LLNL-TR-652027

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

Detecting and Testing for Structural Error in Computer Models with Application to the Community Atmospheric Model

G. Johannesson, D. D. Lucas

March 20, 2014

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Detecting and Testing for Structural Error in Computer Models with Application to the Community Atmospheric Model

Gardar Johannesson and Donald D. Lucas

Lawrence Livermore National Laboratory

LLNL-TR-652027

Abstract

Bayesian calibration/inversion methods for uncertain input parameters to simulation codes have generated an increasing amount of interest in the climate community in the recent years. This approach uses multiple disparate observational datasets related to the output of the simulator of interest to yield a probabilistic constraint on the uncertain inputs (i.e. probability distribution). A major challenge in applying Bayesian calibration methods to simulation codes is related to the possible presence of a large structural error in the output fields of interest; that is, the presence of discrepancy between the observed and the simulated output that is neither explained by the uncertainty in the inputs nor the observations. We present here methods to test for and visualize structural error in an ensemble of input-perturbed simulations. The methods are based on bootstrap (resmpling) test statistics and well-known metrics to quantify the difference between two probability distributions. The proposed methods are demonstrated using selected precipitation output fields of the Community Atmospheric Model (CAM), which are compared to observations provided by the Global Precipitation Climatology Project (GPCP). The methods are found to be useful in identifying regions and seasons with a large structural error. The proposed structural error detection methodology can be used to screen multiple observational datasets of interest for structural error, making it possible to retain only the datasets with a small structural error for use in Bayesian calibration. In addition, information about the amount of structural error can be valuable to model developers.

1 Introduction

Bayesian calibration of computer models (i.e., simulation codes) is a well-established and tested technique for improving the fidelity of computer models that have uncertain and/or tuneable input parameters (e.g., Kennedy and O'Hagan, 2001; Bayarri et al., 2007; Higdon et al., 2008). Not surprisingly, Bayesian calibration methods have been adapted and applied to climate models of various complexity with mixed results (e.g., Sanso and Forest, 2009; Bhat et al., 2012). The general framework behind this approach is relatively simple and straightforward. Using the notation popularized by Kennedy and O'Hagan (2001), let $\eta(x, \boldsymbol{\theta})$ be a simulation (scalar) output quantity of interest, where x characterizes the output (e.g., spatial location, time, and other simulation control parameters) and $\boldsymbol{\theta}$ is a vector of uncertain input parameters to the code. For example, in our application to the Community Atmospheric Model (CAM), the atmospheric component of the Community Earth System Model (CESM), discussed in in Section 4, $\eta(x, \boldsymbol{\theta})$ might be the 5-year average daily precipitation in winter (Dec–Feb) in the 30° zonal band from 0-30°N.

At the core of a Bayesian calibration of a computer model is the following comparison of the computer simulation output $\eta(x, \boldsymbol{\theta})$ and the observation $y(x)$ modeled as:

$$y(x) = \eta(x, \boldsymbol{\theta}) + \delta(x) + \epsilon(x), \tag{1}$$

where $\epsilon(x)$ is the error in the observation and $\delta(x)$ is a potential structural error (bias) in the code that is not explained by the uncertain input parameters. That is, even if there where no uncertain inputs to the code, there may still be a discrepancy between the simulation output $\eta(x)$ and the observation $y(x)$ that is not explained by the observation error $\epsilon(x)$.

The work presented here was carried out under the Climate Science for Sustainable Energy Future (CSSEF) Program and is motivated by questions related to the role of the structural error (δ) versus uncertainty in the input $\boldsymbol{\theta}$ when assessing the prediction accuracy of important precipitation-related metrics in CAM. If important precipitation-related metrics have a large amount of structural simulation error in CAM, then there is little hope of "matching" the observed metrics by tuning uncertain input parameters. What we present here are statistics and tests to shed light on the following question:

> *Can the difference between observations and simulations be explained by the uncertainty in the inputs and the observation error alone?*

To rephrase: is the structural error small enough so that it is “hidden” in the uncertainty induced by the parameters and the observation error?

In the next section we will briefly describe the input-perturbed CAM5 simulations created under CSSEF and the Global Precipitation Climatology Project (GPCP) observations used in this study. We follow with a method section (Section 3) outlining metrics and test statistics we investigate to test for structural error. Finally, in Section 4 we demonstrate these techniques using the CAM5 simulations and the GPCP observations.

2 Simulations and Observations

In the application in Section 4, we will use input-perturbed simulations from CAM (version 5). The quantities of interest are daily precipitation spatial averages at two different spatial scales in two seasons, as will be further described below. The simulation-based precipitation statistics are then compared to analogous aggregates computed using the Global Precipitation Climatology Project (GPCP) dataset.

2.1 The Community Atmospheric Model Version 5 (CAM5) Ensemble of Simulations

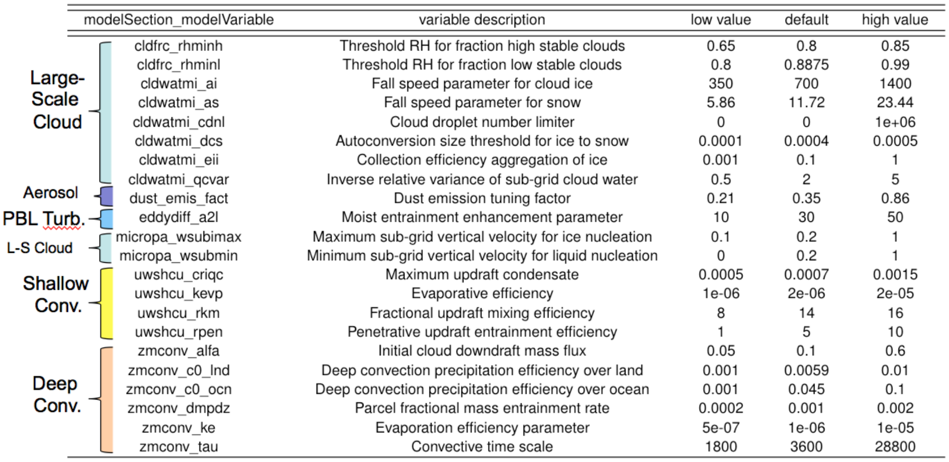
The CAM5 ensemble of simulations consists of 1145 5-years AMIP-style (i.e., prescribed sea-surface temperature) simulations spanning 5 years (2000–2004), generated by perturbing 22 uncertain input parameters. These simulations where carried out at Lawrence Livermore National Laboratory as part of the Climate Science for Sustainable Energy Future (CSSEF) effort.

