Design of coupled atmosphere/ ocean mixed-layer model experi-ments for probabilistic prediction

Hadley Centre technical note 45

David M. H. Sexton and James Murphy

November 2003

Met Office

Hadley Centre Technical Note

Design of coupled atmosphere/ ocean mixed-layer model experiments for probabilistic prediction

David M. H. Sexton and James Murphy

Revision History by Author/s

Author	Revision date	Summary of Changes
David Sexton	16/06/03	First draft
David Sexton	30/07/03	Response to James Murphy's comments, in particular conclusions revised.

Approvals

This document requires the following approvals from the relevant activity manager and contract manager:

Name	Title	Date of Issue	Version
James Murphy	Head Climate Prediction	30/07/03	2

Distribution

This document has been distributed to:

Name	Title	Date of Issue	Version
James Murphy	Head Climate Prediction	16/06/03	1

Design of coupled atmosphere/ ocean mixed-layer model experiments for probabilistic prediction

David M. H. Sexton and James Murphy

Hadley Centre for Climate Prediction and Research, Met Office, Exeter, UK

Corresponding author address: David M. H. Sexton, Hadley Centre for Climate Prediction and Research, Met Office, Fitzroy Road, Exeter, EX1 3PB, UK;
Email: David.Sexton@metoffice.com

Abstract

Probabilistic predictions that account for uncertainty in model physics are sensitive to how the experimental design samples parameter space. Here, two experimental designs are proposed that aim to reduce this sensitivity. The first method is based on a metric which quantifies how realistically a climate model simulates present-day mean climate; this method can easily be modified to provide a strategy for tuning models to particular climates. The second method uses experimental design theory to generate a means of sampling parameter space as efficiently as possible given a limited number of model runs.

Despite these efforts to reduce the effect of sampling of parameter space on probability predictions, this effect cannot be removed completely just by designing the experiment in a suitable way. A method is developed for removing the dependency on the sampling strategy, although this is done at the expense of making an assumption on how the climate system responds to a combination of parameter changes.

1. Introduction

To reliably assess the risks associated with future climate change, it is essential that policy-makers and climate impact scientists have a comprehensive assessment of the uncertainties involved in model predictions of climate change. These uncertainties arise from three sources: uncertainties in projections of the emissions of greenhouse gases and chemicals that produce aerosols, natural climate variability, and the way the climate model represents the climate system. The uncertainty inherent in the climate model arises partly because we do not fully understand climate processes, and partly because the parameters in the climate model, which control the key physical and dynamical processes, are not precisely known or are not measurable in the real world. Modelling uncertainties can arise from atmospheric, oceanic or cryospheric physics or from chemical or ecosystem processes. Climate model experiments have been used to assess the first two uncertainties whereby members of an ensemble differ from each other either in their initial conditions or the emissions scenario (Johns et al. 1997). However, due to limited computer resources, the study of uncertainties in the climate model itself has only recently begun. For instance, the atmospheric component of this uncertainty has only recently started to be systematically explored by the Hadley Centre's Quantifying Uncertainty in Model Predictions (QUMP) project and climate*prediction*.net (Allen 1999; Stainforth et al. 2002). Prior to these two studies, climate projections from several climate models were pooled to produce a so-called 'ensemble of opportunity', which was used to provide an estimate of the uncertainty of future climate change due to modelling errors (Cubasch et al. 2001). However, these ensembles are difficult to interpret, as all models are treated equally and their relative ability to model the climate system is not taken into account.

The long-term aim of QUMP is to provide probabilistic predictions for the 21st century accounting for all these uncertainties. Initially, the QUMP project is focusing on the effect of uncertainties in atmospheric physics on the equilibrium response to doubled CO_2. This is done by running an ensemble of 'slab' models (atmosphere GCM coupled to a 50m ocean mixed-layer) for 1x and 2x pre-industrial CO_2 levels. At present, each ensemble member differs from the Hadley Centre's current standard slab model, HadSM3, by perturbing one of 29 individual parameters to an extreme of their plausible range as specified by experts. The ultimate goal is to run several transient experiments that correspond to a selection of the slab model integrations and use this extra information to provide probabilistic predictions for any time in the 21st century.

Having completed a so-called *physics ensemble* it is straightforward to produce a frequency distribution of the response of global mean temperature to a doubling of CO_2 levels, otherwise known as *climate sensitivity*. Recognising the problems associated with 'ensembles of opportunity' described above, we have defined a Climate Prediction Index (CPI) which quantifies the reliability of climate change predictions according to how well each integration reproduces several aspects of the recent observed mean climate. The CPI can be used to weight the relative contribution of each ensemble member to the frequency distributions.

During the analysis of this first QUMP ensemble it has become apparent that further experiments are needed. For instance, some parameter perturbations do not significantly alter climate sensitivity and so the corresponding ensemble members simply resample the uncertainty of HadSM3, which arises from natural climate variability; this implies that our preliminary frequency distributions are biased towards the standard model. Consequently, the main requirement for the next ensemble of slab models is that it spans parameter space more effectively. To do this, we need to increase the QUMP ensemble to include runs where several parameters are changed simultaneously. The purpose of this technical note is to describe two experimental designs that address this requirement.

In section 2, we review the climate prediction index and outline the deficiencies associated with the frequency distribution of climate sensitivity estimated from the first QUMP ensemble. Section 3 describes another particular advantage of the CPI in that it is possible to reliably predict the CPI for untried combinations of parameters. In section 4 we outline the two strategies for selecting combinations of parameter perturbations, both of which will be run in the near future. One strategy is based on selecting combinations of parameter values that are likely to produce reliable simulations of the present day climate and consequently is of use for those readers who are interested in tuning climate models. In section 5, we present a method for making unbiased estimates from the first QUMP ensemble of the probability distribution of climate change due to doubling CO2 levels, and discuss how the experimental designs described in section 4 may improve our estimates. In section 6, we conclude by discussing the various advantages of the two new experimental designs for the problem of probabilistic prediction.

2. Climate Prediction Index (CPI) and estimating frequency distributions

The CPI used in this study measures how well a climate model reproduces various aspects of the climate system such as atmospheric radiation and clouds, atmospheric dynamics, the

hydrological cycle and surface fluxes (see Table 1 for a list of the climate variables used). For March-May, June-August, September-November and December-February twenty-year modelled seasonal means, each variable is compared against an appropriate observational or reanalysis data set over a region where the data is considered to be reliable. A normalised version of an area-weighted root mean square error (RMSE) (see Eqn. 1)is used because it penalises bias, differences in the spatial variances of the observed and modelled means, and poor pattern correlations. The components of the CPI for each season (j=1,…4) and kth climate variable are defined as

$$CPI_{jk} = -\sqrt{\frac{1}{\sigma^2_{ANN}} MSE}\ , \text{ where } MSE = \frac{1}{n}\sum_{i=1}^{n} w_i (m_i - o_i)^2\ , \qquad \text{Eqn. 1}$$

where m_i and o_i are the modelled and observed data, n is the number of grid points or latitude bands, σ^2_{ANN} is the spatial average of the *modelled* interannual variance used to normalise each component of the CPI; ideally we would also like to include observational estimates of interannual variance but it is not possible for most variables in the CPI as the data sets are not long enough or annual data is not available. The normalisation of *MSE* not only prohibits climate variables with large variance dominating the index but also allows us to include different types of components in future versions. Table 1 also describes which regions of the globe are used for each variable and whether the data is at grid-point, for zonal-means or for latitude-height zonal-mean cross-sections; w_i is the area-weight for latitude-longitude grid-point and zonal-mean data and the area- and mass-weight for zonal-mean and height data. The overall CPI is a weighted average of the CPI_{jk}, where the weights for the various components are shown in Table 1. Currently the ISCCP cloud diagnostics are weighted by 1/3 to reflect the interdependence of the high, medium, and low cloud amounts for each optical thickness. The other components are all given equal weighting of 1, since we currently have no basis for assigning unequal weights for any variables other than the cloud diagnostics.

Based on the CPI, the standard HadSM3 run lies 22nd out of the 53 ensemble members, although only three parameter perturbations show improvements more than 5%. This is very good considering the atmospheric physics in HadSM3 was tuned so that the coupled model could be run without flux corrections as well as on the quality of the simulation of the mean climate. The integration where the fallout speed of ice particles has been halved is the top ranking experiment mainly due to improvements in cloud amounts and the LW radiation budget (see Fig. 1). A few variables dominate the errors e.g. high-top, optically thin cloud

Table 1. Details of components of climate prediction index.

Climate variable	Source	Region used	Type of data used	Weight
1.5m temperature (^oC)	CRU[1]	Land only	Grid-point	1
MSLP (hPA)	ERA[2]	Globe	Grid-point	1
Precipitation (mm/day)	Xie-Arkin [3]	Ocean between 30^{o}S and 30^{o}N and all land	Grid-point	1
Westerly wind (ms^{-1})	ERA	Globe	Lat-height zonal-mean	1
Temperature (^oC)	ERA	Globe	Lat-height zonal-mean	1
Relative humidity (%)	ERA	Globe	Lat-height zonal-mean	1
Outgoing LW radiation at TOA (Wm^{-2})	ERBE[4]	Between 60^{o}S and 60^{o}N	Zonal mean	1
Outgoing SW ratiation at TOA (Wm^{-2})	ERBE	Between 60^{o}S and 60^{o}N	Zonal mean	1
SW cloud forcing (Wm^{-2})	ERBE	Between 60^{o}S and 60^{o}N	Zonal mean	1
LW cloud forcing (Wm^{-2})	ERBE	Between 60^{o}S and 60^{o}N	Zonal mean	1
High-top optically thick cloud (%)	ISCCP[5]	Ocean between 50^{o}S and 50^{o}N and all land	Grid-point	1/3
High-top medium optical thickness cloud (%)	ISCCP	Ocean between 50^{o}S and 50^{o}N and all land	Grid-point	1/3
High-top optically thin cloud (%)	ISCCP	Ocean between 50^{o}S and 50^{o}N and all land	Grid-point	1/3
Medium-top optically thick cloud (%)	ISCCP	Ocean between 50^{o}S and 50^{o}N and all land	Grid-point	1/3
Medium-top medium optical thickness cloud (%)	ISCCP	Ocean between 50^{o}S and 50^{o}N and all land	Grid-point	1/3
Medium-top optically thin cloud (%)	ISCCP	Ocean between 50^{o}S and 50^{o}N and all land	Grid-point	1/3
Low-top optically thick cloud (%)	ISCCP	Ocean between 50^{o}S and 50^{o}N and all land	Grid-point	1/3
Low-top medium optical thickness cloud (%)	ISCCP	Ocean between 50^{o}S and 50^{o}N and all land	Grid-point	1/3
Low-top optically thin cloud (%)	ISCCP	Ocean above 40^{o}S	Grid-point	1/3
Net downward SW flux at surface (Wm^{-2})	SOC[6]	Ocean above 40^{o}S	Zonal mean	1
Net downward LW flux at surface (Wm^{-2})	SOC	Ocean above 40^{o}S	Zonal mean	1
Sensible heat flux (Wm^{-2})	SOC	Ocean above 40^{o}S	Zonal mean	1
Latent heat flux (Wm^{-2})	SOC	Ocean above 40^{o}S	Zonal mean	1
Diurnal temperature range (^oC)	CRU	Globe	Grid-point	1
250hPa velocity potential	ERA	Globe	Grid-point	1
500hPa streamfunction	ERA	Globe	Grid-point	1
Meridional streamfunction	ERA	Globe	Lat-height zonal-mean	1
500hPa transient eddy kinetic energy	ERA	Globe	Grid-point	1
Total runoff efficiency rate (%)	GRDC[7]/CRU	Land points	Grid-point	1
Sea-ice extent	HadISST1[8]	NOAA sea-ice regions	Grid-point	1
Specific humidity	ERA	Globe	Lat-height zonal-mean	1

amounts. This may indeed be due to large model biases. However, it may also be that the normalisation factor $\frac{1}{\sigma^2_{ANN}}$ is too large, as the model significantly underestimates the observed variance.

