Use of Bayesian inference and data to improve simulations of multi-physics climate phenomena

Charles S Jackson
Institute for Geophysics, Jackson School of Geosciences, The University of Texas at Austin
charles@ig.utexas.edu

Abstract. Bayesian inference provides a means to incorporate activities usually associated with post-model development evaluation into the process of model development. Presented is a review of the factors that make this approach challenging, strategies for making the process practical for model development of complex multi-physics phenomena, and suggestions on areas requiring additional research. An analysis is presented of the strategy that was used to determine the joint probability for six non-linearly related parameters important to clouds and convection within the NCAR Community Atmosphere Model version 3.1.

1. Introduction
Research activity concerning climate phenomena, whether observational, theoretical, or simulation is directed at the goal of understanding the physical processes or balance of processes that govern nature. The challenge is that observations are limited, theories tend to address idealizations, and simulations can mimic nature for the wrong reasons. Climate model construction involves the independent development of component models, each of which seem to be ok when forced by observations, but yet can go astray when coupled to each other (e.g. [19]). The experience is much like a fishing expedition in which there is much preparation in setting up the gear, location, and timing, but with very little by way of a guarantee on the success of the catch.

Presented here are the concepts, strategies, and an example application of an alternate, Bayesian inference approach to synthesizing data constraints on choices in climate model development. It has its own set of challenges, but may ultimately provide a more robust avenue to how we use data to inform model development.

2. Use of Bayesian inference in model development
Bayesian inference can be useful for model development by synthesizing information from different sources of data, modeling results, and corresponding error information to inform difficult choices in model development such as the values assigned to multiple, non-linearly related parameters. Bayesian inference involves the stochastic sampling of sources of uncertainty, each sample of which requires a model integration to estimate the sensitivity to the input parameters, which may make it seem impractical for model development. Currently assessments of modeling uncertainty occur after climate model development is complete and is revealed through model inter-comparison projects or specially designed sensitivity experiments that tease out aspects of model predictions that are more robust (e.g. [15], [18], [23]). The problem has been that these activities mainly concern the forward propagation of
errors. But unlike Bayesian inference, these more informal approaches to quantifying uncertainties are not well suited to the task of identifying the sources of uncertainty or providing guidance on choices in model development.

The best opportunity for scientific understanding to enter into the process of model development, within the Bayesian framework, is in the care that is given to the definition of the metric that quantifies model skill. Many different kinds of metrics are being developed in the context of summarizing differences that arise in multi-model projections of climate change (e.g. [3]). Another common example of the use of metrics is to provide weights to multi-model ensemble mean predictions and uncertainty (e.g. [24], [22]). The metrics that have been proposed so far do not have model development in mind. Rather they usually only quantify univariate mean statistics. We need better ways to quantify relationships if we are going to impose this goal in model development.

The other challenge facing model developers is that the volume of solution space, the space where there is a good fit to observational data, is likely a small fraction of the total volume of possibilities. This makes Bayesian inference much more similar to problems in geophysical inversion. The stochastic sampling methods developed for Bayesian inference such as Markov Chain Monte Carlo (MCMC) may not perform well if they cannot focus sampling in an efficient way. However, if we treat complex multi-physics systems as a typical inverse problem, we will only be getting a single solution for a problem where many solutions exist. Least-squares inverse, Maximum a posteriori (MAP), or gradient-based strategies to solving inverse problems target point estimates of the most likely model parameters, rather than full probability densities. This singular objective may be overly sensitive to assumptions concerning the value of collected data or a particular modeled phenomenon. Since, climate phenomena represent a spectrum of relationships, model development should take this into account. What maybe more helpful to the scientific aims of developing a multi-physics model is a Bayesian approach that provides a way to obtain more information about the complete statistical description of the uncertainties inherent in the data and modeling choices.

