

## Error Reduction and Convergence in Climate Prediction

CHARLES S. JACKSON

*Institute for Geophysics, The University of Texas at Austin, Austin, Texas*

MRINAL K. SEN

*Institute for Geophysics, and Department of Geological Sciences, The University of Texas at Austin, Austin, Texas*

GABRIEL HUERTA

*Department of Mathematics and Statistics, University of New Mexico, Albuquerque, New Mexico*

YI DENG\*

*Institute for Geophysics, The University of Texas at Austin, Austin, Texas*

KENNETH P. BOWMAN

*Department of Atmospheric Sciences, Texas A&M University, College Station, Texas*

(Manuscript received 25 June 2007, in final form 21 March 2008)

### ABSTRACT

Although climate models have steadily improved their ability to reproduce the observed climate, over the years there has been little change to the wide range of sensitivities exhibited by different models to a doubling of atmospheric CO<sub>2</sub> concentrations. Stochastic optimization is used to mimic how six independent climate model development efforts might use the same atmospheric general circulation model, set of observational constraints, and model skill criteria to choose different settings for parameters thought to be important sources of uncertainty related to clouds and convection. Each optimized model improved its skill with respect to observations selected as targets of model development. Of particular note were the improvements seen in reproducing observed extreme rainfall rates over the tropical Pacific, which was not specifically targeted during the optimization process. As compared to the default model sensitivity of 2.4°C, the ensemble of optimized model configurations had a larger and narrower range of sensitivities around 3°C but with different regional responses related to the uncertain choice in optimized parameter settings. These results suggest current generation models, if similarly optimized, may become more convergent in their measure of global sensitivity to greenhouse gas forcing. However, this exploration of the possible sources of modeling and observational uncertainty is not exhaustive. The optimization process illustrates an objective means for selecting an ensemble of plausible climate model configurations that quantify a portion of the uncertainty in the climate model development process.

### 1. Introduction

In global climate models (GCMs), unresolved physical processes are included through simplified representations referred to as parameterizations. Parameteriza-

tions typically contain one or more adjustable phenomenological parameters. Parameter values can be estimated directly from theory or observations or by “tuning” the models by comparing model simulations to the climate record. Because of the large number of parameters in comprehensive GCMs, a thorough tuning effort that includes interactions between multiple parameters can be very computationally expensive.

Models may have compensating errors, where errors in one parameterization compensate for errors in other parameterizations to produce a realistic climate simulation (Wang 2007; Golaz et al. 2007; Min et al. 2007; Murphy et al. 2007). The risk is that, when moving to a new climate regime (e.g., increased greenhouse gases),



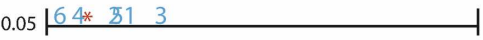



---

\* Current affiliation: School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, Georgia.

---

*Corresponding author address:* Charles Jackson, Institute for Geophysics, The John A. and Katherine G. Jackson School of Geosciences, The University of Texas at Austin, J. J. Pickle Research Campus, Bldg. 196 (ROC), 10100 Burnet Rd. (R2200), Austin, TX 78758-4445.  
E-mail: charles@ig.utexas.edu

TABLE 1. Underconstrained CAM3.1 parameters with default values (asterisk) and ranges of prior distributions and posterior uncertainties represented by the best-performing models within each of the six independent convergence chains (1–6).

Parameter	Definition	Value Ranges
<b>RHMINL</b> [%/100]	Low cloud critical relative humidity	0.80    0.95
<b>RHMINH</b> [%/100]	High cloud critical relative humidity	0.60    0.90
<b>ALFA</b> [fraction]	Initial cloud downdraft mass flux	0.05    0.60
<b>TAU</b> [hours]	Consumption rate of CAPE	0.5    8.0
<b>ke</b> [(kg m <sup>-2</sup> s <sup>-1</sup> ) <sup>-1/2</sup> s <sup>-1</sup> ]	Environmental air entrainment rate	3.0e-6    10.0e-6
<b>c0</b> [m <sup>-1</sup> ]	Precipitation efficiency	3.0e-3    6.0e-3

the errors may no longer compensate. This leads to uncertainty in climate change predictions. The known range of uncertainty of many parameters allows a wide variance of the resulting simulated climate (Murphy et al. 2004; Stainforth et al. 2005; M. Collins et al. 2006). The persistent scatter in the sensitivities of models from different modeling groups, despite the effort represented by the approximately four generations of modeling improvements, suggests that uncertainty in climate prediction may depend on underconstrained details and that we should not expect convergence anytime soon. The question addressed here is whether a more systematic approach to constraining parametric uncertainties would be enough to allow independently developed models to become more convergent in their predictions of global change.

## 2. Optimized tuning and uncertainty quantification

The leading cause of the intermodel differences in sensitivity to CO<sub>2</sub> forcing is related to differences in the treatment of clouds (Cess et al. 1990, 1996; Held and Soden 2000; Colman 2003; Webb et al. 2006). We hypothesize that the primary source of uncertainty for the National Center for Atmospheric Research (NCAR) Community Atmosphere Model version 3.1 (CAM3.1; W. D. Collins et al. 2006) is related to arbitrary aspects of selecting precise values for six parameters associated with the model's parameterization of clouds and convection (Table 1).

Following Jackson et al. (2004), Bayesian inference is

used along with a stochastic importance sampling algorithm, Multiple Very Fast Simulated Annealing (MVFSA), to efficiently identify the regions of model parameter space of CAM3.1 that minimize systematic differences with 15 sets of observational constraints given by regional and seasonal climatologies of satellite and reanalysis data products from January 1990 to February 2001 (Mu et al. 2004). A number of sensitivity experiments have been performed with the parameters in Table 1 and other parameters to establish the importance of each parameter to simulated climates. We select candidate values for each of the parameters from an initially uniform probability prior distribution with the ranges specified to reflect realistic possibilities given sensitivity experiments (Murphy et al. 2004; Stainforth et al. 2005; Mu et al. 2004), the history of values used in climate model development as marked within the source code, and examination of values used within other climate models using the same Zhang and McFarlane (1995) parameterization for convection as CAM3.1.