The 22 input parameters selected for the study are all judged to be potentially important to the hydrology cycle in CAM5 and are shown in Table 1, which identifies the parameters by name, their default value and the range over which they were varied.

The 1145 simulations are made up of five latin-hypercube (LH) sample designs (Mckay et al., 2000), each of size 220, along with one batch of one-at-a-time (low-default-high) simulations of size 45. The two last LH-sampled batches of simulations had altered ranges for one of the 22 parameters.

2.2 The Global Precipitation Climatology Project (GPCP) Data

The precipitation data used were obtained from the Global Precipitation Climatology Project (GPCP). The GPCP data consist of average monthly



	modelSection_modelVariable	variable description	low value	default	high value
Large-Scale Cloud	cldfrc_rhminh	Threshold RH for fraction high stable clouds	0.65	0.8	0.85
	cldfrc_rhminl	Threshold RH for fraction low stable clouds	0.8	0.8875	0.99
	cldwatmi_ai	Fall speed parameter for cloud ice	350	700	1400
	cldwatmi_as	Fall speed parameter for snow	5.86	11.72	23.44
	cldwatmi_cdnl	Cloud droplet number limiter	0	0	1e+06
	cldwatmi_dcs	Autoconversion size threshold for ice to snow	0.0001	0.0004	0.0005
	cldwatmi_eii	Collection efficiency aggregation of ice	0.001	0.1	1
	cldwatmi_qcvar	Inverse relative variance of sub-grid cloud water	0.5	2	5
Aerosol	dust_emis_fact	Dust emission tuning factor	0.21	0.35	0.86
PBL Turb.	eddydiff_a2l	Moist entrainment enhancement parameter	10	30	50
L-S Cloud	micropa_wsubimax	Maximum sub-grid vertical velocity for ice nucleation	0.1	0.2	1
	micropa_wsubmin	Minimum sub-grid vertical velocity for liquid nucleation	0	0.2	1
Shallow Conv.	uwshcu_criqc	Maximum updraft condensate	0.0005	0.0007	0.0015
	uwshcu_kevp	Evaporative efficiency	1e-06	2e-06	2e-05
	uwshcu_rkm	Fractional updraft mixing efficiency	8	14	16
	uwshcu_rpen	Penetrative updraft entrainment efficiency	1	5	10
Deep Conv.	zmconv_alfa	Initial cloud downdraft mass flux	0.05	0.1	0.6
	zmconv_c0_lnd	Deep convection precipitation efficiency over land	0.001	0.0059	0.01
	zmconv_c0_ocn	Deep convection precipitation efficiency over ocean	0.001	0.045	0.1
	zmconv_dmpdz	Parcel fractional mass entrainment rate	0.0002	0.001	0.002
	zmconv_ke	Evaporation efficiency parameter	5e-07	1e-06	1e-05
	zmconv_tau	Convective time scale	1800	3600	28800

Table 1: A summary of the input parameters perturbed in the LLNL ensemble of CAM5 simulations, along with their default value and the assumed uncertainty range.

precipitation at a 2.5° spatial resolution from 1979 to the present. The data are created by fusing together precipitation gauge data, sounding data, and satellite microwave data.

Figure 1 gives a spatial summary of the 5-year average daily precipitation in the 1145 CAM5 simulations and the GPCP data at the 2° grid used in the CAM5 simulations. Figures 2 and 3 compare the average daily precipitation at a coarser spatial scale, in six 30° zonal bands in two seasons (Dec-Jan-Feb and Jun-Jul-Aug) for 440 LH-sampled CAM5 simulations and the GPCP data. As the plots in Figure 2 and the histograms in Figure 3 show, at first glance, for some regions and seasons, the GPCP observation is not "covered" by the spread provided by the ensemble of CAM5 simulations (the two CAM5 LHS are those with a revisited range for one parameter). If that is the case, then one would have no hope of calibrating the uncertain parameters to "match" the observations, hinting at a large structural error in CAM5 for this particular output quantity of interest (or possibly an uncertain input parameter that was not sampled!).

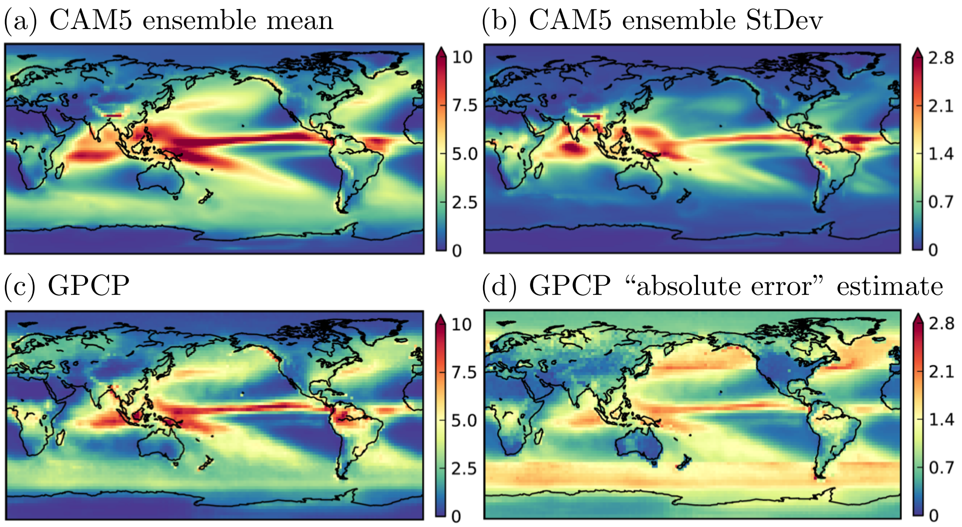


Figure 1: The two top maps summarize average daily precipitation (mm/-day) from 1,100 input-perturbed CAM5 LH-sampled simulations, while the bottom maps summarize the corresponding observed precipitation according the the GPCP data.

