification on the CMIP3

data. Our approach is to set up a statistical model linking true climate, y, to HadCM3, $f(x)$, via CMIP3. This allows us to derive $\mathrm{Var}[z - \mathrm{E}[f(x)]]$, the denominator of our implausibility measure (4), under the interpretation that the resulting discrepancy variance represents a tolerance to error which is consistent with using CMIP3 as representative of the judgements of the climate community regarding what represents an informative climate model.

As a consequence of adopting a formal statistical model relating CMIP3 and HadCM3 to each other and to the true climate, we are able to obtain 'low' and 'high' tolerance alternatives for the discrepancy variance to be used in a sensitivity analysis. In Sect. 4.1 we describe our statistical model for a general MME, the key assumptions underpinning it and its consequences. Section 4.2 describes how the discrepancy variance is obtained using a multi-model ensemble and derives the HadCM3 discrepancy variance using CMIP3 as the MME. A technical derivation and further discussion of the statistical model is available in Appendix A.

4.1 Implausibility via multi-model ensembles

Throughout what follows we use the Bayes linear approach to uncertainty specification (Goldstein and Wooff 2007) described in Sect. 2, so that uncertainty for a collection of quantities is represented through expectations, variances and covariances, without adopting probability distributions.

Let a MME be a collection of climate models $f^1(x^*_{[1]}), \ldots, f^m(x^*_{[m]})$, which we write as $f^1, \ldots, f^m$. A statistical model is required in order to provide a framework that allows the MME to be informative for the real climate. In order to establish this framework we follow the approach of Rougier et al. (2012) and specify a statistical model with two assumptions. Firstly, that the MME members are judged second order exchangeable. This means that prior to observing the MME, we assign each climate model, f^i, the same mean and variance, and each pair of models, f^i, f^j, the same covariances. In other words, prior to seeing the MME, we have no reason to believe that any individual model should be better/different than any other.

Our second assumption is that true climate, y, 'respects second order exchangeability' with the MME, which means that we judge the covariance between y and each f^i to be the same prior to observing the MME. The above assumptions were first introduced to relate CMIP3 to the true climate by Rougier et al. (2012), and detailed discussion can be found there. Rougier et al. (2012) also compare this statistical model with others used to make inference about climate using an MME and argue that other commonly used models are specific cases of this more general statistical model.

Note that both of our assumptions pertain to a property of our beliefs and are therefore subjective. Second order exchangeability, for this type of collection, could not be an intrinsic property of it (for example, any given f^i only has one possible value so has no intrinsic variance). Instead it is a property of our beliefs about the collection, providing a useful statistical modelling construct. Whether or not we are willing to adopt these assumptions depends on the MME and the application. We further justify our assumptions to a general MME in Appendix A, and to CMIP3 in particular in Sect. 4.3.

The first assumption allows each f^j to be expressed as

$$f^j = \mathcal{M} \oplus R_j, \tag{5}$$

where the operator $\oplus$ implies addition of uncorrelated terms as before, $\mathcal{M}$ is the underlying mean of the MME, and the R_js are mean-zero residuals with common variance and no correlation between any pair R_j, R_k. This is a consequence of the second order representation theorem (Goldstein 1986a).

The second assumption gives

$$y = \alpha\mathcal{M} \oplus U \tag{6}$$

where α is a constant and U is a mean zero discrepancy representing the difference between the information available about y from the MME and reality. In what follows we set $\alpha = 1$ as we have no a priori view on the direction of any systematic biases of the MME. This choice leads to tractability of the denominator of the implausibility function in (7) and the estimate for its components in Sect. 4.2. Though for $\alpha \neq 1$ an alternative estimator for the key components of the implausibility calculation would need to be found, an alternative form of (7) is available with a factor of α^2 in front of the last 2 terms. This implies that for $\alpha < 1$, where the MME ensemble mean is believed to over estimate climate y, implausibility is larger than for $\alpha = 1$ for all x_0 (if $\mathrm{Var}[U]$ is unchanged) so that more space is ruled out if all of the terms are known. The converse is also true with $\alpha > 1$ leading to less parameter space being ruled out. A choice of $\alpha \neq 1$ represents a strong view that prior to observing the MME, we believe it has a systematic bias.

Let our climate model of interest, (in our case HadCM3) be f^h, with $h \in \{1, \ldots, m\}$ so that f^h is also an MME member. Then (5) and (6) lead us to derive the denominator of our implausibility function as

$$\mathrm{Var}[z - \mathrm{E}[f^h(x_0)]] = \mathrm{Var}[e] + \mathrm{Var}[U] + \mathrm{Var}[R_h] + \mathrm{Var}[f^h(x_0)], \tag{7}$$

where e is the observation error defined in (3) and $\mathrm{Var}[f^h(x_0)]$ is obtained from our emulator for $f^h(x)$. $\mathrm{Var}[U]$, the variance of the difference between true climate and the underlying mean of the MME, and $\mathrm{Var}[R_h]$, the variance of the difference between any MME member and

the underlying ensemble mean, must be estimated using the MME, in order to complete the uncertainty specification. A derivation for (7) is given in Appendix A.

4.2 Variance estimation

We can derive estimates of Var[U] and Var[R_j], the two remaining variances required to calculate implausibility via (4) and (7) using quantities calculated from the data and the MME. Our estimate for Var[R_j], the variance of the difference between any MME member and the ensemble mean, is S, the variance of the MME members. To estimate Var[U], we first write

$$z - \frac{1}{m}\sum_{j=1}^{m} f^j = (z - y) + (y - \mathcal{M}) + \left(\mathcal{M} - \frac{1}{m}\sum_{j=1}^{m} f^j\right) = e + U - \frac{1}{m}\sum_{j=1}^{m} R_j$$

and, as these three terms are uncorrelated with expectation 0 (see Appendix A),

$$\mathrm{E}\left[\left(z - \frac{1}{m}\sum_{j=1}^{m} f^j\right)^2\right] = \mathrm{Var}[e] + \mathrm{Var}[U] + \frac{1}{m}\mathrm{Var}\left[R_j\right]$$

$$\mathrm{Var}[U] = \mathrm{E}\left[\left(z - \frac{1}{m}\sum_{j=1}^{m} f^j\right)^2\right] - \mathrm{Var}[e] - \frac{1}{m}\mathrm{Var}\left[R_j\right].$$

Our estimate for Var[U] is then

$$\left(z - \frac{1}{m}\sum_{j=1}^{m} f^j\right)^2 - \mathrm{Var}[e] - \frac{1}{m}S.$$

As with any similar variance component decomposition, the observed data can give rise to negative estimates. Large negative estimates have a useful diagnostic function. For example, if the MME were over-tuned to the observations, we would observe small values of $\left(z - \frac{1}{m}\sum_{j=1}^{m} f^j\right)^2$ and S and a larger value of Var[e], leaving us with a negative estimate for Var[U]. Such a negative estimate could also occur if the size of our MME is small and it turned out, by chance, that the mean of the MME members was close to the observations and that Var[e] was large. If we observed a small or negative estimate for Var[U], we may re-visit the second order exchangeability assumption for all models in the collection, or investigate the sensitivity of the estimate to the ensemble size. For example, our estimate itself, which is a random quantity, might have a large variance if the ensemble size is small.

Var[U] is the most difficult quantity to estimate in our statistical model and therefore it is useful to assess the impact of alternative estimates that are 'consistent' with the observational constraints and the data from the MME. For example, since we are treating $f^1, \ldots, f^m$ as a sample from a larger collection of second order exchangeable models, we can generate alternative samples and calculate Var[U] for each of these in order to explore the sampling distribution of Var[U]. This would allow us to set alternative, 'low' and 'high' values of Var[U] that are consistent with the MME as part of a sensitivity analysis and an exploration of the effect of our statistical modelling choices. We describe a procedure for this in Sect. 4.3.

4.3 Linking HadCM3 to climate via CMIP3

In order to use CMIP3 as our MME for history matching HadCM3 using the implausibility defined by (4) and (7), we must choose a subset of CMIP3 that we judge to be a priori second order exchangeable with each other and with HadCM3. As acknowledged by Rougier et al. (2012), because the names of each CMIP3 member tell a climate scientist a great deal about the model output, the models are not second order exchangeable unless we choose to adopt ignorance about the different CMIP3 models.

However, we have already stated the difficulty in obtaining discrepancy judgements from model developers for one model (HadCM3) and the unavailability of these judgements in this case has led us to explore using an MME. Without second order exchangeability we would require similar and separate judgements for each CMIP3 member in order to use the MME at all. Rougier et al. (2012) choose to compromise and only use a subset of CMIP3 that they judge to be "broadly" second order exchangeable, by which they mean that the models in the chosen subset are sufficiently similar by reputation in their judgement to adopt second order exchangeability over the subset prior to observing the CMIP3 data. We take the same approach and choose a subset of CMIP3 containing HadCM3 that we judge to be broadly exchangeable. We are comfortable doing this due to our interpretation of the model discrepancy variance as our tolerance to error and because we intend to conduct a sensitivity analysis for different tolerances.

In order for the chosen subset to be 'broadly' second order exchangeable with HadCM3, we should choose models that we have no a priori reason to think would be any closer to reality, in terms of their resolution and physical description of climate, than HadCM3, and that are not highly correlated with each other. We therefore use at most one model from each research centre, selected to be most like HadCM3 in terms of the parametrization schemes and the resolution.

The CMIP3 members we used were the models {GFDL-CM2.0, INM-CM3.0, BCCR-BCM2.0, IPSL-CM4, PCM,

CGCM3.1(T63), GISS-ER, MIROC3.2(medres), CNRM-CM3, ECHO-G, ECHAM5, CSIRO-Mk3.5, INGV-SXG, MRI-CGCM2.3.2a}. Other models were ruled out as not being second order exchangeable with HadCM3 or because another model from the same group was judged to be more similar to HadCM3 based on its resolution and parameter schemes.

Table 2 shows the estimates for Var[U] and Var[R_j] obtained from the subset of CMIP3 models used. Originally the estimate of Var[U] for SGRAD was -0.171. This may have been because second order exchangeability breaks down for SGRAD or for one of the reasons described in Sect. 4.2. In this case, we had a large value of S and Var[e] compared with the other variables. This suggested that the sampling distribution of the assessment we have used may have had a large variance. This could also have suggested that SGRAD may not be a good quantity to match to as we have a lot of uncertainty regarding the observations, and the climate modelling community (represented by the MME members) has little agreement in their representations of it. The value in the table was obtained by the method described below.