[1] (New et al. 1999)
[2] (Gibson et al. 1997)
[3] (Xie and Arkin 1998)
[4] (Harrison et al. 1990)
[5] (Rossow and Schiffer 1991; Doutriaux-Boucher and Seze 1998)
[6] (Josey et al. 1996)
[7] (Fekete et al. 2002)
[8] (Rayner et al. 2003)

Several improvements could be made to the CPI in its current state. First, the normalisation factor needs more investigation. A further point is that the present index does not account for interdependence within or between the various components. It is also possible to define other diagnostics to evaluate climate processes which might be seen as being more relevant to future climate change than the mean climate of the model. For instance, we would like to check that the climate model is producing a realistic mean climate because it includes the correct climate processes. One way to do this is to evaluate the local relationship between two or more climate variables independently of geographical position e.g. cloud amounts with local SST and vertical velocity (Williams et al. 2003). We should also evaluate the climate model's variability when the observational data is of sufficient quality and the record is long enough. So the CPI presented here is clearly a first stage in a long development process.

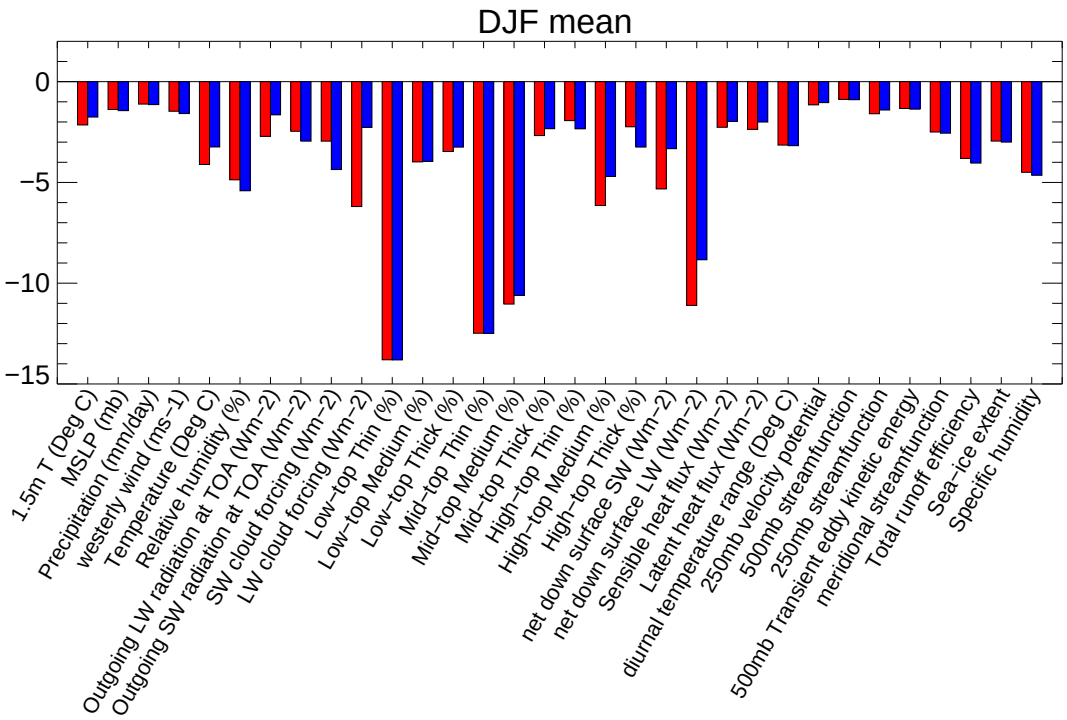

Figure 1. *Components of CPI for the most skilful perturbation run (red) and the standard HadSM3 integration (blue) for December-February. Labels on the x-axis indicate the climate variables used in the CPI. Blue columns which are shorter (longer) than the adjacent red columns indicate where the parameter perturbation is better (worse) than HadSM3.*

However, we feel that due to the extensive number of climate variables in the present CPI, that this provides a robust measure of the model's ability to simulate present-day climate. Tests using other validation statistics like Arcsin Mielke (Watterson 1996) and a version of

the RMSE that allows for the interdependence of data within a particular component show that the conclusions based on this more simple RMSE are robust. Overall, an adequate evaluation of the mean climate provides a useful constraint for climate predictions, though as mentioned above, we believe that this is only a *necessary but not sufficient* condition.

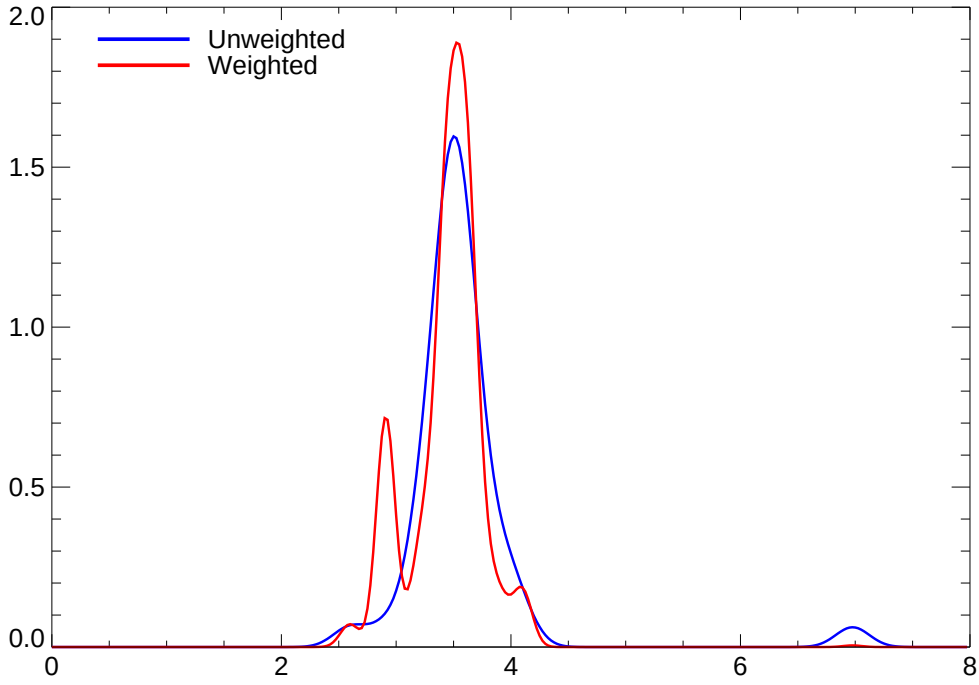

Figure 2. *The unweighted (blue) and weighted (red) frequency distributions of climate sensitivity estimated from the ensemble members. A Gaussian kernel has been used to smooth the distributions.*

One use of the CPI is to weight the frequency distribution towards the more reliable climate model versions. We use $\exp(-CPI^2)$ to estimate the weighting for each ensemble member for the estimation of the PDF of climate sensitivity. This weighting is effectively the likelihood that the observed and modelled data come from the same probability distribution, averaged over all components of the CPI and over all grid-box values. Therefore, although other forms of weighting function may be possible, $\exp(-CPI^2)$ seems a natural choice when estimating probability distributions. The effect of the weighting is to heavily weight down the outlier with high climate sensitivity and increase the probability of climate sensitivities in the range 2.5°C to 3°C and around 3.5°C (see Fig. 2).

However, this histogram is biased by the experimental design of the first ensemble in that it places too much emphasis on the standard model version, HadSM3. This happens for two reasons. Firstly, there should *a priori* be no preferred standard model version in the physics ensemble. This implies that the histogram is under-dispersive because a different choice of standard model might alter the position and shape of the distribution. Secondly, there is a narrow peak centred on HadSM3. This narrowing of the central peak about the HadSM3 climate sensitivity may be misleading, and does not necessarily mean that we have constrained climate sensitivity as accurately as is implied by the distribution. This is because the physics ensemble might include several parameters that do not affect climate sensitivity (they may still be important for regional changes of some climate variables), whatever the choice of standard model. Such parameters should not affect the final shape of the distribution but they affect the estimate here because of the sampling strategy in the first ensemble. That is, runs in which these parameters are perturbed will effectively resample the standard model version again and again. Consequently, this could produce a pronounced peak about the standard model version which gives a false impression of how well we have constrained climate sensitivity.

Therefore, there is a clear requirement to re-assess how the ensemble should sample parameter space. The key aims for designing this experiment are to run another ensemble of integrations that a) span parameter space as much as possible and b) are likely to be 'good' simulators of the present-day climate so that no money is wasted on running poor climate models that will be heavily weighted down in the estimation of the PDF. The first aim implies that we need to perturb several parameters in each ensemble member. The second aim requires the CPI of the current ensemble to predict a CPI for untried combinations of parameter changes, which is the subject of the next section.

3. Prediction of Climate Prediction Index

The CPI in its current form has one major advantage over many other skill scores. That is, we can reliably predict the CPI for an untried combination of parameter values by making a simple assumption that the response to several parameter changes is a linear combination of the responses to the individual parameter changes. Later sections use this result extensively to design the next ensemble of QUMP model integrations (see section 4a) and estimate a probability density function (PDF) rather than a frequency distribution (see section 5). In the next

subsection we outline the theory behind this claim (this can be skipped by the reader if they so wish). In section 3b, we test the predictions to show that they are indeed reliable.

a. Linear prediction of CPI

The prediction of CPI is based on the experimental design of the first QUMP ensemble where we make perturbations to single parameters from a control experiment e.g. HadSM3. First, we consider one component of the CPI. For the climate variable which corresponds to this component of the index, we let $\mathbf{S}_{\mathbf{sig}}^{\mathbf{0}}$ be the population mean of the control experiment and $\mathbf{S}_{\mathbf{sig}}^{\mathbf{i}}$ be the population mean of ith member of the physics ensemble. We run model versions for each parameter perturbation in the ensemble to estimate the climate variable, $\mathbf{S}^{\mathbf{i}}$, so that

$$\mathbf{S}^{\mathbf{i}} = \mathbf{S}_{\mathbf{sig}}^{\mathbf{i}} + \mathbf{N}^{\mathbf{i}} \qquad \text{Eqn. 2}$$

where $\mathbf{N}^{\mathbf{i}}$ is the noise component which would tend to zero as the length of the ensemble members tended towards infinity. Let

$$\mathbf{X}^{\mathbf{i}} = \mathbf{S}^{\mathbf{i}} - \mathbf{S}^{\mathbf{0}} \qquad \text{Eqn. 3}$$

represent the change in the climate variable due to the single parameter perturbation made in the ith ensemble member.

