Obtaining diverse behaviors in a climate model without the use of flux adjustments

Kuniko Yamazaki,[1] Daniel J. Rowlands,[1] Tolu Aina,[1] Adam T. Blaker,[2] Andy Bowery,[1] Neil Massey,[1] Jonathan Miller,[1] Cameron Rye,[1] Simon F. B. Tett,[3] Daniel Williamson,[4] Yasuhiro H. Yamazaki,[5] and Myles R. Allen[1,2]

Received 31 July 2012; revised 17 January 2013; accepted 22 February 2013; published 9 April 2013.

[1] A number of studies have set out to obtain a range of atmosphere and ocean model behavior by perturbing parameters in a single climate model (perturbed physics ensemble: PPE). Early studies used shallow layer slab ocean or flux-adjusted coupled ocean-atmosphere models to obtain a broad range of behavior as characterized by climate sensitivity. A recent study reports a relatively narrow range of sensitivities (2.2–3.2°C) in a PPE of 35 coupled models without flux adjustment, raising the question whether previous broad ranges were an artifact of the use of models that were not in top-of-atmosphere (TOA) energy balance. Moreover, no PPE experiment has reported a large spread of behavior of the ocean compared to that exhibited in a multi-model ensemble (MME) such as Coupled Model Intercomparison Project phase 3 (CMIP3). In this work, we randomly perturb model parameters of a coupled ocean-atmosphere general circulation model using a space-filling design containing 10,000 combinations. The ensemble is run over the distributed computing platform of climateprediction.net under fixed pre-industrial forcing without flux adjustment. We resample a second, 20,000-member, ensemble with perturbations conditioned on the TOA fluxes from the first ensemble to not drift significantly from a realistic base state while targeting a range of behavior. Models within the targeted ensemble show realistic regional control climates when compared to the CMIP3 ensemble, although there is a bias in global mean surface temperature. The range of predicted equilibrium climate sensitivities of the targeted ensemble is substantially smaller than that obtained with flux adjustment, but larger than the range in the CMIP3 ensemble or in the 35-model un-flux-adjusted PPE in a recent study mentioned above. The Atlantic meridional overturning circulation in the targeted ensemble exhibits a spread in strength as wide as that found in the CMIP3 ensemble. We conclude that flux adjustment is not a pre-requisite for obtaining a broad spread of behavior in a perturbed physics ensemble.

Citation: Yamazaki, K., et al. (2013), Obtaining diverse behaviors in a climate model without the use of flux adjustments, *J. Geophys. Res. Atmos.*, *118*, 2781–2793, doi:10.1002/jgrd.50304.

1. Introduction

[2] Perturbed physics ensembles (PPE) have been used in many studies to date with the aim to statistically quantify uncertainty in future projections of climate change. A PPE consists of a number of variants of a single model, usually a General Circulation (or Global Climate) Model (GCM). Variants are constructed by perturbing the values of a relatively small number (10–30) of model parameters that control the sub-grid scale physical processes in the model within plausible ranges determined by expert consultation. The underlying rationale of this approach may be summarized as follows. Sub-grid scale processes are still poorly understood or difficult to observe in nature, and so the parameter values representing these processes have large uncertainties on their bounds. Therefore, even the GCM's "standard" configuration, which has been tuned to reproduce the present climate well, is subject to large uncertainty. There may be other model variants that are statistically as close to the true state of the Earth's climate system, and hence have the ability to simulate the climate of both the present and the future as well as the standard version. The

[1]Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK.
[2]National Oceanography Centre, Southampton, UK.
[3]School of Geosciences, University of Edinburgh, Edinburgh, UK.
[4]Department of Mathematical Sciences, Durham University, Durham, UK.
[5]School of Geography, Politics and Sociology, Newcastle University, Newcastle, UK.
[6]School of Geography and the Environment, University of Oxford, Oxford, UK.

Corresponding author: K. Yamazaki, Atmospheric, Oceanic and Planetary Physics, University of Oxford, Oxford, UK. (kuniko@atm.ox.ac.uk)

2169-897X/13/10.1002/jgrd.50304

idea behind the PPE approach is to include such model variants in producing climate projections.

[3] The PPE approach, however, can be computationally expensive depending on the size of the ensemble; and for this reason in many past studies, the parameters have been varied over only a limited number of discrete values. Therefore, there has always been a concern that some critical corner of the O(10)-dimensional parameter space may have been overlooked. In this study, we present an approach that is both thorough and computationally efficient. We first sweep over the entire parameter space to make random perturbations, then run the perturbed model variants and use a statistical emulator on the simulated results to predict and focus on parameter perturbation combinations that meet certain conditions of interest. Finally, we run simulations with these targeted models, thereby making efficient use of computational resources.

[4] Wide ranges of behavior of the atmosphere have been documented in PPE experiments [e.g., *Collins et al*., 2006; *Rowlands et al*., 2012]. On the other hand, obtaining a large spread of ocean model behavior, expressed in terms of, for example, ocean heat uptake efficiency and Atlantic meridional overturning circulation (AMOC) strength, has so far been unsuccessful (e.g., *Collins et al*. [2007] and *Brierley et al*. [2010], in which ocean parameters related to mixing were perturbed individually). *Brierley et al*. [2010] discover some compensation between the perturbed and unperturbed ocean processes and suggest that this might further reduce the impact of parameter perturbations to ocean heat uptake, which is generally small to start with. For example, since advection and diffusion, two of the key ocean mixing processes, draw on the gradient of the temperature and salinity, it is conceivable that if mixing (of temperature and salinity, although not necessarily density) is increased by perturbation of one process, another mixing process might be forced to contribute less than in the standard configuration. Atmosphere parameter perturbations made in past studies have varied a wider variety of processes than just mixing and therefore the atmosphere might be able to respond more freely.

[5] Flux adjustments have traditionally been applied in climate change simulations using GCMs to prevent modeled oceans from drifting significantly away from a realistic initial base state, by adding artificial heat, freshwater or momentum flux terms at the surface of the ocean model. Its use was phased out as models with improved top-of-atmosphere (TOA) fluxes, hence less drift, were developed [*Randall et al*., 2007]. With the advent of PPE experiments, however, flux adjustments were reintroduced, as perturbations to model physics meant that the small TOA flux imbalance its unperturbed version was no longer guaranteed. *Collins et al*. [2010] report that an unrealistic surface climate causes climate change feedbacks to be different to those for an unbiased surface climate and that flux adjusting for sea surface temperature also improves modeled land temperature.

[6] Flux adjustments, however, affect dynamical ocean processes such as advection and meridional overturning [e.g., *Collins et al*., 2006; *Yamazaki*, 2008], so it would be better not to use it in experiments where ocean dynamics are expected to play an important role. Furthermore, *Marotzke and Stone* [1995] use a simple coupled model to show that although the correct mean state may have been obtained by the additive flux adjustments at the sea surface, the transient behavior of the model is erroneous. Nonetheless, biases in the baseline surface climate are clearly undesirable. We aim to reduce this not by using flux adjustment but by targeting parameter perturbations that have small top-of-atmosphere (TOA) flux imbalance and yet retain the possibility for a wide spread of atmosphere and ocean behaviors, such as manifested in effective climate sensitivity and Atlantic meridional overturning.

[7] *Shiogama et al*. [2012] generate 35 perturbed versions of the Model for Interdisciplinary Research on Climate version 5 (MIROC5) coupled GCM using a set of integrations of the atmosphere-only version of the model to target parameter combinations that were likely to have a substantial impact on radiative forcing or feedback and filtered to give low TOA flux imbalances. They ran each model version for 30 years to evaluate the control climate and computed their climate sensitivities by increasing CO_2 in year 10 and performing parallel integrations for the remaining 20 years. They find a range of climate sensitivities of 2.3–3.2°C. While larger than would be attributable to internal variability, this is clearly very substantially smaller than the ranges obtained with flux-adjusted ensembles. The short lengths of their integrations preclude an extensive discussion of changes in ocean properties.

[8] In this paper, we outline the approach that we use to generate new PPE members to obtain diverse behaviors without the use of flux adjustment which is similar to that of *Shiogama et al*. [2012], with three key differences: (i) we estimate sensitivities from previously-undertaken integrations of the slab versions of the model, exploiting the fact that the climateprediction.net slab ensemble is now so densely sampled, with hundreds of thousands of simulations, that it is possible to accurately predict the climate sensitivity of parameter combinations of interest without performing doubled-CO_2 experiments; (ii) we perform two coupled ensembles, using TOA fluxes diagnosed from the first to target stable versions in the second, rather than using atmosphere-only integrations; and (iii) we extend integrations of the perturbed physics coupled ensemble for up to 120 years, allowing us to investigate the impact of perturbations on ocean behavior. Hereafter, we call the first and the second ensembles the "raw" and the "targeted" ensembles, respectively. We describe the model, experiment design and ensemble design in section 2, compare the global mean time series of the key model properties in the raw and the targeted ensembles, present the estimated spread of effective climate sensitivity in section 3, and give summary and present conclusions in section 4.