To explore the sampling distribution of Var[U] for any constraint, we assume that $f^1,\ldots,f^m$ follow a normal distribution and sample m new values, $f^1,\ldots,f^m$ from N($\mathcal{M},S$), an e from N(0, Var[e]), and a value U from N(0, σ_U^2), for some σ_U^2. We then calculate a sample value $z=\mathcal{M}+U+e$ and compute a sample Var[U] that depends on σ_U. Taking a large sample of such values for a wide range of σ_U gives insight into the distribution of our data-based estimates. In particular, how likely we are to observe negative values for positive σ_U.

Note that there is no "objective" reason to sample $f^1,\ldots,f^m,e$ and U from normal distributions in the way we have suggested. Any method of generating alternative samples from a distribution of interest will lead to a sensitivity study than could be undertaken. We prefer to sample from symmetric distributions as we have no view on any skewness of these quantities.

We performed the sampling experiment described above for SGRAD and took 100,000 samples of Var[U] for each of 100 values of σ_U. We found that the variance of the statistic is so large that even very large values like $\sigma_U^2=3$ still lead to a 20 % chance of observing a negative value less than or equal to the one observed. It turns out that the sampling distribution of Var[U] is heavily skewed (we explain why this is below), which leads to the possibility of negative estimates when the variance is large. In order to choose a value of Var[U] that is consistent with having observed a negative estimate, we select a value of 0.483, so that there is a 50 % chance of observing a negative value and a 39 % chance of observing a value less than or equal to the negative one above.

We use the described sampling method to choose 'low' and 'high' values of Var[U] consistent with the data for each constraint in order to conduct a sensitivity analysis. Our initial idea was to use 'high' and 'low' values of Var[U] by selecting the lowest/highest Var[U] such that the sampled or 'standard' value was in the upper/lower tercile of the sampling distribution. However, the sampling distributions appeared quite skewed so that there was little difference between the standard values and the lower terciles. We therefore decided to use lower quartiles so that the 'low' value was not too extreme with respect to the sampling distribution of Var[U], whilst still allowing us to explore the effect of different tolerances consistent with CMIP3. The different values of Var[U] for each experiment are given in Table 2.

The skewed sampling distributions for Var[U] are due to our sampling method which leads to z samples (z being our observational constraints) being normal. If $\bar{z}$ is the sample mean of z, then it is a well known result from normal distribution theory that $(n_z-1)(z-\bar{z})^2/\mathrm{Var}[z]$ has a chi square distribution with n_z-1 degrees of freedom (where n_z is the number of zs in the sample). This implies that $(z-\bar{z})^2$ has a gamma distribution (see, for example Rice 1995, chapter 6). Here $\frac{1}{m}\sum_j f^j$ is an unbiased estimator for the sample mean of $z,\bar{z}$, and so our calculation leads the sampling distribution to be skewed.

We explore the sensitivity of our conclusions to these alternative values of discrepancy in Sect. 5. We also

Table 2 A table showing the MME mean and the components of the discrepancy variance required for our implausibility calculation obtained using the ensemble

Constraint	Units	$\mathcal{M}$	S	Var[U]	Low	High
SAT	°C	13.1	0.978	0.146	0.071	1.33
SGRAD	°C	28.2	2.18	0.483	0.253	4.00
SCYC	°C	12.5	0.513	0.180	0.122	1.21
PRECIP	kg/m^2/s	3.26×10^{-05}	4.55×10^{-12}	3.00×10^{-12}	1.00×10^{-13}	6.94×10^{-11}

The low and high values correspond to alternative estimates for Var[U] derived by a method outlined in the text below. Values are given to 3 significant figures, but the unrounded values are used to calculate estimates

explore the effect of SGRAD as a constraint. If SGRAD reduces parameter space significantly, it would be worth looking into the second order exchangeability assumption for this constraint more carefully.

5 NROY space

We define NROY space using scalar implausibilities $\mathcal{I}_i(x_0)$ as defined in (4) with implausibility for any parameter choice x_0 given by $\mathcal{I}(x_0) = \max\{\mathcal{I}_i(x_0)\}$. We define NROY space to be $\{x : |\mathcal{I}(x)| \leq 3\}$, using the "3 sigma rule" (Pukelsheim 1994) to set the implausibility threshold. The three sigma rule states that for any continuous unimodal probability distribution, at least 95 % of the probability mass is within three standard deviations of the mean.

5.1 The power of the constraints

In Fig. 2 we compare the implausibilities in the ensemble for each constraint to see which, if any, drive the majority of the reduction of parameter space. For each set of parameter values in the ensemble, $X_{[n]}$, we calculate the implausibility of each output and plot them pairwise. The colours on the figure indicate density of points, with light blue indicating a low density of points and magenta indicating a high density. The red dashed lines indicate the implausibility cut off with points above or to the right of those lines ruled out. We see immediately that SAT is the dominant observational constraint.

The patterns in Fig. 2 show correlations between the implausibilities. This indicates that the constraints themselves may be correlated. As stated in Sect. 3.3, we expected some correlation between PRECIP and SAT (Frieler et al. 2011). However, the observed correlation is higher than expected at 0.96. In addition, both SGRAD and SCYC are negatively correlated with SAT, the former having a correlation of -0.68 with SAT and the latter a strong correlation of -0.89 with SAT. The correlations are about the same in NROY space, indicating that these relationships are not an artefact of 'unphysical' climates in our ensemble.

Previous GCM studies have found that increased temperatures due to increased greenhouse gases result in a greater warming at higher latitudes ('polar amplification') and a greater warming in winter (Manabe and Stouffer 1980; Solomon et al. 2007). It has been suggested that this is a result of a sea-ice albedo feedback where less ice results in a greater absorption of shortwave radiation and hence more warming in the high latitudes, particularly in winter. This would result in a decreased meridional temperature gradient, and a decreased seasonal cycle. In our ensemble, although changes in SAT are a result of different parameter perturbations rather than increasing greenhouse gases, it is likely that the sea ice would respond in a similar way, possibly resulting in the negative correlations found between SAT and SGRAD or SCYC.

SAT is, therefore, the dominant constraint due to the high correlations with the other variables and because of the relatively low discrepancy and observation errors of SAT compared with the other variables in relation to the ensemble ranges for those variables. Though SGRAD is weakly negatively correlated with SAT when compared to SCYC, its observational error variance and discrepancy are much larger in comparison to those for SAT. This means that SGRAD provides only a slight further constraint after accounting for SAT.

It is important to note that although SAT dominates and that, for SCYC and PRECIP at least, no ensemble members that are not ruled out by SAT alone are ruled out by the additional constraints, that does not mean that the additional variables provide no further constraint on parameter space. It simply means that all of our ruled out ensemble members have implausible SAT (or SGRAD in the case of one ensemble member). Ruled out space is a vast subspace of our original 27 dimensional parameter space for which we have a very sparse sample in our ensemble. There may be a subspace of this ruled out space where SAT is not implausible but one or more of the other constraints is implausible. If this subspace is relatively small, it would be highly unlikely to see any of our ensemble members in there and, indeed, either finding points in small subspaces or proving that no such subspace exists is a challenging and open research question.

The power of history matching, versus ensemble filtering as described in 1, is that we can take any point in parameter space and determine whether or not it is in NROY space. We can do this because those ensemble members that are ruled out by history matching have been used to train our emulators and inform us about the regions we want to rule out.

5.2 Looking at NROY space

To explore NROY space and to find out how effective our history match has been, we conduct a large sampling experiment evaluating $\mathcal{I}(x)$ for 1,600,000 different choices of x (described below) within their standard ranges. We found that our history match ruled out 56 % of the standard parameter space of HadCM3 defined in Table 5. NROY space is not defined outside of these ranges. As an aside, non-linear transformations of the parameter space would change this percentage. However, they may also affect the PPE's coverage of parameter space and hence the quality of our emulators and our confidence in our conclusions.

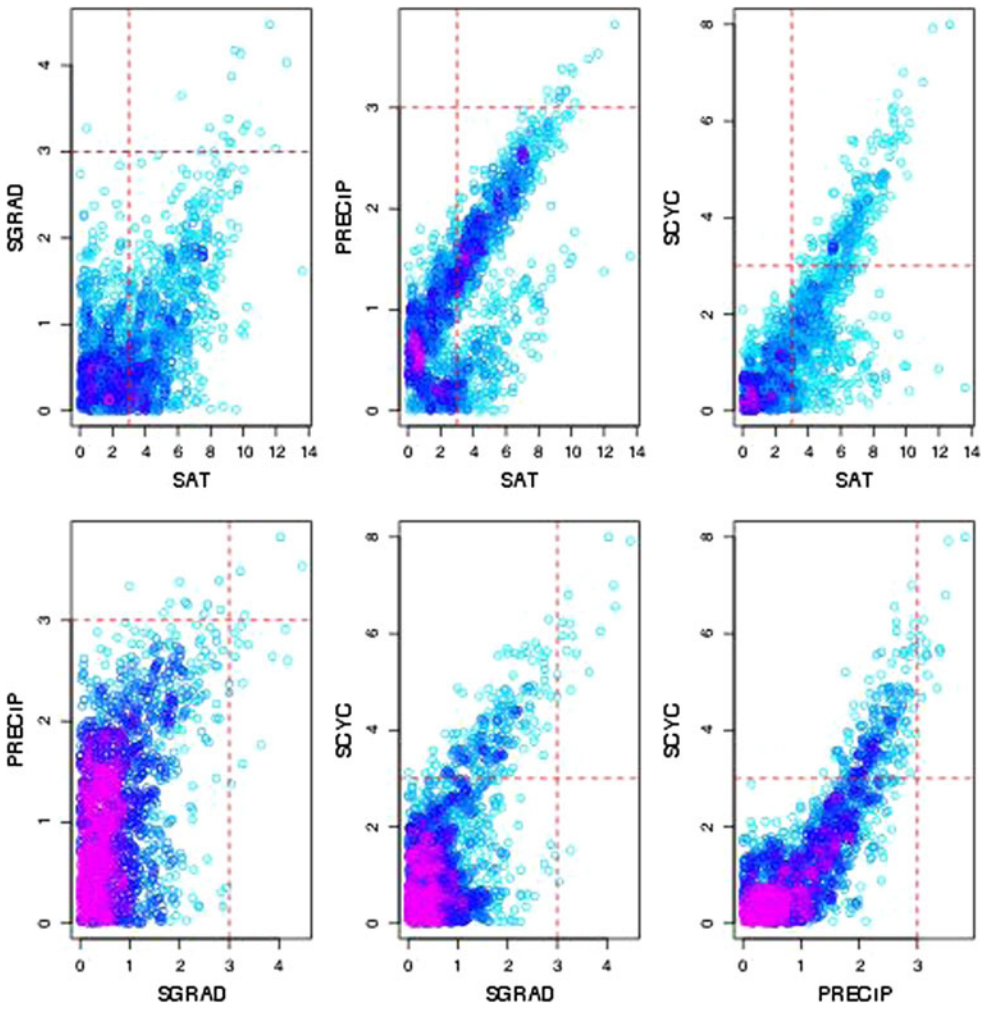


Fig. 2 Comparing the implausibilities for the chosen constraints in our ensemble. The *red lines* show the implausibility cut offs used, with points to the right or above any of the *red lines* ruled out by that constraint. The *colour* shows the density of points in the plot by using a scale linked to the number of points in the plot within a neighbourhood of 0.1 from the point to be coloured. *Light blue* is low density with *magenta* being very high density

It is, therefore, our view that parameter space should be defined prior to designing the PPE and un-changed throughout any analysis using the PPE.

