

## A multivariate empirical-orthogonal-function-based measure of climate model performance

Qiaozhen Mu, Charles S. Jackson, and Paul L. Stoffa

Institute for Geophysics, John A. and Katherine G. Jackson School of Geosciences, University of Texas at Austin, Austin, Texas, USA

Received 29 January 2004; revised 10 May 2004; accepted 25 May 2004; published 5 August 2004.

[1] A measure of the average distance between climate model predictions of multiple fields and observations has been developed that is based on the use of empirical orthogonal functions (EOFs). The application of EOFs provides a means to use information about spatial correlations in natural variability to provide a more balanced view of the significance of changes in model predictions across multiple fields, seasons, and regions. A comparison is made between the EOF-based measure and measures that are normalized by grid point variance and spatial variance for changes in the National Center for Atmospheric Research Community Climate Model, Version 3.10 (CCM3.10), parameter controlling initial cloud downdraft mass flux (ALFA), an important parameter within the *Zhang and McFarlane* [1995] convection scheme. All measures present consistent views that increasing ALFA from its default value creates significant improvements in precipitation, shortwave radiation reaching the surface, and surface latent heat fluxes at the expense of degrading predictions of total cloud cover, near-surface air temperature, net shortwave radiation at the top of the atmosphere, and relative humidity. However, the relative importance of each of these changes, and therefore the average view of the change in model performance, is significantly impacted by the details of how each measure of model performance handles regions with little or no internal variability. In general, the EOF-based measure emphasizes regions where modeled-observational differences are large, excluding those regions where internal variability is small.

*INDEX TERMS:* 3309 Meteorology and Atmospheric Dynamics: Climatology (1620); 3314 Meteorology and Atmospheric Dynamics: Convective processes; 3337 Meteorology and Atmospheric Dynamics: Numerical modeling and data assimilation; 3394 Meteorology and Atmospheric Dynamics: Instruments and techniques; *KEYWORDS:* climate prediction, skill scores, numerical modeling

**Citation:** 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.

### 1. Introduction

[2] Composite measures of model performance or skill scores have traditionally been used within weather prediction to compare different forecast systems and to track forecast improvements over time [e.g., *Murphy*, 1988; *Murphy and Epstein*, 1989]. Taylor diagrams provide an analogous role for depicting climate model performance through the comparison within a polar-coordinate diagram of the standard deviations and spatial correlations of model predictions and observations [*Taylor*, 2001; *Boer and Lambert*, 2001]. Here we consider ways of optimizing these skill scores so that they may be effective within numerical algorithms that automatically navigate the changes that may be required to tune a climate model to reproduce observational data or quantify climate model parameter uncertainties [*Jackson et al.*, 2004]. Ideally, skill scores should reflect a balanced measure of the mismatch between model pre-

dictions and observations over any number of different regions, seasons, and fields. Difficulties arise, however, in that observations do not exist at the same spatial and temporal resolution as the model predictions. Moreover, there usually is a lack of adequate information about observational or model uncertainties that are used to normalize modeled-observational differences. Some of these challenges have been addressed in part by the effort to optimize the “detection and attribution” of climate change signals within the instrumental record of climate [i.e., *North and Stevens*, 1998; *Allen and Tett*, 1999; *Hegerl et al.*, 2000; *Hegerl and Allen*, 2002]. Here we consider one approach that applies these advances to the goal of improving multivariate measures of climate model performance.

### 2. Approaches to Error Normalization

[3] Within the general form of the “cost function” that defines how model errors may be quantified by a mean-square mismatch between model predictions and observations, one may include a normalization factor, the inverse

of the data covariance matrix  $\mathbf{C}^{-1}$ , which provides the primary means to weigh the significance of different aspects of a model's performance. The mathematical form of the cost function for  $N$  different fields  $\mathbf{d}_{\text{obs}}$  (e.g., surface air temperature, precipitation, etc. . .) and model predictions  $g(\mathbf{m})$  at  $M$  points (note that each field may contain a different number of points  $M$ ) and normalized by the inverse of a  $M \times M$  data covariance matrix  $\mathbf{C}^{-1}$  is

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

Equation (1) includes vector  $\mathbf{m}$  of model parameter values, and  $T$  indicates the matrix transpose. The data covariance matrix includes information about potential sources of observational or model uncertainty, including information about uncertainty originating from natural (internal) variability, measurement errors, or theory. This form of the mean-square error  $E$  is the appropriate form for assessing more rigorously the statistical significance of modeled-observational differences when it is known that distributions of model and observational uncertainty are Gaussian. If one assumes that uncertainties are spatially uncorrelated, the data covariance matrix will only contain nonzero elements along the diagonal. When considering uncertainty originating from spatially uncorrelated natural variability, each of these elements is equal to the variance of the natural variability within the corresponding grid point where model predictions are compared to observations. There is a potential danger in comparing observations to model predictions for points where the observed or modeled variance is very small (like rainfall over a desert). Any errors in the estimate of the variance at these points will be easily translated into errors within  $E$  by the singularity that occurs when taking the inverse of the data covariance matrix. Because models tend to underestimate small-scale variability anyway, any difference between grid point differences of models and observations that are normalized by grid point variance tend to be exaggerated. *Taylor* [2001] and *Boer and Lambert* [2001] avoided these singularities by normalizing by a variable's global spatial variance rather than the grid point temporal variance. Although this choice avoids the singularity problem, normalizing by the spatial variance is nonoptimal from two perspectives: First, modeled-observational differences will be more strongly weighted toward regions where the differences are the largest and not necessarily toward those regions where these differences are the most significant. Second, composite measures of model performance will tend to weigh different observable fields unevenly. For instance, fields whose spatial variability is primarily linked to the latitudinal gradient of solar radiation will tend to appear less sensitive to prescribed forcing experiments than other fields [*Williamson*, 1995; *Watterson*, 1996; *Watterson and Dix*, 1999].