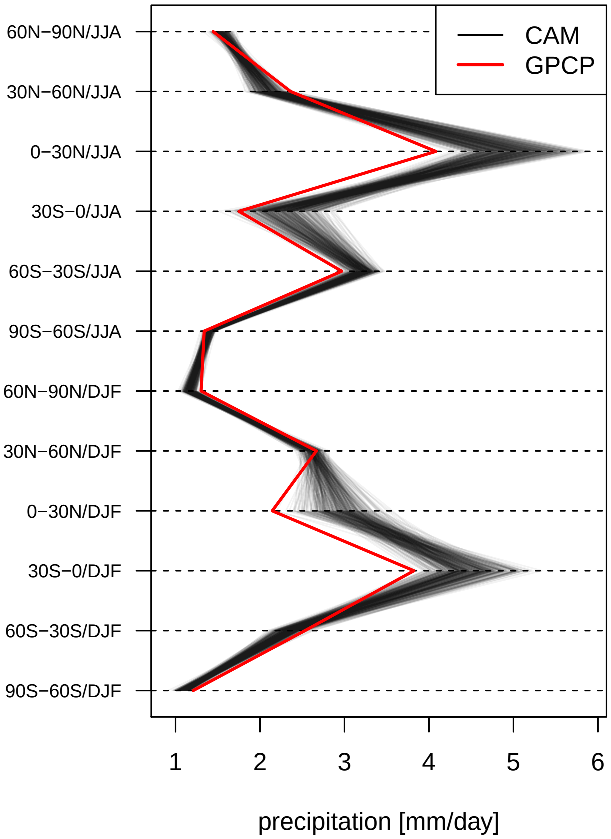


Figure 2: The gray lines show the results of 440 CAM5 simulation average daily precipitation (mm/day) in 6 30° zonal bands in Jun-Jul-Aug (JJA) and Dec-Jan-Feb (DJF). The red line shows the corresponding observations from GPCP.

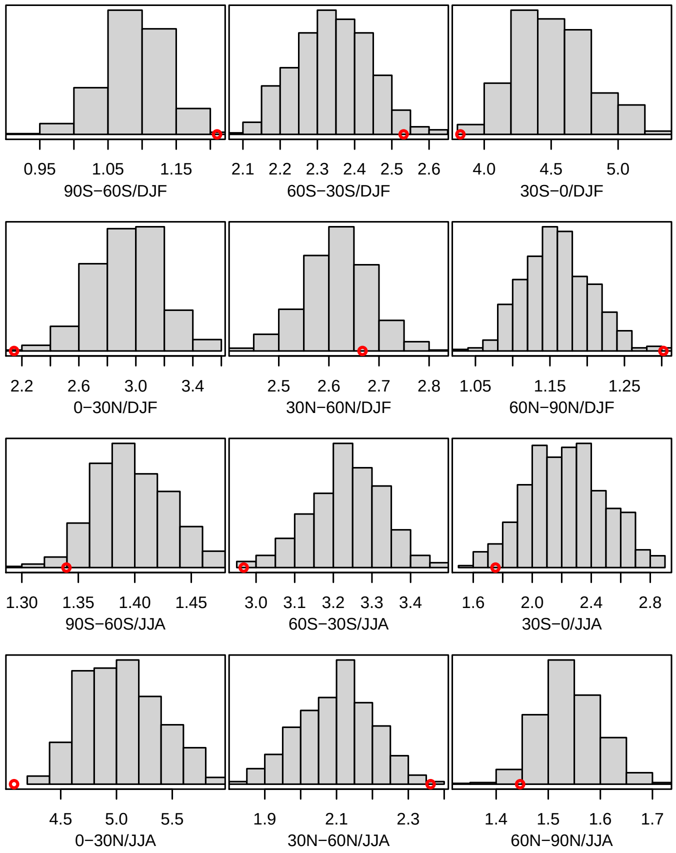


Figure 3: Comparisons of the CAM5 simulations and GPCP precipitation data from Figure 2. Each panel shows the histogram of the CAM5 ensemble data for a given zonal band and season along with the observed GPCP datum (red dot).

3 Methodology

We will introduce two different approaches to construct metrics and tests to assess the size of the structural error. The first approach in inspired by classical statistical test statistics and uses bootstrap methods (resampling methods) to estimate p-values to reject the hypotheses of no (or small) structural error. The second approach draws on metrics that measure the difference between two distributions.

3.1 Bootstrap-based P-values for Structural Error Test Statistics

Bootstrapping is a statistical resampling procedure that allows one to estimate the sampling distribution of nearly any statistic of interest using relatively simple resampling methods (Efron and Tibshirani, 1993). We will develop two flavors of bootstrap-based tests for structural error: (1) a test for the presence of structural error in a single scalar output quantity of interest and (2) a test for the presence of structural error in a given diagnostic statistic. The first family of tests simply focuses on a given scalar output quantity of interest, $y(x)$, and compares it to the empirical distribution of $\eta(x, \boldsymbol{\theta}) + \epsilon(x)$, similar to the comparison visualized in Figure 3. The metric test, on the other hand, considers a particular diagnostic statistic, for example, the expected root mean-squared-error between two spatial maps, and compares the observed diagnostic statistic to the empirical distribution induced by $\eta(x, \boldsymbol{\theta}) + \epsilon(x)$.

Let

$$Y \equiv \eta(\boldsymbol{\theta}) + \epsilon, \tag{2}$$

where we have dropped the dependence on x to simplify notation, but it should be clear that Y, and hence η and ϵ, refer to a particular observation/output of interest. Then Y is a stochastic variable with its probability distribution determined by the function $\eta(\cdot)$, and the two probability distributions $p(d\boldsymbol{\theta})$ and $p(d\epsilon)$, where $p(d\boldsymbol{\theta})$ represents the prior uncertainty in the inputs and $p(d\epsilon)$ the uncertainty in the observations. In what follows, we assume that $\boldsymbol{\theta}$ and ϵ are independent random variables. Given a method to generate realizations from both $p(d\boldsymbol{\theta})$ and $p(d\epsilon)$, we can generate realizations from $p(dY)$. For a computationally expensive $\eta(\cdot)$, as is the case here, we only have a limited number of realizations from $p(d\eta) = p(\eta(d\boldsymbol{\theta}))$, which is our ensemble of simulations carried out for a sample of realizations of $\boldsymbol{\theta}$

(i.e., our ensemble of CAM5 simulations). Let

$$\eta_i \equiv \eta(\boldsymbol{\theta}_i), \text{ for } i = 1, \ldots, n,$$

denote the ensemble of the simulated output quantity of interest for a given set of realizations of $\boldsymbol{\theta}$; $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$. Given $\{\eta_i\}$, then

$$\tilde{p}(d\eta) \equiv \sum_{i=1}^{n} w_i I(\eta_i = \eta), \tag{3}$$

is an empirical approximation to $p(d\eta)$, where $I(x = y) = 1$ if $x = y$ and is otherwise 0, and $w_i = 1/n$. Algorithm 1 outlines how one can generate realizations from $\tilde{p}(d\eta)$.