Figure 3 explores the shape of NROY space via a NROY density plot for the 7 parameters with the biggest effect on implausibility within their standard ranges. For each pixel on the panels on the upper triangle the values of the 2 plotted parameters are fixed and implausibilities for every point in a latin hypercube containing 1,000 points in the other dimensions are calculated. The colour indicates the proportion of those points in NROY space. The resolution of each image is such that each individual panel in Fig. 3 is created using 1,600,000 separate evaluations of our emulators. The panels on the lower triangle show the same plots but each with their own relative colour scale so that interesting patterns in the 2D projections masked by the scale in the key can be observed in detail. Though we can see which parameters have large effects on implausibility by looking at the panels in Fig. 3, these effects cannot be quantified easily as each parameter has a number of non-linear interactions with the other parameters.

There are many noteworthy features in Fig. 3. It is clear that the entrainment parameter entcoef is the most active parameter causing a wide range of behaviours, many of them non-physical. However, the parameter space of HadCM3 is not simple and there is no fixed interval within which entcoef will return not-implausible matches and outside which it will not. Joshi et al. (2010) show that a member of the ensemble from Murphy et al. (2004), with a low value of entcoef and a much higher climate sensitivity than the rest of their ensemble, is unphysical due to an overly high concentration of stratospheric water vapour. From Fig. 3, we can see that even very low values of entcoef can return good matches, and often will when ct is also low, cwland is high or rhcrit is low. We also see that

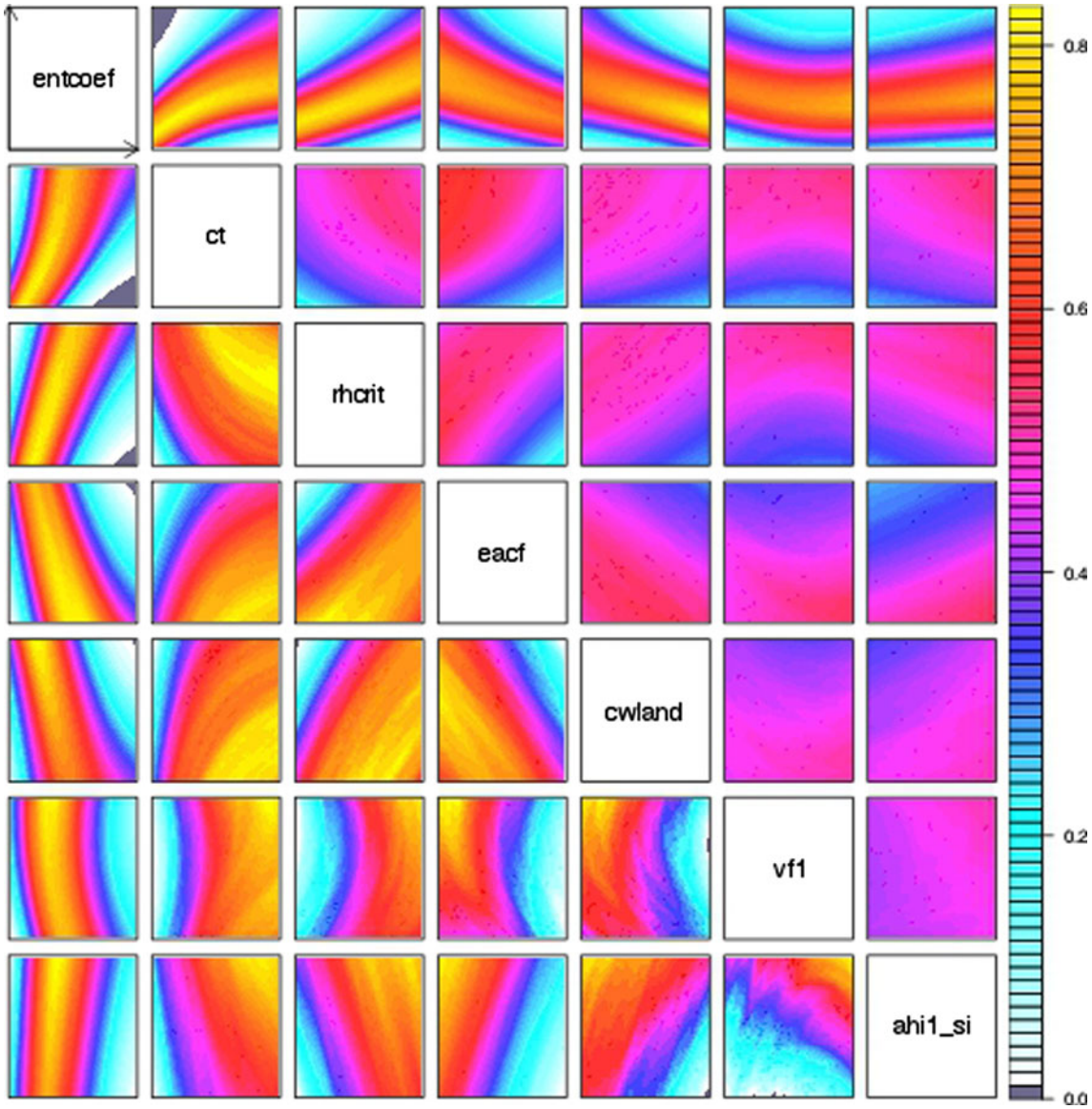


Fig. 3 NROY densities. The *scale* on the *right* applies only to the plots on the *upper triangle* of this *matrix*. Each plot on the *lower triangle* has its own relative scale so that any unusual shapes can be more readily seen and interpreted. Each point in each image represents the density of points in NROY space projected onto the relevant two dimensions within their standard ranges and with the *colour scale* defining the proportion of points in NROY space at that location throughout the other dimensions. The parameter cwland determines another parameter, cwsea (see Appendix C)

there are very few 2D projections that do not return any not implausible matches. The main implausible region is when entcoef is high and and ct is low.

This does not contradict the findings of Joshi et al. (2010), whose low entrainment model is right at the bottom of the range of entcoef plotted in Fig. 3 and so is more likely to be ruled out than not (a low entcoef point chosen randomly in the other 26 dimensions is more likely to be ruled out than not as most of the colours for low entcoef are light blue or white in Fig. 3). Instead, it states that the other parameters can be changed so that low entrainment values lead to climates that are consistent with the observational metrics used. Further, we note that we have not used the concentration of stratospheric water vapour to constrain NROY space here and so there may be many regions of parameter space that are NROY that we would regard as leading to non-physical models with respect to this, and other, metrics.

Five of the parameters in Fig. 3 (ct, rhcrit, eacf, cwland, vf1) affect aspects of the cloud schemes controlling the thickness, distribution and type of clouds formed. Sanderson et al. (2010) find that all these parameters and entcoef dominate the variability in the rate of warming for a given CO_2 increase. They do this through altering the amount and

distribution of both clouds and atmospheric water vapour. This affects how much solar radiation is reflected and how much longwave radiation is emitted, and therefore can have a significant impact on surface temperatures. Since it is SAT that is the strongest constraint on our ensemble (see Fig. 2), it may be that these parameters have a large impact on the implausibility through changing SAT by modifying the radiative properties of the atmosphere.

The final parameter shown is ahi1_si, the isopycnal diffusivity in the ocean which transfers properties such as temperature and salinity along contours of constant density. Where sloping isopycnals exist (e.g at high latitudes) ahi1_si drives an upward heat flux (against the global mean vertical temperature gradient) because the salinity derived density changes dominate (Brierley et al. 2010). Brierley et al. (2010) found that although ahi1_si does have an impact on SAT, perturbations of the vertical diffusivity had slightly bigger impacts on temperature. In this study vertical diffusivity is also perturbed, but it is found to have much less impact. The importance of ahi1_si here may be due to complex interactions between different parameters, that can only be captured using large ensembles. We also note that the timescales over which parameter perturbations take effect may differ by several orders of magnitude. Changes to atmospheric parameters will typically require a few weeks to months of model integration to reach equilibrium, whilst changes to ocean parameters may require decades or even centuries before quasi equilibrium is reached. Therefore our relatively short model integrations may mean we are not capturing the full behaviour of all of the ocean parameters.

We compare our NROY space with the results reported in the Bayesian calibration of HadSM3 in Sexton et al. (2011). HadSM3 has the same atmospheric, land and sea ice components as HadCM3, but uses a mixed layer slab ocean model rather than a dynamic ocean model. Most of their perturbed parameters coincide with those atmosphere parameters perturbed in this study, although they cannot perturb any ocean parameters. In their analysis they find that entcoef imposes the largest constraint and that values at the high end of the prescribed range are extremely unlikely to be x^*.

The amount of posterior density given to high values of rhcrit and ct in their study after using their first two constraints corresponds to the higher density of points in NROY space we see with higher values of ct and rhcrit. However, when further constraints are imposed in their study, their distributions become weighted more towards the standard settings. They state that their first constraint is highly associated with top of the atmosphere radiative balance. Since models with a slab ocean, such as HadSM3, constrain SAT at the expense of altering the top of the atmosphere radiative balance, this is equivalent to a relationship with SAT in a model with a dynamic ocean (Jackson et al. 2012). This suggests a reason for the similarity of our findings. We also find more points in NROY space for higher values of vf1 than lower which is consistent with their findings.