[4] An alternate approach to quantifying modeled-observational differences that does not suffer from these shortcomings is provided, in part, by a number of authors working on the "detection and attribution" of climate change signals in the observational record [*North and Stevens*, 1998; *Allen and Tett*, 1999; *Hegerl et al.*, 2000; *Hegerl and Allen*, 2002]. These authors were mainly concerned with

comparing observational and modeled climate anomalies. There are some additional complications that arise in trying to quantify the large differences that occur between observed and modeled climate means. However, our interests are more akin to the detection and attribution problem insofar as we are primarily interested in quantifying the statistical significance of changes in model performance that occur as one navigates through parameter space. By keeping the large biases within the measure of model performance, one may track how changes to model parameter values either degrade or improve model agreement with observations.

[5] The optimal "detection and attribution" approach relies on the fact that there are significant correlations between grid point variances (the off-diagonal components of the data covariance matrix). Instead of simply comparing point differences between observations and model predictions, one may instead re-express these differences in terms of a linear combination of empirical orthogonal functions, or EOFs, which are maps of the spatial structures of independent expressions of correlated variability derived from time series of long integrations of an unforced climate system model. Each EOF has an associated eigenvalue equivalent to the portion of the total variance accounted for by that EOF. One may re-express equation (1) in terms of a linear sum of  $K$  EOFs,

$$E(\mathbf{m}) \approx \frac{1}{2N} \sum_{i=1}^N \left( \sum_{j=1}^K \frac{a_j^2}{\lambda_j^2} \right)_i, \quad (2)$$

where  $a_j$  are the coefficients of the series of EOFs that reconstruct modeled-observational differences, i.e.,

$$[\mathbf{d}_{\text{obs}} - g(\mathbf{m})] \approx \sum_{j=1}^K a_j \cdot \text{EOF}_j, \quad (3)$$

and  $\lambda_j$  is the variance accounted for by the  $j$ th EOF. The EOFs are ordered such that the first EOF is associated with the largest variance  $\lambda$  and the last EOF is associated with the smallest variance. By recasting modeled-observational differences in terms of a series of EOFs, one may terminate the series at the point when the model becomes inadequate to represent the variability that is present within the observational data. *Allen and Tett* [1999] (hereinafter referred to as AT99) propose one consistency check to detect for model inadequacy. Their method is to examine the running  $E(\mathbf{m})/(k-1)$  statistic, which should ideally be of order unity if the only difference between the observations and model predictions is noise with some allowance for the presence of systematic biases. Because AT99 only compared observed and modeled climate anomalies, their method provided a fairly clear indication of when the  $E(\mathbf{m})/(k-1)$  statistic became excessively different from unity. As we shall explain, this threshold is not so clear when one also includes the large systematic biases that exist between models and observations within the left-hand side of equation (3).

### 3. Experimental Design

[6] We evaluate the performance of the National Center for Atmospheric Research (NCAR) community climate

**Table 1.** Names and Descriptions of Model Quantities Used to Evaluate Model Performance

Model Quantity	Description
CLDTOT	total cloud cover
FLNT	net longwave radiation at the top of the atmosphere
FSNT	net shortwave radiation at the top of the atmosphere
FSDS	downwelling shortwave radiation at the surface
LHFLX	surface latent heat flux
PRECT	total precipitation
PSL	sea level pressure
RELHUM	relative humidity (zonal means at all levels)
SHFLX	surface sensible heat flux
<i>T</i>	temperature (zonal means at all levels)
TREFHT	air temperature at 2-m reference height
<i>U</i>	zonal winds (zonal means at all levels)

model version Community Climate Model, Version 3.10 (CCM3.10), within a number of 19-year-long Atmospheric Model Intercomparison Project II (AMIP-II)–style experiments [Gates *et al.*, 1999]. The model consists of an atmospheric general circulation model at  $\sim 2.8^\circ$  latitude by  $2.8^\circ$  longitude resolution (T42 truncation, 18 levels) with prescribed sea surface temperatures and sea ice extents as observed between December 1977 and December 1997. The first 11 months of the model integration are not used within the climatologies derived from the model integrations. Initial conditions were derived according to AMIP-II conventions, in which the model is first integrated for 5 years with climatological SSTs and ice extent. Because this version of the model does not permit fractional sea ice concentrations, we confine our analysis between the bands  $30^\circ\text{S}$  and  $60^\circ\text{N}$ .