Suppose we are interested in assessing the joint probability for selecting a vector \( \mathbf{m} \) of model parameter values such that the climate model \( g(\mathbf{m}) \) is consistent with observations \( \mathbf{d}_{\text{obs}} \). Assuming Gaussian errors, the solution posterior probability density (PPD), as it is called in Bayesian nomenclature, would be given as

\[
PPD(\mathbf{m} | \mathbf{d}_{\text{obs}}, g(\mathbf{m})) \propto \exp \left[ -\frac{1}{2} (g(\mathbf{m}) - \mathbf{d}_{\text{obs}})^\top C_{\text{noise}}^{-1} (g(\mathbf{m}) - \mathbf{d}_{\text{obs}}) \right] \cdot \text{prior}(\mathbf{m})
\]

(1)

with \( C_{\text{noise}}^{-1} \) as an expression of the uncertainty in comparing model output to data and \( \text{prior}(\mathbf{m}) \) represents prior probability for model parameters \( \mathbf{m} \) independent of \( \mathbf{d}_{\text{obs}} \). The quantity within the square brackets is typically referred to the “cost” function or metric of errors in the model relative to what is observed. Even with the Gaussian assumption for observational and modeling uncertainties, the PPD is generally not Gaussian through the nonlinearities represented in the climate model \( g(\mathbf{m}) \). Because there is generally no closed-form solution to equation (1), it needs to be estimated.

It may be helpful to draw a connection between equation (1), statistical physics, and the solution that Metropolis et al. [16] proposed for estimating its solution in order to highlight the nature of the problem that is being solved. In particular, the problem is analogous to estimating the probability that a physical system have a particular property \( s \) (corresponding to the choices of parameter values \( \mathbf{m} \)) under the constraint that the system has, on sum, an energy \( E \) (corresponding the cost function that quantities modeling errors). This probability is composed of the number of states \( N(s) \) observable to the system with property \( s \), out of all states of the system \( Z \) and related to entropy by \( \sigma = \ln(N(s)) \). In order to estimate \( N(s)/Z \) numerically, it is sufficient to select candidate configurations in proportion to \( N(s)/Z \). Metropolis et al. [16] proposed a relatively simple procedure achieving this that has proven invaluable for simulating complex physical systems on a computer. They proved that the distribution described by \( N(s)/Z \) could be approximated iteratively through by randomly stepping through configuration space with a given step size and with each step applying a set of acceptance and
rejection criteria for each step. Specifically, if a step through state space reduces a system’s energy then that step will always be accepted. If a step increases a system’s energy then that step will only be accepted in proportion to

\[ P = \exp\left(-\frac{\Delta E}{T}\right), \]  

(2)

where \( P \) is the acceptance probability, \( \Delta E = (E_2 - E_1) \) is the change in energy going from energy state \( E_1 \) to \( E_2 \), and \( T \) is the system’s temperature. This transition probability may be re-expressed in terms of \( N(s) \) using the definition for temperature \( T \)

\[ \frac{1}{T} = k_b \left[ \frac{\partial \sigma}{\partial E} \right] \]  

(3)

where \( k_b \) is the Boltzman constant. In discrete form

\[ \frac{\Delta E}{k_b T} = \Delta \sigma. \]  

(4)

Substituting eq. (4) into eq. (2) and redefining the temperature in eq. (2) to include \( k_b \) one finds that this transition probability is

\[ P = \exp(-\Delta \sigma) = \exp\left(-\left(\sigma_2 - \sigma_1\right)\right) = \frac{N(s_1)}{N(s_2)} \]  

(5)

That is, the transition probability is given by ratio of the number of states available to the system with property \( s_1 \) to the number of states available to the system with property \( s_2 \). So while it is very difficult to measure \( N(s)/Z \) directly, the above derivation shows how one may use statistical sampling and calculations of a system’s energy to derive estimates of \( N(s)/Z \). This sampling strategy is an early example of a MCMC. If the solution space is well behaved, the sampling procedure is ergodic (it can visit all possible configurations). However the procedure can be sensitive to details such as the average size of the step that is taken through solution space. The following section reviews different strategies to estimating equation (1) and the various advantages or disadvantages for its applicability to climate model development.