The goal here is to predict the CPI component for an untried combination of the parameter changes based on what we know from the single parameter perturbation ensemble. We assume that the population climate mean for an untried combination of parameter values, $\mathbf{P}_{\mathbf{sig}}$, can be written as the linear combination of the the signals from the individual parameter changes

$$\mathbf{P}_{\mathbf{sig}} = \mathbf{S}_{\mathbf{sig}}^{\mathbf{0}} + \sum_i \alpha_i \mathbf{X}_{\mathbf{sig}}^{\mathbf{i}}, \qquad \text{Eqn. 4}$$

where α_i are coefficients which can be between 0 and 1 inclusive and $\mathbf{X}_{\mathbf{sig}}^{\mathbf{i}} = \mathbf{S}_{\mathbf{sig}}^{\mathbf{i}} - \mathbf{S}_{\mathbf{sig}}^{\mathbf{0}}$. For example, we may want to estimate the CPI component for a physics parameter value of 1 when the value in the control run is zero and we have run a perturbation experiment for a value of 2; in this case, the coefficient α_1 would be 0.5.

If we actually ran a model with this combination of parameter values, we would estimate its climate mean for the climate variable of interest to be

$$\mathbf{P} = \mathbf{P}_{\mathbf{sig}} + \mathbf{N}^{\mathbf{p}} = \mathbf{S}_{\mathbf{sig}}^{\mathbf{0}} + \sum_i \alpha_i \mathbf{X}_{\mathbf{sig}}^{\mathbf{i}} + \mathbf{N}^{\mathbf{p}} \qquad \text{Eqn. 5}$$

The component of the CPI is the root mean square error of the modelled mean for a particular variable compared against a corresponding observational mean, $\mathbf{O}$. We denote a spatial average of a field by an overbar and from now on assume that the noise variance is independent of the parameter perturbations so that $\overline{\mathbf{N^{0^2}}} = \overline{\mathbf{N^{i^2}}} = \overline{\mathbf{N^{p^2}}} = \overline{\mathbf{N}^2}$.

If we denote the model bias of $\mathbf{S}$ and $\mathbf{P}$ as $\mathbf{e_0} = \mathbf{S} - \mathbf{O}$ and $\mathbf{e_P} = \mathbf{P} - \mathbf{O}$ respectively then the mean square error of $\mathbf{P}$ is then

$$\begin{aligned}
\overline{\mathbf{e}_\mathbf{p}^2} &= \overline{(\mathbf{P}-\mathbf{O})^2} \\
&= \overline{(\mathbf{S}-\mathbf{O}+\sum_i \alpha_i \mathbf{X^i_{sig}} + \mathbf{N^p} - \mathbf{N^0})^2} \\
&= \overline{\mathbf{e}_\mathbf{s}^2} + \overline{\mathbf{N^{p^2}}} + \overline{\mathbf{N^{0^2}}} - 2\overline{\mathbf{e_s}.\mathbf{N^0}} + \overline{(\sum_i \alpha_i \mathbf{X^i_{sig}})^2} + 2\sum_i \alpha_i \overline{\mathbf{e_s}.\mathbf{X^i_{sig}}} \\
&= \overline{\mathbf{e}_\mathbf{s}^2} + 2\overline{\mathbf{N}^2} - 2\overline{\mathbf{N^{0^2}}} + \overline{(\sum_i \alpha_i \mathbf{X^i_{sig}})^2} + 2\sum_i \alpha_i \overline{\mathbf{e_s}.\mathbf{X^i_{sig}}} \\
&= \overline{\mathbf{e}_\mathbf{s}^2} + \sum_i \alpha_i^2 \overline{\mathbf{X^i_{sig}}^2} + 2\sum_i \sum_j \alpha_i \alpha_j \overline{\mathbf{X^i_{sig}}.\mathbf{X^j_{sig}}} + 2\sum_i \alpha_i \overline{\mathbf{e_s}.\mathbf{X^i_{sig}}}
\end{aligned} \quad . \qquad \text{Eqn. 6}$$

Eqn.6 needs to be expressed in terms of $\mathbf{X^i}$'s and $\mathbf{X^j}$'s, as $\mathbf{X^i_{sig}}$ and $\mathbf{X^j_{sig}}$ are not measured directly. Combining Eqns. 2 and 3 gives

$$\mathbf{X^i} = \mathbf{X^i_{sig}} + \mathbf{N^i} - \mathbf{N^0}, \qquad \text{Eqn. 7}$$

so that

$$\begin{aligned}
\overline{\mathbf{X^i}\mathbf{X^j}} &= \overline{(\mathbf{X^i_{sig}} + \mathbf{N^i} - \mathbf{N^0}).(\mathbf{X^j_{sig}} + \mathbf{N^j} - \mathbf{N^0})} \\
&= \begin{cases} \overline{\mathbf{X^i_{sig}}.\mathbf{X^j_{sig}}} + \overline{\mathbf{N^{0^2}}} & \text{when i} \neq \text{j} \\ \overline{\mathbf{X^i_{sig}}^2} + \overline{\mathbf{N^{0^2}}} + \overline{\mathbf{N^{i^2}}} & \text{when i = j} \end{cases} .
\end{aligned} \qquad \text{Eqn. 8}$$

When $i = j$, $\overline{\mathbf{X^i_{sig}}^2} > 0$ must hold. Therefore, if $\overline{\mathbf{X^i}\mathbf{X^j}} < 2\overline{\mathbf{N}^2}$ when $i = j$ then we set

$$\overline{\mathbf{X^{i^2}}} = 2\overline{\mathbf{N}^2}. \qquad \text{Eqn. 9}$$

Using Eqns. 8 and 9, Eqn. 7 can be rewritten as

$$\overline{\mathbf{e}_\mathbf{p}^2} = \overline{\mathbf{e}_\mathbf{s}^2} + \sum_i \alpha_i^2 (\overline{\mathbf{X^{i^2}}} - 2\overline{\mathbf{N}^2}) + 2\sum_i \sum_j \alpha_i \alpha_j (\overline{\mathbf{X^i}.\mathbf{X^j}} - \overline{\mathbf{N}^2}) + 2\sum_i \alpha_i (\overline{\mathbf{e_s}.\mathbf{X^i}} + \overline{\mathbf{N}^2}), \qquad \text{Eqn. 10}$$

to give the final form for our prediction of CPI. To predict the full CPI, each component needs to be predicted, normalised as with the actual CPI and then summed with the relevant weights for each component.

b. Tests of prediction of CPI

To test the prediction of the CPI, we have run several test cases which were originally designed to try out various hypotheses about the effects of perturbing several parameters at once and were not designed to be cases where the prediction of the CPI worked well. The first two runs combine parameter changes that have already been tried individually in the first ensemble. The third run was designed to produce a model with low climate sensivity. The fourth run was predicted to produce a better present-day climate than HadSM3, as measured by an earlier version of the CPI (essentially no components based on ISCCP). The fifth run was an attempt to sample the interior of parameter space, changing parameters to halfway between the values used in the ensemble of single parameter perturbations. The sixth was a run predicted to produce a reasonably good climate.

Table 2. *Comparison of predicted and actual CPI for test runs with several parameter changes from HadSM3.*

Parameter perturbations	**Actual CPI**	**Predicted CPI**
vf1=2, ct=0.0004, rhcrit=0.9, cwland=0.002, cwsea=0.0005	-4.436	-4.544
vf1=0.5, rhcrit=0.6, cwland=0.0001, cwsea=2e-5, minsia=0.65, ice_tr=2, cape=2	-3.614	-3.705
ct=5e-5, cwland=0.002, cwsea=0.0005, ent=9, cape=1, eacfbl=0.7, eacftrp=0.6	-6.918	-6.330
vf1=0.52, ct=0.000176, rhcrit=0.62, cwland=0.000171, cwsea=4.13e-5, minsia=0.636, ice_tr=2.525, ent=2.381, icesize=29.9	-3.451	-3.331
vf1=1.5, ct=7.5e-5, rhcrit=0.8, cwland=0.001, cwsea=0.00025, minsia=0.54, ice_tr=7.143, ent=1.8, icesize=33, cape=1.5	-4.776	-4.762
vf1=0.58239, ct=0.000276, rhcrit=0.80735, cwland=0.00108, cwsea=0.00027, minsia=0.54613, ice_tr=6.705, ent=2.38935, icesize=33.316, cape=1.95, g0=8.6302, charnock=0.0127, asymptotic_length_scale=0.18377, conv_rough_length=0.00324, dyndiff=6.539, eacfbl=0.51486, eacftrp=0.50743, k_gwd=14400, k_lee=216000	-3.783	-3.777

Table 2 shows that the prediction of CPI works very well and only the third example is moderately different; however, the predicted CPI for this third example certainly picks out that this run is not expected to produce a very good simulation of present-day climate as measured by the CPI. Encouraged by these results we use Eqn. 10 extensively in the design of the runs where several parameter values are changed at once (see section 4a) and in the unbiased estimation of the PDF of climate sensitivity (see section 5).

4. Design of multiple parameter perturbation runs

a. QUMP parameters

Before discussing the design of the multiple parameter perturbation runs, it is necessary to describe the nature of the physics parameters in QUMP as these affect the statistical techniques we can use for predicting the response of the model at untried combinations of parameters. There are three kinds of parameters in QUMP. The first kind are parameters that take values in a continuous range e.g. the fallout speed of ice particles, VF1, can take any value between $0.5 ms^{-1}$ and $2 ms^{-1}$, as specified by an expert in the model's large-scale cloud scheme. The second kind are parameters that take a finite number of values such as on/off switches which take values of 0 or 1. Each unique value of a factor is called a *level*. Forest roughness length, which is implemented by prescribing one of four vegetation ancillary files, is treated as a four-level parameter. We shall distinguish between these first two types of parameters by referring to the latter as *factors*. To complicate the issue especially from the point of view of having to design an efficient experiment, there is a third type of parameter in the QUMP experiment which we will call a *hybrid parameter*. These hybrid parameters e.g. CAPE closure time scale are like on/off factors but become continuous when they are on. There are three more hybrid parameters in QUMP. The anvil factor and convective updraught factor are only used when the convective anvil cloud scheme is switched on. Rhcrit must also be regarded as parameter, because it becomes redundant when the Rhcrit parameterisation scheme is switched on.

b. The ensemble of 'tuned' model versions

In section 2, we noted that the estimated frequency distribution was biased by the experimental design of the first QUMP ensemble. For instance, if a single parameter perturbation had no significant effect on the climate variable of interest then we would effectively be sam-

pling the response of HadSM3 again. Therefore, in the first QUMP ensemble the frequency distribution may be biased towards the response of the standard run. Even if an ensemble of this size was designed to be unbiased, it probably does not sample the parameter space of the climate model well and the frequency distribution might be considered unrepresentative. Therefore, there is a need to sample the parameter space more thoroughly. However, as we weight the contribution of the ensemble members to the frequency distribution, we want to avoid sampling areas of parameter space which provide relatively poor simulations of the present day climate. Here we describe an experimental design for an ensemble of runs where several parameters are perturbed simultaneously that samples parameter space as efficiently as possible given a finite number of runs that are all likely to provide good simulations of the present day climate. This design increases the chances that all ensemble members will contribute significantly to the frequency distribution.