5.3 Sensitivity analysis

We now compare the NROY space that we obtained using the 'low' and 'high' values of Var[U], that component of our discrepancy assessment that is the variance of the difference between the true climate and the underlying mean of the MME. We perform the same experiment used to obtain the data plotted in Fig. 3 on these two alternative discrepancy cases and investigate the percentage of parameter space remaining under all three discrepancies.

The results are shown in Fig. 4, plotted against the implausibility cutoff value. Implausibility comparison plots for these alternative cases were very similar to those for the standard case. We can see from Fig. 4 that the lower value of discrepancy allows one to remove approximately 1 % more parameter space than the standard value. The high discrepancy value retains approximately 13 % more parameter space than the standard value. The small change between the standard and low estimated values of Var[U] is due to the fact that S, the part of discrepancy calculated as the variance of the CMIP3 models, dominates the denominator of (4) for these values of Var[U] (see

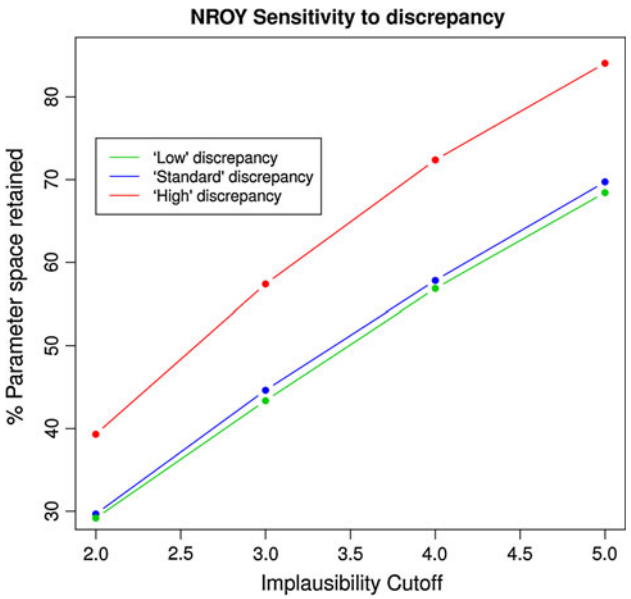


Fig. 4 A plot showing the percentage of parameter space remaining following a history match with the three levels of discrepancy obtained by setting Var[U] to the 'low', 'standard' and 'high' values quoted in Table 2. The percentage of parameter space remaining is plotted against the implausibility cutoff value, which we have taken to be 3 in the figures in this paper

Table 2). Note that even the high discrepancy case removes more than 40 % of the 'standard' parameter space using just the 4 observational constraints described in Sect. 3.3.

Due to the additive nature of the variances in (7) this sensitivity analysis could also apply to any of the other uncertainties we have quantified. In particular, consider the observation error variances that we obtained, which do not formally account for the uncertainty in the twentieth century change and could be seen as underestimates. Although adding the extra uncertainty to the discrepancy does mean that we rule out a lot less space, the characteristics of NROY space and the driving constraints remain the same, though we do not include the figures to save space. It is also worth noting that our high NROY space discrepancy corresponds to multiplying Var[U] by 10. Var[U] is greater than Var[e], the observational error variance, in all cases other than PRECIP. It is unlikely that we have underestimated the observation errors by an order of magnitude, hence we can assume that our conclusions are not as sensitive to this underestimate as they are to our discrepancy assessment.

5.4 The extended ranges

We repeat this sampling experiment within the extended ranges described in Table 5. We found that 46 % of the extended space that we defined was removed. Most notably, NROY density plots revealed that no part of any parameter's extended range had been classified implausible using our 4 constraints. This result illustrates the importance of careful consideration of parameter ranges in the design of PPEs, because any analysis done with a PPE is only valid within the parameter space it was designed for.

In general, where possible, ranges should be chosen to reflect the analyst's careful judgement regarding the feasible range of the parameters within which they are physically meaningful. In the absence of such judgment, either the boundaries of parameter space should be chosen to be as wide as possible, with the implication that anything outside the boundaries is deemed a priori implausible, or an iterative design process geared towards exploring these boundaries should be undertaken.

6 The AMOC and MHT in NROY space

Figure 5a shows the AMOC plotted against SAT and coloured by whether each ensemble member is in NROY space or not. It reveals a nonlinear relationship between SAT and AMOC across the whole parameter space, but an approximately linear relationship within NROY space. Jackson et al. (2012) also found this strong relationship between AMOC and SAT, however the nonlinearities at cold temperatures are only revealed with the larger ensemble size in this study.

As SAT is the dominant constraint, we know that our history match has essentially ruled out models that are either too cold or too hot. For warm models, greater relative warming at high latitudes and accelerated hydrological cycle result in a weaker AMOC. However, it is also true that the very cold models have a weak AMOC, with the maximum AMOC strength occurring at around 10 °C. At very cold temperatures there is enhanced sea ice coverage which insulates the ocean from the atmosphere, reducing the area over which convection can occur. We believe it is this that causes the weak overturning at cold temperatures.

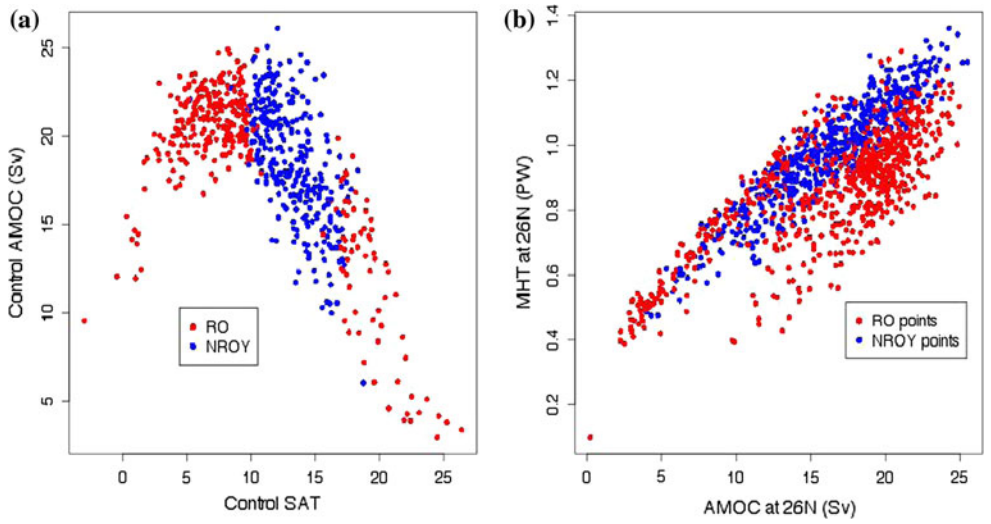


Fig. 5 **a** The AMOC plotted against SAT. Ruled out points appear in *red*, where as points in NROY space are in *blue*. **b** The meridional heat transport (MHT) at 26°N in petawatts plotted against the AMOC at the same latitude for the mean of the last decade. *Red points* have been ruled out, whereas the *blue points* are in NROY space

Investigating this mechanism is a subject of future research.

There is also a nonlinear relationship between the AMOC and MHT strength in the full parameter space (Fig. 5b) in that a quadratic fit of MHT on AMOC explains more variability than a linear one. However, the relationship could also be viewed as linear with two potential trajectories. This "bimodality" is clearer when looking at only the ruled out ensemble members. The relationship between AMOC and MHT is approximately linear in NROY space and is consistent with only one of the trajectories followed by the RO runs. The bimodality we have described is related to the nonlinear relationship between SAT and AMOC in the full space (Fig. 5a) with colder models having a lower MHT than warmer RO models with the same AMOC strength.

Figure 6 shows the transient and control AMOC projections in NROY space from our ensemble. By history matching on simpler (i.e. univariate and easier to emulate) quantities such as global mean SAT, we have removed many ensemble members with unrealistically weak control AMOC strengths and removed a substantial part of the burden faced in emulating a complex quantity such as the AMOC time series over the whole parameter space by just focussing on that part of the parameter space that we are unable to easily rule out on the basis of observations. An emulator for this time series in NROY space is the subject of Williamson and Blaker (2013). However, we now show some results that illustrate the simplification of the emulation problem for the AMOC, and the effect of making inferences about the AMOC within NROY space as compared with the whole parameter space.

The MHT and AMOC change across the ensemble is plotted against the percentage of CO_2 increase in Fig. 7 with the points coloured by whether or not they are in NROY space. MHT/AMOC change is calculated across the ensemble by subtracting the mean of the last 20 years of the control run from the mean of the last 20 years of the corresponding transient run where both are available. Subtracting the control value removes the effect of drift from continued spin up (we make the assumption that the adjustments due to spin up and from increasing CO_2 can be linearly separated). The MHT/AMOC decrease after 70 years of increasing CO_2 appears to respond linearly to the rate of CO_2 increase in NROY space. However the relationship appears to be more complex in the unconstrained parameter space.

The figure also suggests that the AMOC in NROY space is more consistently sensitive to CO_2 than those in ruled out (RO) space, with greater reduction of AMOC strength in NROY space for a given rate of CO_2 increase. This would imply that the sensitivity of the AMOC in HadCM3 depends in some way on the model parameters. For example, if the AMOC response to increasing CO_2 depends on the control strength of the AMOC (as seen in Jackson et al. 2012) and the control strength depends on the model parameters.