3. Sampling strategies appropriate for climate model development

Among the challenges to use Bayesian inference in the development of a climate model are (1) the computational expense for testing choices, (2) the potentially complex structure of the response surface that describes how a cost function changes as a function of sources of uncertainty, (3) the large number of uncertain parameters, or (4) that the volume of parameter space in which viable solutions reside is may be small. Climate model development is also primarily interested in quickly defining a single working model (this does not negate the interest in finding many working models!). Therefore any method that is of practical use must make inferences with relatively few modeling experiments. These objectives and constraints do not map well onto the spectrum of strategies that have been developed for Bayesian inference, which typically assume an infinite capacity to obtain samples. Discussed below are several approaches to mitigating these challenges that, when combined with high performance computing (HPC) resources, may prove sufficiently practical for climate model development.

3.1. Multiple Very Fast Simulated Annealing (MVFSA)

MVFSA balances a number of the objectives described above and does so adaptively such that it can provide efficient estimates of complex response surfaces that may reside in small volumes of solution space [12]. The strategy uses the rules of the Metropolis sampler to statistically seek out regions of parameter space where the cost function is minimized. However over the course of the sampling, the average step size is reduced and centered on the last accepted model configuration with a Cauchy distribution the width of which is determined by the experiment number [9]. Also over time, the cost
function is scaled to exaggerate the response surface through an algorithm control parameter analogous to temperature in equation (2). Any single Markov Chain is not guaranteed to find the global minimum. Rather through multiple convergence attempts, one may increase the probability of identifying regions of parameter space that allow the model to track the observational constraints. This approach works well for HPC resources because it allows one to distribute the workload across the multiple chains. Although Simulated Annealing approaches have the connotation of a technique that is useful for global optimization, not uncertainty quantification, the MVFSA strategy outlined in [12] is deployed so that it makes use of these optimization features to enhance the efficiency of estimating the PPD.

Among the shortcomings of this approach is that this strategy does not have ergodic properties. The PPD tends to have long tails from the fact that sampling starts from a random position in parameter space and may require a number of iterations before finding regions of interest [26]. This may be considered a feature for strategies that depend on statistical emulation (discussed below).

3.2. Adaptive Metropolis (AM)
A class of MCMC strategies have been developed which can adapt to solution characteristics, making them more robust and much more efficient when compared to the Metropolis/Gibbs sampler [4]. These AM methods work by making better use of samples taken to update the proposal distribution for new samples. The single component AM (SCAM) [5] and the delayed rejection AM [25] has shown sampling efficiencies comparable to MVFSA with the advantage that these approaches have been shown to be ergodic [26]. One of the surprising results of the tests of the AM strategies is the large sample autocorrelations (a measure of sample mixing) for the full component AM which would make this choice non-competitive relative to other options. Another issue that needs to be explored further is the trouble many sampling strategies have, particularly MCMC, in sampling multi-modal distributions. It would be good to know if we should expect multi-modal PPDs for the climate problem.

3.3. Statistical emulation
The point of statistical emulation is to minimize the number of expensive physics-based modeling experiments that would be required to estimate uncertainties by predicting what the output would be at untried parameter settings [14] [7]. However there has only been limited experience in developing statistical emulators of climate, almost exclusively focusing on univariate approaches with a few variables (e.g. [18], [20]). Thus as interest grows to construct more sophisticated metrics of model output that take into account relationships in space, time, and across fields such as in Smith et al. [22], statistical emulation will need to similarly evolve to model these relationships as well.