For the second QUMP ensemble, the computer resources available to us are going to restrict the size of the ensemble to 100-150 members, some of which will be used to test the experimental design in section 4b. To illustrate the method, we will pick 50 combinations of parameter values but the procedure allows this to be easily extended. The procedure we use is a three-step process:

1. We use a Monte Carlo algorithm to sample parameter space. We assume *a priori* that the parameters are independent and each have a uniform probability distribution, so that it is equally likely to select any combination of these parameter values. For the first iteration, we use a uniform distribution to randomly select a value for each parameter within the range defined by the first QUMP ensemble. For this set of parameter values we predict the CPI using Eqn 10. As the number of parameters increases, this method can become very inefficient at locating areas of parameter space that are predicted to simulate the present day climate as realistically as the standard model. For 8 parameters, we found that 10% of randomly-generated runs were predicted to be better than HadSM3. For 20 parameters, only a few runs out of a million were predicted to be better than HadSM3. For the full 29 parameters, no randomly-generated runs were predicted to be better than HadSM3 and the procedure had to be repeated 3.6 million times to find several hundred runs that were predicted to be only slightly worse than HadSM3.

2. One way to improve on this situation might be to increase the number of iterations in the procedure described above but again this is very inefficient. To make the experimental design algorithm more efficient we used the fact that the randomly-generated runs were un-

unlikely to be locally optimal. Furthermore, it is very likely that there is a better model than the randomly-selected one in the small region of parameter space where the continuous parameters can change but the factors remain the same. In this second stage a "Downhill simplex" method (Press et al. 1992) is used to find a locally better solution for each of the top 500 combinations of parameter values. To restrict the optimal solution to a local region of parameter space, the downhill simplex algorithm searched the region of parameter space where the continuous parameters were allowed to change within 5% of the original parameter value but remain within the original parameter space of the first QUMP ensemble. As the algorithm is only guaranteed to find a better rather than optimal solution, we ran four iterations where the start-point of each iteration was the end-point of the previous iteration. After this second stage we have a set of 500 possible combinations of parameters that are likely to produce good simulations of present-day climate, if not better than HadSM3.

3. In the final stage we aim to select a subset of the 500 possible combinations of parameters, which we can afford to run on the computer. The main criterion for our algorithm to select this subset was that we spanned parameter space as efficiently as possible. We did this by first of all selecting the combination of parameters that was predicted to provide the best model. This combination of parameter values is the starting point for the set of models to run, $\Re$. The set of runs we chose for the final design is called E. At this stage, E only contains the run in $\Re$ that is predicted to be have the best CPI score. Then, in normalised parameter space where the values range from 0 to 1, we calculated the distance, D_j, between this first combination in $\Re$ and the other j=1,…499 combinations not in E using

$$D_j = \sum_{i \in E} \sum_{p \in P} \frac{(\alpha_{jp} - \alpha_{ip})^2}{6(\alpha_{jp} - 1/2)^2} \quad \text{for } j \notin \Re, \qquad \text{Eqn. 11}$$

where P is the set of parameters and the *pth* parameter value for the ith Monte Carlo run is α_{ip}. The weighting factor is designed to counteract the fact that parameter values at the edges are more likely to be further away. This is particularly important for factors which take values 0,1/2, and 1 because it avoids the experimental design being biased towards the 0 and 1 values. The next combination to be chosen to be included in $\Re$, was that which was furthest apart from the first combination, that is with the largest D_j. This combination was then added to the set $\Re$. Subsequent combinations were chosen to maximise the sum of the distance from the previous choices in set $\Re$. The algorithm has the advantage that if we wish to increase the

size of our experiment from N_1 to N_2 members, we guarantee that the first N_1 members of the second ensemble are the members of the first ensemble.

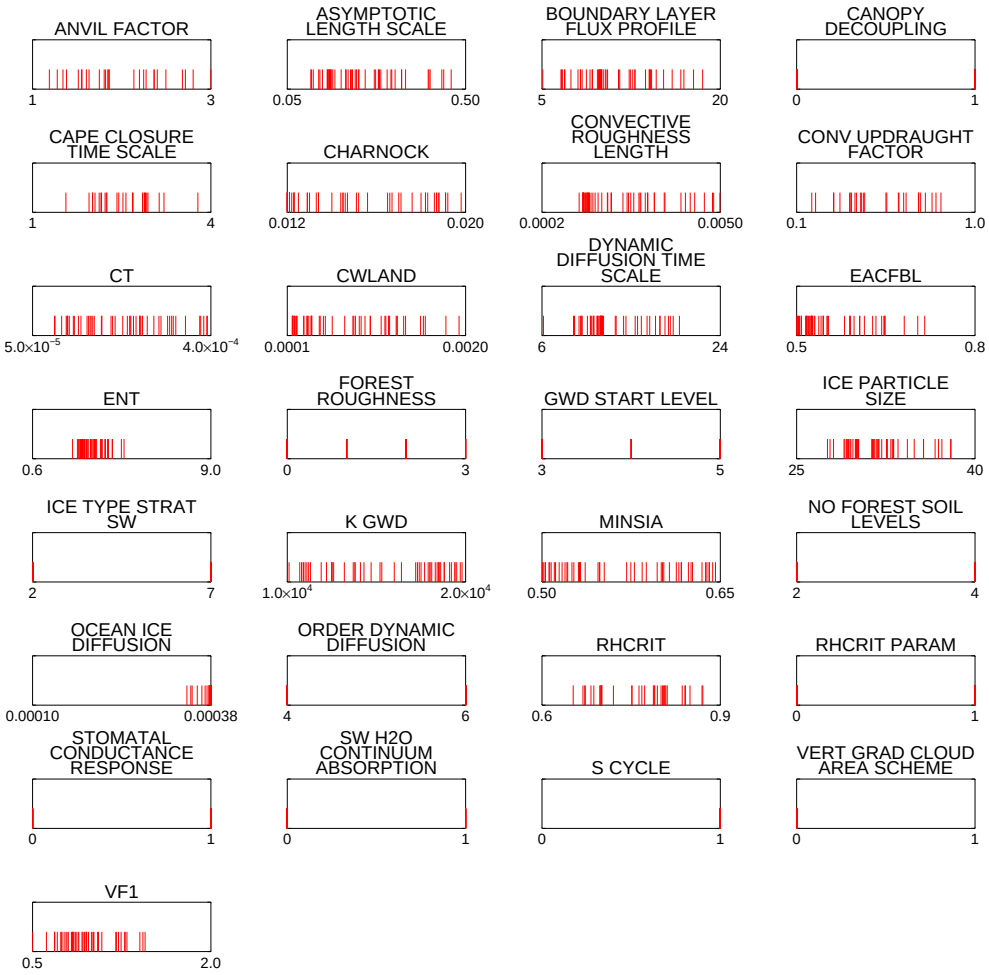

Figure 3. *Distribution of values for each parameter in the 50-member experiment. Red dashes indicate values of parameters in the new design. The sulphur cycle will be included in all runs so S-CYCLE is set to 1 only.*

Fig. 3 shows how the values chosen for each parameter. The algorithm generally selects values which span the full range for each parameter. There are four exceptions to this. For entrainment rate (ENT) and ocean-ice diffusion, the algorithm has restricted the range of values to avoid producing runs that are likely to produce low CPI scores. For the number of forest soil levels, the algorithm has rejected any runs where there were 3 forest soil levels. Fig. 4 shows that there were very few runs in the subset of 500 `good` runs that had the number of forest soil levels set to three. Finally, the vertical gradient cloud area scheme was not chosen for any members of this experimental design. Runs with HadAM3H, a closely-related varia-

tion of HadAM3, indicate that this scheme can interact with other schemes to improve the climate simulation. This illustrates a potential problem with the prediction of the CPI and is presumably because the linear assumption behind the CPI prediction cannot account for such beneficial nonlinear interactions between schemes.

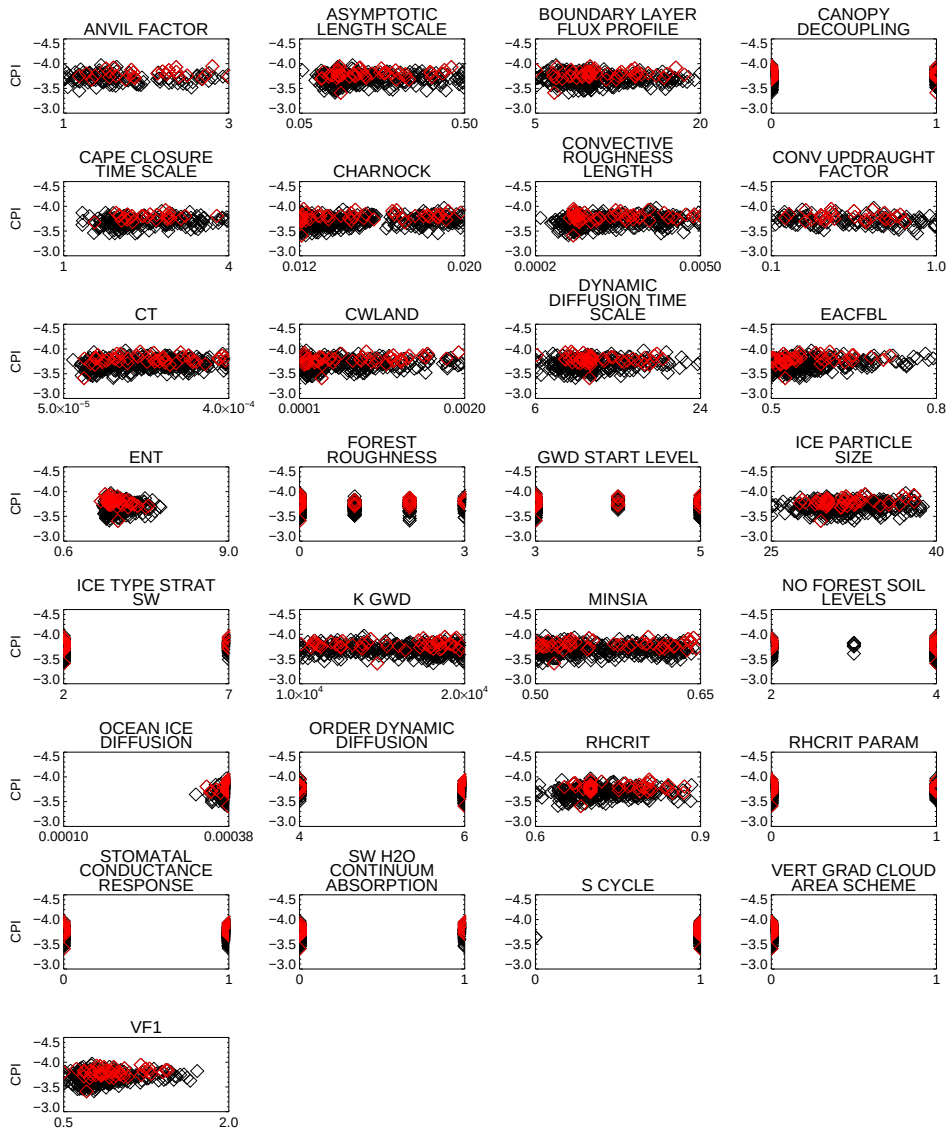

Figure 4. *Plot of CPI against parameter values for each parameter from the 500 `good` runs (black diamonds). The 50 runs chosen in the third part of the procedure are marked by red diamonds.*

Fig. 4 also shows that part three of the procedure did not select runs with necessarily the best predicted CPI scores. Indeed the run with the best predicted skill stands out from the

other selections. This may indicate that the better runs in the 500 `good` runs may have been relatively close together in parameter space and so were not selected. More of these runs would have been picked if the predicted CPI had been included in the cost function, D_j. However, this was not done, as the CPI prediction is not more accurate than the range of CPI scores covered by the top 500 runs; that is, the top 500 runs are all likely to be cost-effective.