We investigate this further in Fig. 8 where we compare the MHT and AMOC change due to around 1 and 2 % annual CO_2 increase in the whole ensemble with those RO

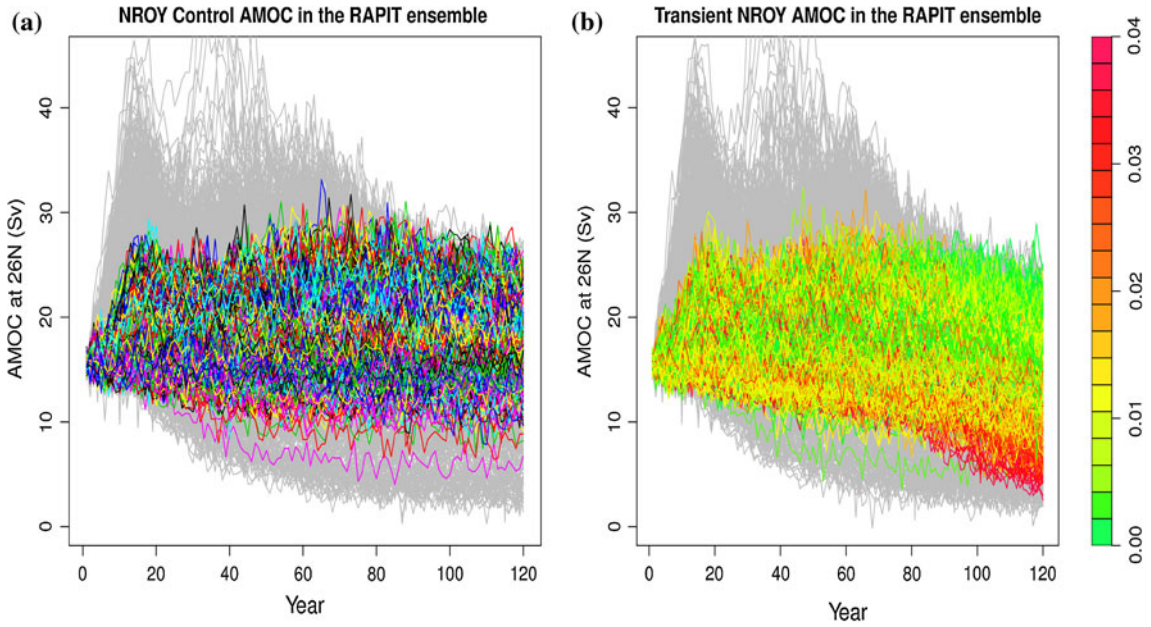
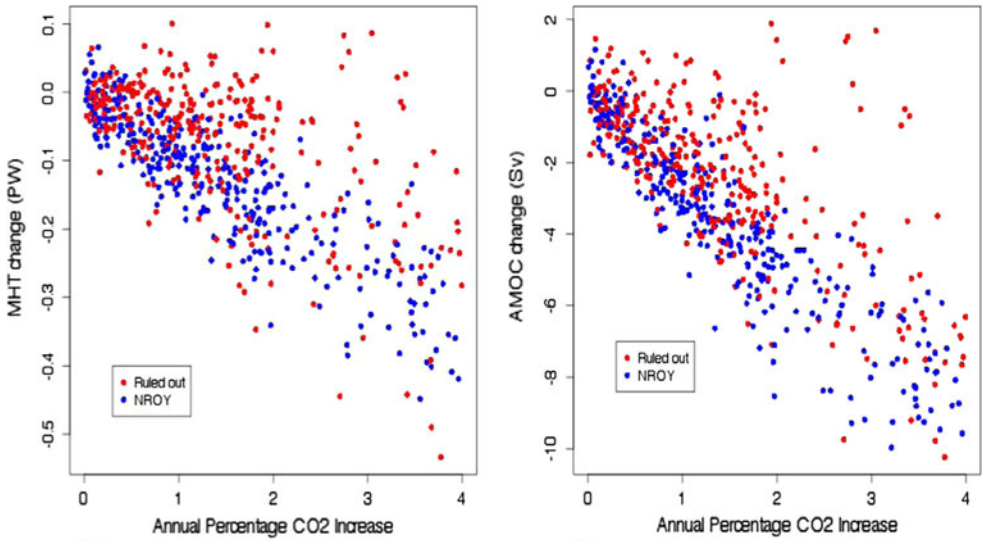


Fig. 6 **a** Control AMOC at 26°N in NROY space in the RAPIT ensemble. The *colours* serve to highlight the variability of individual runs in NROY space. **b** Transient AMOC at 26°N in NROY space in the RAPIT ensemble coloured by Δ, the annual rate increase of CO_2 forcing applied after year 50. The *grey lines* in both figures are the ruled out runs



(a) MHT change against CO_2 increase **(b)** AMOC change against CO_2 increase

Fig. 7 **a** MHT change plotted against the annual percentage increase in CO_2 and coloured by whether or not the points are ruled out (RO) or in NROY space. **b** AMOC change plotted against the annual percentage increase in CO_2 and coloured by whether or not the points are ruled out (RO) or in NROY space. The MHT/AMOC change is calculated as the difference between the mean MHT/AMOC over years 101–120 in the transient and matching control simulations

changes and those in NROY space. We take the subsets of those runs that are forced by between 0.9 and 1.1 % as those at '1 %' and those forced between 1.8 and 2.1 % (because not enough ensemble members within 1.9−2.1 % had completed) as those at '2 %' and draw boxplots for all the runs in the sub-ensemble together and for the RO and NROY components of the sub-ensemble.

From Fig. 8 we note that the variability in our projections is reduced in NROY space when compared with the unconstrained parameter space. It is also noteworthy that the NROY projections show a larger mean AMOC and MHT decrease than those that are ruled out. This implies that our unconstrained perturbed physics ensemble hides the extent of projected AMOC weakening due to CO_2 forcing and that history matching has "shifted" the distribution of the AMOC response as well as reducing the spread. Looking at Fig. 6, we may suspect that this is due to weaker control AMOCs being ruled out. However, by re-performing the analysis for only those ensemble members in which the AMOC in the control simulation is stronger than 10Sv, we find that the difference in the location of the distribution of RO and NROY responses is still present and is unchanged. The reduced sensitivity of the AMOC to increasing CO_2 in RO space instead comes from the colder models and may be related to the positive gradient (increased AMOC strength for increased SAT) seen in the relationship between the control values of SAT and AMOC in Fig. 5a. Inspection of the timeseries of the cold models reveals weaker decreasing trends than the models in NROY space for the 70 years of increased CO_2, and no evidence of nonlinear transient behaviour. This means that the AMOC in NROY space is, on average, more responsive to CO_2 forcing than in the ruled out space, and that the sensitivity of the AMOC depends on the model parameters through the SAT.

7 Discussion

We have used history matching to constrain the parameter space of HadCM3 using a small selection of pre-industrial global and hemispheric averages of climatic variables. Our analysis rules out 56 % of the original parameter space explored here and in previous studies (Murphy et al. 2004; Sexton et al. 2011). We used our large computing resource to test extended ranges of a number of the parameters of HadCM3 and discovered that no parts of the extended ranges could be ruled out using our constraints. NROY (not ruled out yet) space is not defined outside of the original parameter space or under alternative versions of HadCM3 with transformed parameters (though our emulators may be robust to certain transforms).

The final NROY parameter space represents the region of our original parameter space where all four of our constraints are predicted to be within a tolerance of the true historical values for these constraints. This tolerance is set using the error on the observations and discrepancy variance information derived from CMIP3, whilst accounting for the uncertainty in our emulator-based predictions. Though a number of assumptions are required in order to establish a statistical model to help derive discrepancy variance, these assumptions are both weaker than the sorts

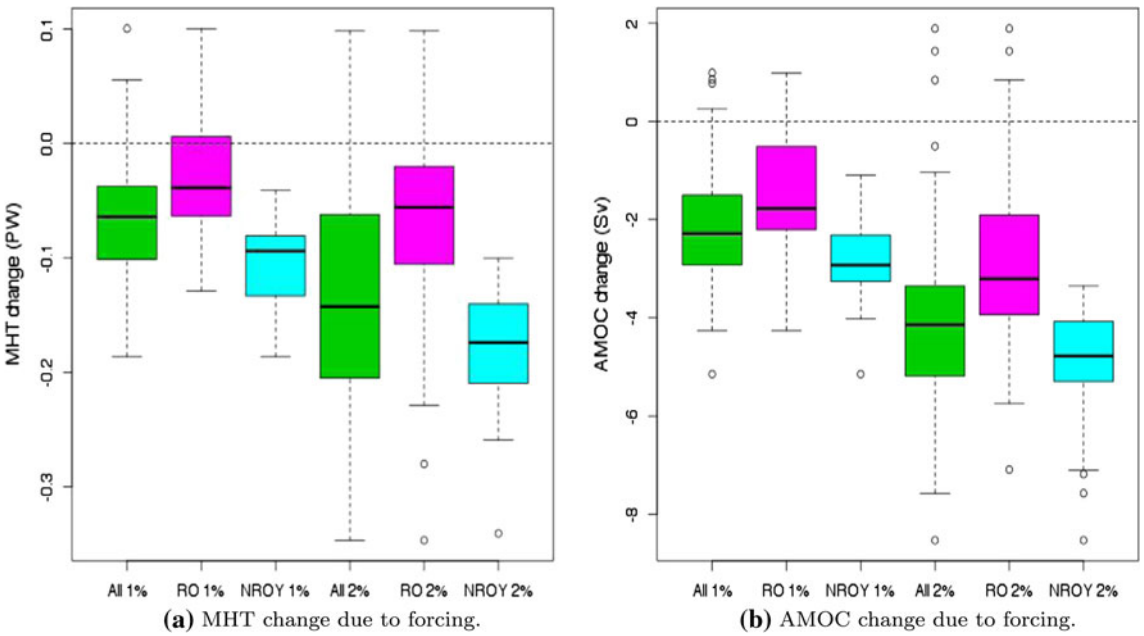


Fig. 8 *Boxplots* of the absolute MHT (*left image*) and AMOC (*right image*) change due to around 1 and 2 % CO_2 forcing in the RAPIT ensemble. MHT/AMOC change is calculated as the difference in means for the last 20 years of the transient and control runs. The distribution of changes across the whole ensemble is shown and compared with that of the points ruled out (RO) and with those in NROY space. The bounds of the *boxes* are the upper and lower quartiles of the data, the *horizontal centre line* is the median, the *lines* represent the extent of the distribution after any outliers are removed. The outliers, those points lying more than 1.5 times the interquartile range from the upper and lower quartiles, appear as separate points away from the *box*

of assumptions usually used when using climate models to learn about climate. The potential impact on our conclusions is assessed using a sensitivity analysis showing that even increasing the estimated discrepancy variance by an order of magnitude only results in around 10 % less parameter space removed by history matching.

We found that the dominant observational constraint was the global mean surface air temperature, with the other observations hardly contributing to the resulting NROY space. We discovered that this was due to high correlations between the global mean surface air temperature and the other constraints and to larger uncertainties for those constraints. We found seven parameters had a large impact on NROY space, most of which were parameters controlling cloud processes. One, however, was an ocean mixing parameter which might affect the heat uptake by the ocean and hence the atmospheric temperature.