2. Model and Methods

2.1. Model and Experiment Design

[9] The model we have used in this study is a version of HadCM3, a Hadley Centre Global Climate Model [*Gordon et al*., 2000]. The atmosphere component is a hydrostatic model and has $3.75° \times 2.5°$ (longitude, latitude) resolution in the horizontal and 19 levels in the vertical with hybrid vertical coordinates. The ocean component is a version of the *Cox* [1984] ocean model and has $1.25° \times 1.25°$ horizontal resolution and 20 vertical levels, with finer resolution near the

surface. A thermodynamic sea ice model is included. The standard configuration of HadCM3 is well known for its small TOA flux imbalance and thus can be run without flux adjustment, apart from a very small freshwater flux adjustment in the vicinity of land ice, to account for freshwater flux into the ocean from iceberg calving, which is not explicitly modeled in HadCM3. In perturbed physics experiments, however, the perturbed model parameters may alter the physical processes so TOA fluxes are no longer in balance. An interactive sulphur cycle is activated in this experiment, which brings TOA fluxes in the unperturbed model to a small negative value.

[10] The multi-thousand member ensemble of models are sent out over the climateprediction.net (CPDN) distributed computing platform and are computed on PCs on idle processor time that has been donated by the participants from across the world. To enable the models to run on as many PCs as possible and to increase speed and reduce data size, we converted the original full-resolution, 64-bit, multi-processor HadCM3 configuration to run in 32-bit on a single processor. Due to truncation errors, the ocean model was initially numerically unstable and did not conserve heat and salt, but doubling the numerical precision in key routines, including those that solve the barotropic and the tracer diffusion equations, eliminated the instability, and restored heat and salt conservation.

[11] A schematic diagram of the experiment design is shown in Figure 1. All perturbed control simulations have been initialized from the same ocean and atmosphere states. The states have been initiated from year 100 of the original HadCM3 control run [*Jackson et al*., 2011], then "spun up" under the 32-bit, single processor configuration for a further 100 years. Each perturbed model-version was run as a control simulation for 120 years, in 40-year segments, under a fixed pre-industrial external forcing of the year 1900, with a seasonal cycle. After a 40-year task is completed on a participant's PC, the CPDN server automatically prepares and distributes the experiment with the same parameter perturbation for the next 40 years. The atmosphere and ocean states obtained in the control experiment are used to initialize the next stage of the experiment, one forced with idealized CO_2 concentrations and another forced with observed or estimated 20th century natural and anthropogenic forcing. This paper will focus on the control experiment.

2.2. Ensemble Design

[12] The complete list of parameters and their standard values is given in Tables 1 and 2. We perturb 33 model parameters, 22 of which are associated with the atmosphere model, six with the ocean model and five with the external forcing associated with aerosols. The atmosphere parameters are a subset of those perturbed in *Murphy et al*. [2004], and the ocean parameters are identical to those perturbed in *Collins et al*. [2006]. Our experiment differs from these studies in that all the parameters are perturbed not singly but simultaneously. These parameters are selected as they are considered to have large uncertainties. The parameters and their ranges have been chosen by consulting the experts who originally designed and implemented HadCM3 (Table 3).

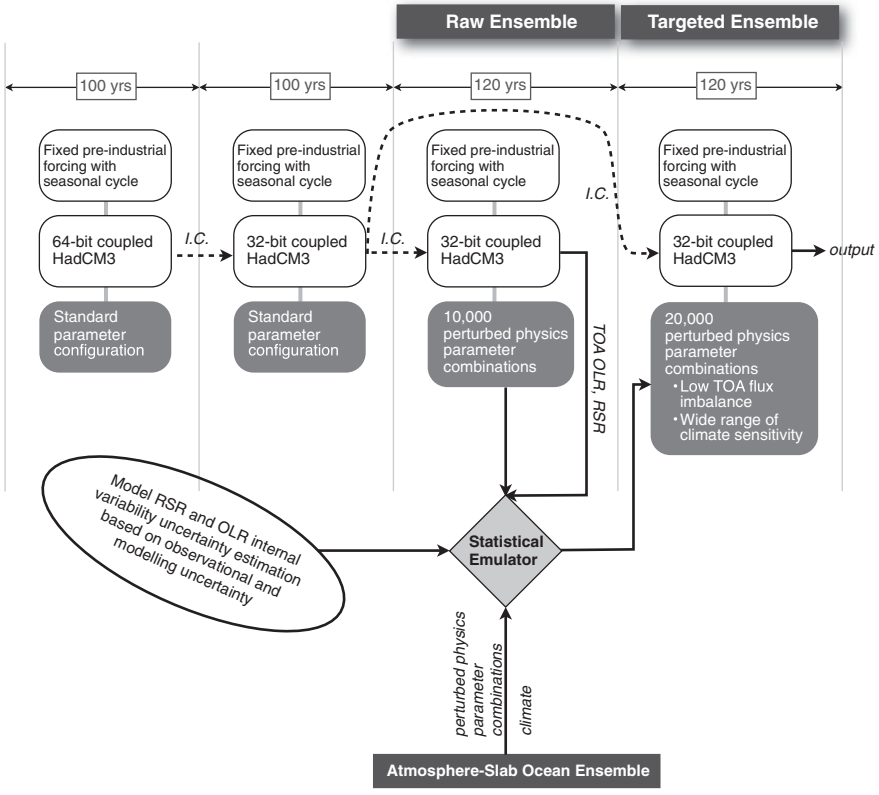

Figure 1. Schematic diagram of experiment design. "I.C." indicates initial conditions (i.e., initial climate states) for the atmosphere and the ocean model components.

Table 1. Atmospheric Parameters Used in Physical Parameterization Schemes in HadCM3[a]

Description	Component	Default Value	Name
Rate at which cloud liquid water is converted to precipitation	Cloud	1×10^{-4} (s^{-1})	CT
Threshold cloud liquid water content over sea	Cloud	5×10^{-5} ($kg\,m^{-3}$)	CW_SEA
Threshold cloud liquid water content over land	Cloud	2×10^{-4} ($kg\,m^{-3}$)	CW_LAND
Empirically adjusted cloud fraction	Cloud	[b]	EACF
Critical relative humidity for cloud formation	Cloud	[c]	RHCRIT
Ice fall speed	Cloud	1 (ms^{-1})	VF1
Entrainment rate coefficient	Convection	3	ENTCOEF
Albedo at melting point of sea ice	Radiation	0.5	ALPHAM
Temperature range over which ice albedo varies	Radiation	10(°C)	DTICE
Ice particle size	Radiation	30×10^{-6} (m)	ICE_SIZE
Horizontal diffusion coefficient for temperature and wind speed	Dynamics	[d]	DIFF_COEFF
Horizontal diffusion coefficient for water vapor	Dynamics	[e]	DIFF_COEFF_Q
Order of horizontal diffusion for temperature and wind speed	Dynamics	[f]	DIFF_EXP
Order of horizontal diffusion of water vapor	Dynamics	[g]	DIFF_EXP_Q
Surface gravity wave drag: typical wavelength	Dynamics	2×10^4 (m)	KAY_GWAVE
Surface gravity wave trapped lee wave constant	Dynamics	3×10^5 ($m^{-3/2}$)	KAY_LEE_GWAVE
Lowest model level for gravity wave drag	Dynamics	3	START_LEVEL _GWDRAG
Number of soil levels from which water can be extracted	Land surface	4,4,3,3 7	R_LAYERS
Vertical distance over which air parcels travel before mixing with their surroundings	Boundary layer	0.15	ASYM_LAMBDA
Constant in Charnock formula for calculating roughness length for momentum transport over sea	Boundary layer	1×10^{-2}	CHARNOCK
Used in calculation of stability function for heat, moisture, and momentum transport	Boundary layer	10	G0
Roughness length for free heat and moisture transport over the sea	Boundary layer	1×10^{-3} (m)	Z0FSEA

[a]The columns show the description of the parameter, the component of the model in which the physical scheme is located, values the standard configuration of HadCM3, unit and the name in the Unified Model code.
[b]0.5 in 19 levels.
[c]0.95, 0.9, 0.85, 0.7 in 16 levels.
[d]5.47×10^8 in 18 levels, 4×10^6.
[e]5.47×10^8 in 13 levels, 1.5×10^8 in 5 levels, 4×10^6.
[f]3 in 18 levels, 1.
[g]3 in 13 levels, 2 in 5 levels, 1.