Bayesian inference with the MVFSA stochastic sampling algorithm estimates a “posterior” joint probability distribution for the uncertain parameter sets  $\mathbf{m}$  given a “prior” probability for selecting reasonable values for  $\mathbf{m}$ . We include in our inferences of parametric uncertainty a parameter  $S$ , which determines, in part, which model configurations may be deemed acceptable. The parameter  $S$  has its own prior that provides information about observational and other uncertainties that are hard to quantify within the metric of model skill  $E(\mathbf{m})$ :

$$\text{posterior}(\mathbf{m}, S) = \frac{\exp[-SE(\mathbf{m})]}{\int \int \exp[-SE(\mathbf{m})] \times \text{prior}(\mathbf{m}, S) d\mathbf{m} dS} \text{prior}(\mathbf{m}, S). \quad (1)$$

We refer to the metric of model skill  $E(\mathbf{m})$  as a cost function. It provides a weighted measure of mean squared differences between model predictions and a set of observational constraints. This cost function weights different sources of uncertainty through an inverse of the data covariance matrix  $\mathbf{C}^{-1}$ ,

$$E(\mathbf{m}) = \frac{1}{2N} \sum_{i=1}^N [[\mathbf{d}_{\text{obs}} - g(\mathbf{m})]^T \mathbf{C}^{-1} [\mathbf{d}_{\text{obs}} - g(\mathbf{m})]]_i \quad (2)$$

where  $\mathbf{d}_{\text{obs}}$  is the set of observations that may be compared directly with model predictions  $g(\mathbf{m})$  and superscript T indicates a matrix transpose. Note that index  $i$  in Eq. (2) applies to the whole expression in the brackets such that model–observational data comparisons are made from  $N$  separate regions and seasons (see section 2c). Thus  $E(\mathbf{m})$  does not explicitly account for the potentially significant correlations among these observational constraints. We discuss below how we take these correlations into account for limiting regions of acceptability through our choice of scaling factor  $S$ , which, in effect, is a modifier of the data covariance. This form of the cost function is the appropriate form for assessing more rigorously the statistical significance of modeled–observational differences when it is known that sources of model and observational uncertainty are Gaussian. We have plotted the distribution of errors that arise from internal model variability and confirmed that the resulting distributions are consistent with the Gaussian assumption (not shown).

The MVFSA sampling algorithm works by taking random steps in parameter space and at each step running an 11-yr climate model integration, quantifying the differences between simulated and observed climate in terms of a scalar skill score or “cost,” and reselecting parameter values based on the skill score so that the algorithm progressively moves toward regions of the global parameter space that minimize modeling errors. Candidate parameter values are initially chosen from a uniform prior. However, as sampling progresses, candidate parameter values are chosen from a Cauchy distribution whose width becomes increasingly focused on the last accepted model (Ingber 1989). This convergence chain may be repeated numerous times starting from randomly chosen points in parameter space to make inferences about uncertainty. With sufficient sampling, MVFSA provides a computationally tractable approximation to a posterior joint probability distribution of uncertainties comparable to Markov Chain Monte Carlo (MCMC) algorithms (Sen and Stoffa 1996; Jackson et al. 2004; Villagran et al. 2008).

One of the challenges in attaining optimal efficiency in some MCMC sampling strategies such as the Metropolis–Hastings version of the Gibbs’ sampling algorithm (Hastings 1970) is the choice of the step size that is taken through parameter space. It is often not possible to know what this optimal step size should be. MVFSA uses a range of step sizes to enable it to focus on sampling only those regions that are relevant to representing uncertainties. Our experience suggests this flexibility also enables the algorithm to be useful and efficient for a broad range of problems. There also exist other adaptive sampling algorithms with correct ergodic properties that have been shown to be flexible and efficient for estimating unimodal posteriors. These algorithms automate and optimize for the ideal step size and result in superior performance over the more traditional, Metropolis–Hastings-type Gibbs’ samplers (e.g., Haario et al. 2006; Villagran et al. 2008).

MVFSA selects a distribution of model configurations consistent with prior estimates of sources of uncertainty. In our case we consider the uncertainty that comes from representing climate from a relatively short 11-yr time series as well as structural and observational uncertainties contributing to the systematic biases that exist between the model and the selected observational targets. Considered here are six convergence chains of a single model and we examine the configurations with the best skill scores after  $\sim 41$  steps within each chain. Because we are only considering relatively few chains, the present analysis is limited to discussing the uncertainty in identifying the optimal parameter settings. However, the results show that the selected ensemble of six optimized model configurations is more broadly representative of estimates of the posterior distribution. The analysis is meant to model the uncertainty in climate model development whose goal is the creation of a single model that best represents observed climate. The uncertainties in observational, modeling, and structural uncertainty that are included or represented within the cost function will impact the algorithms ability to discern what model is best and, with sufficient sampling, provides an objective basis for selecting an ensemble of plausible climate model configurations that represents the full range of model development uncertainty given the uncertainties considered.

#### a. Experiment design

Each experiment testing the sensitivity of CAM3.1 to combined changes in select parameters follows an experimental design in which the model is forced by observed sea surface temperatures (SST) and sea ice for an 11-yr period (March 1990 through February 2001).

The model includes 26 vertical levels and uses an approximately  $2.8^\circ$  latitude by  $2.8^\circ$  longitude (T42) resolution. For the experiments testing the sensitivity to a doubling of atmospheric  $\text{CO}_2$  concentrations, CAM3.1 is coupled to a slab ocean with prescribed heat flux adjustments, calculated separately for each configuration, such that each model reproduces the observed monthly climatological sea surface temperatures without explicitly accounting for ocean dynamics. Thus the CAM3.1–slab ocean model may only represent the thermodynamic and not the dynamic response of the ocean to changes in  $\text{CO}_2$  forcing. A control simulation of modern climate is made from a 40-yr-long integration of the CAM3.1–slab ocean model. Doubled  $\text{CO}_2$  experiments are also integrated for 40 yr. We use the final 20 yr for analysis. This process was repeated for the default and six alternate configurations.

### b. Observational constraints

Observational constraints include satellite, instrumental, and reanalysis data products. The fields selected were chosen because of the existence of corresponding instrumental or reanalysis data products; they provide good constraints on top of the atmosphere and surface energy budgets, and they are fields that are commonly used to evaluate model performance. The segments from observations were chosen to overlap the years and months of the experiment. The exception was the Earth Radiation Budget Experiment (ERBE) measurements of top of the atmosphere radiative balances which included years 1985–89 (Barkstrom et al. 1989). We also included a term constraining the global net radiative balance at the top of the atmosphere. We had intended to give this a target of  $0.3 \text{ W m}^{-2}$  in order to compensate for the approximately  $0.3 \text{ W m}^{-2}$  brightening that typically occur for the model when it is coupled to a slab ocean (P. Rasch, NCAR, 2004, personal communication). However we mistakenly imposed this constraint without area weighting, resulting in optimized configurations that are  $4\text{--}7 \text{ W m}^{-2}$  out of balance. The size of the imbalance is comparable to observational uncertainty, but model development efforts typically try to keep this number small in order to minimize long term trends in deep water temperature when the atmosphere is coupled to an ocean GCM. We explore the implications of this error on our results within the discussion section. Below is a list of fields that were included within the cost function and the corresponding data targets. All fields, including the data constraints, were seasonally averaged [December–February (DJF), March–May (MAM), June–August (JJA), September–November (SON)] over the time interval indicated.