4. Application to the Community Atmosphere Model
The objective here is to evaluate the utility of MVFSA for model development based on the results of Jackson et al. [11] whose primary objective was to estimate a joint PPD for six parameters important to clouds and convection within the NCAR Community Atmosphere Model version 3.1 (CAM3.1) [1]. In particular, how many experiments are required to get useful results? How might the sampling process affect the distribution of errors for individual fields? What can we infer from the results about how appropriate MVFSA is for quantifying uncertainties and as a tool for model development? The Jackson et al. [11] exercise included over 800 11-year long model integrations from 1990 to 2001 in which MVFSA selected different combinations of parameter values and compared the model output to observations. Six independent chains were run in parallel with each experiment using 64 processors of a Intel Xeon-based linux cluster with high-speed InfiniBand interconnect. With each experiment requiring about 10 hours, all experiments were completed in two months.

An analysis of the computational costs of MVFSA for estimating the PPD was previously completed with a surrogate climate system model [12]. This analysis has been extremely useful in designing new experiments and predicting the computational requirements for a wide range of
problems including systems involving 50 parameters (e.g. [13], [27]). That analysis predicts that one would need ~3000 experiments to estimate statistically stable estimates of 15-bin marginals of the PPD, thus the 800 experiments provides only a very rough estimate of these probabilities.

The cost function considered in [11] is much more typical of what we may find in model development than the version in [12] which only considers a perfect model analysis of temperature. The [11] cost function quantifies model discrepancies with instrumental, satellite, and reanalysis data including radiative balances at the top and bottom of the atmosphere, precipitation, winds, humidity, sea level pressure, and clouds. The cost function did not represent relationships among these fields. However there was an attempt to normalize the cost function by the effects of internal model variability on individual fields. The correlations among errors among multiple fields were accounted for through a renormalization process through the choice of a scaling factor ‘S’ [17]. Jackson et al. [11] took a more Bayesian approach to the section of ‘S’ by treating its selection as being condition on the choice of the other model parameters and allowing it to scale inversely by the mean bias that exists between models and predictions. The point of this step was to introduce some additional adaptivity in the acceptance/rejection of model configurations with the logic that modelers need to be more accepting of non-optimal configurations when there is significant biases. A full discussion of this feature is beyond the objectives of the present analysis. However the treatment of ‘S’ illustrates some of the flexibility that Bayesian inference can provide to this type of analysis that incorporate the scientific sensibilities that would be helpful to model development.

4.1. Number of experiments needed
If the goal is to identify a single model configuration, one must perform multiple sampling chains and evaluate the configuration with the lowest cost value. Figure 1 shows the path to convergence among the six chains performed in [11]. Shown is the fractional change in the cost function relative to the standard configuration of cam3.1. Any value less than 1 indicates a model configuration whose errors across all fields in the cost function sum to a number less than cost of the standard model. That does not imply that the model is better scientifically. Any of the selected model configurations may include large compensating errors. An evaluation of this point will be considered in the following section. Each sampling chain indicated in figure 1 shows the lowest cost value achieved up to a given experiment number. Six out of six chains were able to identify a model that was better than the default model configuration by experiment number 10. There were exponential reductions in the cost function of the best performing configurations, with an approximate e-folding scale of 40 experiments, a similar rate to what was found in [12]. The overall reduction in cost values amounted to 9% which was similar to that found in the perturbed physics ensemble of [23] derived from a 2578 Monte Carlo (random) sampling among six uncertain model parameters of the Hadley Center atmospheric climate
model. The analysis of the model configurations after 41 experiments within each line is given by [11]. The present analysis shows that there was an additional 3% reduction in the cost function that occurred in subsequent sampling such that all models had nearly identical reductions in cost by experiment 120. Presumably one could run more lines in parallel an increase the likelihood that an optimal configuration could be found by experiment 20. There will be some additional analysis of this last point in a following section. Again, these numbers are appropriate for a six-parameter problem, but are very much in line with expectation established by [12] that the number of experiments required to get 66% of the way to convergence is by ~6 times the number of uncertain parameters. The same analysis also shows that a six-parameter problem would need ~125 experiments to fully converge, which is also in line with the present analysis.