[7] Fifteen standard AMIP-II experiments were conducted with different initial conditions. These experiments test the uncertainty in our ability to measure modeled-observational differences owing to the existence of internal variability. In addition, we conducted six additional experiments with varying values of a single model parameter ALFA, which controls the initial downdraft mass flux for convective cloud systems as formulated by Zhang and McFarlane [1995]. The model is quite sensitive to different choices in this parameter and therefore useful for demonstrating measures of model performance for different model configurations.

[8] We compare model predictions of 12 quantities with observational or reanalysis data (Table 1): Nine of the fields are defined on a latitude-longitude grid (with the CCM3.10 variable name in capitals): total cloud cover CLDTOT, net longwave radiation at the top of the atmosphere FLNT, net shortwave radiation at the top of the atmosphere FSNT, downwelling shortwave radiation at the surface FSDS, surface latent heat flux LHFLX, total precipitation PRECT, sea level pressure PSL, surface sensible heat flux SHFLX, and air temperature at 2-m reference height TREFHT. Three quantities are defined as zonal means on a latitude height grid: zonal winds *U*, relative humidity RELHUM, and temperature *T*. Table 2 lists the observational or reanalysis data that were used to evaluate these model quantities. These quantities were selected because of the existence of a corresponding instrumental or reanalysis data set; 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. In fact,

because of this need, the observational data products exist at the same or nearly the same resolution as model output.

## 4. An EOF-Based Multivariate Measure of Model Performance

### 4.1. Covariance Matrix

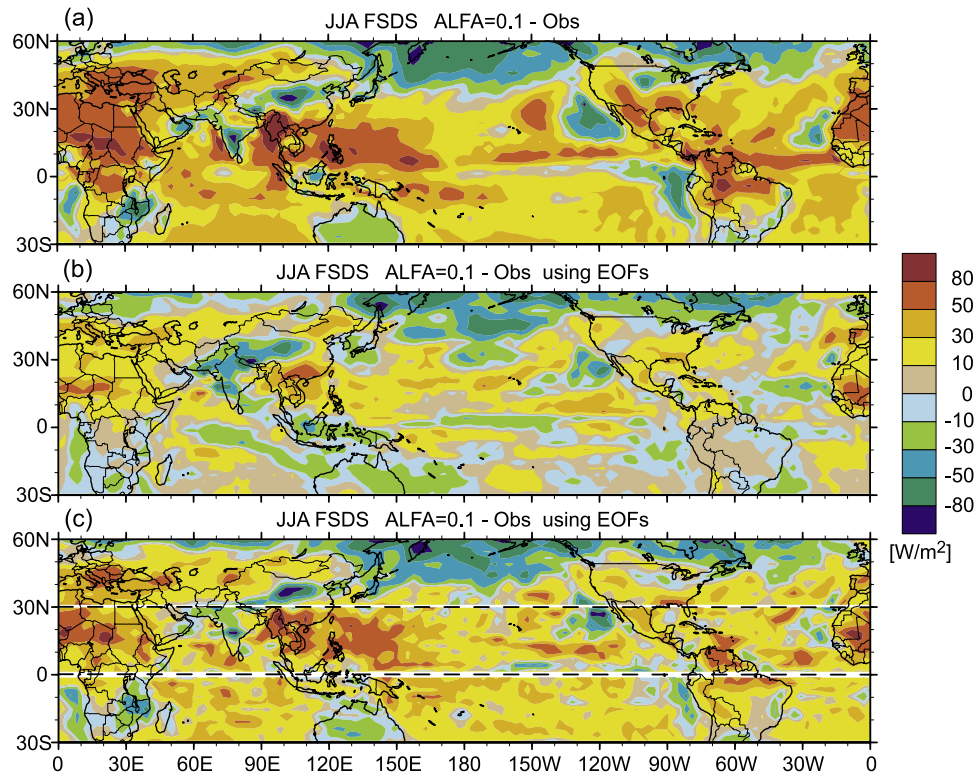
[9] Elements of the covariance matrix were calculated using a 290-year-long control integration of Community Climate System Model, Version 2.0 (CCSM2.0), a version of the coupled atmosphere-ocean climate system model developed at NCAR. Ideally, one would like a much longer integration so that one may accurately quantify the covariance between all model grid points that occurs within different 19-year segments, the length of the CCM3.10 AMIP-style experiments. The inverse covariance matrix based on variability of 19-year means of the control integration was singular, as was the inverse covariance matrix defined from the coupled model's interannual variability. These singularities simply reflect the fact that the number of degrees of freedom expressed within control integration is less than the rank of the covariance matrix.