Assuming that the probability distribution of ϵ is known (e.g., a Gaussian distribution with known mean and variance) and a method to generate realizations of ϵ is available, we can generate realizations from $\tilde{p}(dY)$, an empirical approximation of $p(dY)$, as outlined in Algorithm 1. These are the fundamental building blocks we use to estimate p-values for test statistics to assess the presence of a significant structural error.

Finally, we note that the general approach outlined for the scalar quantity of interest can be generalized to an m-dimensional observation vector $\mathbf{y}$ and the corresponding model output vector $\boldsymbol{\eta}(\boldsymbol{\theta})$ and observation error vector $\boldsymbol{\epsilon}$. A potential additional complication that may arise in generating a realization of $\mathbf{Y} \equiv \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$ is that the observation errors may be correlated (e.g. $\boldsymbol{\epsilon} \sim \text{Gau}(\boldsymbol{\mu}_m, \Sigma_{m\times m})$ and the $m \times m$ covariance matrix $\Sigma_{m\times m}$ is not diagonal and m is large).

Algorithm 1 Generate realizations from $\tilde{p}(d\eta)$ and $\tilde{p}(dY)$

```
procedure GENERATE $\eta^* \sim \tilde{p}(d\eta)$
    Generate $i^* \in \{1, \ldots, n\}$ with weights $w_1, \ldots, w_n$
    return $\eta^* = \eta_{i^*}$
end procedure

procedure GENERATE $y^* \sim \tilde{p}(dY)$
    Generate $\eta^* \sim \tilde{p}(d\eta)$
    Generate $\epsilon^* \sim p(d\epsilon)$
    return $y^* = \eta^* + \epsilon^*$
end procedure
```

Scalar Test

The scalar test is relatively straightforward. Under the H_0 (null) hypothesis, the observation of interest is a realization from the stochastic process $Y = \eta(\boldsymbol{\theta}) + \epsilon$. That is,

$$H_0 : \quad y \sim p(dY) \quad \text{where} \quad Y = \eta(\boldsymbol{\theta}) + \epsilon. \tag{4}$$

To assess the validity of H_0, we estimate the probability $p_l \equiv \text{Prob}\{Y \leq y\}$ under this hypothesis; that is, the probability of encountering a realization of Y that is smaller or equal to y, the actual observation. As we are generally concerned with observing y in the "tails" of $p(dY)$, that is having either very low p_l or $1 - p_l$ ($= \text{Prob}\{Y > y\}$), we define

$$p_0 \equiv \min(p_l, 1 - p_l).$$

Then p_0 can be used to reject H_0 using a given "α" level, for example, reject H_0 for any output that yields $p_0 < 0.01$.

The resampling algorithm to estimate p_l is trivial and given in Algorithm 2.

Algorithm 2 P-value for a scalar quantity of interest.

```
procedure SCALARPVAL($y$)
    Initialize $c \leftarrow 0$
    for $b \leftarrow 1, \ldots, B$ do
        Generate $y^* \sim \tilde{p}(dY)$                    ▷ (see Algorithm 1)
        if $y^* \leq y$ then increment $c$ by 1
    end for
    $p_0 \leftarrow \min(c, B - c)/B$
    return $p_0$
end procedure
```

Diagnostic Statistic Test

The scalar test is well-suited for the setting with a single output quantity of interest. For a multivariate output, such as a spatial map or a zonal profile of precipitation, one could in principle compute the p-value for each grid cell or zonal band and inspect the resulting "map" of p-values. While such a map might be informative for regional/seasonal variation in the role of the structural error term, it is non-trivial to combine the multiple p-values to form a single overall "metric" for the structural error for that particular multivariate output.

An alternative approach is to define a single diagnostic test statistics that measures the discrepancy between two output vectors, $\mathbf{y} = (y_1, \ldots, y_m)^T$ and $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_m)^T$,

$$T(\mathbf{y}, \hat{\mathbf{y}}) = \text{a diagnostic discrepancy statistic},$$

where $\mathbf{y}$ is typically derived from the observations and $\hat{\mathbf{y}}$ from the model output. Two examples of such diagnostic statistics are:

$$\begin{aligned} T_{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) &\equiv \sqrt{\frac{1}{m} \sum_{j=1}^{m} v_j (y_j - \hat{y}_j)^2}, \\ T_{ABS}(y, \hat{y}) &\equiv |y - \hat{y}|. \end{aligned} \tag{5}$$

The first metric above is the commonly used weighted root mean-squared-error metric with weights v_i's while the second one is a simple metric for the absolute (L1) difference between two scalar quantities, which could be used in place of the scalar test introduced previously.

For a given observed $\mathbf{y}$, we define the expected $T(\mathbf{y}, \cdot)$ metric as

$$T(\mathbf{y}) \equiv E_{\mathbf{Y}}[T(\mathbf{y}, \mathbf{Y})|\mathbf{y}] = \int T(\mathbf{y}, \mathbf{Y}) p(d\mathbf{Y}),$$

where the distribution of $\mathbf{Y}$ is induced by the model $\mathbf{Y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}$. Hence, $T(\mathbf{y})$ is simply the average value of the diagnostic statistic $T(\mathbf{y}, \hat{\mathbf{y}})$, averaged over possible values of $\hat{\mathbf{y}}$, assuming $\hat{\mathbf{y}} \sim p(d\mathbf{Y})$ (i.e., no structural error). The null hypothesis of interest H_0 is therefore given by

$$H_0: \quad T(\mathbf{y}) \sim p(T(d\mathbf{Y})) \quad \text{where} \quad \mathbf{Y} = \boldsymbol{\eta}(\boldsymbol{\theta}) + \boldsymbol{\epsilon}.$$

The test of this hypothesis is equivalent to answering the question: is the observed $T(\mathbf{y})$ a realization from a distribution of $T(\mathbf{Y})$ that assumes no structural error? We can test the validity of H_0 by computing $p_0 = \text{Prob}\{T(\mathbf{Y}) > T(\mathbf{y})\}$, which is the probability of observing more extreme (larger) value than $T(\mathbf{y})$ when we assume no structural error (H_0). Algorithm 3 gives a resampling algorithm to estimate $T(\mathbf{y})$ and p_0.