Table 2. Same as in Table 1 but for Ocean and Forcing

Description	Component	Default Value	Name
Ocean parameters			
Isopycnal diffusion of tracer at surface	Dynamics	1000 ($m^2\,s^{-1}$)	AHI1_SI
Background vertical diffusion of tracer at surface	Dynamics	1×10^{-5} ($m^2\,s^{-1}$)	KAPPA0_SI
Increase of background diffusion of tracer with depth	Dynamics	3×10^{-8} (ms^{-1})	DKAPPA_DZ_SI
Background vertical diffusion of momentum (viscosity)	Dynamics	1×10^{-5} ($m^2\,s^{-1}$)	FNUB_SI
Decay of wind mixing energy with depth	Mixed layer	100 (m)	DELTA_SI
Wind mixing energy scaling factor	Mixed layer	0.7	LAMDA
Forcing Parameters			
Scaling factor for emission from anthropogenic sulphate aerosols	Chemistry	1000	ANTHSCA
Sulphate mass scavenging parameter L0	Chemistry	7×10^{-5} (s^{-1})	L0
Sulphate mass scavenging parameter L1	Chemistry	3×10^{-5} (s^{-1})	L1
Model level for SO2 (high level) emissions	Chemistry	3	SO2_HIGH_LEVEL
Scaling factor for emission from natural (volcanic) emissions	Chemistry	1	VOLSCA

Table 3. Ranges of Estimated Climate Sensitivity, TOA Imbalance, Reflected Shortwave Radiation (RSR), and Outgoing Longwave Radiation (OLR) for Different Model Ensembles

	Raw Ensemble	Targeted Ensemble		CMIP3
		20% Confidence Region	99% Confidence Region	
Climate sensitivity (K)	2.00–9.63	2.44–7.28	1.99–8.64	2.0–4.5
TOA imbalance (W/m^2)	–23.3–19.7	–0.84–1.35	–4.75–5.25	–0.74–4.62
RSR (W/m^2)	70.8–154	95.6–99.3	89.2–106	98.5–111
OLR (W/m^2)	197–262	242–245	238–250	231–242

2.2.1. Raw Ensemble

[13] The raw ensemble has been designed to double as the foundation ensemble for this work and as the primary ensemble for a separate study investigating abrupt changes of AMOC under increase in CO_2 [*Williamson et al*., 2012]. The parameters are perturbed within plausible ranges elicited from model developers in a space-filling Latin hypercube design containing 10,000 parameter combinations. Together with 40 models containing standard HadCM3 parameter values (hereafter "standard physics" values), a total of 10,040 models have been distributed using CPDN. Hereafter, this ensemble will be called the raw ensemble. The details of the design of the "raw" ensemble is found in *Williamson et al*. [2012] In total, 8052 of the original set of simulations have returned valid data to the CPDN servers.

2.2.2. TOA Flux Uncertainties

[14] Figure 2 shows the year 1–10 TOA flux components from the raw ensemble, with horizontal and vertical lines denoting the standard physics values. Given the random sampling used by the space-filling design, the raw ensemble shows a wide range in both components compared to that observed across the CMIP3 ensemble (black dots). To gain a more quantitative estimate, we have estimated uncertainties in the components of TOA imbalance using observed estimates from satellite observations.

[15] We estimate uncertainties in the individual components of outgoing radiation, for both reflected solar radiation (RSR) and outgoing longwave radiation (OLR), focusing on both observational and modeling uncertainties that affect outgoing radiation. Uncertainty estimates are made for the 2001–2005 period, based on the following sources and assuming all are independent Gaussian distributions: satellite measurement uncertainty for the individual components of OLR and RSR [*Loeb et al*., 2009], uncertainty in forcing that leads to changes in the top of atmosphere radiation, uncertainty in natural aerosols, uncertainty in the observed imbalance, uncertainty in the incoming radiation, and uncertainty arising from internal variability.

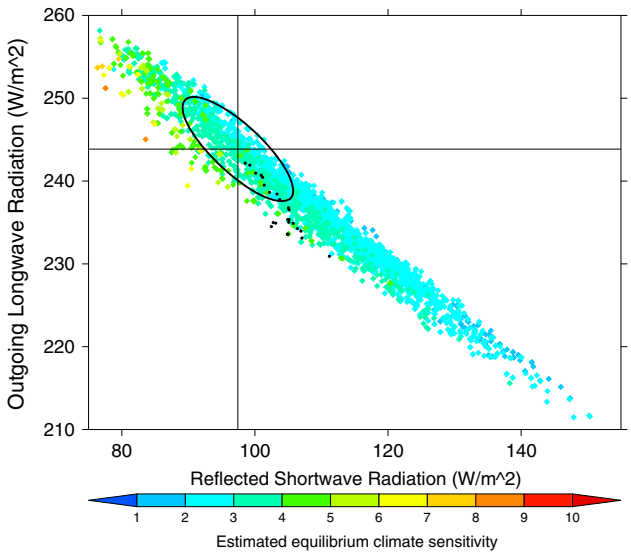

Figure 2. Global decadal mean reflected shortwave radiation (RSR) and outgoing longwave radiation (OLR) from the random ensemble, averaged over the first 10 years of the control spin up and colored by estimated equilibrium climate sensitivity. Intersection of the horizontal and vertical lines denotes our standard 32-bit, single-processor HadCM3 configuration, which differs from the CMIP3 HadCM3. Black dots denote comparable values from CMIP3 models/runs, obtained as the average of the final 50 years of the pre-industrial control experiment. The ellipse indicates the 99% confidence region.

[16] Total outgoing radiation (RSR + OLR) is computed from the incoming minus the observed imbalance, and its uncertainty is computed by combining uncertainty in the incoming solar radiation and in the net imbalance. Uncertainty in the total solar irradiance measurements is estimated to be 0.5 W/m^2 [*Kopp and Lean*, 2011]. Uncertainties in the energy imbalance from two recent estimates are 0.86 ± 0.12 W/m^2 ([*Willis et al*., 2004] for the upper 750 m) and 0.55 to 0.73 W/m^2 [*Lyman et al*., 2010]. A value of 0.75 ± 0.25 W/m^2 includes both and we estimate the uncertainty in the net flux as 0.25 W/m^2. The difference between RSR and OLR is obtained by assuming that the difference is a normal distribution with mean −40% (corresponding to an albedo of 0.3) and standard deviation of 10% of the incoming radiation. This covariance matrix was then combined with the covariance estimate for the individual observations for OLR and RSR (see *Tett et al*. [2012] for more detail).

[17] Uncertainty arising from internal climate variability is estimated from a separate, initial condition ensemble of 118 coupled simulations of the standard configuration of HadCM3 each of 120 years and almost negligible compared to other sources of uncertainty. Uncertainty in radiative forcing was computed by scaling based on atmospheric simulations (see *Tett et al*. [2012] for details) the uncertainties in the LW and SW forcings. These were computed from Table TS.5 of *Solomon et al*. [2007] to give 1σ uncertainties of 1 and 0.20 W/m^2 for RSR and OLR. We computed uncertainty in natural aerosols by using three simulations from *Penner et al*. [2006]. After correcting the three contemporary simulations to the same RSR value that the range in pre-industrial RSR was 1 W/m^2, which we used as our 1σ uncertainty. These sources of uncertainty were combined with the observational covariance derived above.

[18] Combining these uncertainties gives individual standard errors of 1.65 and 0.94 W/m^2 in RSR and OLR respectively, and a correlation between them of

$$\begin{pmatrix} 2.7497660 & -0.65229269 \\ -0.65229269 & 0.87466201 \end{pmatrix}$$

[19] An ellipse of the 99% confidence region shown in Figure 2, containing 99% of (RSR,OLR) combinations generated from this covariance, corresponds to a TOA imbalance of 5 W/m^2 relative to the standard physics configuration.

[20] Estimation of uncertainties in the components of the TOA imbalance described above requires a set of subjective choices and judgments. We are mindful, therefore, to treat this as an estimate; and in section 3.3, we investigate the sensitivity of our results to an alternative formulation of the uncertainty estimate.

2.2.3. Targeted Ensemble

[21] Given the very wide range of TOA flux components observed in the raw ensemble, many simulations drift away from the initial state taken from a long simulation of the standard HadCM3 configuration. Since the initial TOA flux

imbalance represents how much the model climate must adjust to restore radiative balance, we can use this quantity to target regions of parameter space that will not drift significantly from the standard physics base state. The key assumption we make is that the standard physics configuration is a realistic base state to target (although the methodology presented here can be applied to targeting any property of the model).

[22] As Figure 2 indicates, the range of TOA flux components from the raw ensemble vastly exceeds the estimated uncertainty estimate, shown by the ellipse indicating the 99% confidence region. We find that approximately 29% of the raw ensemble lies within the 99% confidence region. Our goal here is to refine the ensemble design and increase the sampling in regions of parameter space producing simulations within the 99% confidence region.

[23] We add a further constraint that the ensemble should explore a wide range of atmosphere and ocean properties: this is to avoid producing many model-versions that are virtually identical to the standard physics version, which is clearly not useful for representing uncertainty in the climate response to anthropogenic forcing. To guard against this, we also target model versions showing a wide spread of equilibrium climate sensitivity, although in principle this could be any property of the model (or combination thereof).