[10] Following the example of comparisons of model predictions with observational records developed for the optimal detection and attribution of climate change (section 1), we attempted to express modeled-observational differences in terms of a series of EOFs defined from interannual variations of the 290-year-long control integration. However, we found that the EOFs were not able to sufficiently represent modeled-observational differences (Figures 1a and 1b). We attribute this to the fact that model means have large systematic biases relative to observations. By defining EOFs over limited regions, however, one may increase the ability of EOFs to represent separate portions of these large-scale differences. For fields that are defined on a latitude-longitude grid, we split the domain into  $30^\circ$  latitudinal bands ( $30^\circ\text{S}$  to  $0^\circ\text{N}$ ,  $0^\circ$ – $30^\circ\text{N}$ , and  $30^\circ$ – $60^\circ\text{N}$ ), which separates spatial structures of variability within the tropics from that of the middle to high latitudes. The idea is that spatial structures of variability would tend to be coordinated within each  $30^\circ$  band. Each latitude band contains approximately 1450 grid points. For fields that are defined as zonal

**Table 2.** Description of Observational or Reanalysis Data Used to Evaluate Model Quantities

Observational or Reanalysis Data Set <sup>a</sup>	Description	Model Quantities Evaluated
CMAP (Xie-Arkin)	1979–1998 instrumental record of precipitation	PRECT
ECMWF	1979–1993 reanalysis data	FSDS
ERBE	1985–1989 satellite observations of radiative fluxes	FSNT, FLNT
ISCCP	1983–1999 satellite observations of clouds	CLDTOT
Legates	1920–1980 instrumental record of air temperature at 2 m	TREFHT
NCEP	1979–1998 reanalysis data	RELHUM, <i>U</i> , <i>T</i> , SHFLX, PSL, LHFLX

<sup>a</sup>References: CMAP, Xie and Arkin [1996, 1997]; ECMWF, Gibson *et al.* [1997]; ERBE, Barkstrom *et al.* [1989]; ISCCP, Rossow *et al.* [1991]; Legates, Legates and Willmott [1990]; NCEP, Kalnay *et al.* [1996] and Kistler *et al.* [2001].

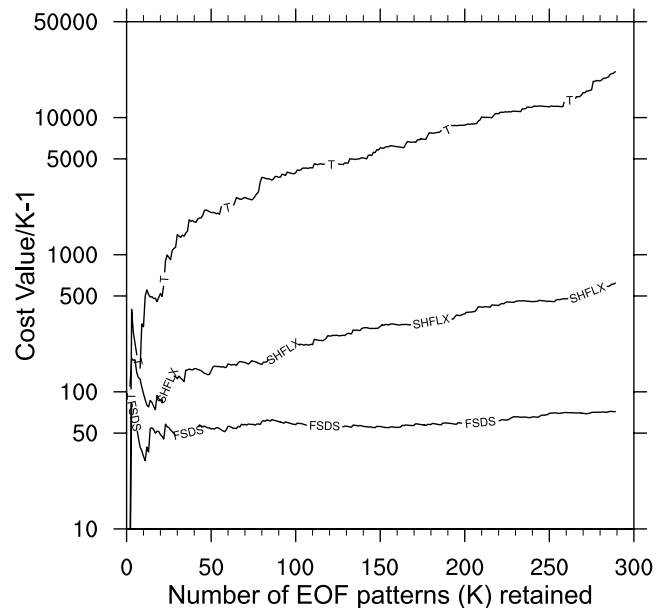


**Figure 1.** June through August mean difference between CCM3.10 and observations (ECMWF reanalysis “data”) of shortwave radiation downwelling at the surface FSDS. Panels show modeled-observational differences according to (a) grid point differences, (b) a series of 58 EOFs defined between 30°S and 60°N, and (c) a series of EOFs defined in 30° latitude bands. The number of EOFs retained for Figures 1b and 1c are determined by a selection criterion as described in section 4.2. Although we present truncated representations of modeled-observational differences, nontruncated representations of Figures 1b and 1c show similar differences in ability to capture the primary features of Figure 1a.

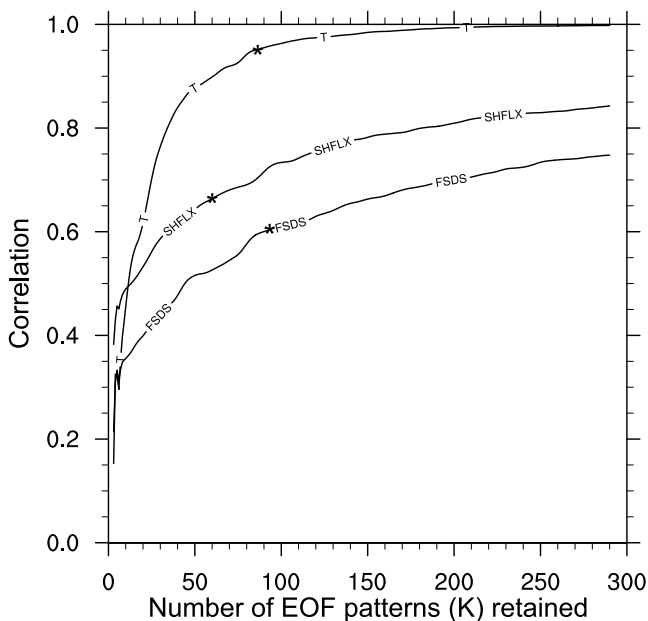
mean height cross sections, we did not need to subdivide the fields into latitudinal bands. For these upper air diagnostics, the number of grid points is 561 for  $T$  and  $U$  and 336 for RELHUM. The EOF reconstructed modeled-observational differences based on subregions improved the fidelity of the reconstruction for all fields considered (Figure 1c).