- 1) Low-level clouds, 1990–2001, International Satellite Cloud Climatology Project (ISCCP) satellite observations (Rossow et al. 1991).
- 2) Midlevel clouds, 1990–2001, ISCCP satellite observations (Rossow et al. 1991).
- 3) High-level clouds, 1990–2001, ISCCP satellite observations (Rossow et al. 1991).
- 4) Shortwave radiation to surface, 1990–2001, National Centers for Environmental Prediction (NCEP) reanalysis data (Kalnay et al. 1996; Kistler et al. 2001).
- 5) Net shortwave top, 1985–89, ERBE satellite observations (Barkstrom et al. 1989).
- 6) Net longwave top, 1985–89, ERBE satellite observations (Barkstrom et al. 1989).
- 7) Two-meter air temperature, 1990–2001, NCEP reanalysis data (Kalnay et al. 1996; Kistler et al. 2001).
- 8) Surface sensible heat flux, 1990–2001, NCEP reanalysis data (Kalnay et al. 1996; Kistler et al. 2001).
- 9) Surface latent heat flux, 1990–2001, NCEP reanalysis data (Kalnay et al. 1996; Kistler et al. 2001).
- 10) Relative humidity (zonal mean), 1990–2001, NCEP reanalysis data (Kalnay et al. 1996; Kistler et al. 2001).
- 11) Air temperature (zonal mean), 1990–2001, NCEP reanalysis data (Kalnay et al. 1996; Kistler et al. 2001).
- 12) Zonal winds (zonal mean), 1990–2001, NCEP reanalysis data (Kalnay et al. 1996; Kistler et al. 2001).
- 13) Sea level pressure, 1990–2001, NCEP reanalysis data (Kalnay et al. 1996; Kistler et al. 2001).
- 14) Precipitation, 1990–2001, Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP) instrumental record (Xie and Arkin 1996, 1997).

### c. Definition of cost function

The cost function used to evaluate model skill [Eq. (2)] follows the treatment of Mu et al. (2004) in which squared differences between model predictions and observations are projected onto a truncated set of empirical orthogonal functions (EOFs) representing larger spatial regions of correlated year-to-year variability. The EOFs are generated from the seasonal interannual variability contained within the 900-yr-long time series from the b30.004 control integration of NCAR Community Climate System Model, version 3 (CCSM3; W. D. Collins et al. 2006). Each EOF eigenvalue is a measure of the variance for each mode of variability and provides a way to weight the significance of any discrepancies between observations and model predic-

tions. Because these discrepancies need to be represented by a sum of EOFs, the 11 fields defined on a latitude–longitude grid have been subdivided into six 30° nonoverlapping latitude bands. The three zonally averaged fields were subdivided by hemisphere. Moreover, the analysis is performed with seasonal means (DJF, MAM, JJA, and SON) so that the cost function can constrain the amplitude of the seasonal cycle. The only exception to regional and seasonal components of the cost function is the constraint on global mean (annual mean) net radiative balance at the top of the atmosphere. Therefore there are a total of 289 components of the cost function that include 11 latitude–longitude fields over 6 regions and 4 seasons, 3 latitude–height fields over 2 regions and 4 seasons, and one field representing the global radiative balance. The total cost function is mostly a straight averaging of these 289 components with the exception of the three cloud levels which were weighted as a single cloud field. The fields with the largest (smallest) cost values also tend to be fields with the smallest (largest) variability (Fig. 2, component cost values for the default model is shown in parentheses). Thus model-data distance for each field is expressed in terms of the size of interannual variability for that field.

#### d. Renormalization factor “ $S$ ”

To select candidate model configurations that represent the intended uncertainties, one needs the cost function to be normalized with respect to these uncertainties. That is, the MVFSA algorithm is designed to search through candidate parameter sets that are within a certain cost function distance from the global minimum, passing over places that are notably badly performing and searching more thoroughly where the performance is acceptable. A proper normalization is insured through a two-step process: First, for each field, region, and season, the components of the cost function are scaled in such a way that the effects of natural variability resulted in the same 1-unit standard deviation range in cost values over a large number of control experiments that only differed by their initial conditions. For this purpose 83 control experiments of the default configuration CAM3.1 were run with different initial conditions. The second step in the normalization process is to consider correlations that exist among the cost components themselves. These correlations were found to be quite significant leading to a reduction in the effects of natural variability on the cost function from  $\sigma = 1$  unit to only  $\sigma = 0.12$  cost units. One may compensate for this omission in the cost function by rescaling the cost function using scaling factor  $S = \sigma^{-1}$  (i.e.,  $S = 8.33$ ; see Gelman et al. 2004), which has the

effect of focusing sampling around the regions of maximum likelihood. However, rather than make  $S$  a constant factor, we allow the prior distribution of  $S$  to scale inversely with model skill as measured by  $E(\mathbf{m})$ . If the model could provide a perfect match to observations, then the effects of internal variability would be the main source of uncertainty in discerning the goodness of fit between one model configuration and another. However, all climate models show significant compensating errors and systematic biases that give rise to an irreducible component of  $E(\mathbf{m})$  (McWilliams 2007). Therefore by scaling  $S$  inversely with  $E(\mathbf{m})$  the regions of acceptability are inevitably broadened. The convenient functional form of the prior for  $S$ , being a modifier of the inverse of the data covariance matrix within Eq. (1), is a gamma distribution function. During the course of sampling,  $E(\mathbf{m})$  is allowed to modify the mean and variance of the gamma distribution according to

$$\bar{S} = \frac{\alpha}{\beta + E(\mathbf{m})}, \quad (3)$$

$$\text{var}(S) = \frac{\alpha}{[\beta + E(\mathbf{m})]^2}. \quad (4)$$

Here,  $\alpha$  and  $\beta$  are parameters that control the mean and variance of the gamma distribution using information about the effects of natural variability on  $E(\mathbf{m})$ . Specifically, if  $E(\mathbf{m})$  were properly normalized with respect to this variability,  $\alpha$  and  $\beta$  would be equal, with a variance in  $S$  determined from the uncertainty in estimating the mean value of  $S$  from a limited number of control experiments that we have run (number = 83). Because the emphasis in the present analysis concerns the uncertainty of near-optimal choices of model parameters from a limited sampling of the posterior distribution (i.e., the uncertainty in the optimization), we defer to future work to discuss the details of the treatment of  $S$  in our approach to quantifying the effects of observational and other sources of uncertainty in our estimates of parametric uncertainties.