[24] The climate sensitivities are estimated using a statistical emulator [*Breiman*, 2001] and results from the original CPDN HadSM3 ensemble [*Stainforth et al*., 2005] as discussed in *Rowlands et al*. [2012]. Figure 3a shows predictions of climate sensitivity from the statistical emulator in a 10-fold cross-validation experiment. The predictions explain over 95% of the variance in simulated climate

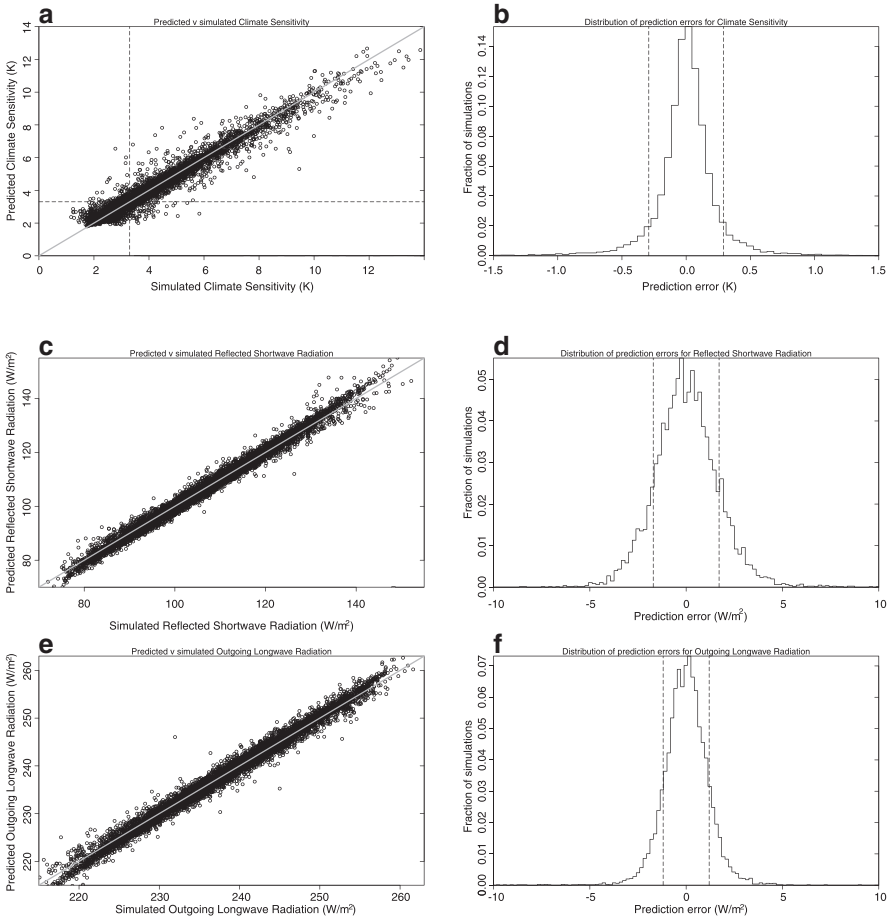

Figure 3. Validation of the random forecast statistical algorithm in predicting: (a) Climate Sensitivity (b) Year 1–10 Reflected Shortwave Radiation, and (c) Year 1–10 Outgoing Longwave Radiation. Figures 3a, 3c, and 3e show predicted values from the random forest against simulated values from HadCM3/HadSM3, and Figures 3b, 3d, and 3f show the distribution of prediction errors. All results are out-of-sample predictions which are generated automatically by the algorithm, with similar results found in a 10-fold cross validation. Results in Figure 3a correspond to a 14,001 member ensemble of HadSM3 slab-ocean experiments, while Figures 3b and 3c are from the raw HadCM3 ensemble consisting of 8052 members. Horizontal and vertical dashed lines in Figures 3a, 3c, and 3e correspond to the simulated values for the standard physics configuration. Dashed vertical lines in Figures 3b, 3d, and 3f correspond to the root mean square prediction error. In all cases, the random forest predictions explain over 95% of the variance in simulated values.

sensitivity values, and the root-mean-square prediction error of approximately 0.3 K (Figure 3c) is consistent with estimates of uncertainty arising from internal variability.

[25] The estimation of climate sensitivity only uses the atmospheric portion of the parameter space, and, while the slab model climate sensitivity will not fully quantify the response of the coupled model, we are confident that it will be a good indicator, given the dominant role that uncertainties in atmospheric parameters have on uncertainty in the climate response [e.g., *Collins et al*., 2010; *Rowlands et al*., 2012].

[26] We use a similar statistical emulator to fit the relationship between the full input parameter space and OLR and RSR from raw ensemble. Figures 3b and 3c show predictions from the emulator against simulated values in another 10-fold cross validation. In both cases, the predicted values explain over 95% of the variance in simulated quantities. Importantly, the variances of prediction errors (Figures 3d and 3e) are smaller than uncertainties discussed above (by a factor of 3), indicating that the emulator is able to provide information when interpolating within the 99% confidence region.

[27] Given an accurate emulator of the model response for these quantities of interest, we are now in a position to estimate the model response at an arbitrary point in parameter space. Thus, with a large candidate set of parameter configurations and sampling design, we can select configurations with a wide range of climate sensitivities that are predicted to be close to the standard physics configuration in the OLR/RSR space.

[28] The choice of sampling design is dependent on the probabilistic interpretation of the ensemble. Here, we present a traditional Bayesian importance sampling approach often used in similar studies [e.g., Rougier and Sexton, 2007] and a threshold sampling case, which is more closely tied to a frequentist interpretation. The importance sampling approach could be interpreted as generating a distribution of possible models corresponding to a probability distribution of possible behavior, subject to problematic issues regarding the specification of a prior "probability density" to different regions of parameter space. The threshold sampling approach simply aims to explore the range of behavior accessible to the model while satisfying a given goodness-of-fit threshold with respect to an observational constraint (in this case the TOA fluxes), and hence does not admit a probabilistic interpretation. It should be noted that much of this section attempts to highlight the methodology we have taken rather than justifying the particular choices made.

[29] For both cases, we start with the same 10^6 member candidate ensemble, produced through a Latin-hypercube sampling of parameter space of parameters, which can be interpreted as a joint uniform prior distribution on all input parameters. This design implicitly ensures that all factor/switch variables have an equal number of cases for each value. A 10,000 member ensemble of candidate parameter combinations is produced for each approach as follows:

[30] 1. Importance Sample: To ensure a wide distribution of climate sensitivity, we attach a prior weight to each candidate ensemble member such that the resulting prior distribution of estimated climate sensitivity is uniform. Second, we estimate a likelihood weight for each candidate ensemble member based on the predicted year 1–10 OLR and RSR from the emulator. Specifically, we evaluate the likelihood based on a Gaussian distribution with a mean given by the standard physics OLR/RSR and covariance matrix of the previous section. We then combine the two weights to give a posterior weight for each candidate, from which we sample 10,000 ensemble members according to this weight.

[31] 2. Threshold Sample: We first subset the candidate ensemble members, selecting those predicted to lie within the 99% confidence region around the standard physics values of OLR and RSR. To ensure a wide spread of climate sensitivity, we then attach a weight to each of the subset to ensure that the predicted distribution of climate sensitivity was approximately uniform (it will not be precisely uniform since we do not allow duplicates), and sample 10,000 members according to this weight.

[32] We discuss the properties of the two samples in the next section, although in the rest of the paper we simply combine the two, since our objective is to assess the range of behavior that is accessible to non-flux-adjusted models rather than to produce a distribution of models with a particular probabilistic interpretation. The combined distribution can be thought of as threshold sampling with somewhat denser sampling in regions of high likelihood, provided by the importance sample. Overall, the threshold sampling interpretation is more straightforward, given the difficulty in defending a particular prior distribution for model parameters and also our inability to guarantee that all simulations can be returned from CPDN. This is not as significant a problem for threshold sampling as for importance sampling, where the distribution of returned models is interpreted probabilistically.

3. Results

3.1. Effectiveness of the Targeting Method

[33] In the importance sampling approach, the proportion of ensemble members contributing 99% of the total weighted likelihood is 25% for raw and 80% for the targeted ensemble. In the threshold sampling approach, the proportion of ensemble members within the 99% confidence region is 29% for raw and 78% for targeted. As designed, the threshold sample ensemble achieves this improved sampling of low imbalance regions while still showing a wider range of climate sensitivities.

[34] The histograms of quantities such as the RSR, OLR, TOA imbalance and predicted climate sensitivity (not shown) show that, apart from some tails on either side of the predicted boundary, the simulated and the predicted distributions are very similar. This shows that both ensemble targeting methods have successfully resampled parameter perturbations resulting in the desired TOA flux imbalance.

[35] Since the emphasis of this paper is purely on assessing the range of behavior that can be found in perturbed-physics models without the use of flux adjustment, from now on we simply combine the importance-sampled and threshold-sampled ensembles to provide a single targeted ensemble. This provides greater sampling density in the region where likelihoods are maximum (provided by the importance-sampled ensemble) together with an artificially inflated spread

to ensure that all parts of parameter-space that yield models consistent with observations above a given threshold are adequately sampled.