3.2 Discrepancy Measures Between Two Probability Distributions

The second approach we investigate to assess the presence of significant structural error in a quantity of interest is centered on techniques to compare two (univariate) probability distribution functions (PDFs). The two

Algorithm 3 P-value for a diagnostic statistics of interest.

```
procedure diagnosticPval($\mathbf{y}$)
    $t_0 \leftarrow$ getAveT($\mathbf{y}$)                                  ▷ (see below)
    Initialize $c \leftarrow 0$
    for $b \leftarrow 1, \ldots, B$ do
        Generate $\mathbf{y}^* \sim \tilde{p}(d\mathbf{Y})$                  ▷ (see Algorithm 1)
        $t^* \leftarrow$ getAveT($\mathbf{y}^*$)
        if $t^* > t_0$ then increment $c$ by 1
    end for
    $p_0 \leftarrow c/B$
    return $p_0$
end procedure

function getAveT($\mathbf{y}$)
    for $b \leftarrow 1, \ldots, B$ do
        Generate $\mathbf{y}^* \sim \tilde{p}(d\mathbf{Y})$                  ▷ (see Algorithm 1)
        $t^{(b)} \leftarrow T(\mathbf{y}, \mathbf{y}^*)$
    end for
    $t \leftarrow (\sum_{b=1}^{B} t^{(b)})/B$
    return $t$
end function
```

PDFs we compare represent both information about the true underlying unobserved process of interest, $\xi(x)$, which we attempt to measure with $y(x)$ and simulate (i.e., predict) with $\eta(x, \boldsymbol{\theta})$, using the framework of Kennedy and O'Hagan (2001):

$$\begin{aligned} y(x) &= \xi(x) + \epsilon(x), \\ \xi(x) &= \eta(x, \boldsymbol{\theta}) + \delta(x). \end{aligned} \tag{6}$$

Under the hypotheses that the structural error $\delta(x) = 0$, we have that the "prior" distribution for $\xi(x)$ is simply given by $p(d\eta(x))$, that is,

$$\pi_o(d\xi(x)) \equiv p(d\xi(x)) \propto \int I(\xi(x) = \eta(x, \boldsymbol{\theta}))p(d\boldsymbol{\theta}),$$

which we simply approximate by the empirical distribution of the ensemble of simulations, $\tilde{p}(d\eta(x))$ in (3). The second distribution is the "posterior" distribution of $\xi(x)$ given $y(x)$ when assuming a "flat" prior and can be expressed as

$$\pi(d\xi(x)) \equiv p(d\xi(x) \,|\, y) \propto \int I(\xi(x) = y + \epsilon)p(d\epsilon).$$

To measure the discrepancy between π_o and π we investigate three well-known discrepancy measures:

Kullback-Leibler (KL) Divergence. The KL divergence measure is given by Kullback and Leibler (1951)

$$D_{KL}(P||Q) \equiv \int \log\left(\frac{p(x)}{q(x)}\right) p(x)dx,$$

where $p(\cdot)$ is the "true" PDF and $q(\cdot)$ is its estimate. Note that if $q(x) = p(x)$ (almost surely with respect to $p(\cdot)$), then $D_{KL}(P||Q) = 0$;n so the lower the value of this measure the better. Since we do not have a notion of the "true" PDF in this setting, we consider the symmetric version of the KL divergence, given by

$$\bar{D}_{KL}(P, Q) = (D_{KL}(P||Q) + D_{KL}(Q||P))/2 \tag{7}$$

Bhattacharyya Coefficent (BC). The BC is given by (Bhattacharyya, 1943)

$$D_B(p, q) = \int \sqrt{p(x)q(x)}dx. \tag{8}$$

Note that if there is no overlap between $p(\cdot)$ and $q(\cdot)$, then $D_B(p, q) = 0$ and if $q(x) = p(x)$ then $D_B(p, q) = 1$, so the higher the value of the coefficient the better.

Earth Mover's Distance (EMD). The EMD can be described as the smallest effort needed to move/transform one pile of 1D dirt, with its profile given by $p(\cdot)$, into another pile, given by $q(\cdot)$. There is a well-established algorithm to compute the EMD (the Hungarian algorithm to solve the instant transportation problem). In our case, with two PDFs of equal mass over the domain of interest, the EMD is equivalent to the 1st Mallow's distance and the 1st Weisserstein distance, and in general, the EMD is known as the Weisserstein distance in mathematics (`http://en.wikipedia.org/wiki/Earth_mover's_distance`).

In the application to CAM5 in Section 4, all three distance metrics described above are estimated using realizations from the two PDFs to be compered: π_0 and π. This is done by using the same "bins" (break points) when "binning" the realizations, that is, by using the same break points when creating the empirical histogram of the realizations from the two target PDFs. For example, if p_j is the empirical density of the j-th bin for $p(x)$ (i.e., the number of realizations in the j-th bin divided by the total number) and similarly for $q(x)$, then

$$D_B(p,q) \approx \sum_j \sqrt{p_j q_j}$$

4 Application to CAM5

We now give a demonstration of the methods presented in the previous section to CAM5, using the ensemble of simulations and observations described in Section 2.

4.1 Bootstrap Test Statistics for Structural Error

To demonstrate the value of the scalar-based test statistics for structural error and the more general diagnostic-based test statistics, we will use the 5-year average daily precipitation observed and simulated in 6 30° zonal bands in Dec–Feb (DJF) and Jun–Aug (JJA), resulting in a total of 12 quantities of interest (see Figures 2 and 3).

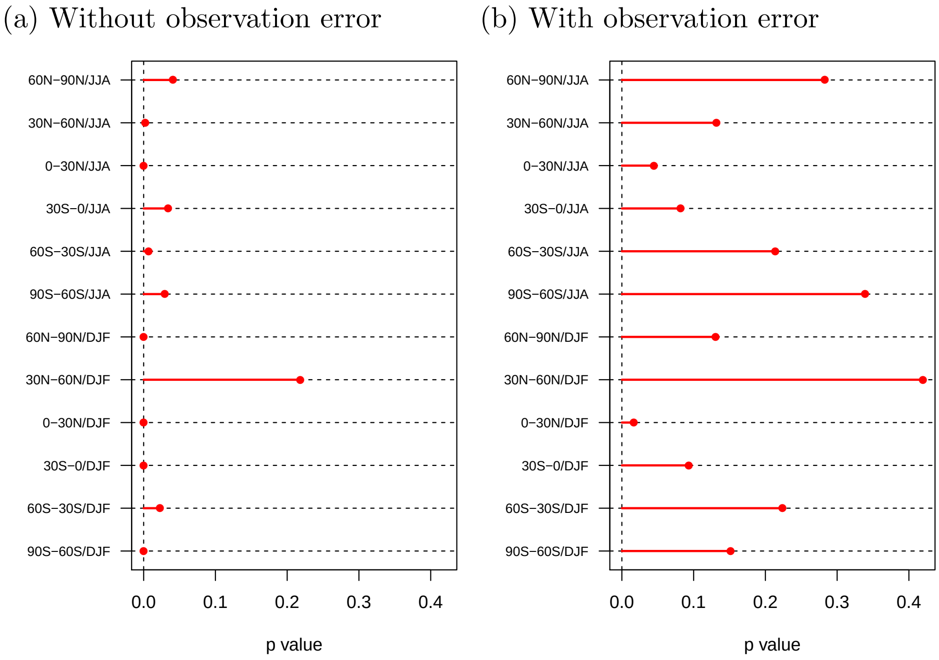
We present two cases. First, we assume that the error in the observations is minimal relative to the spread in the ensemble of simulations and therefore ignore it; that is, assume that $\epsilon = 0$ or very small in (2). In the second case we assume non-zero ϵ with a known error distribution, which we take to be Gaussian with standard deviation equal to 10% of the expected value; that is,