### 3. Results

Of the 518 experiments that were completed, 332 achieved skill scores that were smaller than the default case and include a broad range of parameter values. Six configurations were selected for further analysis, one from each convergence chain based on the maximum skill after  $\sim 41$  experiments (Table 1). Sampling then continued in order to consider the degree to which these six samples were representative of the 332 sample posterior distribution of parametric uncertainties (Fig. 1). The six parameter sets are broadly representative of the posterior distribution with particularly wide-

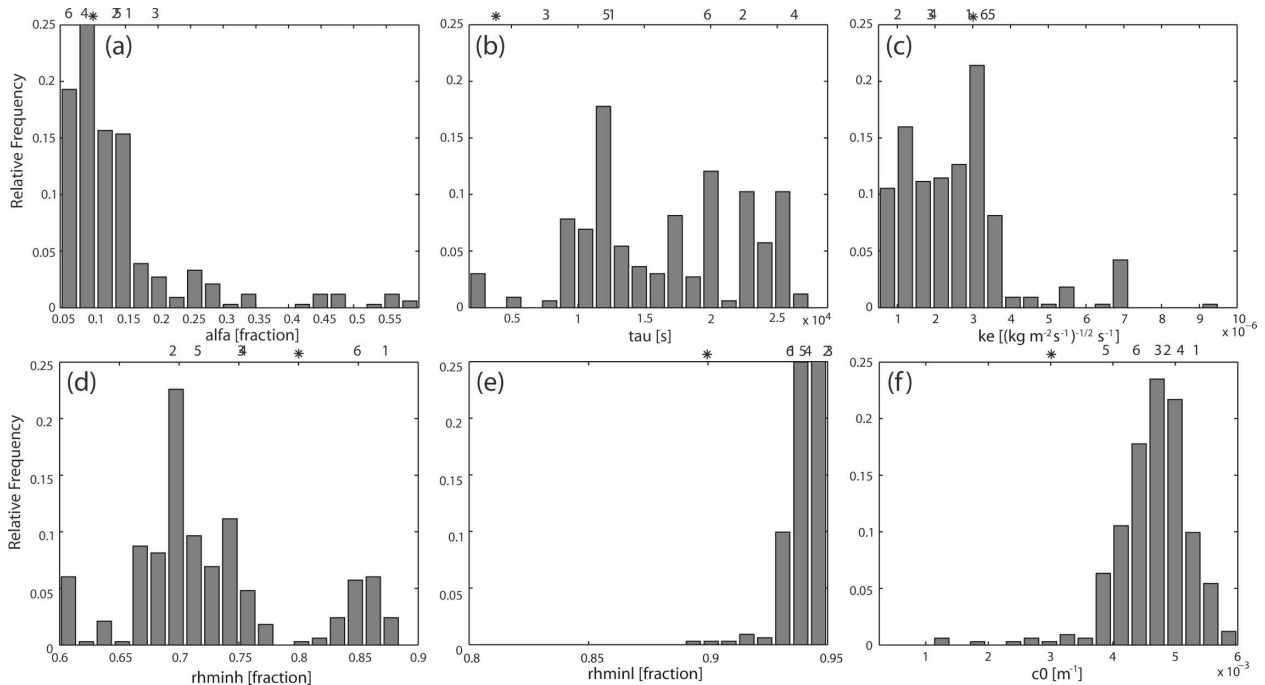


FIG. 1. Posterior probability for 6 parameters of CAM3.1 important to clouds and convection (see Table 1): (a) ALPHA, (b) TAU, (c) ke, (d) RHMINH, (e) RHMINL, and (f) c0. The histograms are derived from the 332 experiments whose cost values were the same or showed an improvement over the default model configuration. The parameter values of the default model are given by an asterisk (\*). The values of the top performing 6 parameter sets are labeled by the particular line number that produced them.

ranging values selected for parameters TAU and RHMINH and more tightly constrained values for ALFA, ke, RHMINL, and c0.

Relative to the default configuration, systematic errors of the optimized model configurations were reduced by an average of 7% (Fig. 2a). There were consistent reduction in errors related to low-level clouds (averaging 3% improvement, Fig. 2b), shortwave radiation reaching the surface (averaging 14% improvement, Fig. 2e), surface latent heat flux (averaging 4% improvement, Fig. 2k), zonal mean air temperature (averaging 4% improvement, Fig. 2m), zonal winds (averaging 6% improvement, Fig. 2n), and sea level pressure (averaging 5% improvement, Fig. 2o), and precipitation (12% improvement, Fig. 2p). Some fields became worse such as midlevel clouds (averaging -10% degradation, Fig. 2c), high-level clouds (averaging -6% degradation, Fig. 2d), and net shortwave radiation at the top of the atmosphere (averaging -7% degradation, Fig. 2f). There were mixed results or relatively minor changes in skill for the remaining fields. Thus, the similar cost values achieved for all six optimal model configurations is achieved through different compromises in model skill for predicting constrained fields.

The optimization process also provided unantic-

ipated performance gains in the frequency distribution of hourly rain rates (Fig. 3). The default configuration of CAM3.1 as well as many other climate models typically drizzles too often with little ability to simulate observed heavy rainfall events (Deng et al. 2007; Wilcox and Donner 2007). Five of the six optimized CAM3.1 configurations were able to capture the observed distribution of heavy and light rainfall events of the tropical Pacific ITCZ region. Because the variability of rainfall rates is not targeted in the model skill scores, these improvements could have only been achieved indirectly through the long-term seasonal mean constraints that were included. There also appears to be a correlation between model configurations with larger values of the rate at which clouds consume available potential energy (TAU), and the emergence of extreme rainfall rates.

Tests were performed to evaluate the extent to which the parametric uncertainties remaining among the optimized configurations would affect the model's equilibrium response to a doubling of atmospheric CO<sub>2</sub> concentrations. The default CAM3.1 configuration sensitivity of 2.4°C near-surface global mean annual mean air temperature change is on the lower end of sensitivities relative to the scatter among two generations of models (Fig. 4). However, after optimization, five of the