We have shown the benefit of history matching with easy to emulate, yet physically important, observational constraints before analysing more complex processes using PPEs. We observed that the distribution of transient behaviour of the AMOC and its relationship with the meridional heat transport (MHT) was considerably simpler inside NROY space to that in the whole ensemble. We found that the average rate of AMOC weakening predicted due to CO_2 forcing was slower in RO space than in NROY space and found that this was due to a non linear relationship between AMOC and SAT in the full parameter space. These results show that if we were to use the whole ensemble to make inferences about these aspects of the climate, we would come to different conclusions than if we had first constrained the ensemble.

The history match we present here is the first step in a quantification of parametric uncertainty that must go on to include the emulation of quantities of interest under forcing within NROY space and, ideally, further constraint, either in the form of more history matching and refocussing, as described in Sect. 2, or in the form of a Bayesian calibration within NROY space. The constraints may become more complex and may include fields of observations over time (as in Sexton et al. 2011) depending on the scientific judgments pertaining to what behaviours it is important for a simulation to be able to capture before projections using that simulation are trusted. However, complex constraints require sophisticated statistical emulators that are valid throughout NROY space in order to impose them. We have shown that these emulators would be easier to fit within NROY space thus emphasising the value of this initial history match.

Parametric uncertainty may be a major source of uncertainty in predictions on decadal to centennial time scales using even the most advanced climate models. This uncertainty must be quantified and reported if we are to have

confidence in the reported uncertainties for future projections. The only way to do this is through the use of large perturbed physics ensembles. For more advanced climate models that require greater computer resources, such as those with higher resolution, it may not be possible to obtain sufficiently large ensembles. However, AOGCMs of different resolution can be linked statistically (Williamson et al. 2012) so that large PPEs using coarser-resolution models and a small PPE using a very expensive model can be used together to emulate the advanced model and quantify its parametric uncertainty in the ways we have described.

The specification of a statistical model relating a climate model to the real climate system is important in any model-based quantification of uncertainty or risk assessment in the real world. The most popular statistical model for this defines the real climate to be the sum of the model evaluated at the 'best input' x^*, and an uncorrelated model discrepancy term via Eq. (2) that requires detailed expert scientific judgement from climate modellers to specify. We have introduced an alternative statistical model (Sect. 4.1) for the case when detailed expert judgements are not available. Under this model we view discrepancy variance as our tolerance to climate model error. We use a multi model ensemble, CMIP3, to derive tolerances to error that are consistent with using CMIP3 as representative of the judgements of the climate community regarding what represents an informative climate model.

Our statistical model is based on a transparent set of assumptions and leads to a discrepancy variance estimate that can be readily computed. However, there is considerable variability in the distribution of the variance estimates we can obtain from a MME like CMIP3. Though our statistical model allows tractable discrepancy variance estimates to be readily obtained, it represents just one possible way of obtaining these judgements. Further innovations in statistics may lead to improvements in this model or improved alternative models.

An ideal analysis, and perhaps the only way to truly quantify this uncertainty, must involve expert judgement regarding the deficiencies in the mathematical representations of the physics in the model and the ways they might be improved in order to capture these deficiencies in future generations of models (see Goldstein and Rougier 2009, for further discussion).

Acknowledgments This research was funded by the NERC RAPID-RAPIT project (NE/G015368/1), and was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (GA01101). We would like to thank the CPDN team, in particular Andy Bowery, for their work in submitting our ensemble to the CPDN users. We'd also like to thank Richard Allan for helpful discussions regarding precipitation estimates, and all of the CPDN users around the world who contributed their spare computing resource as part of the generation of our ensemble. Finally, we'd like to thank both referees for their thoughtful and detailed comments.

Appendix A: The statistical model derivation and justification for model discrepancy

We begin this technical Appendix A by deriving Eq. (7) using the statistical model in Sect. 4.1 given by Eqs. (5) and (6). Under our statistical model, supposing $f^h(x)$ is our climate model and is part of the MME and given the relation between the observations and the true climate to be as in (3), if x_0 were x^* we would have

$$\begin{aligned} z - \mathrm{E}[f^h(x_0)] &= (z - y) + (y - \mathcal{M}) + (\mathcal{M} - f^h(x_0)) \\ &\quad + (f^h(x_0) - \mathrm{E}[f^h(x_0)]) \\ &= e + U - R_h + \xi(x_0), \end{aligned}$$

The observation error, e, is uncorrelated with the other terms by definition, and U, the discrepancy between y and the expectation of the MME, and R_h, the discrepancy between f^h and the expectation of the MME, are uncorrelated as a consequence of the representation theorem. We judge $\xi(x_0)$ to be uncorrelated with the other terms as this represents the error in our emulator based on a large ensemble at x_0, which should have no relationship to these other errors. Hence, taking variances, Eq. (7) follows.

Our statistical model makes the explicit assumption that $f^h(x^*_{[h]})$, our climate model run at its best input is second order exchangeable with the MME. If the MME members are second order exchangeable, this would be fine so long as they represented different climate models run at their best inputs. However, this is not the case. Each MME member is a 'tuned run' of a different climate model run at some setting of its parameters $x^t_{[j]}$ where this setting is highly unlikely to be the best input, $x^*_{[j]}$, because, if it were, we would not be interested in the parametric uncertainty of the climate model. The usually adopted statistical model assumes $y = f^j(x^*_{[h]}) + \eta^j$, where we cannot learn about η^j using our climate model (Rougier 2007). If $x^t_{[j]}$ were equivalent to $x^*_{[j]}$, under this model, a PPE tells you nothing more about climate than you already knew from $f^j(x^*_{[h]})$. Hence, usually, we would expect $f^h(x^*_{[h]})$ to be closer to y than $f^h(x^*_{[h]})$. The implication of this assumption is to have derived a larger discrepancy than is, perhaps, required.

To illustrate, suppose that we did view the climate models that contribute to the MME at their best inputs as second order exchangeable so that $f^j(x^*_{[j]}) = \mathcal{M} \oplus R_j(f^j(x^*_{[j]})) = \mathcal{M} \oplus R^*_j$. Further, suppose that we have no a priori reason to believe that any of the MME members have been tuned closer to their best inputs than any other, so that

$$f^j(x^t_{[j]}) = f^j(x^*_{[j]}) \oplus R^t_j,$$

where R^t_j is a mean-zero residual and $\mathrm{Cov}[R^t_j, R^t_k] = 0$ for $j \neq k$.

Then

$$f^j(x^t_{[j]}) = \mathcal{M} \oplus R^*_j \oplus R^t_j,$$

which, by ignoring tuning as we have in (5), implies that $R_j = R^*_j + R^t_j$. This means that our discrepancy term has additional error due to tuning built in.

In practice it would be very difficult to assess the individual contribution of tuning error R^t_j to the discrepancy without expert judgement from the climate modellers, even if the illustrative framework above were adopted.

However, by having a larger discrepancy we rule out less parameter space. This cautious approach is consistent with a first-wave history match, with the aim of removing those regions of parameter space that correspond to extremely unphysical climates. Though some of the remaining error can be attributed to tuning, this statistical model does not specifically account for the fact that the MME values we get from CMIP3 are not ten year means from climate models with a short spin up phase as our ensemble members are. This factor leads us to choose a statistical model that leads to larger discrepancy variance and less parameter space ruled out than might be obtained using expert judgement.

Appendix B: Constraint emulators

In order to fit each emulator mean function, we used a stepwise selection method to add or subtract functions of the model parameters to our vector $g(x)$. The allowed functions were linear, quadratic and cubic terms with up to third order interactions between all parameters. Switch parameters were treated as factors (variables with a small number of distinct possible "levels") and factor interactions with all continuous parameters were allowed.

First we perform a forward selection procedure where we permit each variable to be added to $g(x)$ in its lowest available form. So if ct is not yet in $g(x)$, ct can be added, but ct^2 cannot be added. If ct is already in $g(x)$ then ct^2 can be added, but we simultaneously add all first order interactions with the other variables in $g(x)$ so that the resulting statistical model will be robust to changes of scale (see Draper and Smith 1998, for discussion). We do similar for third order interactions. The term(s) added at each stage is the one that reduces the residual sum of squares from the regression the most.

When it becomes clear that adding more terms is not improving the predictive power of the emulator (a judgement made by the statistician based on looking at the proportion of variability explained by the regression and at plots of the residuals from the fit) we begin a backwards elimination algorithm. This removes terms, one at a time, with the least contribution to the sum of squares explained, without compromising the model. Lower order terms are not permitted to be removed from $g(x)$ whilst higher order terms remain. We stop when the removing the next term chosen by the algorithm leads to a poorer statistical model. For more details on forwards selection methods and backwards elimination, see Draper and Smith (1998).

The emulators contained between 76 terms in $g(x)$ (in the case of SAT) and 44 terms (in the case of PRECIP). For the SAT emulator, it was found that the response was easier to model as a function of log(entcoef), though this

Table 3 A table indicating which terms are in $g(x)$ for our emulator of SAT in Eq. (1)

	ent	ct	rhcrit	eacf	cwland	vf1	AH1	dyndiff	g0	minS	ice	dynd	r_lay
ent	2	1	1	1	1	1	1	0	1	1	1	1	1
ct	**ent**	3	1	1	1	1	1	0	1	1	1	0	0
rhcrit	**eacf**	**ct**	1	1	1	0	1	0	1	1	1	0	0
eacf	**ct**	**ct**	**ct**	1	1	1	1	1	1	0	1	0	0
cwland	**ent**	**ct**	**ent**	**ct**	1	1	1	0	1	0	1	0	0
vf1	**vf1/ent**				**ct**	2	0	1	0	0	0	0	0
AH1							1	0	0	1	1	0	0
dyndiff								1	0	0	0	1	0
g0		**ct**							1	0	0	0	0
minS										1	0	0	1
ice					**ct**						1	0	0
dynd												1	0
r_lay													1

The column and row names refer to the parameter names in Table 4, shortened in an obvious way in order to save space. The upper triangle labels which interaction pairs are present. The diagonal indicates the order of the highest order term in that variable. The lower triangle indicates which three way interactions are included

transformation produced inferior mean functions for each of the other 3 outputs. The terms for the SAT emulator can be seen in Table 3. Each header in the table refers to one of the parameters from Tables 5 and 6, shortened in an obvious way to save space. Numbers on the diagonal refer to power terms in $g(x)$ in each of the relevant parameters. The number 1 on the diagonal implies only a linear term was included in $g(x)$. The number 2 implies that both quadratic and linear terms were in $g(x)$. The number 3 implies cubic, quadratic and linear terms.