$$\epsilon_j \sim \mathrm{Gau}(0, (\sigma\eta_j)^2), \quad j = 1, \ldots, 12, \tag{9}$$

where $\sigma = 0.1$ and $\eta_j = \eta(x_j)$ is the true precipitation assuming no structural error. Hence, for a given ensemble member, $\eta(x_i, \boldsymbol{\theta}^*)$, the expected error distribution for the observation is $\mathrm{Gau}(0, (\sigma\eta(x_j, \boldsymbol{\theta}^*))^2)$. This yields a median standard deviation of 0.23 mm/day in $\{\sigma\eta(x_j, \boldsymbol{\theta}_i)\}$, with the 1st and 3rd quantiles at 0.14 and 0.31, respectively.

4.1.1 Tests Based on a Single Quantity of Interest

The procedure outlined in Algorithm 2 was used (with $B = 5000$) to obtain separate p-values for each of the 12 zonal precipitation averages when assuming (a) no or little observation error and (b) 10% observation error (as in (9)). The resulting p-values are shown in Figure 4. As expected, the p-values without any observation error (left) are smaller than those with observation error (right). In fact, when 0 observation error is assumed (left), 5 of the 12 p-values are identically equal to 0; that is, none of the 440 simulated precipitation values is more extreme than the observed GPCP value. This is not the case for any of the 12 p-values when we assume 10% observation error (the smallest p-value is 0.016). Note that the smallest p-values are found in the tropics (30°S–30°N).


(a) Without observation error

(b) With observation error

Figure 4: Estimated p-values for rejecting, independently, the H_0 hypotheses of no structural error in 6 30° zonal (aka, latitudinal) bands for two seasons, DJF and JJA. The left panel assumes no observation error, while the right panel assumes 10% error (as in (9)).

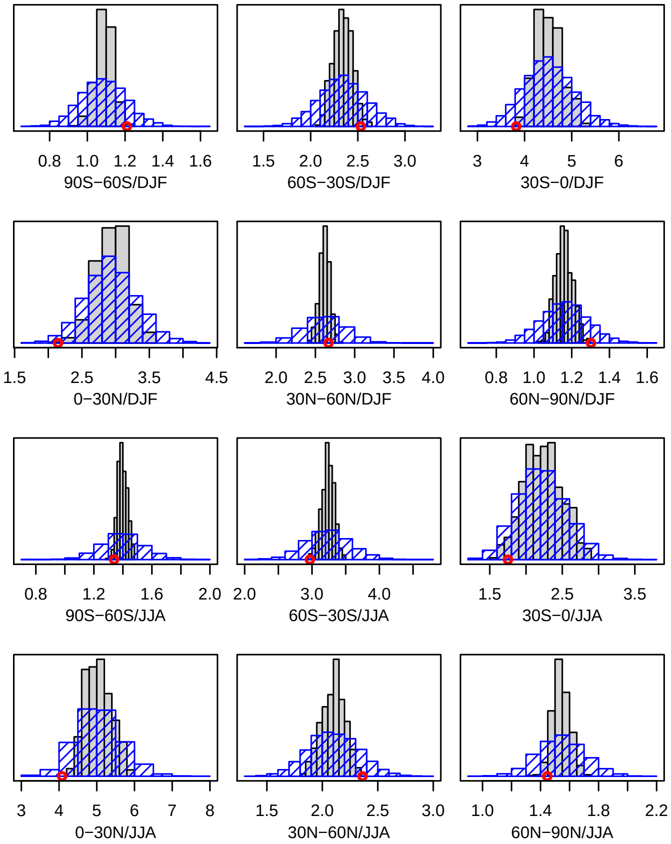


Figure 5: The empirical densities of $\eta(x, \boldsymbol{\theta})$ (gray) and $Y = \eta(x, \boldsymbol{\theta}) + \epsilon$ (blue) for the 6 zonal bands in two seasons. The GPCP observation is shown as a red circle.

Figure 5 compares the empirical densities of $\eta(x, \boldsymbol{\theta})$ and $Y(x)$, that is $\tilde{p}(d\eta(x))$ and $\tilde{p}(dY(x))$. Note that here x identifies the zonal band and season that η and Y correspond to. This clearly shows the impact of adding the observation noise to the ensemble spread, yielding a distribution with larger support, and potentially making what initially appears to be an implausible observation a plausible one.

4.1.2 Test Based on Diagnostic Statistics

The main motivation for the diagnostic test statistics was to test for structural error in a multivariate output of interest (a vector), but at the same time, one can develop diagnostic statistic for a scalar quantity of interest, for example, using the T_{ABS} diagnostic in (5). Figure 6 shows the p-values estimated by Algorithm 3 (with $B = 5000$) for the scalar diagnostic statis-