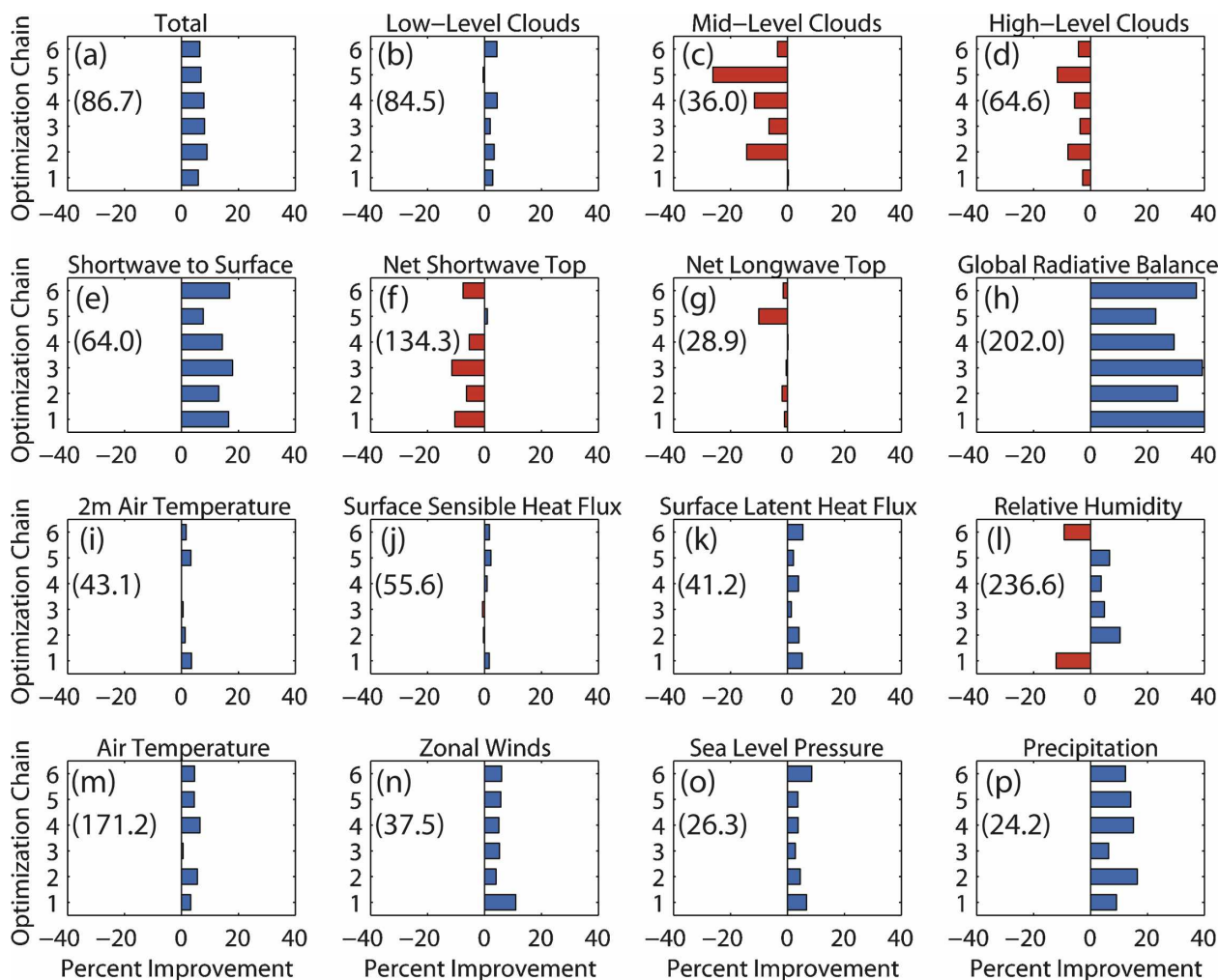


FIG. 2. Impact of stochastic optimization on CAM3.1 skill in predicting seasonal and regional climatologies of 15 fields commonly used for climate model development. (a)–(p) Each bar represents the change in skill of one of six optimized configurations of CAM3.1 as compared to the default configuration. The size of each bar indicates the percent improvement (blue) or degradation (red) in skill score for (a) a weighted average total and (b)–(p) the 15 observational targets. The skill score for the default configuration is given in parentheses.

six optimized configurations increased in sensitivity to around  $3^{\circ}\text{C}$  and  $3.1^{\circ}\text{C}$  sensitivity with the remaining optimized configuration having an even larger sensitivity at  $3.4^{\circ}\text{C}$ . The uncertainty in evaluating a model's sensitivity from the 20-yr-long experiments is less than  $0.1^{\circ}$ . Therefore, the shift in the model's sensitivity is significant. The narrow spread in sensitivities among the six-member ensemble, despite the wide range in parameter values considered, suggests that either the observational constraints that were placed on the selection of parameter values were informative enough to constrain the global balance of internal feedbacks that control the model's response to the change in radiative forcing, or the selected parameters are not the primary sources of uncertainty that contribute to the  $2^{\circ}$ – $6^{\circ}\text{C}$

range in sensitivities seen in multimodel intercomparisons (LeTreut and McAvaney 2000; Cubasch et al. 2001; Alley et al. 2007).

The convergent predictions on a global scale occurred with slightly different physical balances, resulting in a significant spread of predictions at regional scales (Fig. 5). Many of the regional differences among the model configurations occurred within the tropics where the parameters considered have their largest influence. The parameters also affect changes in the model's response in the mid-high latitudes. Most notably, the  $\sim 25\%$  uncertainty in near-surface air temperatures southwest of Greenland is associated with large changes in surface wind stress among the different model versions. These wind stress changes appear to be

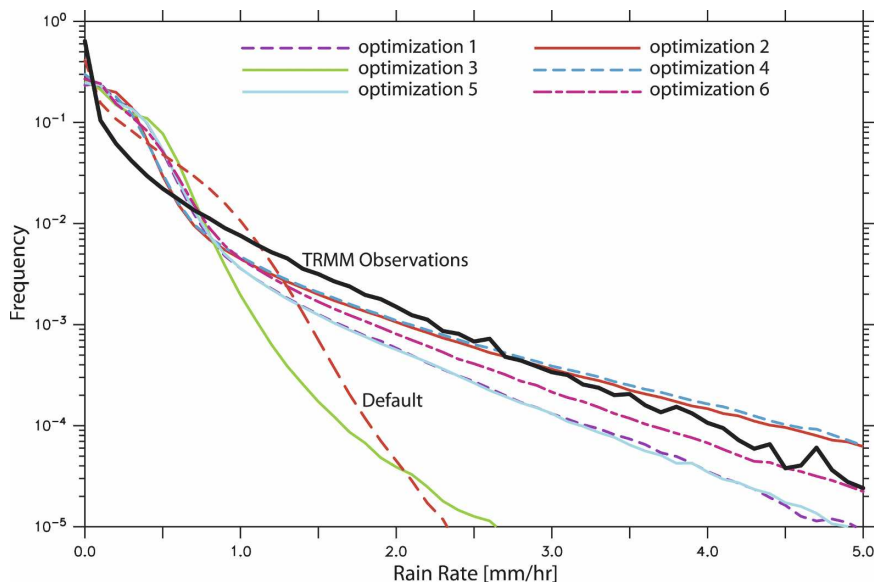


FIG. 3. Histograms of hourly rain rates from the tropical Pacific ITCZ ( $12^{\circ}\text{N}$ – $0^{\circ}$ ,  $130^{\circ}\text{E}$ – $90^{\circ}\text{W}$ ). The default configuration of CAM3.1 fails to capture the heavy rainfall rates seen in the TRMM observations (Deng et al. 2007). Most of the optimized versions of CAM3.1 provided a much closer match to observations despite the fact that these observations of high-frequency rain variability were not included in the stochastic inversion cost function.