Numbers on the upper triangle refer to the inclusion or not of interactions between the two relevant variables. For example, reading from the table, the term (vf1*dyndiff) is included in $g(x)$, but the term (vf1*g0) is not. Variables indicated in bold on the lower triangle refer to three way interactions that are present in $g(x)$. For example, the terms (ent*vf1^2) and (vf1*ent^2) are both included in $g(x)$. Note that dynd and r_lay are both handled by the model as factors with pre-specified levels. Our emulator does contain a constant term, so the vector $g(x)$ includes the element 1.

Having fitted a mean function, the next step in emulation is to provide a statistical model for the residual $\epsilon(x)$. Upon investigation, the residuals from the fitted mean functions appeared to behave like white noise. Hence we decided that it was not worth modelling $\epsilon(x)$ in Eq. (1) as a weakly stationary Gaussian process (where $\epsilon(x)$ and $\epsilon(x')$ are correlated via a covariance function $R(|x - x'|)$ that depends on the distance between points in parameter space (See, for

Table 4 Parameter descriptions and model section

Parameter	Description	Section
vf1	Ice fall speed (m/s)	Cloud
ct	Cloud droplet to rain conversion rate (/s)	Cloud
CWland	Cloud droplet to rain threshold over land (kg/m^3)	Cloud
CWsea	Cloud droplet to rain threshold over sea (kg/m^3)	Cloud
RHCrit	Relative humidity threshold for cloud formation	Cloud
eacfbl	Boundary layer cloud fraction at saturation	Cloud
entcoef	Convective cloud entrainment rate coefficient	Convection
MinSIA	Albedo at ice melting point	Sea ice
dtice	Ocean ice diffusion coefficient	Sea ice
Icesize	Ice particle size (μm)	Radiation
k_gwd	Surface gravity wavelength (m)	Dynamics
lay_lee_gwave	Trapped lee wave constant for surface gravity waves (m$^{3/2}$)	Dynamics
start_level_gwdrag	First level for gravity wave drag	Dynamics
dyndiff	Diffusion efolding time (hours)	Dynamics
dyndel	Order of diffusion operator	Dynamics
asym_lambda	Asymptotic neutral mixing length parameter	Boundary
charnock	Charnock constant	Boundary
cnv_rl	Free convective roughness length over sea (m)	Boundary
flux_g0	Boundary layer flux profile parameter	Boundary
r_layers	No. of soil levels for evaporation	Land surface
L	SO_2 wet scavenging rate (/s)	Sulphur cycle
volsca	Scaling for volcanic SO_2 emissions	Sulphur cycle
anthsca	Scaling for anthropogenic SO_2 emissions	Sulphur cycle
so2_high_level	Model level for SO_2 emissions	Sulphur cycle
vb	Background vertical viscosity (m^2/s)	Ocean
kb	Background vertical diffusivity (m^2/s)	Ocean
dkb/dz	Background vertical diffusivity gradient (m/s)	Ocean
AH1_SI	Isopycnal diffusivity (m^2/s)	Ocean
lambda	Wind mixing energy scaling factor	Ocean
delta_si	Wind mixing energy decay depth (m)	Ocean

CWland determines CWsea, MinSIA determines dtice and k_gwd determines kay_lee_gwave

example, Williamson et al. 2012). We therefore opted to model $\epsilon(x)$ as mean zero uncorrelated error as is done by Sexton et al. (2011) and Rougier et al. (2009). We let the variance of $\epsilon(x)$ be equal to the variance of the residuals from the four regressions.

To validate the emulators, we kept 100 randomly chosen ensemble members back as a validation set. The emulators were not trained using these members. We then predict the value of the constraints in each ensemble member using the emulator and examine the residuals (the difference between our prediction and the truth). The residuals are plotted against the predicted values in Fig. 9.

The emulator's performance is acceptable as Fig. 9 shows that our emulator predictions are consistent with our uncertainty specification. If anything, the validation suggests that we have over specified our variance (at least in the case of SCYC and PRECIP), because no points fall outside 3 standard deviations. However, this will only lead to less parameter space being ruled out, and we should be cautious in terms of what parts of parameter space are ruled out and what is retained. We make the four emulators available to download along with our implausibility code so that our results may be replicated, and their sensitivity to any of our judgements explored.

Appendix C: Tables

Tables 4, 5 and 6 give descriptions and ranges for the parameters and the settings of switches used in our ensemble. Some parameters have relationships with other model parameters that were given to us by the Met Office so that a change in one leads to a derivable value for the other. CWland also determines CWsea, the cloud droplet to rain threshold over sea (kg/m^3), MinSIA also determines

Table 5 Ranges for each of the continuous parameters varied in our 10,000 member ensemble

Parameter	Section	Standard lower	Standard higher	New lower	New higher
vf1	Cloud	0.5	2	0.15	2.35
ct	Cloud	5×10^{-05}	4×10^{-04}	*	5.625×10^{-04}
CWland	Cloud	1×10^{-04}	2×10^{-03}	*	*
RHCrit	Cloud	0.6	0.9	*	*
eacfbl	Cloud	0.5	0.8	*	*
entcoef	Convection	0.6	9	*	*
MinSIA	Sea ice	0.5	0.65	*	*
Icesize	Radiation	2.5×10^{-05}	4×10^{-05}	2×10^{-05}	8×10^{-05}
k_gwd	Dynamics	$1 \times 10^{+04}$	$2 \times 10^{+04}$	*	*
dyndiff	Dynamics	6	24	*	*
asym_lambda	Boundary	0.05	0.5	0.01	0.61
charnock	Boundary	0.012	0.02	0.012	0.024
cnv_rl	Boundary	2×10^{-04}	5×10^{-03}	2×10^{-04}	6.2×10^{-03}
flux_g0	Boundary	5	20	2.5	22.5
L	Sulphur cycle	0.33	0.33	*	*
volsca	Sulphur cycle	1	3	0.5	3.5
anthsca	Sulphur cycle	0.5	1.5	0.25	1.75
vb	Ocean	5×10^{-06}	8×10^{-05}	1×10^{-06}	1.1×10^{-04}
kb	Ocean	5×10^{-06}	2×10^{-05}	1×10^{-06}	3.1×10^{-05}
AH1_SI	Ocean	200	2,000	100	2,500
dkb/dz	Ocean	7×10^{-09}	9.8×10^{-08}	*	*

* we don't change the standard range in the exploratory sub ensemble. We don't give values for dependent parameters CWsea, dtice and kay_lee_gwave as these are calculated from CWland, MinSIA and k_gwd respectively via a one to one mapping

Table 6 Switch parameters and their settings in our 10,000 member ensemble

Parameter	Section	Setting 1	Setting 2	Setting 3
so2_high_level	Sulphur cycle	3	5	*
start_level_gwdrag	Dynamics	3	4	5
r_layers	Land surface	[2, 1]	[3, 2]	[4, 3]
dyndel	Dynamics	4	6	*
lamda/delta_si	Mixed layer	[0.3, 100]	[0.5, 50]	[0.7, 100]

* indicates that there are only 2 settings of a switch

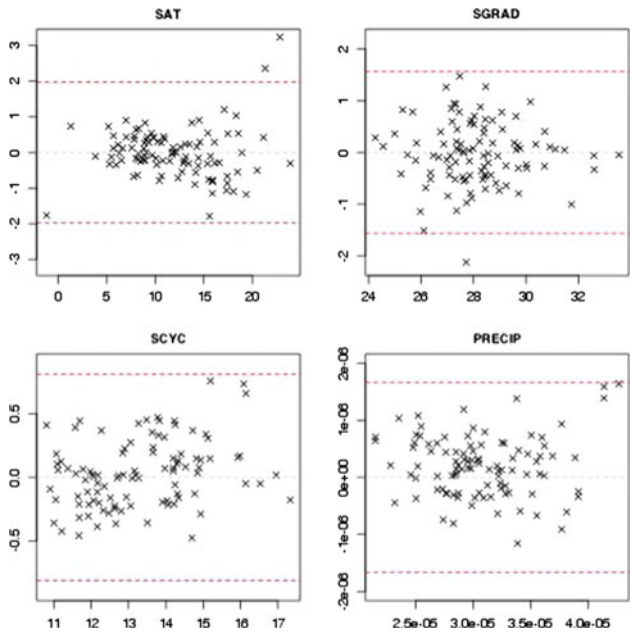


Fig. 9 Validation plots for each of the 4 emulators fitted to the observational constraints. For each variable stated in the plot title, the predicted value of the emulator for that variable (x-axis) is plotted against the difference between the truth and the emulator prediction for the data in the validation set. The *red dashed lines* are ±3 standard deviations of the predictions. The validation set was not used when building the emulators

dtice (the ocean ice diffusion coefficient) and k_gwd also determines kay_lee_gwave (the trapped lee wave constant for surface gravity waves $m^{3/2}$).