3.2. Behavior of Radiatively Balanced Models in the Targeted Ensemble

[36] Table 3 shows the ranges of estimated climate sensitivity, TOA imbalance, RSR, and OLR for the raw, targeted and the CMIP3 ensembles. The range of climate sensitivities in the targeted members are substantially smaller than that obtained with flux adjustment, which spans 1.9–11.5 K [*Stainforth et al.*, 2005] but still larger than the range in the CMIP3 ensemble of opportunity [*Solomon et al.*, 2007].

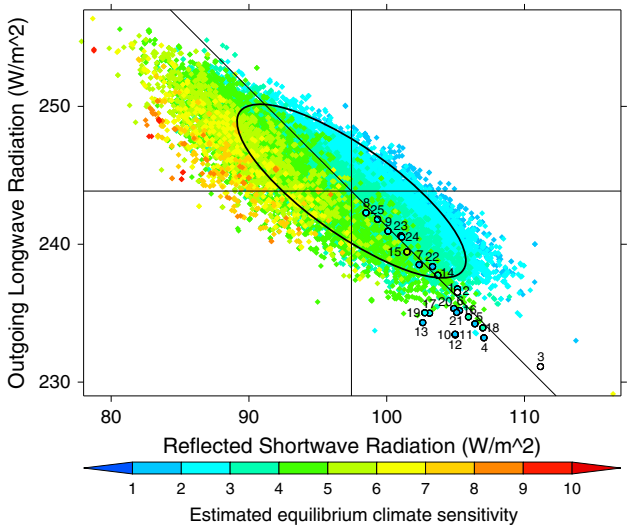

Figure 4. Distribution of global decadal mean reflected shortwave radiation (RSR) and outgoing longwave radiation (OLR) averaged over the first 10 years of the control spin-up and colored by estimated equilibrium climate sensitivity. Diamonds denote the targeted ensemble, colored by estimated equilibrium climate sensitivity. Colored circles with black outlines denote models from the CMIP3 ensemble, values of which were obtained as the average of the final 50 years of the pre-industrial control experiment. Intersection of the horizontal and vertical lines denotes the default HadCM3 configuration. The diagonal line indicates the line of zero net radiative heat flux. The ellipse indicates the 99% confidence region. The numbers by the circles with black outlines indicate CMIP3 models as follows: 1 = CCCMA_CGCM3_1_T47 run01, 2 = CCCMA_CGCM3_1_T63 run01, 3 = CNRM_CM3 run01, 4 = CSIRO_MK3_0 run01, 5 = CSIRO_MK3_0 run02, 6 = GFDL_CM2_0 run01, 7 = GFDL_CM2_1 run01, 8 = GISS_Model_E_H run01, 9 = GISS_Model_E_R run01, 10 = IAP_FGOALS1_0_G run01, 11 = IAP_FGOALS1_0_G run02, 12 = IAP_FGOALS1_0_G run03, 13 = INMCM3_0 run01, 14 = IPSL_CM4 run01, 15 = MIROC3_2_HiRes run01, 16 = MIROC3_2_MedRes run01, 17 = MIUB_ECHO_G run01, 18 = MPI_ECHAM5 run01, 19 = MRI_CGCM2_3_2a run01, 20 = NCAR_CCSM3_0 run01, 21 = NCAR_CCSM3_0 run02, 22 = NCAR_PCM1 run01, 23 = UKMO_HadCM3 run01, 24 = UKMO_HadCM3 run02, 25 = UKMO_HadGEM1 run01.

[37] Figure 4 shows the distribution of simulated global decadal mean RSR and OLR from the targeted ensemble (colored diamonds) over the first 10 years of the control spin-up. Each filled circle indicates a model simulation and is colored by estimated climate sensitivity. All circles but a few are located very close to the predicted 99% confidence region defined by RSR and OLR. This demonstrates the ability of the targeting method to identify parameter perturbations that simulate desired TOA fluxes. Furthermore, higher estimated climate sensitivities appear where RSR is small and the net downward radiative flux is positive (below the line of zero net radiative heat flux), and lower estimated climate sensitivities emerge where OLR is large and the net downward radiative flux is negative (above the line of zero net radiative heat flux), consistent with the results of *Sanderson et al.* [2008]. This suggests that climate sensitivities estimated from the slab ensemble [*Stainforth et al.*, 2005] are an adequate indicator of those of the coupled ensemble. We remind the readers once again that all simulations described in this paper are control runs, under fixed external forcing of year 1900, with a seasonal cycle. Our standard version of HadCM3 gives different TOA fluxes because we use a fully interactive sulphur cycle, as compared with the CMIP3 HadCM3 configuration in which the sulphur cycle is parameterized from an offline calculation. Note that the original HadCM3 model was tuned with the offline sulphur cycle, but we have used the original model parameters as our standard version without retuning.

[38] Figure 5 shows the time series of global area-weighted mean air temperature at 1.5 m in the targeted ensemble. The radiatively balanced models are generally

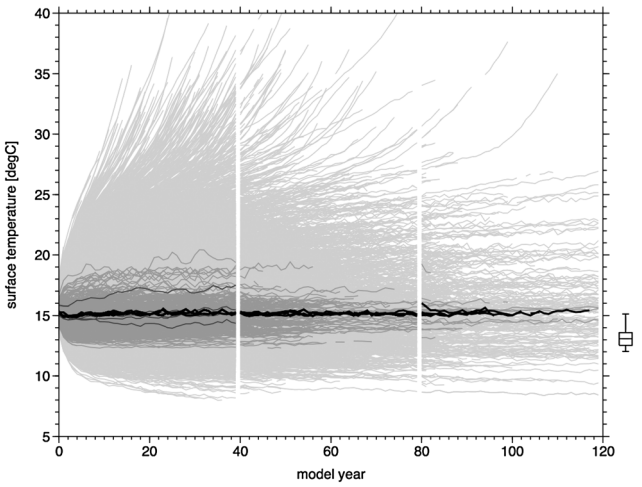

Figure 5. Time series of surface air temperature in the targeted ensemble. Light gray lines indicate the results from the entire ensemble. Gray lines indicate models whose simulated OLR and RSR are within the 20% confidence region, with dark gray lines showing random samples. Thick black lines indicate standard HadCM3 parameter configuration. Box-whisker plot indicates the range of the same property in 25 CMIP3 models in the pre-industrial control runs at minimum, 25%, median, 75%, and maximum.

warmer than the CMIP3 models and exhibit double the spread of the range of the models in the CMIP3 ensemble.

[39] Figure 6 is the same as Figure 5 but for TOA flux imbalance. It is interesting that the flux imbalance is positively greater in the CMIP3 ensemble than the radiatively balanced targeted ensemble. The fact that the global mean air temperature at 1.5 m is generally warmer in the targeted ensemble despite the negative or smaller positive TOA flux imbalance suggests that, on average, the ocean component of the CMIP3 models must take up more heat than the radiatively balanced targeted models.

[40] Ocean model behavior is often measured in terms of how efficiently the ocean transports the excess heat generated by radiative forcing under climate change from the ocean surface to the deep ocean. Metrics such as ocean heat uptake efficiency [*Raper et al*., 2002] are defined and used for this purpose. In the present study based on control simulations, however, it is not meaningful to assess this quantity because radiative forcing is zero by construction in balanced models, and hence there is no excess heat for the ocean to take up in the annual mean and global mean sense. However, the ocean's uptake/release of heat from/into the atmosphere is non-zero in the regional sense in control simulations as well. Thus we exploit a metric reflecting regional ocean behavior: the strength of the Atlantic meridional overturning circulation (AMOC) to quantify the behavior of the ocean.

[41] Figure 7 shows the time series of the strength of the simulated AMOC in the targeted ensemble, measured as the maximum strength between 500–2000 m at 26°N. After an initial drift over the first 30 years or so, in the models subject to the constraint that TOA fluxes are close to observations, the AMOC appears to be steady, with natural variability. The range of the spread in the strength of AMOC is 8–28 Sv among the radiatively balanced members of the targeted ensemble. It is slightly wider than the range in the CMIP3 models of 9.8–25 Sv [*Jackson et al*., 2011]. Since there is a strong correlation between the AMOC strength in a control experiment and the change of strength under increasing CO_2 concentrations [*Gregory et al*., 2005], the spread in our ensemble suggests a wide spread in the change in the AMOC strength under climate change.

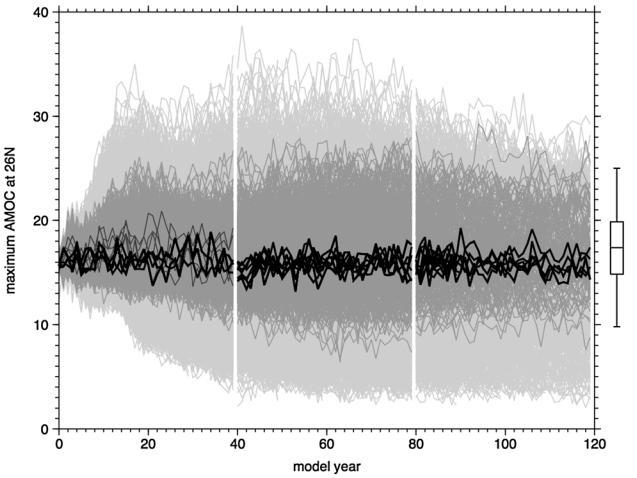

Figure 7. As in Figure 5 but for maximum Atlantic meridional overturning circulation (AMOC) at 26°N in the targeted ensemble. Box-whisker plot indicates the range of the same property in 25 CMIP3 models in the pre-industrial control runs at minimum, 25%, median, 75%, and maximum (Jackson et al. 2010).