4.2. Distribution of component errors in optimization solution
The question that is addressed here is whether the evolution of errors favors particular fields over others on the path to optimization. The answer to this question may affect the number of experiments that are required to obtain satisfactory results. If, for instance, a particular field dominates the cost function, one may expect, as I did, that progress on reducing errors among the “less important” fields would be delayed. Figure 2 shows the inter-chain average cost among the optimal configurations as a function of experiment number. Figure 2 shows that the above expectation is wrong. From the fields that dominate the cost function in panel (a) to the minor actors in panel (c), there is a similar evolution to the change in errors (with a value of zero for the standard model configuration). The reason for this is likely because these fields are not independent from one

Figure 2. Fractional change in cost (in percent) for each component of the cost function shown in figure 1.
another. They are all physically linked, such that a reduction in errors in one field should give rise to a corresponding reduction of errors among related fields, if the model is an adequate description of the system. This appears to fail with the model depiction of clouds. In particular the improvements that occurred with surface variables, came at the expense of top of the atmosphere shortwave radiative fluxes and clouds.

After approximately 40 experiments, the errors that occur across many of the fields appear to evolve randomly with improvements and correspondingly sized degradations. Thus, the overall march to convergence that is apparent in figure 1 is subtle feature when viewed component by component. It may be important to realize that any of these model configurations are acceptable and the jumping around represents a fundamental uncertainty in competing choices of how to balance the multiple objectives.

4.3. Characteristics of the response surface
Presented next is two-dimensional projections of the response surface estimated as a weighted average over all accepted experiments defined as models with equal or better cost metric to the default model. The inference is binned into smaller paired subregions of parameter space $p_1$ and $p_2$. All experiments that fell into that subregion contribute to the weighted average through

$$\text{cost}(p_1, p_2) = \frac{\sum_{i=1}^{n_{\text{exp}}} \text{cost}(p_1, p_2) \cdot \exp[-S \cdot (\text{cost}(p_1, p_2) - \text{cost}(p_1, p_2)_{\text{min}})]]}{\sum_{i=1}^{n_{\text{exp}}} \exp[-S \cdot (\text{cost}(p_1, p_2) - \text{cost}(p_1, p_2)_{\text{min}})]]} \quad (6)$$

where $\text{cost}_{\text{min}}$ refers to the minimum value of the cost function over all sub-regions of parameter space, and $S$, as previously discussed, is determined by Bayesian inference to be the range in cost values that are deemed acceptable. As it turns out, $S$ selects a similar range in cost values than from the $\text{cost}_{\text{min}}$ to the cost of the standard model. Thus the decision to filter experiments by only those models that have a cost value equal or less than the default model does not appreciably affect the results. Figure 3 shows the inferences of the response surface for the six model parameters which are initial cloud downdraft mass flux (alfa) and consumption rate of CAPE (tau) in panel (a) which are parameters important to convection, environmental air entrainment rate (ke) and precipitation efficiency (c0) which are important to cloud micro-physics, and low and high cloud critical relative humidity (rhminl and rhminh) which are important to cloud macro-physics. In each panel, the lowest cost values mostly appear to reside in a particular region of parameter space (the distribution does not appear to be multi-modal). The response surface becomes much more varied away from these favored regions, however the number of samples (indicated by the white and black ‘+’ marks of the accepted and rejected configurations, respectively) become less dense precisely where the colors are more varied. Thus there are not likely enough samples to make robust inferences of the response surface.

Also presented in figure 3 are the optimal parameter choices that were selected by experiment 20. There can be considerable variation parameter values selected, however the location where the cost value is smallest also tends to be close to global optimized model configuration over all lines and experiments. Given the large variations in the cost values for each of the model components illustrated in figure 2 shows how important even small variations in parameter values have on modeling results.