The experimental design is dependent on the formulation of the CPI. Therefore, it is comforting to see that the experimental design algorithm can still cover most of parameter space. On the other hand, this indicates that there are many small, disparate regions of parameter space that are likely to provide relatively good simulations of the present-day climate. Whether this is a consequence of the way CPI is predicted or is indeed a real property of the model over its parameter space, can only be tested by running the ensemble of multiple parameter perturbations.

c. Alternative experimental design for the first stage

The complexities of the algorithm described in section 4a are necessary to solve the problem of how to sample the parameter space as efficiently and as cost-effectively as possible when the number of runs allowed is of the same order as the number of parameters. That design is suitable for efficiently estimating frequency distributions. The method is equally viable for larger ensembles but does rely on the availability of a single parameter perturbation ensemble as a first stage. However, as we increase the number of ensemble members that can be run, a number of alternative experimental designs become available to us. The experimental design outlined below would ideally be used when the size of the ensemble that we are allowed to run is about 10+ times the number of parameters. However, it can also be used when fewer ensemble members can be run. Then, this design provides an alternative to the single parameter perturbation ensemble as a first stage for the QUMP experiments but has two clear advantages. Firstly, the sampling yields an unbiased estimate of the frequency distribution. Secondly, it is possible to incorporate nonlinear interactions between parameters in the prediction of the response and CPI at untried combinations of parameter values. Therefore, this design is more suitable than our current first QUMP ensemble as a basis for a tuning procedure and as a design used to estimate unbiased probability distributions of the response to doubling CO_2 levels. The reasons for this last point are discussed below.

i. Estimation of response at untried parameter values

The statistical principles behind the estimation of the response of any climate variable at untried combinations of parameter values depends on the nature of parameters themselves, whether they are factors or parameters or hybrid parameters. Below we show that any prediction of the response at any combination of parameter and factor values has two components: the prediction based on the factor values plus a prediction interpolated from the error terms based on the parameters. First, we consider the prediction of the response due to changes in factor values.

It is necessary to estimate the response of a climate variable for each different level of a factor. These responses, often called *effects*, are usually estimated using a regression equation like Eqn. 12 where $\mathbf{y}$ is an N-element vector of the response for a particular climate variable from each of the N members in the ensemble. The effects of the factors at each level (a total of p effects, say) are estimated by p regression coefficients stored in the p-element vector $\boldsymbol{\beta}$. For example, if there three 2-level factors and one 5-level factor, p would be 3x2+5=13. $\mathbf{X}$ (a standard notation for regression not to be confused with $\mathbf{X}$ in section 3) has an N-element column for each of the p regression coefficients. Each element of $\mathbf{X}$, X_{ij}, stores a 0 or 1 depending on whether the corresponding level of factor to the ith regression parameter is included in the jth ensemble member.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \qquad \text{Eqn. 12}$$

It is also straightforward to consider additional effects by two or more factors interacting with each other. These additional effects are called *interactions*[9]. Each interaction can be included in the estimation by adding another regression coefficient to $\boldsymbol{\beta}$ and adding a corresponding column to $\mathbf{X}$. This extra column is calculated as the product of the columns in $\mathbf{X}$ of the main effects that contribute to the interaction.

The effect of each level of each factor, $\hat{\boldsymbol{\beta}}$, is estimated by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \qquad \text{Eqn. 13}$$

and is only possible if $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ is invertible. The standard error of each estimate is the measured by the square root of the diagonal elements of the matrix

[9] Interactions are named according to the number of factors involved so that they are called two-way interactions, three-way interactions etc.

$$\mathrm{var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y}\,. \qquad \text{Eqn. 14}$$

For estimation of the response of a climate variable to different values of a continuous parameter, traditional regression techniques can be used. However, this approach is relatively inflexible in the way it can deal with nonlinear interactions between two or more parameters. Sacks et al. (1989) have developed an interpolation technique (similar to a technique called *kriging*) for this purpose which can be used to estimate the response in the presence of parameters and factors. The statistical model is like Eqn. 12 but a smooth response surface is fitted to the error terms, $\boldsymbol{\varepsilon}$ at the same time the effects of the factors are estimated. The work behind this technique is a matter of statistically determining the smoothness of the response by fitting correlation functions for each continuous parameter. Sacks et al. (1989) discuss this is more detail. The technique is a type of interpolation because it guarantees that the prediction at combinations of parameter and factor values for which a model has already been run, will be identical to the response from the model runs. The technique can be modified to include uncertainty due to natural variability in each model run (e.g. Craig et al. 2001).

ii. The Latin hypercube and D-optimal design

An advantage of the Sacks et al. (1989) interpolation technique is that the response to changes in factors and in parameter values can be treated as independent. Consequently, the experimental design could be split into a design for the factors and a design for the continuous parameters. However, the presence of hybrid parameters precludes this. Therefore, we first describe how to design a experiment with no hybrid parameters and then adapt the design to cope with their inclusion.

One suitable technique for continuous parameters is the *Latin hypercube*, which has been often used in several scientific fields but only on a few occasions in climate studies (e.g. Bowman et al. 1993; Gough and Welch 1994). In a Latin hypercube experiment, we want to investigate P continuous parameters with a given number of model integrations, N where $N>=P+1$ to ensure that the estimate of the response to the different parameters can be uniquely determined. The number of model integrations determines how well each parameter is sampled because for each parameter, its range is split into N intervals which are typically evenly spaced[10]. The combinations of the N-member ensemble are then selected randomly. For the first member, one of the N intervals is randomly selected for each of the parameters

separately. For subsequent members and for each parameter, intervals are randomly selected from those which have not been used for that parameter in the previous ensemble members. In this way no single interval is selected more than once for each parameter and the ensemble is guaranteed to be sample every interval once for each parameter. Fig. 5 shows examples of good and bad Latin hypercube designs for two parameters using 10 ensemble members.

However, it is possible to generate bad Latin hypercube experiments by chance. For instance, the first ensemble member might sample the first bin of each parameter, the second member samples the second bin for each parameter and so on. From this Latin hypercube, it would be impossible to identify which parameter might be causing the different responses across the ensemble members as the values of each pair of parameters are perfectly correlated across the ensemble. This is an extreme and very unlikely example but it illustrates the point that to effectively identify which parameters are responsible for various aspects of the response, we require the parameter values to be as uncorrelated with each other as possible. Iman and Conover (1982) provide an algorithm which can be used to design a Latin hypercube experiment so that any desired level of correlation between the parameters is achieved.

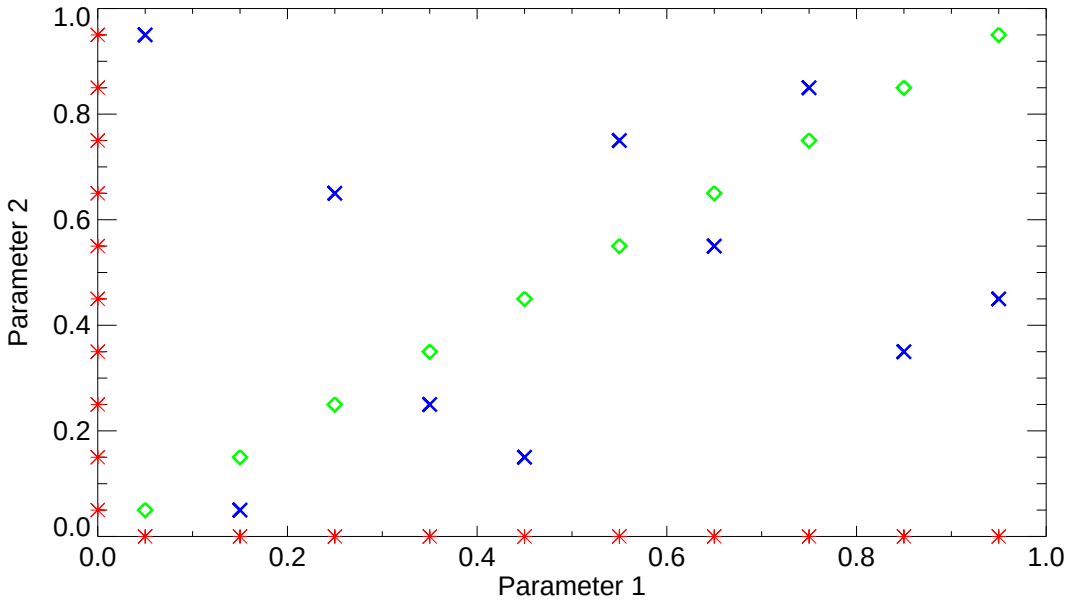

Figure 5. *Blue crosses indicate a Latin hypercube 10-member experiment for two parameters, where the two parameters are uncorrelated across the ensemble. Green diamonds indicate a perfectly correlated and not very good Latin hypercube. The red stars indicate where the two parameters have been sampled.*

---

[10] Sometimes the parameter values may be transformed e.g. logarithmically prior to the binning procedure.

The Latin hypercube is more efficient at spanning parameter space and very easy to determine. It is also a more efficient design for estimating the response of the climate variable of interest to doubled CO2 levels at untried combinations of parameter values. This is because there is always a model run that samples any parameter in the interval(s) adjacent to the one of interest.

For factors, as with the continuous parameters, we would ideally like our columns of **X** to be uncorrelated with each other. One solution is an experimental design where there is an ensemble member for each combination of factors, the so-called *full factorial design* (e.g. Fisher 1935). This experimental design allows us to not only estimate the main effects of each factor but also all additional interactions between two or more factors. In practice, we are limited to run fewer ensemble members than possible combinations of the factors and so at best we require that there is the least amount of correlation between each column of **X** as possible. One possible solution is to restrict the problem to only estimating main effects and/or interactions between pairs of factors, only requiring $N>=p$.