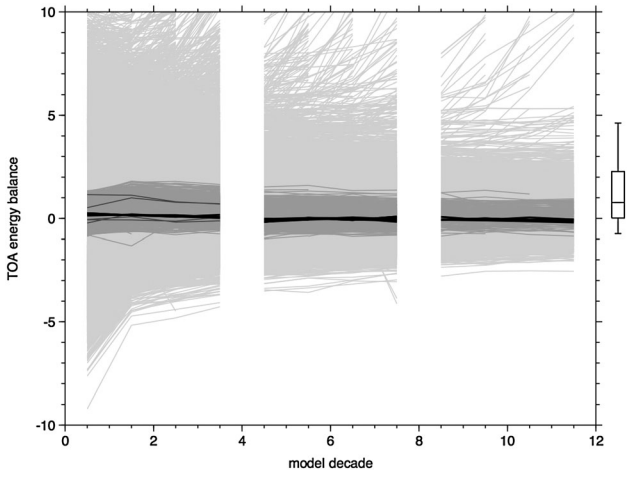

Figure 6. As in Figure 5 but for net TOA radiative flux in the targeted ensemble. Box-whisker plot indicates the range of the same property in 25 CMIP3 models in the pre-industrial control runs at minimum, 25%, median, 75%, and maximum.

[42] Direct comparison of the range of the AMOC strength with that in flux-adjusted PPEs is not straightforward, as flux adjustment tends to cause the AMOC to weaken (e.g., approximately 2 Sv over 100 years in *Yamazaki*, 2008). The spread of the AMOC strength in a flux-adjusted PPE in a control experiment using HadCM3L, a reduced-ocean resolution version of HadCM3, is found to be 7 Sv (width between two standard deviations) [*Yamazaki*, 2008]. Therefore, it might be possible to say that we have obtained a wider range of behavior in ocean models without using flux adjustment.

[43] Next, we examine diversity in model behavior with regards to El Niño-Southern Oscillation (ENSO). The range of ENSO behavior in a smaller coupled ensemble was previously discussed in *Philip et al*. [2009] and *Toniazzo et al*. [2008], but here, for the first time, we have a much larger ensemble without flux adjustment, which has been shown to adversely affect ENSO behavior [*Neelin and Dijkstra*, 1995]. Figures 8a and 8b show two examples of the time series of Southern Oscillation Index (SOI), which describes the frequency and the amplitude of the fluctuations of atmospheric pressure over the Equatorial Pacific Ocean which is deeply tied to the sea surface temperature fluctuations in the El Niño-La Niña events through atmospheric convection. During El Niño, the pressure is high over the western Pacific and low over the eastern Pacific and vice versa during La Niña. SOI is designed to reflect the pressure difference between Tahiti (149°W, 17°S) and Darwin (130°E, 12°S), so positive (negative) values indicate the occurrence of El Niño (La Niña). The model, shown in Figure 8a, shows much greater frequency and amplitude than the model shown in Figure 8b. The diversity in model

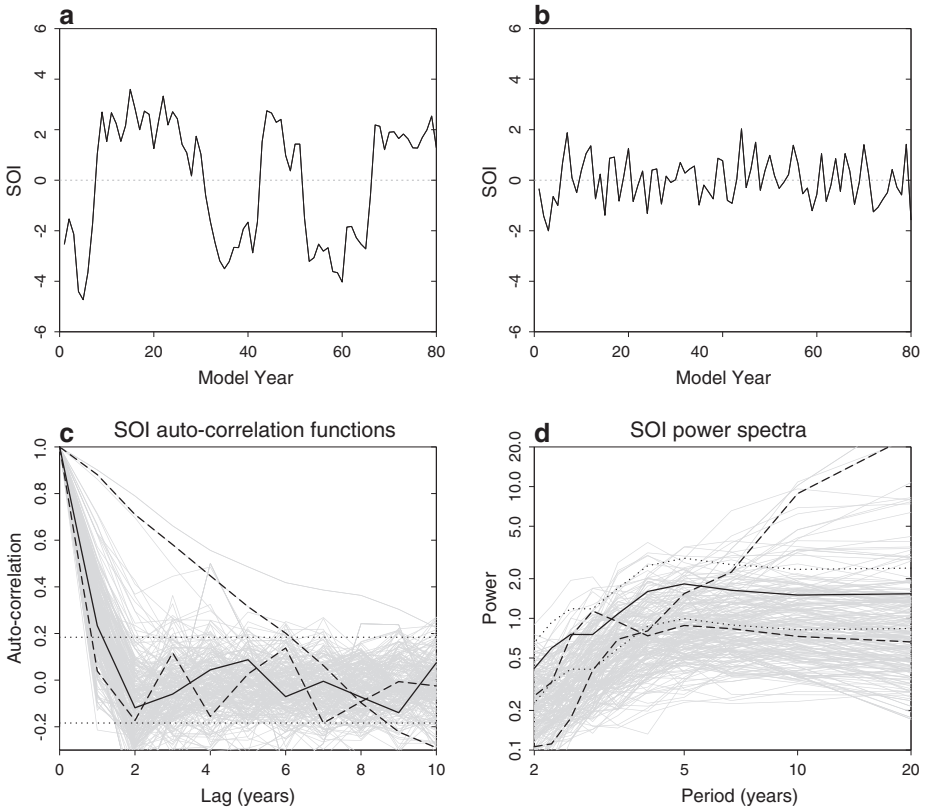

Figure 8. Southern-Oscillation Index (SOI) simulated in the targeted ensemble members within the 99% confidence region of the TOA fluxes: (a,b) Annual mean time series over the 80-year spin-up from two ensemble members showing diverse SOI behavior. (c) SOI auto-correlation function from the members of the targeted ensemble (light gray lines), along with auto-correlation function from observed SOI over 1936–2011 (solid black). Horizontal dotted black lines show the estimated 5–95% confidence interval on estimated correlation coefficients. (d) As Figure 8c showing the SOI power-spectra. Dotted black lines show the estimated 5–95% confidence interval on the observed power-spectra. Models highlighted in Figures 8a and 8b are shown by dashed black lines in Figures 8c and 8d. All SOI values are shown after linear de-trending (although this makes little practical difference) and are normalized by the standard-deviation of the observed SOI.

behavior is reflected in the plots of auto-correlation (Figure 8c) and the power spectra (Figure 8d), in which dashed black lines denote the two examples, solid black lines the observed SOI over 1936–2011 and gray lines the results from models within the 99% confidence region for TOA fluxes.

[44] Now, we compare the ranges of simulated surface temperature and precipitation in the targeted ensemble with the CMIP3 models by how much they differ from observations. The reader is reminded that the models are forced with atmospheric CO_2 concentrations from the year 1900, so we would expect sea surface temperatures to be around 0.2–0.3°C cooler than present day, with some reflection of this appearing in the surface air temperature [*Houghton et al.*, 1996]. Figure 9a shows surface air temperature from the ERA-Interim reanalysis data averaged over 1979–2011. Figures 9b and 9c show the maximum and minimum deviations, respectively, from (Figure 9a) the 40-year mean surface temperature from twenty CMIP3 models. In each grid-cell, the temperature in the twenty CMIP3 models is compared with that of the ERA-Interim data, and the grid is colored with the maximum temperature deviation. Over large regions, this quantity is negative (albeit small), indicating the entire CMIP3 ensemble lies below the ERA-Interim reanalysis. The strong negative anomalies at high latitudes, in particular over the Nordic Seas, are likely due to changes in the ice cover in that region compared with observations.

[45] Figure 9d shows the 40-year average surface temperature from a model in the targeted ensemble with minimum global mean root-mean-square error, and Figure 9e shows the deviation of Figure 9d from Figure 9a. Figures 9f and 9g are as Figures 9d and 9e for a model in a subset of the targeted ensemble with estimated climate sensitivity exceeding 4.5°C. If we take Figures 9b and 9c as the upper and lower range in surface temperature defined by the CMIP3 models, then Figures 9e and 9g appear to lie within this range. This suggests that the targeted ensemble succeeds in producing control climates that can be considered realistic as far as the CMIP3 models are concerned. Similar conclusions can be drawn for annual mean accumulated precipitation, shown in Figure 10.