References

Acreman DM, Jeffery CD (2007) The use of Argo for validation and tuning of mixed layer models. Ocean Model 19:53–69

Berliner LM, Kim Y (2008) Bayesian design and analysis for superensemble-based climate forecasting. J Clim 21(9):1981–1910

Brierley CM, Collins M, Thorpe AJ (2010) The impact of perturbations to ocean-model parameters on climate and climate change in a coupled model. Clim Dyn 34:325–343

Broecker WS (1987) The biggest chill. Nat Hist Mag 97:74–82

Brohan P, Kennedy JJ, Harris I, Tett SFB, Jones PD (2006) Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. J Geophys Res 111:D12106

Challenor P, McNeall D, Gattiker J (2009) Assessing the probability of rare climate events. In: O'Hagan A, West M (eds) The handbook of applied Bayesian analysis, chap. 10. Oxford University Press, Oxford

Collins M, Brierley CM, MacVean M, Booth BBB, Harris GR (2007) The sensitivity of the rate of transient climate change to ocean physics perturbations. J Clim 20:23315–2320

Collins M, Booth BBB, Bhaskaran B, Harris GR, Murphy JM, Sexton DMH, Webb MJ (2011) Climate model errors, feedbacks and forcings: a comparison of perturbed physics and multi-model experiments. Clim Dyn 36:1737–1766

Craig PS, Goldstein M, Seheult AH, Smith JA (1996) Bayes linear strategies for matching hydrocarbon reservoir history. In: Bernado JM, Berger JO, Dawid AP, Smith AFM (eds) Bayesian statistics 5. Oxford University Press, Oxford, pp 69–95

Craig PS, Goldstein M, Seheult AH, Smith JA (1997) Pressure matching for hydrocarbon reservoirs: a case study in the use of Bayes linear strategies for large computer experiments. In: Gatsonis C, Hodges JS, Kass RE, McCulloch R, Rossi P, Singpurwalla ND (eds) Case studies in Bayesian statistics vol III. Springer, New York, pp 36–93

Craig PS, Goldstein M, Rougier JC, Seheult AH (2001) Bayesian forecasting for complex systems using computer simulators. J Am Stat Assoc 96:717–729

Cumming JA, Goldstein M (2010) Bayes linear uncertainty analysis for oil reservoirs based on multiscale computer experiments. In: O'Hagan A, West M (eds) The Oxford handbook of applied Bayesian analysis. Oxford University Press, Oxford, pp 241–270

de Finetti B (1974) Theory of probability, volume 1. Wiley, New York

de Finetti B (1975) Theory of probability, volume 2. Wiley, New York

Dickson RR, Brown J (1994) The production of North Atlantic deep water: sources, rates, and pathways. J Geophys Res Oceans 99:12,319–12,341

Draper NR, Smith H (1998) Applied regression analysis, 3rd edn. Wiley, New York

Edwards NR, Cameron D, Rougier JC (2011) Precalibrating an intermediate complexity climate model. Clim Dyn 37:1469–1482

Frieler K, Meinshausen M, Schneider von Deimling T, Andrews T, Forster P (2011) Changes in global-mean precipitation in response to warming, greenhouse gas forcing and black carbon. Geophys Res Lett 38:L04702. doi:10.1029/2010GL045953

Furrer R, Sain SR, Nychka D, Meehl GA (2007) Multivariate Bayesian analysis of atmosphere–ocean general circulation models. Environ Ecol Stat 14:249–266

Goldstein M (1986) Exchangeable belief structures. J Am Stat Assoc 81:971–976

Goldstein M (1986) Prevision. In: Kotz S, Johnson NL (eds) Encyclopaedia of statistical sciences, vol 7. pp 175–176

Goldstein M, Rougier JC (2004) Probabilistic formulations for transferring inferences from mathematical models to physical systems. SIAM J Sci Comput 26(2):467–487

Goldstein M, Rougier JC (2009) Reified Bayesian modelling and inference for physical systems. J Stat Plan Inference 139:1221–1239

Goldstein M, Wooff D (2007) Bayes linear statistics theory and methods. Wiley, New York

Gordon C, Cooper C, Senior CA, Banks H, Gregory JM, Johns TC, Mitchell JFB, Wood RA (2000) The simulation of SST, sea ice

extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. Clim Dyn 16:147–168

Gregory JM, Dixon KW, Stouffer RJ, Weaver AJ, Driesschaert E, Eby M, Fichefet T, Hasumi H, Hu A, Jungclaus JH,Kamenkovich IV, Levermann A, Montoya M, Murakami S, Nawrath S, Oka A, Sokolov AP, Thorpe, RB (2005) Subannual, seasonal and interannual variability of the North Atlantic meridional overturning circulation. Geophys Res Lett 32. doi:10.1029/2005GL023209

Huffman GJ, Adler RF, Bolvin DT, Gu G (2009) Improving the global precipitation record: GPCP Version 2.1. Geophys Res Lett 36:L17808. doi:10.1029/2009GL040000

Jackson L, Vellinga M, Harris G (2012) The sensitivity of the meridional overturning circulation to modelling uncertainty in a perturbed physics ensemble without flux adjustment. Clim Dyn. doi:10.1007/s00382-011-1110-5

Johns WE, Baringer MO, Beal LM, Cunningham SA, Kanzow T, Bryden HL, Hirschi JJM, Marotzke J, Meinen C, Shaw B, Curry R (2011) Continuous, array-based estimates of Atlantic Ocean heat transport at 26.5N. J Clim 24:2429–2449

Jones PD, New M, Parker DE, Martin S, Rigor IG (1999) Surface air temperature and its changes over the past 150 years. Rev Geophys 37(2):173–199

Joshi MM, Webb MJ, Maycock AC, Collins M (2010) Stratospheric water vapour and high climate sensitivity in a version of the HadSM3 climate model. Atmos Chem Phys 10:7161–7167

Kraus EB, Turner J (1967) A one dimensional model of the seasonal thermocline II. The general theory and its consequences. Tellus 19:98–106

Legates DR, Willmott CJ (1990) Mean seasonal and spatial variability in global surface air temperature. Theor Appl Climatol 41:11–21

Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. J R Stat Soc Ser B 63:425–464

Kuhlbrodt T, Griesel A, Montoya M, Levermann A, Hofmann M, Rahmstorf S (2007) On the driving processes of the Atlantic meridional overturning circulation. Rev Geophys 45. doi:10.1029/2004RG000166

Liu C, Allan RP (2012) Multisatellite observed responses of precipitation and its extremes to interannual climate variability. J Geophys Res 117:D03101. doi:10.1029/2011JD016568

Manabe S, Stouffer RJ (1980) Sensitivity of a global climate model to an increase of CO2 concentration in the atmosphere. J Geophys Res 85:5529–5554

McManus JF, Francois R, Gherardi JM, Keigwin LD, Brown-Leger S (2004) Collapse and rapid resumption of Atlantic meridional circulation linked to deglacial climate changes. Nature 428:834–837. doi:10.1038/nature02494

Meehl GA, Covey C, Delworth T, Latif M, McAvaney B, Mitchell JFB, Stouffer RJ, Taylor KE (2007) The WCRP CMIP3 multimodel dataset: a new era in climate change research. Bull Am Meteorol Soc 88:1383–1394

Morris MD, Mitchell TJ (1995) Exploratory designs for computational experiments. J Stat Plan Inference 43:381–402

Murphy JM, Sexton DMH, Barnett DN, Jones GS, Webb MJ, Collins M, Stainforth DA (2004) Quantification of modelling uncertainties in a large ensemble of climate change simulations. Nature 430:768–772

Murphy JM, Sexton DMH, Jenkins GJ, Booth BBB, Brown CC, Clark RT, Collins M, Harris GR, Kendon EJ, Betts RA, Brown SJ, Humphrey KA, McCarthy MP, McDonald RE, Stephens A, Wallace C, Warren R, Wilby R, Wood R (2009) UK Climate Projections Science Report: Climate change projections. Met Office Hadley Centre, Exeter, UK. http://ukclimateprojections.defra.gov.uk/images/stories/projections_pdfs/UKCP09_Projections_V2.pdf

Pope VD, Gallani ML, Rowntree PR, Stratton RA (2000) The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3. Clim Dyn 16:123–146

Pukelsheim F (1994) The three sigma rule. Am Stat 48:88–91

Rhines PB, Häkkinen S (2003) Is the oceanic heat transport in the North Atlantic irrelevant to the climate in Europe? ASOF Newsl 13–17

Rice JA (1995) Mathematical statistics and data analysis, 2nd edn. Duxbury Press, Wadsworth Publishing Company, Belmont, California

Rougier JC (2007) Probabilistic inference for future climate using an ensemble of climate model evaluations. Clim Change 81:247–264

Rougier JC, Sexton DMH, Murphy JM, Stainforth D (2009) Emulating the sensitivity of the HadSM3 climate model using ensembles from different but related experiments. J Clim 22:3540–3557

Rougier JC, Goldstein M, House L (2012) Second-order exchangeability analysis for multi-model ensembles. J Am Stat Assoc (to appear)

Rowlands DJ, Frame DJ, Ackerley D, Aina T, Booth BBB, Christensen C, Collins M, Faull N, Forest CE, Grandey BS, Gryspeerdt E, Highwood EJ, Ingram WJ, Knight S, Lopez A, Massey N, McNamara F, Meinshausen N, Piani C, Rosier SM, Sanderson BJ, Smith LA, Stone DA, Thurston M, Yamazaki K, Yamazaki YH, Allen MR (2012) Broad range of 2050 warming from an observationally constrained large climate model ensemble. Nat Geosci, published online. doi:10.1038/NGEO1430

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. Stat Sci 4:409–435

Sanderson BM, Shell KM, Ingram W (2010) Climate feedbacks determined using radiative kernels in a multi-thousand member ensemble of AOGCMs. Clim Dyn 35:1219–1236

Santner TJ, Williams BJ, Notz WI (2003) The design and analysis of computer experiments. Springer, New York

Sexton DMH, Murphy JM, Collins M (2011) Multivariate probabilistic projections using imperfect climate models part 1: outline of methodology. Clim Dyn. doi:10.1007/s00382-011-1208-9

Solomon S, Qin D, Manning M, Chen Z, Marquis M, Averyt KB, Tignor M, Miller HL (eds) (2007) Contribution of working group I to the fourth assessment report of the intergovernmental panel on climate change, 2007. Cambridge University Press, Cambridge

Stephens GL, Wild M, Stackhouse Jr PW, L'Ecuyer T, Kato S, Henderson DS (2012) The global character of the flux of downward longwave radiation. J Clim 25:2329–2340

Trenberth KE, Fasullo JT, Kiehl J (2009) Earth's global energy budget. Bull Am Meteorol Soc 90:311–323. doi:10.1175/2008BAMS2634.1

Vernon I, Goldstein M, Bower RG (2010) Galaxy formation: a Bayesian uncertainty analysis. Bayesian Anal 5(4):619–846, with Discussion

Whittle P (1992) Probability via expectation, 3rd edn. Springer texts in statistics. Springer, New York

Williamson D, Goldstein M, Blaker A (2012) Fast linked analyses for scenario based hierarchies. J R Stat Soc Ser C 61(5):665–692

Williamson D, Blaker AT (2013) Evolving Bayesian emulators for structured chaotic time series, with application to large climate models. SIAM J Uncertain Quantification (resubmitted)

Zaglauer S (2012) The evolutionary algorithm SAMOA with use of design of experiments. In: Proceeding GECCO companion '12. ACM, New York, pp 637–638