#### 4.2. EOF Truncation

[11] One advantage of using EOFs is the ability to identify and quantify the spatial structures of correlated variability at which models can be compared to data. AT99 argue that there should exist an EOF truncation number beyond which models would underrepresent the variability that exists within observational data. Including EOFs that account for too little variability relative to observations will cause a possibly misleading inflation of the measure of the mismatch. Although AT99 found a fairly clear indication of where this truncation should be, we found that the appropriate choice of the EOF truncation number can be obscured by systematic differences that exist between models and data at all scales. Figure 2 shows the value of the cost function (equation (2) normalized by the truncation number  $K$ ) with truncation number for variables  $T$ , SHFLX, and FSDS. If modeled-observational differences were only from natural variability, the value of the cost function should be of order unity until the point at which the model underrepresents observational variability. What Figure 2 shows



**Figure 2.** Change in the cost function value (normalized by the number of EOFs retained) as a function of the number of EOFs retained for JJA zonal mean temperature  $T$ , surface sensible heat flux SHFLX, and shortwave radiation downwelling at the surface FSDS for the 0°–30°N latitudinal band.



**Figure 3.** Spatial correlation between the biases within CCM3.10 model predictions as compared to observations and the reconstructed biases as a function of the number of EOF patterns retained for JJA zonal mean temperature  $T$ , surface sensible heat fluxes SHFLX, and shortwave radiation downwelling at the surface FSDDS for the  $0^{\circ}$ – $30^{\circ}$ N latitudinal band. The asterisk in each curve indicates the truncation determined by the criterion described in section 4.2.

instead is that the differences that exist between models and data are much greater than natural variability for all spatial scales. The matter of identifying an objective EOF truncation number is not well defined in this case.

[12] We propose an alternate way to determine an appropriate EOF truncation number. Rather than focus on whether the model adequately represents observational variability, a notion that may be too refined for the gross differences that exist between model and observational means, we choose to focus on determining the truncation number beyond which additional EOFs no longer contribute significantly to representing modeled-observational differences. Figure 3 shows a smooth curve of the improved spatial correlation between modeled-observational differences of  $T$ , SHFLX, and FSDDS with increasing number of EOFs retained in equation (2). The smoothed correlation curve was created by applying, 5 times, an 11-point Trenberth filter to the correlation versus number of EOFs retained curve [Trenberth, 1984]. Three-point and two-point smoothing functions were used near the lowest and highest truncation numbers. Assuming that each additional EOF corresponds to an effective degree of freedom of the actual variance represented by the modeled-observational differences, one may evaluate at what point the rate of improved correlation provided by the slope of the smoothed correlation curve is no longer significant according to standard significance tests of the Pearson correlation for a given number of effective degrees of freedom. A similar approach was taken by North and Wu [2001], although they primarily concentrated on the total amount of variance accounted for by