[46] Figures 9 and 10 demonstrate that the errors in base climates of many members of the targeted ensemble are not substantially worse than typical errors in members of CMIP3,

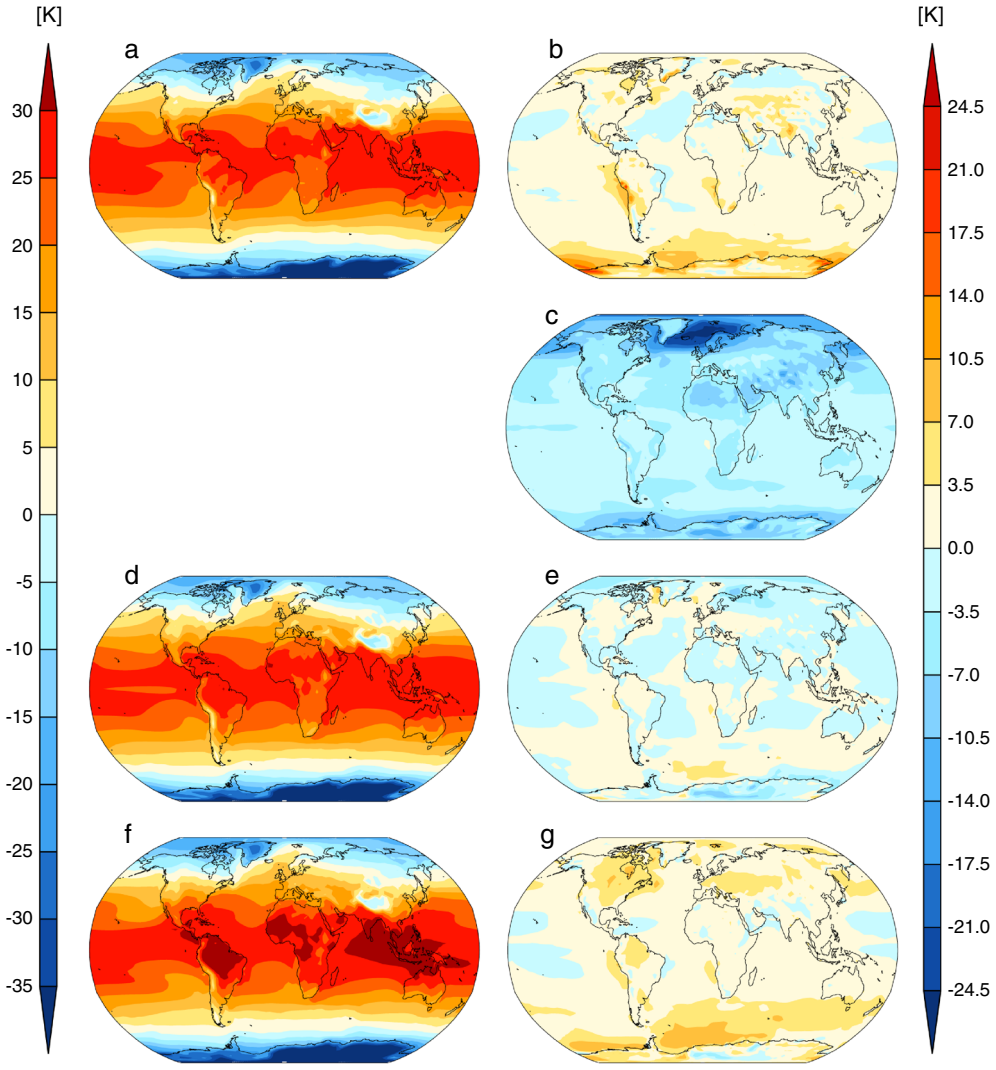

Figure 9. Maps of surface air temperature. (a) Average of ERA-Interim data over 1979–2011, (b) maximum difference between CMIP3 models from un-flux-adjusted pre-industrial control runs and Figure 9a, (c) same as Figure 9b but minimum difference, (d) targeted model averaged over year 40–80 with minimum global mean RMS error, (e) difference between Figures 9d and 9a, (f and g) same as Figures 9d and 9e but minimum within models with estimated climate sensitivity of higher than 4.5 K.

even if we restrict attention to models with climate sensitivity exceeding 4.5°C, often regarded as "anomalously high". Hence, we conclude it is possible to generate models with climate sensitivities outside the 2–4.5°C considered "likely" by the Intergovernmental Panel on Climate Change in 2007 [*Solomon et al.*, 2007], although clearly the range of sensitivities in Table 3 is very substantially smaller than the range obtained in flux-adjusted ensembles [*Stainforth et al.*, 2005].

3.3. Sensitivity to Uncertainty in TOA Components

[47] As discussed in section 2.2.2, the estimation of uncertainty in the components of the TOA flux relies on a set of subjective choices. We therefore performed a separate evaluation, using a second estimate of uncertainties in the covariance, which is somewhat smaller.

[48] For the importance sample to estimate the distributions as though this second confidence region had been used, we take the models returned in the importance sample and attach a weight to each based on the ratio of the likelihood with the new confidence region to that with the old confidence region. This set is then resampled with replacement. For the threshold sample, we simply take the subset of the returned models from the threshold sample that lie within the 99% confidence region as the sample. The resulting distribution (not shown) demonstrates that while the uncertainties for RSR/OLR are smaller, owing to the smaller covariance term, the uncertainty in TOA imbalance is somewhat similar to the original distribution. Interestingly, the distributions of climate sensitivity are very similar. Another reason for the similarity is that the estimation errors in RSR/OLR from the emulator are now comparable to the size of the confidence region, and so the emulators can no longer distinguish inside the confidence region. Hence, the sample produced is probably quite similar to that using the larger confidence region.

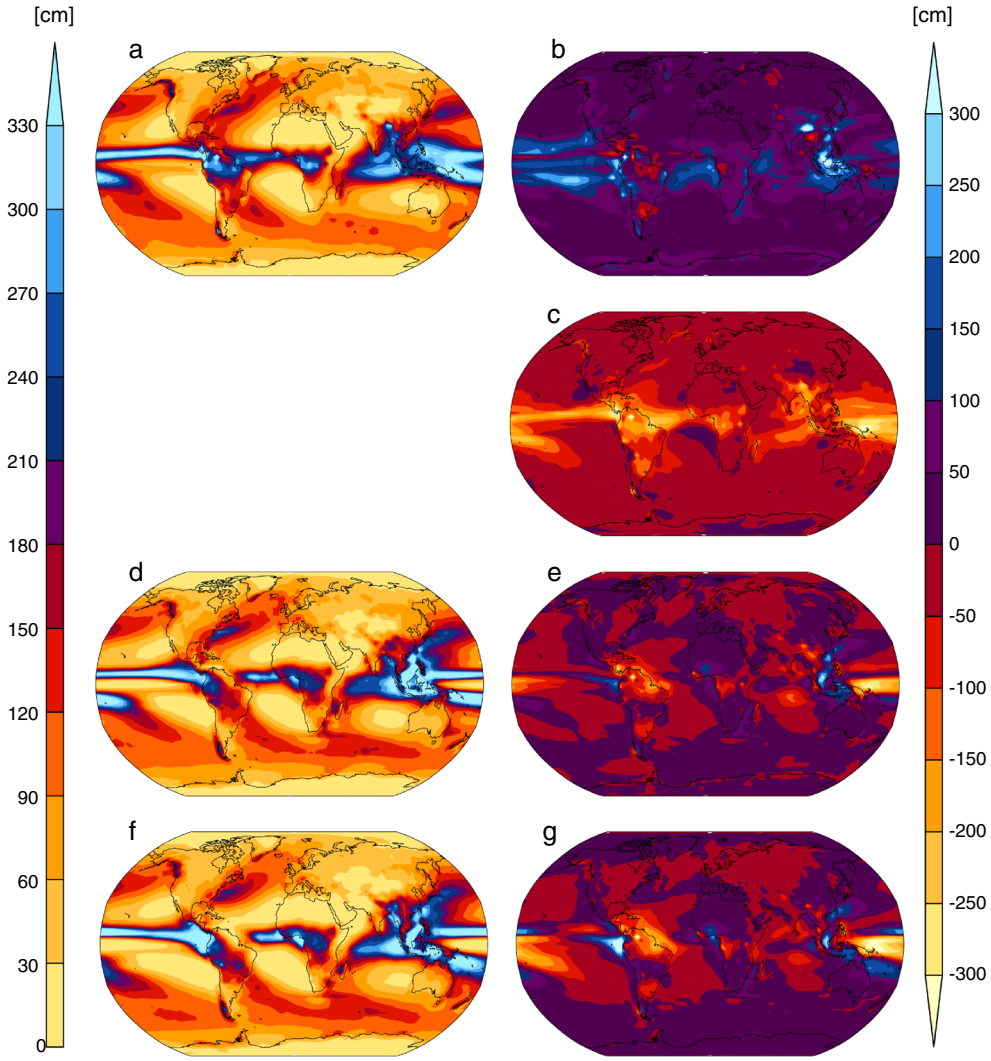

Figure 10. Same as Figure 9 but for annual mean total precipitation.