In some special cases there are several designs available such as fractionally factorial experiments, Plackett-Burman designs and Box-Behnken experiments (NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/index.htm, 10/06/03). However, in the QUMP ensemble we have nine 2-level factors (the on/off switches), two 3-level factors (number of soil levels and the start level for gravity wave drag) and one 4-level factor (forest roughness length). The most flexible method for designing the experiment for factors is to use a method called D-optimality (e.g. Pukelsheim 1993). D-optimality uses the principle that we chose $\mathbf{X}$ such that $\mathbf{X}^T\mathbf{X}$ has the maximum determinant $D=\det(\mathbf{X}^T\mathbf{X})$ to minimise the overall precision with which we estimate the effects of the factors (see Eqn. 13). This works because the volume of the confidence ellipsoid around any estimate is inversely proportional to $\det(\mathbf{X}^T\mathbf{X})$ and we require this volume to be as small as possible.

As the factors and parameters are treated independently, an efficient design could be a combination of a Latin hypercube for the continuous parameters (using the Iman and Conover algorithm) and a D-optimal design for the factors (see Matlab routine in Appendix A) to do this.

iii. Including hybrid parameters in the design

The inclusion of hybrid parameters in the experiment complicates the design because the factors and parameters are no longer independent. Standard regression techniques can be used to analyse what are sometimes termed as *incomplete treatment structures* (e.g. Mead 1990). However, we are not aware of any adaptations to the techniques of Sacks et al. (1989) to cope with hybrid parameters. Here we propose to use a factor and a parameter for the cape closure time scale and RHcrit parameters, and one factor and two parameters for the convective updraught factor and anvil factors. That is, three factors and four parameters in all. Therefore, when we use the RHcrit parameterisation scheme, that factor is set to 1, and we have to set the rhcrit parameter to a default value. Therefore, the factor does not measure the effect of the parameterisation scheme; instead it quantifies the difference the scheme makes compared to the effect of having RHcrit=0.7.

The experimental design depends on the available number of runs. There are 8 possible combinations of the three hybrid factors. If the number of runs available was greater than 10 times the number of parameters for each permutation of hybrid factors, it would be feasible to design a joint D-optimal and Latin hypercube, as described in the previous subsection, for each permutation of hybrid factors and combine these.

In QUMP we have already committed at least 50 runs to the ensemble of 'tuned' model versions. Therefore, we will run a small D-optimal and Latin hypercube experiment to test this design. This second ensemble will have 40 members. It has the advantage that it can be used to increase the size of the ensemble of 'tuned' model versions.

The first stage is to design the D-optimal experiment. First, there is a column in **X** to measure the baseline effect which is combined effect when each factor is set at level 1. The effect at every level of each factor other than the first is measured relative to the baseline effect and there is a column in **X** for each of these. Finbally, there are additional columns for the three two-way interactions between the hybrid parameters. A MATLAB® program (see Appendix) was used to design the following 40-member D-optimal experiment.

The second stage is to use Iman and Conover's algorithm to generate a Latin hypercube where pairs of parameters are as uncorrelated as possible. To determine the extent to which a Latin hypercube is uncorrelated, we calculate the determinant of its Spearman rank correlation coefficient matrix. As this measure approaches 1, the parameters in the Latin hypercube

become more uncorrelated with each other. After several thousand iterations of Iman and Conover's procedure, the highest determinant was ~0.84.

The final stage is to merge the two designs together bearing in mind that when the three hybrid parameters are off i.e. 0 in the D-optimal design, the corresponding elements of the Latin hypercube should be set to some default number, which were chosen to be the median value of 0.5. The effect of the Latin hypercube is to make the parameters in the design more correlated and therefore lower the determinant of the rank correlation matrix. The main work of this first stage is to randomly combine the D-optimal and Latin hypercube designs and re-iterate until a suitably large determinant is found. For the D-optimal design above we found the highest determinant was ~0.37, which reflects the inefficiency in the Latin hypercube due to the hybrid parameters.

5. Estimating probability density functions (PDFs)

The aims of the ensemble of 'tuned' model versions are to sample parameter space as efficiently as possible given a limited number of model runs. Despite these efforts, our estimated frequency distributions may still be regarded as being susceptible to sampling error, mainly because parameter space is so huge. To overcome this problem, we assume that the response can be predicted reasonably well using linear theory. Therefore, we generate a PDF that is conditional on the underlying structure and physics of the model, and the formulation of the CPI.

A PDF of climate sensitivity $p(\Delta T_{2\times}|\mathbf{O})$ constrained by some observational data $\mathbf{O}$ can be written as

$$p(\Delta T_{2\times}|\mathbf{O}) = \int_{\mathbf{x}\in\chi} p(\Delta T_{2\times}|\mathbf{x}).p(\mathbf{x}|\mathbf{O})d\mathbf{x} \qquad \text{Eqn. 15}$$

where χ is the parameter space, $\mathbf{x}$ is an element of parameter space. $p(\mathbf{x}|\mathbf{O})$ can be viewed as the relative likelihood that $\mathbf{x}$ is the set of parameter values that best model the observed present day climate $\mathbf{O}$ assuming that all combinations of parameter values $\mathbf{x}$ are *a priori* equally likely (Leroy 1998). $p(\mathbf{x}|\mathbf{O})$ was set to $\exp(-CPI^2)$. $p(\Delta T_{2\times}|\mathbf{x})$ is the probability that the climate sensitivity will be $\Delta T_{2\times}$ given a set of parameter values, $\mathbf{x}$, and is information that can be obtained from the model estimates of $\Delta T_{2\times}$ and its uncertainty by running the model. Clearly, the values of $p(\Delta T_{2\times}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{O})$ are known for values of $\mathbf{x}$ for which we

have run models. The key to estimating $p(\Delta T_{2\times})$ in Eqn. 15 is that we can predict the CPI and the response and hence $p(\Delta T_{2\times}|\mathbf{x})$ and $p(\mathbf{x}|\mathbf{O})$ at untested values of $\mathbf{x}$. From section 3a, we can already predict the CPI for untried combinations of parameter values. For $p(\Delta T_{2\times}|\mathbf{x})$, the experimental design of the first QUMP ensemble is not optimal for predicting $\Delta T_{2\times}$ for any $\mathbf{x}$ [11], but by making a linear assumption it is possible.

Climate sensitivity, $\Delta T_{2\times}$, is often written as

$$\Delta T_{2\times} = Q_{2x} / \lambda . \qquad \text{Eqn. 16}$$

where $Q_{2\times}$ is the radiative forcing due to doubling CO2 concentrations and λ is a feedback parameter.

For the ith ensemble member we estimate $\Delta\lambda_i$, the change in feedback parameter relative to λ for the standard model version. $\Delta\lambda_i$ is then set to zero if the climate sensitivity of the ith ensemble member is not significantly different at the 5% level to the control climate sensitivity estimated from the long 600-year HadSM3 run. This prevents the situation where there a large number of parameters which have small non-significant effects but these can combine linearly to produce a large erroneous predicted response.

For a given $\mathbf{x}$, we determine the α_i's as in section 3a, and estimate

$$\Delta T_{2\times} = \frac{Q_{2x}}{\sum_i \alpha_i \lambda_i} . \qquad \text{Eqn. 17}$$

Using the runs in Table 2, we find our predictions have an error with a standard deviation of about 0.4. A straightforward way to estimate the PDF would then be to run a Monte Carlo experiment to randomly sample the parameter space χ assuming each parameter was independent on the others. However, the method is very sensitive to the control $\Delta T_{2\times}$ which has a mean of 3.46°C with a standard deviation of 0.07°C, as estimated from a 600-year HadSM3 run. Therefore the method has been adapted in the following way to allow for this sensitivity to the control $\Delta T_{2\times}$.

[11] A Latin hypercube would provide predictions with the least uncertainty because it samples parameter space more efficiently for predictions at untried parameter combinations, and allows for nonlinear interactions between two or more parameters.

1. Run a Monte Carlo simulation to generate *N* combinations of parameter values and use Eqn. 10 to predict the CPI and $p(\mathbf{x}|\mathbf{O}) = \exp(-CPI^2)$.

2. Use the 600-year long HadSM3 to estimate the climate sensitivity of the standard run, ΔT_{SM3} and the standard deviation of natural variability, σ_{SM3}. Divide the standard climate sensitivity range $\Delta T_{SM3} - 2\sigma_{SM3}$ to $\Delta T_{SM3} + 2\sigma_{SM3}$ into *M* equally spaced intervals.

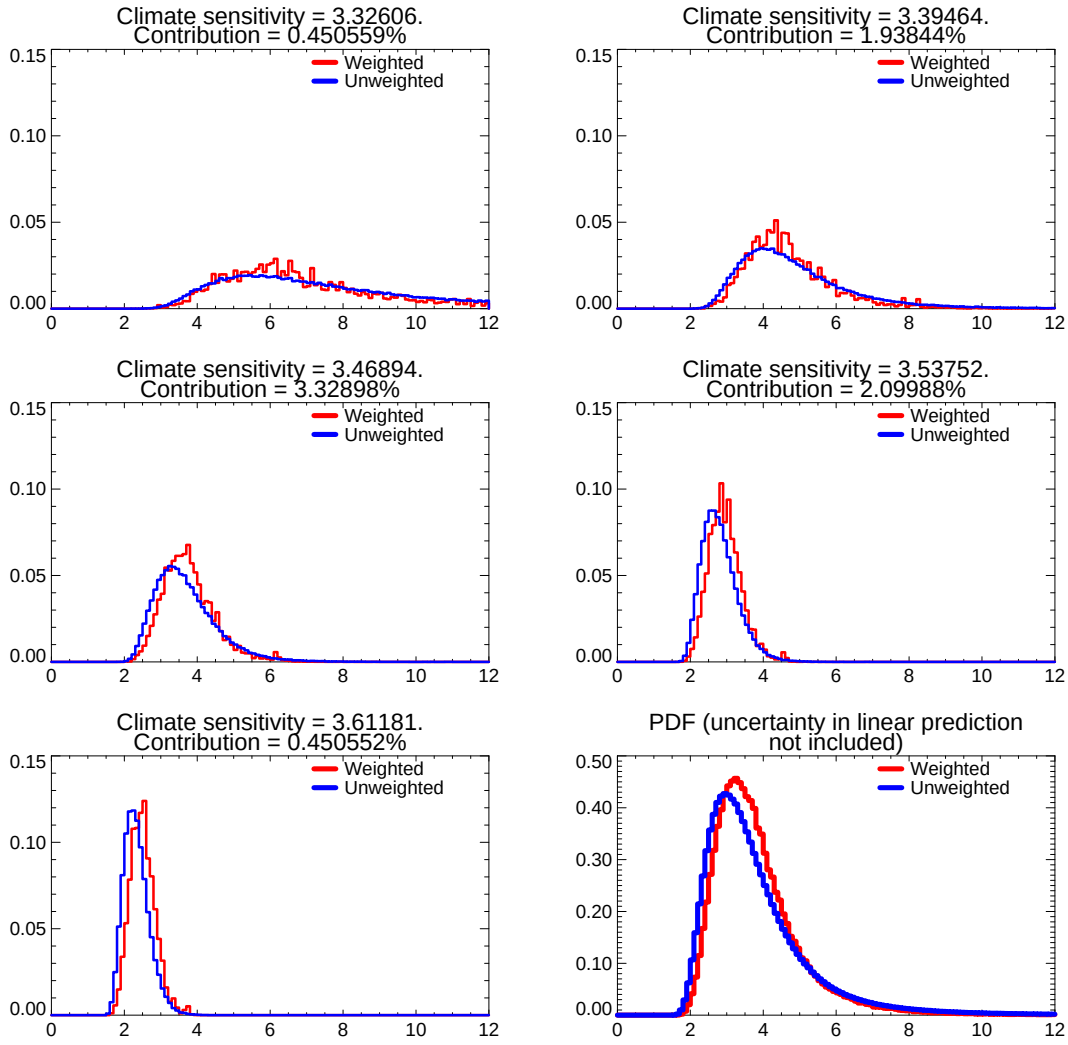

Figure 6. *a-e) Weighted (red) and unweighted histograms for 5 of the M intervals in the control $\Delta T_{2\times}$ range, each centred on the climate sensitivity in the plot title. Contribution is the probability of each interval; f) PDF which does not allow for additional uncertainty due to natural variability and the linear prediction methodology.*