4.4. Volume of solution space
The models that were deemed acceptable fill a 2% volume of the total six-dimensional volume. This is remarkably small and underscores the need to have sampling strategies that can aggressively find and sample these regions densely. This estimate of the total volume is only approximate and is dependent on how one infers regions of acceptability around the samples that have been completed. The 2% estimate comes from an accounting of the fraction of squares that have been deemed acceptable in 25
by 25 binned segments from the parameter minimum to maximum. Note that this region includes a broader range than indicated by figure 3, which focuses only on regions that contained at least a single accepted experiment. The probability is then the product of the fraction for the three pairs of variables indicted in figure 3 (0.35 for panel (a) times 0.31 for panel (b) times 0.2 for panel (c)). The estimate of the volume goes up to 8% if one subdivides parameter space into 15 bins for each parameter. The hit rate for MVFSA for sampling within the acceptable region is 70%, with a slightly lower percentage at the beginning.

5. Discussion and conclusion
The merit of using Bayesian inference in the development of numerical models of multi-physics systems such as a climate model is that it formalizes the process of model evaluation and uncertainty quantification for evaluating choices in model development. With so many degrees of freedom, such models can reproduce realistic behaviors for the wrong reasons. With Bayesian inference, the scientist can define meaningful metrics and make use of innovative sampling strategies for selecting model configurations consistent with data. The result is an ensemble of model configurations that provides a more rigorous assessment of uncertainties that remain.

The example application is an atmospheric general circulation model where features of interest could be gleaned from integrating the model for a few years with a given change in parameter values. However, some of the most challenging aspects of climate model development are in getting realistic simulations of coupled phenomena like ENSO. Coupled models take centuries to equilibrate and it may take a few centuries or millennia to fully characterize a model’s ENSO characteristics [28]. One strategy for tackling this problem is to focus sampling on only a few parameters that affect the coupled phenomena and use Bayesian inference to select the appropriate combination of parameter values. A second strategy would be to make use of the results of a more elaborate quantification of uncertainties in each sub-system and use those configurations as a prior for the coupled model. One could then systematically test a limited number of coupled model configurations.

Figure 3. Estimate of cam3.1 response to six parameters important to clouds and convection. The average cost within each box follows equation (6) with the ‘+’ marks indicating an experiment whose cost is less (white) or greater (black) than the standard model configuration. Numbers refer to the optimal model configuration after experiment 20.
configurations with longer model integrations, knowing that each component configuration has been filtered for its realism.

The number of parameters that can be explored with Bayesian inference is realistically limited to only a handful, perhaps less than 20. The limitation is that the selection process assumes all parameter choices are correlated and the number of samples required to quantify these relationships become very large. There are innovative strategies for using multi-scale techniques to reduce the apparent dimensionality of a large set of parameters [8].

The issue that would most solidify the utility of Bayesian inference for model development is whether it aids in the effort to learn the physical processes or balance of processes that govern nature. The path for this to happen is to first identify models that capture observed climate change for the right reasons and then to incorporate its key elements into a cost function so that we do not lose the ability to maintain this level of skill in future model development efforts.

References

[1] Collins, W.D., and Coauthors, 2006: The Formulation and Atmospheric Simulation of the Community Atmosphere Model Version 3 (CAM3). J. Climate, 19, 2144–2161.

[2] Gleckler, J. J. Hnilo, S. M. Marlais, T. J. Phillips, G. L. Potter, B. D. Santer, K. R. Sperber, K. E. Taylor, D. N. Williams, 1999: An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I). Bulletin of the American Meteorological Society 80(1), 29-56.

[3] Gleckler, P. J., K. E. Taylor, and C. Doutriaux (2008), Performance metrics for climate models, J. Geophys. Res., 113, D06104, doi:10.1029/2007JD008972.

[4] Haario, H. (2001). “An Adaptive Metropolis algorithm.” Bernoulli , 7: 223–242.

[5] Haario, H., Laine, M., Lehtinen, M., Saksman, E., and Tamminen, J. (2004) Markov Chain Monte Carlo methods for high dimensional inversion in remote sensing. Journal of Royal Statistical Society , 66: 591–607.