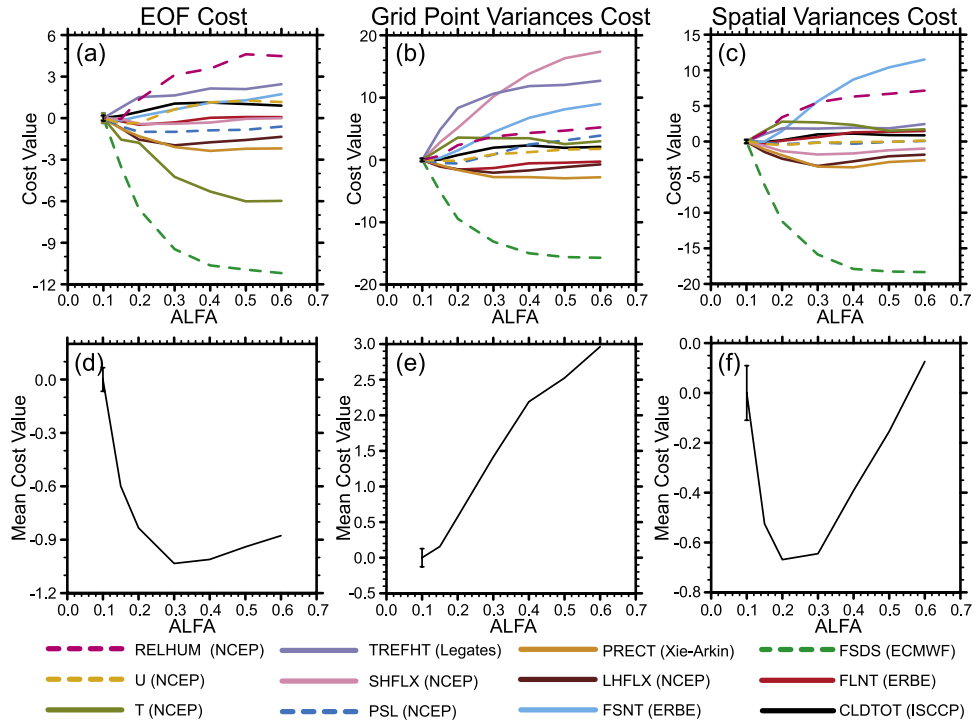
the retained EOFs. For example, for 30 EOFs, the critical slope is 0.007 correlation units for every additional degree of freedom at the 0.01 significance level. For 60 EOFs the critical slope is 0.0026 correlation units for every additional degree of freedom. We have defined our EOF truncation number to be the first point at which the slope of the smoothed correlation curve falls below 0.004 for potential truncation numbers below 50. If the potential truncation number is greater than 50, the threshold slope is 0.0014. (We have deliberately chosen not to be too rigorous with our treatment of the truncation number criterion as we do not want to convey the impression that this is an exact science despite our use of statistical thresholds.) Because the smoothed correlation curve may be negative when the number of EOFs is small (typically less than 10), we also stipulate that the correlation curve be smoothly varying at the point the truncation number is selected. We have tested the sensitivity of the cost function to 25% changes in the truncation number and found that the results are not significantly altered. The truncation numbers for  $T$ , SHFLX, and FSDDS are indicated within Figure 3. We determine a truncation number for every region and three-month season (DJF, MAM, JJA, SON) where EOFs are separately determined. The average truncation number for all fields, regions, and seasons for the latitude-longitude gridded data is 73 with a standard deviation of 21. While this number of retained EOFs may be considered high, it represents far fewer degrees of freedom than the approximately 1450 potential degrees of freedom in each of the  $30^{\circ}$  latitude bands.

### 4.3. Renormalization

[13] A natural way to compensate for the fact that modeled-observational differences within each region and season may be represented by a different number of EOFs is to normalize the cost function by the number of EOFs used to evaluate model performance in each of these regions and seasons. Even so, it is natural for there to be large differences in cost function values for different quantities as some quantities are easier to model than others. The real test as to whether these different quantities have been appropriately normalized is to consider whether the range in cost function values that arises from internal model variability and that has been separately determined for each region, season, and quantity is the same. Differences in this range may arise from the fact that the EOFs may have varying degrees of success in representing modeled-observational differences and therefore represent a different amount of the total variability. Moreover, EOFs that are derived from a model may under-represent the variability that is present in observational data. As our definition of the cost function does not change from experiment to experiment, we felt that it was appropriate to renormalize each region and season for every quantity by the range in cost function values that occurs from natural variability. We use the 15 model integrations of the standard model configuration with different initial conditions to estimate the effect of natural variability on the cost function.

## 5. Measures of Model Performance as Functions of ALFA

[14] We evaluate the performance of the Community Climate Model CCM3.10 in three ways according to the



**Figure 4.** Changes in cost function values relative to the mean of the 15 control experiments shown as a function of ALFA for 12 model fields defined by EOFs (Figure 4a), grid point variances (Figure 4b), and spatial variances (Figure 4c) and their respective means over the 12 fields (Figures 4d, 4e, and 4f). The range in the 15 control integrations is indicated by the vertical bar at ALFA = 0.1. The observational data set used to evaluate each field is indicated in parentheses. Note that cost values presented here are averaged over cost values obtained for each of the four seasons and three regions.

method of data normalization: a normalization by spatial variances, a normalization by grid point temporal variances, and a normalization based on EOFs. All measures of model performance or “costs” are of the form of equation (1), which quantifies modeled-observational differences in terms of a normalized square of modeled-observational differences. The only difference between cost definitions is in the way the data covariance matrix  $\mathbf{C}$  is defined. The normalization by spatial variances expresses matrix  $\mathbf{C}$  with the scalar  $c$  along its diagonal elements and is given by the spatial variances that exist within the 290-year control integration of CCSM2,

$$c = \frac{1}{N_i + N_j} \sum_{i=1}^{N_i} \sum_{j=1}^{N_j} \left( \bar{X}^t(i, j) - \bar{X}^{jt} \right)^2. \quad (4)$$

In equation (4) the  $i$  and  $j$  indices refer to grid points either on a two-dimensional latitude-longitude grid or latitude-height grid. The overbar with either the  $i$ ,  $j$ , or  $t$  superscript denotes an average over the respective grid dimension or time axis. For the normalization by grid-point variances, each component of the diagonal of matrix  $\mathbf{C}$  is defined separately as

$$c(i, j) = \frac{1}{N_t} \sum_{t=1}^{N_t} \left( X(i, j, t) - \bar{X}^t \right)^2. \quad (5)$$