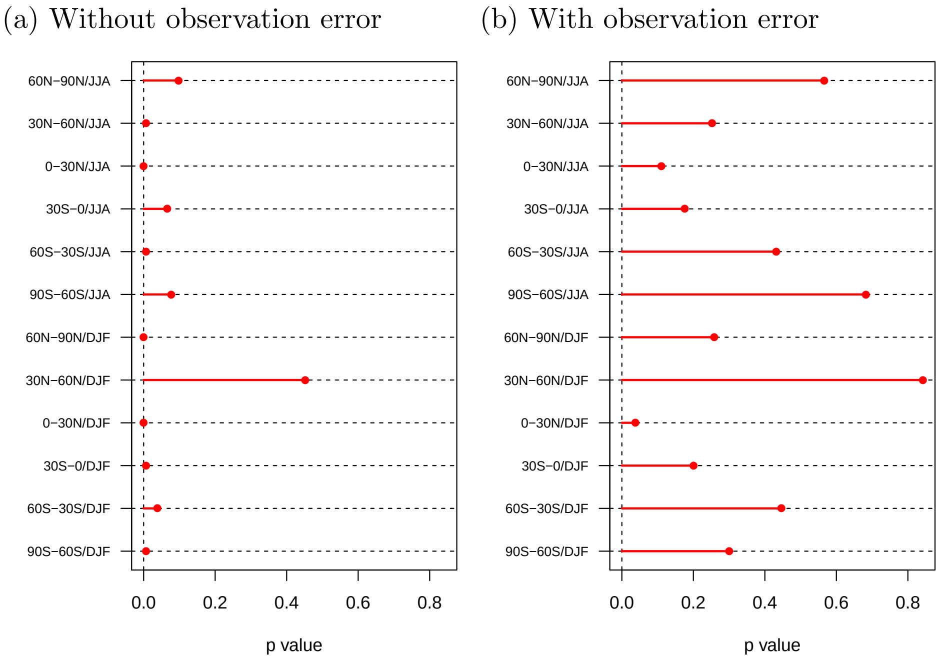
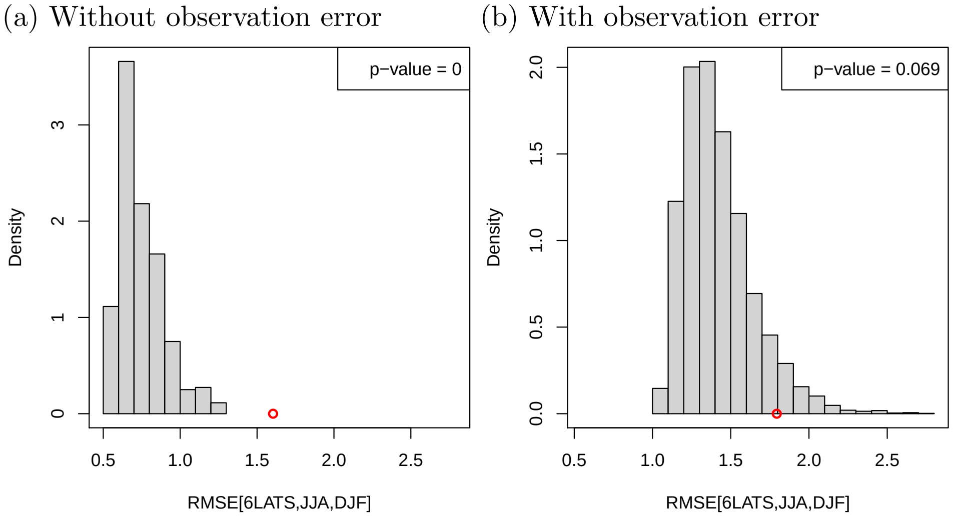


Figure 6: P-values for average daily precipitation in 6 30° zonal bands in two seasons (DJF and JJA) based on the T_{ABS} diagnostic statistics.

tics T_{ABS} for precipitation in the 6 zonal bands in DJF and JJA. The T_{ABS} p-values can be compared to those in Figure 4. As expected, they yield an identical pattern, but with p-values roughly twice as big since the p-values based on the T_{ABS} test statistics are "two sided", while the p-values derived with Algorithm 2 are "one-sided" (i.e., testing for observing something that is more extreme than is actually observed in just one direction).

The main benefit of the diagnostic statistic method is that it provides a measure of the discrepancy between two vectors, one corresponding to the simulation output and another to the observations. Figure 7 shows the H_0 distribution of the $T_{RMSE}(\cdot)$ under H_0 and the observed value of the statistics for $\mathbf{y}$ that is a vector of length 12 containing the daily average precipitation in 6 30° zonal bands in seasons DJF and JJA (with the zonal bands weighted according to area—i.e., the v_j's in (5)). Without any observation error, the observed RMSE diagnostic statistics is outside the support of its distribution under H_0 (i.e. p-value $= 0$ as estimated by Algorithm 3). However, as a result of adding observation error noise of 10%, the observed RMSE statistics become plausible albeit somewhat in the upper tail of the


(a) Without observation error

(b) With observation error

Figure 7: P-values for the T_{RMSE} diagnostic statistics that measure the difference in the daily precipitation in 6 30° zonal bands in two seasons (DFJ and JJA).

null distribution.

4.2 Metrics for Comparing PDFs

To demonstrate the value of the PDF discrepancy measures introduced in Section 3.2, consider the 2° gridded map of the annual average daily precipitation described in Section 2 and shown in Figure 1. The "prior" distribution (π_0) is given by the empirical distribution from the CAM5 ensemble of 1100 simulations for each gridbox, while the "posterior" is obtained from the distribution of the GPCP observations and the observation error process $\epsilon(x)$. We simply take $\epsilon(x) \sim \text{Unif}(-e(x), e(x))$, where $e(x)$ is the reported "absolute error" estimate in the GPCP (see Figure 1(d)). Figure 8 gives two examples of the empirical distribution of the annual average daily precipitation in two grid cells, as observed and simulated over 5 years (2000–2004). The CAM5 empirical distribution (π_0) is simply given by the 1100 CAM5 simulations, while π is obtained by generating 1100 realizations from $y(x) + \text{Unif}(-e(x), e(x))$.

The three PDFs discrepancy measures were applied to the empirical estimates of π_0 (CAM5) and π (GPCP) in each gridbox of the CAM5 grid. The resulting estimates are shown in Figure 9. There is good agreement