[6] Haario, H., Saksman, E., and Tamminen, J. (1999). “Adaptive proposal distribution for random walk Metropolis algorithm.” Computational Statistics, 14: 375–395.

[7] Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafco, and R. D. Ryne (2004) Combining field data and computer simulations for calibration and prediction, SIAM Journal on Scientific Computing, 26, 448-466.

[8] Hong, T., and M. K. Sen (2009) A new MCMC algorithm for seismic waveform inversion and corresponding uncertainty analysis, Geophys. J. Int. doi:10.1111/j.1365-246X.2008.04052.x

[9] Ingber, L. (1989). “Very fast simulated re-annealing.” Mathematical Computational Modelling , 12: 967–973.

[10] Jackson, C. S., M. K. Sen, G. Huerta, Y. Deng, and K. P. Bowman (2008) Error Reduction and Convergence in Climate Prediction. Journal of Climate, 21(24), 6698-6709. DOI: 10.1175/2008JCLI2112.1

[11] Jackson, C., M. Sen, and P. Stoffa (2004) An Efficient Stochastic Bayesian Approach to Optimal Parameter and Uncertainty Estimation for Climate Model Predictions, Journal of Climate, 17(14), 2828-2841.

[12] Jackson, C., Y. Xia, M. Sen, and P. Stoffa (2003) Optimal parameter and uncertainty estimation of a land surface model: A case example using data from Cabauw, Netherlands, Journal of Geophysical Research, 108 (D18), 4583 10.1029/2002JD002991.

[13] Kennedy, M. C., and A. O’Hagan (2001) Bayesian calibration of computer models, Journal of the Royal Statistical Society. Series B, Statistical Methodology, 63, 425-464.

[14] Meehl, G.A., G.J. Boer, C. Covey, M. Latif, R.J. Stouffer, 2000: The Coupled Model Intercomparison Project (CMIP). Bulletin of the American Meteorological Society 81(2), 313-318.

[15] Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E. (1953) Equations of State Calculations by Fast Computing Machines. Journal of Chemical Physics, 21: 1087–
[17] Mu, Q., C. S. Jackson, P. L. Stoffa (2004) A multivariate empirical-orthogonal-function-based measure of climate model performance, *J. Geophys. Res.*, 109, D15101, doi:10.1029/2004JD004584.

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

[19] Richter, I., and S.P. Xie (2008) On the origin of equatorial Atlantic biases in coupled general circulation models. Clim. Dyn. (31) 587-598. (doi:10.1007/s00382-008-0364-z)

[20] Rougier, J. C. 2007 Probabilistic inference for future climate using an ensemble of simulator evaluations. Clim. Change 81, 247–264. (doi:10.1007/s10584-006-9156-9)

[21] Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns (2009) Bayesian Modeling of Uncertainty in Ensembles of Climate Models. *Journal of the American Statistical Association*, 104(485), doi:10.1198/jasa.2009.0007

[22] Stainforth, D. A., and Coauthors, 2005: Uncertainty in predictions of the climate response to rising levels of greenhouse gases, *Nature* 433, 403-406 (27 January 2005) doi:10.1038/nature03301.

[23] Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O. (2005) Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multi-model Ensembles. *Journal of Climate*, 18, 1524-1540 (corrigendum p. 3405).

[24] Tierney, L. and Mira, A. (1999). “Some Adaptive Monte Carlo Methods for Bayesian Inference.” Statistics in Medicine, 18: 2507–2515.

[25] Villagran, A., G. Huerta, C. S. Jackson, M. K. Sen (2008) Parameter Uncertainty Estimation in Climate Models. Bayesian Analysis 3(4), 823-850. DOI:10.1214/08-BA331

[26] 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.

[27] Wittenberg, A. T., 2009: Are historical records sufficient to constrain ENSO simulations? Geophys. Res. Lett., in press, May 2009