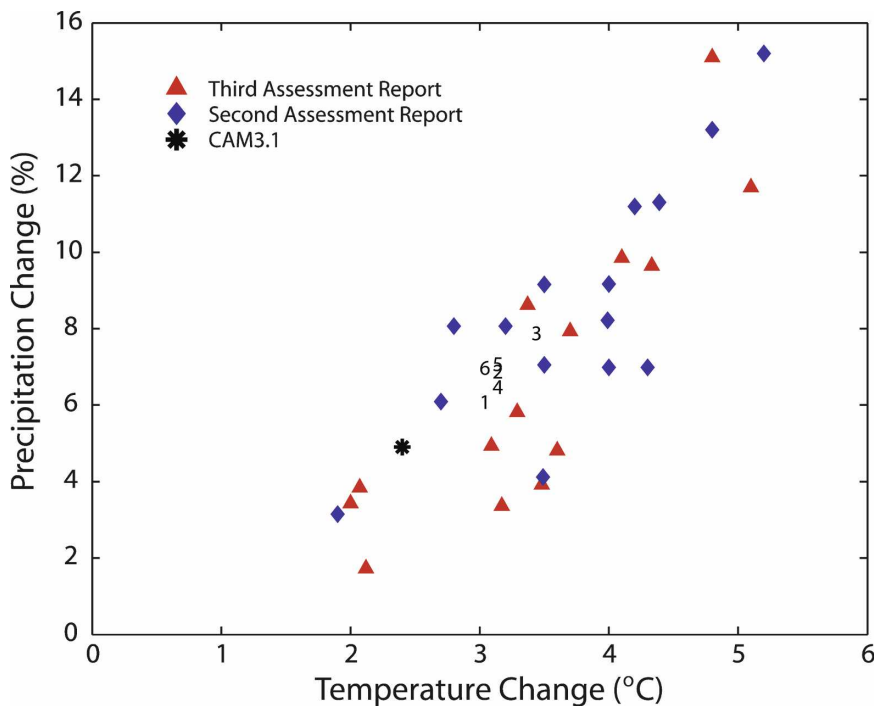


FIG. 4. Equilibrium global and annual mean response of surface air temperature and precipitation change to a doubling of atmospheric  $\text{CO}_2$  concentrations. Blue diamonds and red triangles show response of atmospheric models used within the Second and Third Assessment Reports of the Intergovernmental Panel on Climate Change (IPCC; data reproduced from Fig. 9.18 of Cubasch et al. 2001). The corresponding sensitivity of the CAM3.1 model is indicated by the asterisk. Each of the six optimized CAM3.1 model sensitivities are indicated by number.



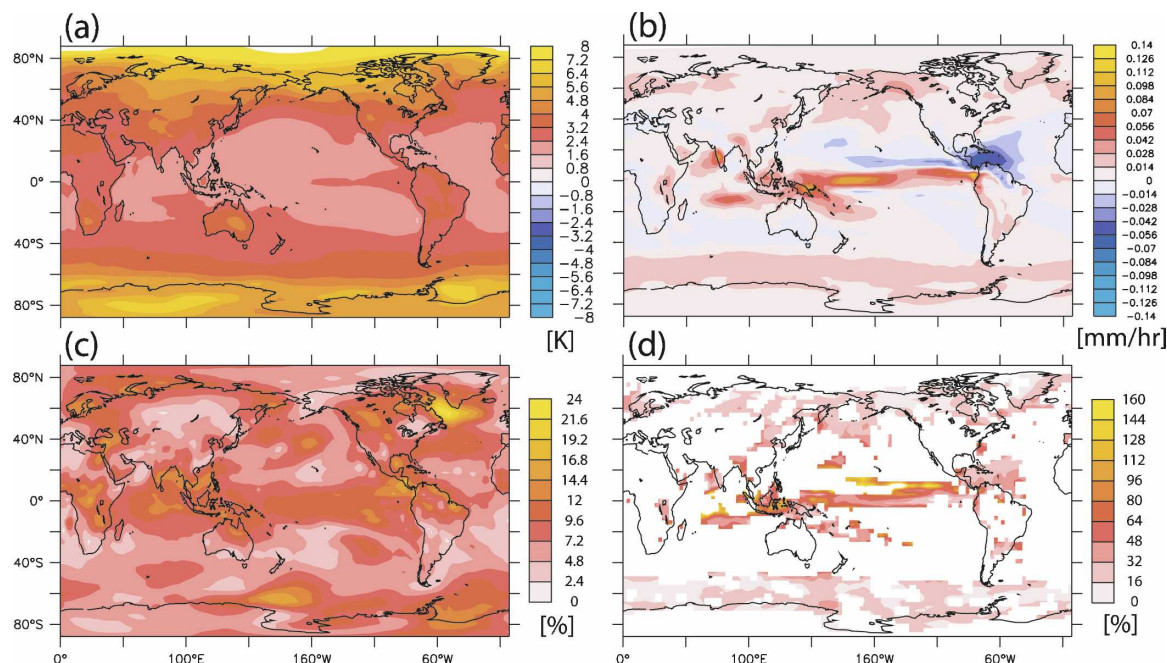


FIG. 5. Regional response and uncertainties of predictions from an ensemble of optimized CAM3.1 configurations included within Fig. 3 to a doubling of atmospheric  $\text{CO}_2$  concentrations. (a) Response of annual mean surface air temperature. (b) Response of annual mean precipitation. (c) Regional uncertainties in predictions of annual mean surface air temperature calculated from the range in regional responses divided by the ensemble mean. (d) Regional uncertainty in predictions of annual mean precipitation. Uncertainties are expressed in term of the percent range in responses that occurred among the ensemble members to the ensemble average response. Areas with no color in panels (c) and (d) indicate forced response is more than 1.2 times smaller than a signal that can be explained by natural variability.

affecting the production and export of sea ice from the Labrador Sea and their respective radiative feedbacks. The largest uncertainties are associated with predictions of tropical rainfall where the ensemble spread accounts for upward of 160% uncertainty in the predicted shifts in the north–south position of the intertropical convergence zone.