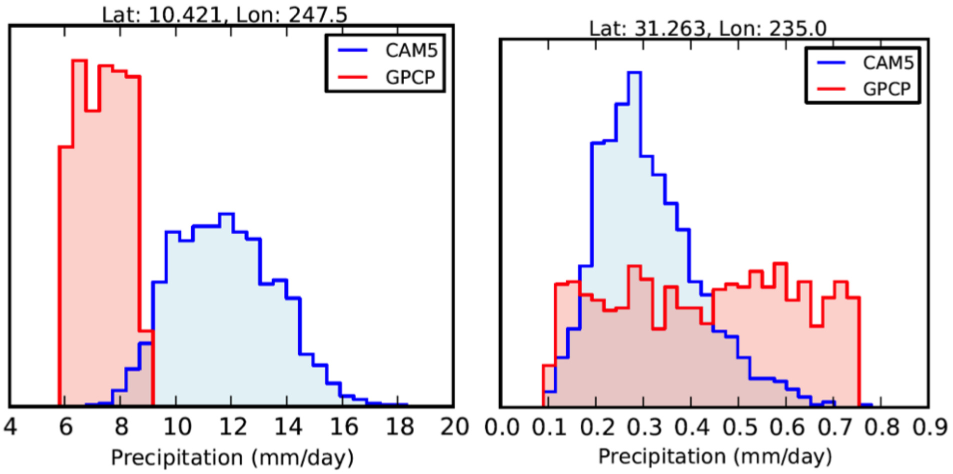
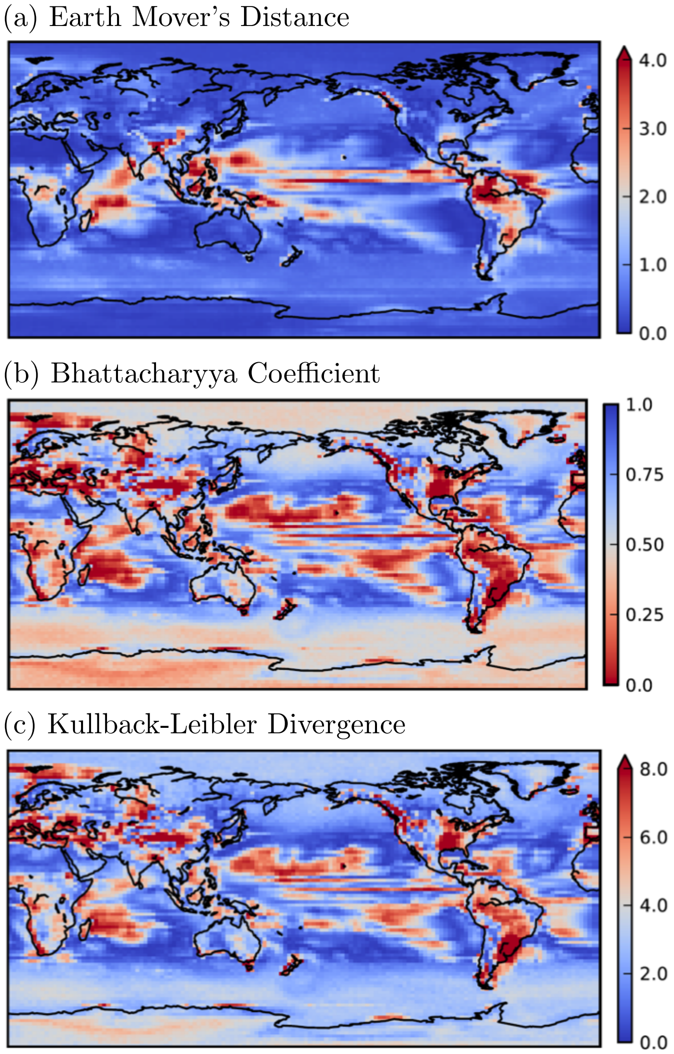


Figure 8: Two examples of CAM5- and GPCP-derived PDFs (π_0 and π) for precipitation in two different grid cells. The empirical CAM5 distribution is given by 1100 simulations, while the empirical GPCP distribution is derived from 1100 realizations from $y(x) + \text{Unif}(-e(x), e(x))$.

between the three distance measures, although the Kullback-Leibler and the Bhattacharyya visually appear to agree more. We observe the largest discrepancy in the tropics.


(a) Earth Mover's Distance

(b) Bhattacharyya Coefficient

(c) Kullback-Leibler Divergence

Figure 9: The three discrepancy metrics discussed in the text for comparing the PDFs of annual average daily precipitation at the grid box level. In all three plots, red denotes larger discrepancy between the PDFs.

5 Conclusion

We have presented multiple methods to detect structural error in simulating output metrics of interest, given an ensemble of simulations created by sampling uncertain input parameters. All the methods investigated here also account for the observation error in the quantities of interest, in addition to the input uncertainty. The first family of methods is based on resampling techniques to test a statistical hypothesis, while the second family of methods is based on metrics that quantify the difference between two PDFs. All the methods presented yield both qualitative visual diagnostic plots and p-values/metrics for a quantitative comparison.

All the investigated methods show a larger structural error in the tropics. The manner in which the observation error is accounted for appears to be important across all methods, as is seen in Figure 4.

Structural error tests can be used to screen a collection of important output metrics for use in Bayesian calibration/tuning of uncertain input parameters. Such screening would both focus on output metrics with small structural error, as judged by the methods presented here, as well as on output quantities that show large variation in the ensemble of simulations compared to the observation error. In addition, information about the shape and significance of the structural error can be valuable to model developers for determining the source of the structural error.

6 Acknowledgment

This work was funded by the Climate Science for a Sustainable Energy Future (CSSEF) Program under the Biological and Environmental Research (BER) Office in the U.S. Department of Energy. The authors specifically acknowledge the Program Manager, Dorothy Koch, and the leadership of CSSEF, including Stephen Klein, William Collins, and David Bader. The CAM5 ensemble of simulations was generated at LLNL by Don Lucas, John Tannahill, and Jim Boyle. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

References

M.J. Bayarri, J.O. Berger, R. Paulo, J. Sacks, J.A. Cafeo, J. Cavendish, C. Lin, and J. Tu. A framework for validation of computer models. *Technometrics*, 49:138–154, 2007.

K.S. Bhat, M. Haran, R. Olson, and K. Keller. Inferring likelihoods and climate system characteristics from climate models and multiple tracers. *Environmetrics*, 23(4):345–362, 2012.

A Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bull. Calcutta Math. Soc.*, 35:99–109, 1943.

CESM. The community earth system model (CESM). http://www2.cesm.ucar.edu/models.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993.

GPCP. Global precipitation climatology project (GPCP). http://www.gewex.org/gpcpdata.htm.

D. Higdon, J. Gattiker, B. Williams, and M. Rightley. Computer model calibration using high dimensional output. *Journal of the American Statistical Association*, 108:570–583, 2008.

Marc C. Kennedy and Anthony O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B, Methodological*, 63:425–464, 2001.

Solomon Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.

M. D. Mckay, R. J. Beckman, and W. J. Conover. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, 42(1):202–208, 2000.

B. Sanso and C.E. Forest. Uncertainty quantification: Statistical calibration of climate system propertie. *Journal of the Royal Statistical Society: Series C*, 58(4):485–503, 2009.