3. Loop through each of *M* intervals in the control $\Delta T_{2\times}$ range and determine the probability p_i of the standard climate sensitivity actually coming from the *ith* interval. For each interval, predict $\Delta T_{2\times}$ for each Monte Carlo run using the centre of the interval as the standard value of climate sensitivity. Figs. 5a-e show how sensitive the histogram of $\Delta T_{2\times}$ is to the value of $\Delta T_{2\times}$ from the standard model. When the control $\Delta T_{2\times}$ is low (see Fig. 6a), the majority of ensemble members are effectively positive changes in feedback compared to the standard model, and so the resulting histogram has very large values for $\Delta T_{2\times}$. The opposite happens for high values of standard $\Delta T_{2\times}$ (see Fig. 6e). Using the CPI, estimate the weighted histogram of climate sensitivity (red curves in Figs. 6a-e) for this interval in the climate sensitivity range. Fig. 6f shows the effect of summing up the histograms across the *M* intervals in the control $\Delta T_{2\times}$ range, weighting by the probability p_i. This estimate does not account for natural variability or uncertainty in the linear prediction of $\Delta T_{2\times}$ at untried parameter combinations. Step 4 allows for these additional uncertainties.

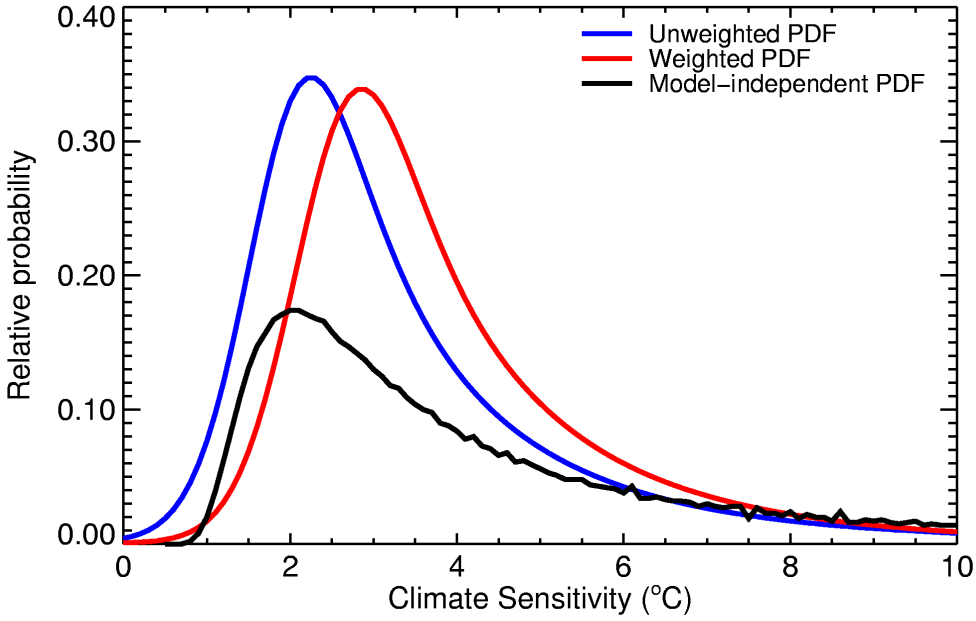

Figure 7. *Comparison of unweighted (blue), weighted (blue) and observationally-constrained (black) PDFs of climate sensitivity. 40% of the observational PDF lies to the right of 10°C.*

4. Predict $\Delta T_{2\times}$ for the test ensemble of multiple parameter perturbations using the value of standard model climate sensitivity in step 3. Calculate the standard error of the prediction, which is a measure of the suitability of our linear assumption but also encompasses natural

climate variability. For each point of the PDF, using the standard error of the prediction, estimate the total probability of the point occurring in the histogram.

5. Multiply the PDF from step 4 by the probability p_i from step 3 and add to the final PDF (see red curve in Fig. 7). This PDF is not biased by sampling but its validity depends on the success of the linear assumption in predicting $\Delta T_{2\times}$ at untried parameter combinations. The final PDF takes into account natural variability, and our ability to predict the response at untried combinations of parameters, allowing for the uncertainty to climate sensitivity of the control member. The use of the CPI can be omitted to produce an unweighted PDF (see blue curve in Fig. 7) using

$$p(\Delta T_{2\times}) \propto \int_{\mathbf{x}\in\chi} p(\Delta T_{2\times}|\mathbf{x}) d\mathbf{x}\,. \qquad \text{Eqn. 18}$$

The weighted and unweighted PDFs in Fig. 7 are similar in shape but the effect of the weighting is to shift the mode by 0.6°C and to change the 95% confidence interval from 1.3-8.6°C to 1.8-8.3°C. The two model PDFs are also compared with one produced by Gregory et al. (2002), which uses a simple physical constraint based on the relationship between climate sensitivity and observational estimates of radiative forcing and ocean heat uptake. The model-dependent PDFs are confined to a much smaller range than the model-independent PDF (black curve), which gives a 40% chance of having climate sensitivities greater than 10°C.

6. Conclusions

Ensemble climate prediction in the QUMP project is a two-stage process. The first stage involves running an ensemble designed to explore the sensitivity of the equilibrium response to doubled CO_2 concentrations to various parameter perturbations from a standard slab model, HadSM3. Section 3 shows such an ensemble provides a good basis for determining other models within parameter space that would simulate the observed present day climate well. However, the sampling strategy places too much emphasis on the standard model. Only by making the assumption of linearity were we able to remove the influence of the standard model to provide an *unbiased* estimate of the probability distribution function (PDF) of climate sensitivity. Therefore, the PDF produced in section 5, is not only conditional on the observations used in the metric of climate model performance, called the Climate Prediction Index (CPI), the underlying structure of the climate model, and the choice of parameter space, but also on this restrictive assumption of linearity.

For the second stage, a second ensemble is generated so that parameter space is sampled as efficiently as possible in a way that is not biased to any particular combination of parameters. The first stage is used to infer this second ensemble where each member has changes to several parameters from those values used in HadSM3. The algorithm described in section 4b also uses information from the first ensemble to select combinations of parameter values that are likely to simulate the present day climate as well if not better than HadSM3. This makes the second ensemble very cost-effective in terms of the computer resources needed to complete it.

Both the design of the second ensemble and the unbiased estimate of the PDF of climate sensitivity rely on our ability to predict the CPI and the equilibrium response to doubled CO2 levels at *untried* combinations of parameter values. So far this has required the assumption of linearity. Although tests in section 3 imply this might be a reasonable assumption for obtaining good predictions of the CPI, it is unlikely to work for the response of climate variables at sub-global or sub-hemispheric spatial scales. The reason we had to assume linearity was because of the sampling in the first ensemble. Ensembles where several parameters are perturbed simultaneously have greater potential for being able to predict responses on regional scales. There are two reasons for this. First, perturbing several parameters simultaneously enables the statistical methodology to incorporate nonlinear interactions between two or more parameters into the prediction. Second, the ensemble itself can be used to test the procedure by trying to predict the response from one member based on the response from the other members. This cross-validation technique can be then used for each member in turn to calculate an overall prediction error, which can be included in the final PDF. The usefulness of this PDF then depends on how large the prediction error is and this will vary with region and climate variable. It is not possible to cross-validate in this way with a single parameter perturbation ensemble.

The second ensemble where several parameters at perturbed at once from HadSM3 in any particular run, does provide the scope to explore nonlinear interactions between two or more parameters with more sophisticated statistical techniques. However, the design of the second stage requires a first stage. In section 4c, a combination of a D-optimal and latin hypercube design (with some modification to cope with a few awkward parameters) was used to provide an ensemble design which sampled the whole parameter space efficiently and in a way that was not biased towards any particular model. This ensemble will be started after the completion of the second ensemble described above. We will test the benefits of including nonlinear

interactions between parameters in the design of a new second-stage ensemble based on this new first-stage design. Any benefits will not only be good for ensemble climate prediction but also will improve the prospects for objectively tuning a climate model to the present day climate. Another advantage of having an unbiased first-stage design is that it can be used to augment the second ensemble used for the prediction of the PDF of the equilibrium response to doubled CO_2 levels. There will be an element of luck in how useful this will be because it depends on the proportion of members in the first-stage ensemble with relatively good skill at simulating the present-day mean climate.

In section 4c, standard techniques from experimental design theory were applied to design this new first-stage ensemble. In doing this, it became very apparent that a large amount of work was needed to incorporate *hybrid parameters* (parameters that became continuous only when a switch has been set). Therefore, we recommend that for future projects, these hybrid parameters are simply treated as on and therefore become relatively straightforward to deal with.

Another result from section 4c is that it is harder to efficiently sample parameters that take a set of discrete values (*factors*) rather than those which are continuous. This has consequences for future work where we explore the uncertainty of climate change predictions due to the underlying structural assumptions made in the climate model. Structural changes in climate models are often a matter of switching one scheme off and replacing it with another, which makes their treatment very similar to that used for factors. Therefore, large ensembles will be required to explore structural uncertainty.

Overall, we expect the ensemble of multiple parameter perturbations to greatly improve the estimation of the PDF. However, the acid test is actually running these two ensembles and checking that they deliver the expected results. Whilst they are running, other important areas of work are to implement the interpolation procedure of Sacks et al. (1989) used to predict the response at untried parameter values and to develop a method for predicting the CPI from runs where several parameters have been perturbed at once.