We have explored the sensitivity of the results to changes in the constraint on the global radiative balance. Without the area weighting on the global mean radiative balance, the six model configurations analyzed were 4–7  $\text{W m}^{-2}$  out of balance. We reran three convergence chains (250 additional experiments) with area weighting of the global radiative balance. We selected three of the top performing models, one from each chain, for further analysis. The results are broadly similar insofar as we found 1) a 7% maximum reduction in the cost function, 2) a wide range in parameter value combinations, 2) a range of both improvements and degradations in particular components of the cost function, 3) dramatic improvements in capturing observed rain-rate extremes over the tropical Pacific, and 4) a narrow spread in sensitivities to a doubling of atmospheric  $\text{CO}_2$  concentrations. However, in detail, the

global radiative balance constraint alters particular results. For instance, marginal profiles of the posterior probability derived from this new ensemble were shifted, selecting model configurations that were previously deemed unlikely. Aside from zonal mean air temperatures, precipitation, and net shortwave radiation at the top of the atmosphere, which all show significant  $\sim 10\%$  reductions in component cost values, there is less consistency from our previous results as to which fields improved or degraded. The three new configurations' sensitivity to a doubling of atmospheric  $\text{CO}_2$  concentrations are 2.6°, 2.7°, and 2.8°C, or slightly smaller than the predominantly 3°C sensitivity for configurations previously discussed.

#### 4. Discussion and conclusions

The MVFSA sampling strategy to quantify uncertainties differs in some potentially important ways from previous or proposed approaches that make different assumptions about the smoothness and linearity of the climate model response to uncertain parameter choices (Murphy et al. 2004; Annan and Hargreaves 2004; Stainforth et al. 2005; M. Collins et al. 2006; Murphy et

al. 2007; Annan and Hargreaves 2007). The smoothness of the response surface itself will depend on the treatment of uncertainties between model predictions and observational data within the cost function, which itself can be a matter of scientific judgment. The gulf between the two is large enough to call into question the assumption that the relative likelihood of any given model configuration can be measured by an exponential function of the cost function [e.g., Eq. (1)] (Frame et al. 2007; Stainforth et al. 2007). The MVFSA sampling strategy has the capacity to resolve limited regions of parameter space that may be missed by strategies that depend on interpolation or emulation from a limited number of experiments (Annan and Hargreaves 2007; Murphy et al. 2007; Rougier and Sexton 2007). It is not clear from the present results that are based on a limited number of convergence chains whether other strategies would have been sufficient. From a model development perspective that is perhaps most interested in identifying points of maximum likelihood (minima in the response surface), it was quite difficult to find parameter combinations that improved model skill ( $<20\%$ , in the case with the constraint on global radiative balance specified correctly) and that was with a focused sampling strategy.

Currently all climate models have large systematic modeling errors with respect to observations consistent with those shown in Fig. 2. Stochastic sampling of six parametric uncertainties related to clouds and convection was only able to improve the metric of model skill by 7%–10% although the optimized model did capture observed extreme rain rates over the tropical Pacific, which was not specifically targeted. Thus there remains a significant portion of model–observational data difference that may be considered “irreducible” without fundamental improvements in the parameterizations of subgrid-scale processes (McWilliams 2007). These errors do not necessarily imply that climate models provide poor predictions of forced change if the errors do not affect the balance of processes that control the strength or structure of the forced response and feedback patterns. However, it has been always been difficult to know how significant existing systematic errors are to affecting the prediction problem. For at least the global scales, the result of constraining parametric uncertainties with observations suggests that current generation climate models may result in more convergent measures of sensitivity to  $\text{CO}_2$  forcing. Such a convergence, however, would not necessarily imply that model predictions are correct, especially without agreement of change at regional scales. The apparent convergence we found may be an artifact of undersampling sources of uncertainty that were not included in the present

analysis or the limited number of convergence chains considered.

Although the convergence in Fig. 4 seems to contradict results of parametric uncertainty studies with the Hadley Centre Atmospheric Model version 3 (HadAM3; Stainforth et al. 2005), which shows the possibility for wide ranges in sensitivity between  $1.5^\circ$  and  $11^\circ\text{C}$ , both CAM3.1 and HadAM3 show that the model configurations with the smallest systematic errors have sensitivities within a half a degree of  $3^\circ\text{C}$ . The most improved models in both cases occur with a similar 10% reduction in skill score values. The limited sampling of the CAM3.1 is not meant to be representative of the range of uncertainties admitted by observational and structural uncertainties. However, the difference in extreme sensitivities reflected in these two parametric uncertainty studies illustrates the importance of the subjectively determined skill score in determining whether a given nonoptimal configuration is deemed acceptable. For instance, according to the analysis of Knutti et al. (2006), HadAM3 configurations with the highest sensitivities also have unrealistically large seasonal cycles, a feature that was missed by the choice to focus model evaluations on annual mean quantities. Another difference in skill score definitions that may be significant is our choice to include a constraint on the global radiative balance at the top of the atmosphere. Other groups have chosen to not impose this constraint because they felt it would be too limiting of the possible plausible configurations (Annan and Hargreaves 2004; Murphy et al. 2007). We also find that the global radiative balance is extremely sensitive to the selected parameter choices, yet the range in plausible parameter choices remains broad. We note that many climate models are tuned to have a global radiative balance, but yet still have large differences in their measure of sensitivity to  $\text{CO}_2$  forcing. Although this constraint limits the parameter combinations that we would deem plausible, we do not suspect that this one constraint is solely responsible for the shift and narrow range in sensitivities among the optimized configurations of CAM3.1. The fact that these sensitivities were within a degree of optimized versions of the HadAM3 model sensitivity despite the many differences in the models’ designs and the ways the models are compared to observations may suggest that this apparent approach to convergence is a robust result. Testing the level of convergence among other similarly optimized models would provide further support for the generality of this inference.

*Acknowledgments.* We thank Dmitri Matrosov for his help in developing software tools for managing model experiments and analysis. Support to C.S.J.,

M.K.S., and G.H. was provided by NSF under the Collaboration in Mathematical Geosciences program Grant OCE-0415738. Support to Y.D. was provided by the Jackson School of Geosciences Initiative Program.