The notation in equation (5) is similar to equation (4) except that  $c(i, j)$  refers to the element of the diagonal of matrix  $\mathbf{C}$

corresponding to grid point  $(i, j)$ . The normalization factor for the definition based on EOFs is the square of the eigenvalue ( $\lambda^2$ ) of each corresponding eigenvector of the covariance matrix (the EOFs). We shall refer to these different definitions simply as spatial variances cost, grid point variances cost, and EOF cost. All definitions only apply within the region 30°S to 60°N. In all cases we have renormalized each season, region, or field by the range in cost function values that were obtained in the 15 control experiments that differ by internal variability (see section 4.3). Figure 4 shows how CCM3.10 is affected by increasing parameter ALFA from 0.1 to 0.6. Figures 4a–4c show the change in model performance (one panel for each definition of the cost function) of individual fields, with the vertical bar centered on ALFA = 0.1 indicating the uncertainty by the range in cost values for the 15 control experiments. The range in cost values for these experiments is approximately equal to 1 cost unit. It differs slightly from unity when averaged over each season and region. For each field the mean cost of the 15 control experiments has been subtracted so that it is easier to compare how individual fields are affected by changes in ALFA.

[15] Figures 4d–4e show the average cost value among the 12 fields considered. One may contend that the equal weighting among an arbitrarily chosen number of observational constraints does not provide a completely objective means to measure model performance. One idea suggested by an anonymous reviewer is to use information about the correlations between the different fields to assess the information that is truly independent. In principle, one could derive a correlation matrix of cost values for each field from

a large number of control experiments. The principle components of this matrix define a new set of independent “factors.” One may select a reduced number of factors that are sufficient to explain the significant degrees of freedom exhibited by the fields considered. This idea has not been implemented here because the 15 control experiments are not sufficient to establish the statistical independence of the principle components of a  $12 \times 12$  correlation matrix.

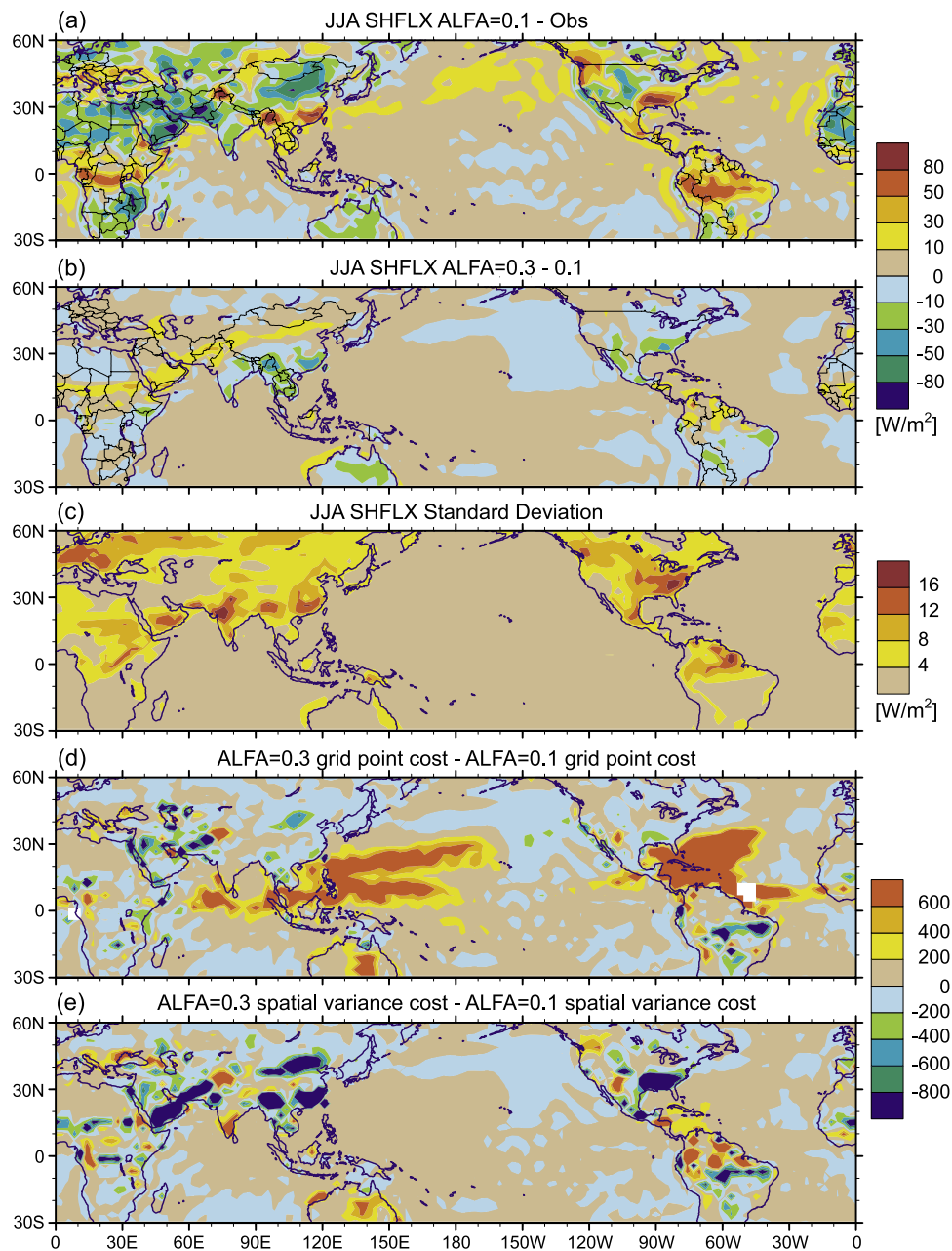
[16] The three alternate measures of model performance give different impressions of how ALFA affects model climate and modeled-observational biases. Increasing ALFA has the effect of strengthening convection, especially in the tropics. In particular, increasing ALFA reduces the double Intertropical Convergence Zone precipitation (PRECT) bias with respect to the Xie-Arkin observational data set in the western tropical Pacific, as well as biases throughout the Indian Ocean, Southeast Asia, the central Pacific, the Caribbean, Saudi Arabia, and Africa, but increases slightly the biases in South America and the eastern Pacific (not shown). It is interesting to note that although the effect of ALFA on precipitation is highly structured, many of these structures map very well onto the modeled-observational biases. This suggests that nonideal values of ALFA could be a leading explanation for modeled-observational biases in precipitation in this region. The improvement in precipitation shows up as a reduction in the cost values for the solid orange line in Figures 4a–4c. Associated with the increase in convection, there are significant increases in low- and high-level clouds (cost function only considers total cloud amounts CLDTOT) and significant decreases in the amount of shortwave radiation downwelling at the surface (FSDS). Although the reduction in FSDS going from ALFA = 0.1 to ALFA = 0.3 reduces the surplus radiation by 45%, model cloud amounts are already in excess of what is observed in ALFA = 0.1, and changing to ALFA = 0.3 only makes this bias worse. Consistent with these biases and changes in shortwave radiation and total cloud amounts, model simulations of air temperature at 2-m reference height (TREFHT) that contained a zonally averaged  $1^{\circ}$ – $4^{\circ}$ C cold bias in the tropics at ALFA = 0.1 were made worse at ALFA = 0.3 by  $\sim 10\%$ . The increase in cloud amounts also reduces further net shortwave radiation at the top of the atmosphere (FSNT), which at ALFA = 0.1 was already smaller than is observed. Increasing ALFA also increases relative humidity (RELHUM) through much of the tropical and midlatitude midtroposphere. This reduced some of the model biases in the subtropics; however, this change also made worse the excess relative humidity in the midlatitudes.