Appendix

MATLAB® program to determine D-optimal design.

```
function status=qump_expt_design()

%QUMP experiment has seven 2-level factors, three of which are
%hybrid factors,
%two 3-level factors and one 4-level factor.

%Design the full factorial for these
f=fullfact([2,2,2,2,2,2,2,3,3,4]);

%Calculate the design matrix and remove the degnerate columns.
c=dummyvar(f);
c(:,[3,5,7,9,11,13,15,18,21])=[];

%set the first column to be the mean
c(:,1)=1;

%set the last three columns to be the interactions between the
%three hybrids.
c(:,16)=c(:,2).*c(:,3);
c(:,17)=c(:,2).*c(:,4);
c(:,18)=c(:,3).*c(:,4);


%After some testing it seems that the candexch function is not
%guaranteed to produce the D-optimal design. Therefore we do
%100 iterations and use the best D-optimal design

max_det=0.0;
min_tr=1.0e+30;
n=32;
max_det_rows=1:n;
min_tr_rows=1:n;

iter=100;
dets=1:iter;
trs=1:iter;

for i=1:iter
  i
%candexch is the MatLab procedure that does all the hard work.
  rows=candexch(c,n,'display','off');
  e=c(rows,:);
  trs(i)=sum(diag(inv(e'*e)));
%  diag(e'*e)
  dets(i)=det(e'*e);

  if dets(i) > max_det
    max_det=dets(i);
    max_det_rows=rows;
  end
```

```
end


dets
trs
max_det

%set e to be the optimal design
e=c(max_det_rows,:)
det(e'*e)

e'*e

diag(inv(e'*e))


%save variable e
%save C:\mydata.txt e -ASCII
dlmwrite('C:\mydata.txt',e,' ')
```

The design produced is

A B C D E F G H I J K L M N O P Q R S T

[1,1,1,1,0,0,1,1,1,1,0,1,0,0,0,0,1,1,1,1]
[1,0,1,0,0,0,0,0,1,1,1,0,0,0,0,0,1,0,0,0]
[1,1,1,0,1,0,1,1,0,0,1,0,0,0,0,1,0,1,0,0]
[1,1,1,1,1,1,1,1,0,0,0,1,1,1,0,0,0,1,1,1]
[1,1,1,0,0,1,0,1,1,1,1,0,1,1,0,0,0,1,0,0]
[1,1,0,1,1,0,1,0,1,0,0,1,0,1,0,0,0,0,1,0]
[1,1,0,0,1,1,1,1,0,1,0,0,0,0,0,1,0,0,0,0]
[1,1,1,0,1,1,1,0,0,0,1,0,1,0,0,0,1,1,0,0]
[1,0,1,1,1,1,1,1,1,1,0,1,0,0,0,0,1,0,0,1]
[1,1,0,0,0,0,0,0,1,0,0,0,1,1,0,0,0,0,0,0]
[1,0,0,0,1,0,1,0,0,1,0,0,1,0,0,1,0,0,0,0]
[1,1,1,1,0,1,0,0,1,0,0,0,1,0,0,1,0,1,1,1]
[1,0,1,0,0,1,1,0,0,0,0,0,1,0,1,0,0,0,0,0]
[1,1,0,0,0,0,1,1,1,1,0,1,1,0,0,1,0,0,0,0]
[1,0,0,0,1,0,1,1,0,1,1,0,0,0,1,0,0,0,0,0]
[1,0,1,1,1,1,0,1,1,0,1,0,1,1,0,0,0,0,0,1]
[1,0,0,1,0,1,1,0,1,1,1,0,0,1,0,0,0,0,0,0]
[1,1,1,0,1,0,0,0,1,1,0,1,1,0,1,0,0,1,0,0]
[1,1,0,1,0,1,1,1,1,1,1,0,1,0,1,0,0,0,1,0]
[1,0,1,0,0,0,1,1,0,1,1,0,1,1,0,0,0,0,0,0]
[1,0,1,0,0,0,1,1,1,0,0,0,0,0,0,0,1,0,0,0]
[1,0,0,1,0,0,1,0,0,1,0,0,0,1,0,0,0,0,0,0]
[1,0,0,1,1,0,0,1,0,1,0,0,1,0,0,0,1,0,0,0]
[1,0,0,0,0,0,1,1,1,0,1,0,1,0,0,0,1,0,0,0]
[1,0,0,0,1,0,0,0,1,1,0,1,0,1,0,0,0,0,0,0]
[1,1,1,1,0,1,0,0,1,1,1,0,0,0,0,1,0,1,1,1]
[1,0,0,1,1,1,1,1,1,0,0,1,0,0,1,0,0,0,0,0]
[1,0,1,0,0,1,0,0,0,0,0,1,1,0,0,1,0,0,0,0]
[1,0,0,1,0,0,0,1,1,1,0,1,1,0,0,1,0,0,0,0]
[1,0,0,0,0,0,1,0,1,0,0,0,1,1,0,1,0,0,0,0]
[1,0,1,1,0,0,1,0,0,0,1,0,1,0,1,0,0,0,0,1]
[1,1,0,0,0,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0]
[1,1,0,1,0,1,0,1,1,0,0,1,0,1,1,0,0,0,1,0]
[1,0,0,0,1,1,0,0,0,0,0,1,0,1,0,0,0,1,0,0]
[1,1,1,0,0,0,0,0,1,0,1,0,0,0,0,1,0,0,1,0]
[1,1,0,0,0,1,0,0,0,0,0,1,0,0,0,0,1,0,0,0]
[1,0,1,1,1,0,1,0,1,1,0,1,1,0,0,1,0,0,0,1]
[1,1,1,1,0,1,0,1,0,0,0,1,0,1,0,0,0,1,1,1]
[1,1,0,1,1,0,0,1,1,0,0,0,1,0,0,1,0,0,1,0]
[1,0,1,0,1,0,0,1,0,0,1,0,0,1,0,0,0,0,0,0]]

The columns are labelled as follows:
A: Mean
B: Cape closure time scale
C: Rhcrit parameterisation scheme
D: Convective anvil scheme
E: Order of dynamic diffusion

F: Non-spherical cloud ice particles
G: Cloud area scheme
H: Canopy decoupling scheme
I: Stomatal conductance response to CO_2 off
J: SW water vapour continuum absorption
K: Number of accessible forest levels =2
L: Number of accessible forest levels =3
M: Gravity Wave Drag start level=4
N: Gravity Wave Drag start level=5
O: Forest roughness length type II
P: Forest roughness length type III
Q: Forest roughness length type IV
R: Interaction between B and C
S: Interaction between B and D
T: Interaction between C and D

References

Allen, M. R., 1999: Do-it-yourself climate prediction. *Nature* **401,** 642.

Bowman, K. P., J. Sacks, and Y.-F. Chang, 1993: Design and analysis of numerical experiments. *J, Atmos Sci* **50,** 1267-1278.

Craig, P. S., M. Goldstein, J. C. Rougier, and A. H. Seheult, 2001: Bayesian forecasting for complex systems using computer simulations. *J Amer Stat Assoc* **96,** 717-729.

Cubasch, U., and Coauthors (2001) Projections of future climate change. *Climate Change 2001: The Scientific Basis.* J. T. Houghton, et al. , Eds.Cambridge University Press, 881 pp.

Doutriaux-Boucher, M., and G. Seze, 1998: Significant changes between the ISCCP C and D cloud climatologies. *Geophys Res Lett* **25,** 4193-4196.

Fekete, B. M., C. J. Vorosmarty, and W. Grabs, 2002: High-resolution fields of global runoff combining observed river discharge and simulated water balances. *Global Biogeochem Cycles* **16,**

Fisher, R. A., 1935: *Design of Experiments.* Oliver and Boyd, 252 pp.

Gibson, J. K., P. Kallberg, S. Uppala, A. Noumura, A. Hernandez, and E. Serrano, 1997: ERA Description. ECMWF Re-Analysis Project Report Series, 1. Reading, UK. 77 pages.

Gough, W. A., and W. J. Welch, 1994: Parameter space exploration of ocean general circulation model using an isopycnal mixing parameterisation. *J Marine Res* **52,** 773-796.

Gregory, J. M., R. J. Stouffer, S. C. B. Raper, P. A. Stott, and N. A. Rayner, 2002: An observationally based estimate of the climate sensitivity. *J Clim* **15,** 3117-3121.

Harrison, E. P., P. Minnis, B. R. Barkstrom, V. Ramanathan, R. D. Cess, and G. G. Gibson, 1990: Seasonal variation of cloud radiative forcing derived from the Earth Radiation Budget Experiment. *J Geophys Res* **95,** 18687-18703.

Iman, R. L., and W. J. Conover, 1982: A distribution-free approach to inducing rank correlation among input variables. *Commun Statist-Simula Computa* **11,** 311-334.

Johns, T. C., R. E. Carnell, J. F. Crossley, J. M. Gregory, J. F. B. Mitchell, C. A. Senior, S. F. B. Tett, and R. A. Wood, 1997: The Second Hadley Centre Coupled Ocean-Atmosphere GCM: Model Description, Spinup and Validation. *Clim Dyn* **13,** 103-134.

Josey, S. A., E. C. Kent, D. Oakley, and P. K. Taylor (1996) A new global air-sea heat and momentum flux climatology. International WOCE Newsletter 24 : 3-5

Leroy, S., 1998: Detecting climate signals: Some Bayesian aspects. *J Climate* **11,** 640-651.

Mead, R., 1990: *Design of experiments: Statistical principles for practical applications*. Cambridge University Press, 634 pp.

New, M., M. Hulme, and P. Jones, 1999: Representing twentieth-century space-time climate variability. Part I: Development of a 1961-90 mean monthly terrestrial climatology. *J Clim* **12,** 829-856.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran: the Art of Scientific Computing*. Cambridge University Press, 963 pp.

Pukelsheim, F., 1993: *Optimal Design of Experiments*. John Wiley and Sons, 454 pp.

Rayner, N. A., D. E. Parker, E. B. Horton, C. K. Folland, L. V. Alexander, D. P. Rowell, E. C. Kent, and A. Kaplan, 2003: Global analyses of SST, sea ice and night marine air temperature since the late nineteenth century. *J Geophys Res* **108,**

Rossow, W. B., and R. A. Schiffer, 1991: ISCCP cloud data products. *Bull Am Meterorol Soc* **72,** 2-20.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn, 1989: Design and analysis of computer experiments. *Statistical Science* **4,** 409-435.

Stainforth, D., J. Kettleborough, M. A. M, M. Collins, A. Heaps, and J. Murphy, 2002: Distributed computing for public-interest climate modeling research. *Comput Sci Eng* **4,** 82-89.

Watterson, I. G., 1996: Non-dimensional measures of climate model performance. *Int J Climatol* **16,** 379-391.

Williams, K., M. Ringer, and C. Senior, 2003: Evaluating the cloud response to climate change and current climate variability. *Clim Dyn* **20,** 705-721.

Xie, P., and P. A. Arkin, 1998: Global monthly precipitation estimates from satellite-observed outgoing longwave radiation. *J Clim* **11,** 137-164.