## REFERENCES

- Alley, R. B., and Coauthors, 2007: Summary for policymakers. *Climate Change 2007: The Physical Science Basis*, S. Solomon et al., Eds., Cambridge University Press, 1–18.
- Annan, J. D., and J. C. Hargreaves, 2004: Efficient parameter estimation for a highly chaotic system. *Tellus*, **56A**, 520–526, doi:10.1111/j.1600-0870.2004.00073.x.
- , and —, 2007: Efficient estimation and ensemble generation in climate modelling. *Philos. Trans. Roy. Soc. London*, **A365**, 2077–2088, doi:10.1098/rsta.2007.2067.
- Barkstrom, B. R., E. Harrison, G. Smith, R. Green, J. Kibler, R. Cess, and The ERBE Science Team, 1989: Earth Radiation Budget Experiment (ERBE) archival and April 1985 results. *Bull. Amer. Meteor. Soc.*, **70**, 1254–1262.
- Cess, R. D., and Coauthors, 1990: Intercomparison and interpretation of climate feedback processes in 19 atmospheric general circulation models. *J. Geophys. Res.*, **95**, 16 601–16 615.
- , and Coauthors, 1996: Cloud feedback in atmospheric general circulation models: An update. *J. Geophys. Res.*, **101**, 12 791–12 794.
- Collins, M., B. Booth, G. Harris, J. Murphy, D. Sexton, and M. Webb, 2006: Towards quantifying uncertainty in transient climate change. *Climate Dyn.*, **27**, 127–147, doi:10.1007/s00382-006-0121-0.
- Collins, W. D., and Coauthors, 2006: The formulation and atmospheric simulation of the Community Atmosphere Model Version 3 (CAM3). *J. Climate*, **19**, 2144–2161.
- Colman, R., 2003: A comparison of climate feedbacks in general circulation models. *Climate Dyn.*, **20**, 865–873.
- 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, 525–582.
- Deng, Y., K. P. Bowman, and C. Jackson, 2007: Differences in rain rate intensities between TRMM observations and community atmosphere model simulations. *Geophys. Res. Lett.*, **34**, L01808, doi:10.1029/2006GL027246.
- Frame, D. J., N. E. Faull, M. M. Joshi, and M. R. Allen, 2007: Probabilistic climate forecasts and inductive problems. *Philos. Trans. Roy. Soc. London*, **A365**, 1971–1992, doi:10.1098/rsta.2007.2069.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 2004: *Bayesian Data Analysis*. 2nd ed. Chapman and Hall, 668 pp.
- Golaz, J.-C., V. E. Larson, J. A. Hansen, D. P. Schanen, and B. M. Griffin, 2007: Elucidating model inadequacies in a cloud parameterization by use of an ensemble-based calibration framework. *Mon. Wea. Rev.*, **135**, 4077–4096.
- Haario, H., M. Laine, A. Mira, and E. Saksman, 2006: DRAM: Efficient adaptive MCMC. *Stat. Comput.*, **16**, 339–354.
- Hastings, W., 1970: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.
- Held, I. M., and B. J. Soden, 2000: Water vapor feedback and global warming. *Annu. Rev. Energy Environ.*, **25**, 441–475.
- Ingber, L., 1989: Very fast simulated re-annealing. *Math. Comput. Modell.*, **12**, 967–973.
- Jackson, C., M. Sen, and P. Stoffa, 2004: An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions. *J. Climate*, **17**, 2828–2841.
- Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471.
- Kistler, R., and Coauthors, 2001: The NCEP–NCAR 50-Year Reanalysis: Monthly means CD-ROM and documentation. *Bull. Amer. Meteor. Soc.*, **82**, 247–267.
- Knutti, R., G. A. Meehl, M. R. Allen, and D. A. Stainforth, 2006: Constraining climate sensitivity from the seasonal cycle in surface temperature. *J. Climate*, **19**, 4224–4233.
- LeTreut, H., and B. J. McAvaney, 2000: A model intercomparison of equilibrium climate change in response to CO<sub>2</sub> doubling. Note du Pole de Modelisation de l'IPSL Number 18, Institut Pierre Simon LaPlace, 24 pp.
- McWilliams, J. C., 2007: Irreducible imprecision in atmospheric and oceanic simulations. *Proc. Natl. Acad. Sci. USA*, **104**, 8709–8713.
- Min, S.-K., D. Simonis, and A. Hense, 2007: Probabilistic climate change predictions applying Bayesian model averaging. *Philos. Trans. Roy. Soc. London*, **A365**, 2103–2116, doi:10.1098/rsta.2007.2070.
- Mu, Q., C. S. Jackson, and P. L. Stoffa, 2004: A multivariate empirical-orthogonal-function-based measure of climate model performance. *J. Geophys. Res.*, **109**, D15101, doi:10.1029/2004JD004584.
- 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**, 768–772, doi:10.1038/nature02771.
- , B. B. Booth, M. Collins, G. R. Harris, D. Sexton, and M. Webb, 2007: A methodology for probabilistic predictions of regional climate change from perturbed physics ensembles. *Philos. Trans. Roy. Soc. London*, **A365**, 1993–2028, doi:10.1098/rsta.2007.2077.
- Rossow, W. B., L. C. Garder, P. J. Lu, and A. W. Walker, 1991: International satellite cloud climatology project (ISCCP) documentation of cloud data. World Meteorological Organization Tech. Doc. WMO/TD 266, 76 pp.
- Rougier, J., and D. S. Sexton, 2007: Inference in ensemble experiments. *Philos. Trans. Roy. Soc. London*, **A365**, 2133–2143, doi:10.1098/rsta.2007.2071.
- Sen, M., and P. L. Stoffa, 1996: Bayesian inference, Gibbs' sampler and uncertainty estimation in geophysical inversion. *Geophys. Prospect.*, **44**, 313–350.
- Stainforth, D. A., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, **433**, 403–406, doi:10.1038/nature03301.
- , M. R. Allen, E. R. Tredger, and L. Smith, 2007: Confidence, uncertainty and decision-support relevance in climate predictions. *Philos. Trans. Roy. Soc. London*, **A365**, 2145–2161, doi:10.1098/rsta.2007.2074.
- Villagran, A., G. Huerta, C. S. Jackson, and M. K. Sen, 2008: Computational methods for parameter estimation in climate models. *Bayesian Anal.*, in press.
- Wang, F., 2007: Investigating ENSO sensitivity to mean climate in an intermediate model using a novel statistical technique. *Geophys. Res. Lett.*, **34**, L07705, doi:10.1029/2007GL029348.
- Webb, M. J., and Coauthors, 2006: On uncertainty in feedback mechanisms controlling climate sensitivity in two GCM en-

- sembles. *Climate Dyn.*, **27**, 17–38, doi:10.1007/s00382-006-0111-2.
- Wilcox, E. M., and L. J. Donner, 2007: The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *J. Climate*, **20**, 53–69.
- Xie, P., and P. Arkin, 1996: Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *J. Climate*, **9**, 840–858.
- , and ——, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.*, **78**, 2539–2558.
- Zhang, G. J., and N. A. McFarlane, 1995: Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian Climate Centre general circulation model. *Atmos.–Ocean*, **33**, 407–446.