4. Summary and Conclusion

[49] Within the perturbed physics framework, we aimed to generate an ensemble of atmosphere-ocean coupled GCMs that showed a wide spread of behavior and yet balanced enough to have stable surface climate without applying flux adjustment. Flux adjustment is to be avoided because it alters the dynamics of the model ocean, which is undesirable when the model is used to estimate uncertainty in the ocean heat transfer processes, in which ocean dynamics such as the AMOC is expected to play a large role. We aimed to do this by first generating an O(10^6) parameter perturbation combinations by filling the space of 33 model parameters and performing control simulations. Conditioning on the results from the first ("raw") ensemble, we generated a second ("targeted") ensemble using a statistical emulator, targeting parameter combinations with constrained OLR and RSR and as wide a range as possible of estimated climate sensitivity.

[50] Targeted ensemble members successfully matched the distribution of OLR and RSR to those predicted. The ensemble members exhibited a wide range of behavior in both the atmosphere and the ocean. The range of climate sensitivities in the radiatively balanced members was substantially smaller than that obtained with flux adjustment [*Stainforth et al*., 2005], which spans 1.9–11.5°C (not shown), but is still as large or larger than the range in an ensemble of opportunity [*Solomon et al*., 2007], which has a range of 2–4.5°C. The range of AMOC strength was slightly larger than that in the ensemble of opportunity [*Jackson et al*., 2011]. We conclude that flux adjustment is not a prerequisite for obtaining a broad spread of behavior in a perturbed physics ensemble.

[51] **Acknowledgments.** We thank all participants in the climateprediction.net experiment, as well as the academic institutions and individuals who have helped make the experiment possible. We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy.

References

Breiman, L. (2001), Random forests, *Mach. Learn.*, *45*(1), 5–32.

Brierley, C. M., M. Collins, and A. J. Thorpe (2010), The impact of perturbations to ocean-model parameters on climate and climate change in a coupled model, *Clim. Dyn.*, *34*, 325–343, doi:10.1007/s00382-008-0486-3.

Collins, M., B. B. B. Booth, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb (2006), Towards quantifying uncertainty in transient climate change, *Clim. Dyn.*, *27*(2–3), 127–147, doi:10.1007/s00382-006-0121-0.

Collins, M., C. M. Brierley, M. MacVean, B. B. B. Booth, and G. R. Harris (2007), The sensitivity of the rate of transient climate change to ocean physics perturbations, *J. Clim.*, *20*(10), 2315–2320, doi:http://dx.doi.org/10.1175/JCLI4116.1.

Collins, M., B. B. B. Booth, B. Bhaskaran, G. R. Harris, J. M. Murphy, D. M. H. Sexton, and M. J. Webb (2010), Climate model errors, feedbacks and forcings: A comparison of perturbed physics and multi-model ensembles, *Clim. Dyn.*, *36*, 1737–1766, doi:10.1007/s00382-010-0808-0.

Cox, M. D. (1984), A primitive equation, 3 dimensional model of the ocean. GFDLOcean Group Tech. Rep. 1, GFDL/NOAA, Princeton University, Princeton NJ, USA.

Gordon, C., C. Cooper, C. A. Senior, H. Banks, J. M. Gregory, T. C. Johns, J. F. B. Mitchell, and R. A. Wood (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*(2–3):147–168, doi:10.1007/s003820050010.

Gregory, J. M. et al. (2005), A model intercomparison of changes in the Atlantic thermohaline circulation in response to increasing atmospheric CO_2 concentration. *GRL*, VOL. *32*, L12703, doi:10.1029/2005GL023209.

Houghton, J. T., L. G. Meira Filho, B. A. Callander, N. Harris, A. Kattenberg, and K. Maskell (1996), Climate Change 1995. The Science of Climate Change. Cambridge University Press, Cambridge, UK.

Jackson, L. C., M. Vellinga, and G. R. Harris (2011), 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.

Kopp G., and J. L. Lean (2011), A new, lower value of total solar irradiance: Evidence and climate significance, *Geophys. Res. Lett.*, L01706, doi:10.1029/2010GL045777.

Marotzke, J., and P. H. Stone (1995), Atmospheric transports, the thermohaline circulation, and flux adjustments in a simple coupled model, *J. Phys. Oceanogr.*, *25*(6), 1350–1364, doi:http://dx.doi.org/10.1175/1520-0485(1995)025 < 1350:ATTTCA > 2.0.CO;2.

Murphy, J. M., D. M. H. Sexton, D. N. Barnett, G. S. Jones, M. J. Webb, M. Collins, and D. A. Stainforth (2004), Quantification of modelling uncertainties in a large ensemble of climate change simulations, *Nature*, *430*(7001), 768–772, doi:10.1038/nature02771.

Loeb, N. G., B. A. Wielicki, D. R. Doelling, G. L. Smith, D. F. Keyes, S. Kato, N. Manalo-Smith, and T. Wong (2009), Toward optimal closure of the Earth's top-of-atmosphere radiation budget, *J. Clim.*, *22*, 748–765, doi:10.1175/2008JCLI2637.1.

Lyman, J. M., S. A. Good, V. V. Gouretski, M. Ishii, G. C. Johnson, M. D. Palmer, D. M. Smith, and J. K. Willis (2010), Robust warming of the global upper ocean. *Nature*, *465*(7296), 334–337, doi:10.1038/nature09043.

Neelin, J. D., and H. A. Dijkstra (1995), Ocean–atmosphere interaction and the tropical climatology. Part I: The dangers of flux correction, *J. Climate*, *8*, 1325–1342, doi:http://dx.doi.org/10.1175/1520-0442(1995)008<1325:OAIATT>2.0.CO;2.

Penner, J. E., J. Quaas, T. Storelvmo, T. Takemura, O. Boucher, H. Guo, A. Kirkevag, J. E. Kristjansson, and O. Seland (2006), Model intercomparison of indirect aerosol effects, *Atmos. Chem. Phys.*, *6*, 3391–3405, doi:10.5194/acp-6-3391-2006.

Philip, S. J., M. Collins, G. J. van Oldenborgh, and B. J. J. M. an den Hurk (2009), The role of atmosphere and ocean physical processes in ENSO, *Ocean Sci.*, *6*, 441–459, 2010, doi:10.5194/os-6-441-2010.

Randall, D. A. et al. (2007), Climate Models and Their Evaluation, in Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, p. 607.

Raper, S. C. B., J. M. Gregory, and R. J. Stouffer (2002), The role of climate sensitivity and ocean heat uptake on AOGCM transient temperature response. *J. Climate*, *15*, 124–130, doi:http://dx.doi.org/10.1175/1520-0442(2002)015<0124:TROCSA>2.0.CO;2.

Rougier, J., and D. M. H. Sexton (2007), Inference in ensemble experiments. *Phil. Trans. R. Soc. A*, *365*(1857), 2133–2143, doi:10.1098/rsta.2007.2071.

Rowlands, D. J. et al. (2012), Broad range of 2050 warming from an observationally constrained large climate model ensemble, *Nature Geosci.*, *5*, 256–260, doi:10.1038/ngeo1430.

Shiogama, H. et al. (2012), Perturbed physics ensemble using the MIROC5 coupled atmosphere–ocean GCM without flux corrections: Experimental design and results, *Clim Dyn.*, doi:10.1007/s00382-012-1441-x.

Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor, and H. L. Miller (eds.) (2007). Climate change 2007: The physical science basis, in Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, 2007 Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Sanderson, B. M., C. Piani, W. J. Ingram, D. A. Stone, M. R. Allen (2008), Towards constraining climate sensitivity by linear analysis of feedback patterns in thousands of perturbed-physics GCM simulations. *Clim. Dyn.*, *30*, 175–190, doi:10.1007/s00382-007-0280-7.

Stainforth, D. A. et al. (2005), Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, *433*:403–406, doi:10.1038/nature03301.

Tett, S. F. B., D. J. Rowlands, M. J. Mineter, and C. Cartis (2012), Can Top Of Atmosphere Radiation Measurements Constrain Climate Predictions? Part 2: Climate Sensitivity. Submitted.

Toniazzo, T., M. Collins, and J. Brown (2008), The variation of ENSO characteristics associated with atmospheric parameter perturbations in coupled model, *Clim. Dyn.*, *30*, 643–656, doi:10.1007/s00382-007-0313-2.

Williamson, D., M. Goldstein, L. Allison, A. T. Blaker, P. Challenor, L. Jackson, and K. Yamazaki (2012), History matching for exploring and reducing climate model parameter space using observations, *Clim. Dyn.*, submitted.

Willis, J. K., D. Roemmich, and B. Cornuelle (2004), Interannual variability in upper ocean heat content, temperature, and thermosteric expansion on global scales, *J. Geophys. Res.-Oceans*, *109*(12), doi:10.1029/2003JC002260.

Yamazaki, K. (2008), Exploring the Impact of Ocean Representation on Ensemble Simulations of Climate Change, DPhil Thesis, University of Oxford, pp 205.