[17] All measures of model performance in Figures 4a–4c indicate that increasing ALFA improves simulations of FSDS, PRECT, and LHFLX while making worse simulations of CLDTOT, TREFHT, FSNT, and RELHUM. There were different assessments of whether or not there were improvements in zonally averaged air temperature  $T$ , sensible heat flux SHFLX, sea level pressure PSL, zonal winds  $U$ , and net longwave radiation at the top of the atmosphere FLNT. Even when these different measures agree in sign on how ALFA affects model performance, each measure assigns a different level of significance to how modeled changes affect model biases.

[18] To evaluate further what accounts for these differences, we consider similarities or differences in the change

in cost function measures for the JJA seasonal average for the surface sensible heat flux SHFLX and zonal mean air temperature  $T$ . In the first case, the normalization by grid point variance suggests that simulations of SHFLX became worse with increasing ALFA ( $\Delta\text{cost} = 17.25$  in contrast to  $-1.18$  for the EOF-based measure and  $-1.79$  for the normalization by spatial variances). Figure 5a shows the difference between observations and model predictions of SHFLX for ALFA = 0.1. The largest biases occur over land with a surplus of SHFLX in Southeast Asia, southeastern North America, and northern South America and a deficit over parts of central Asia, the Middle East, and northern Africa. All these biases are reduced as one changes ALFA to 0.3 (Figure 5b), except for northern South America. Figure 5c shows that the grid point variance is larger over the continents than over the ocean. Accordingly, the change in the cost function based on normalization by grid point variance (Figure 5d) has its largest values over the oceans. In contrast, the normalization by spatial variances emphasizes regions where modeled-observational differences are the largest irrespective of grid point variances. In this case, the improvements over continental regions allowed for the cost measure to be reduced as ALFA increased from 0.1 to 0.3. Unlike the normalization by grid point variances, the EOF-based measure of model performance emphasizes regions where variances are the largest, as it is in these regions that the EOF approach will do the best job in representing modeled-observational biases. The EOFs representing SHFLX are primarily representing changes over the continents as it is in these regions where the grid point variances are the largest. Accordingly, the EOF-based measure of model performance is mostly reflecting the improvements that are occurring over the continents.

[19] The vertical profile of zonally averaged air temperature  $T$  provides another good example to contrast the different ways that these three approaches to quantifying model performance work. Moreover, the JJA season contains the largest differences between the three approaches to quantifying climate model performance. In the case of  $T$ , the EOF cost measure suggests that the simulation with ALFA = 0.3 is significantly improved from simulations with ALFA = 0.1 ( $\Delta\text{cost} = -5.12$ ). In contrast, both the normalization by grid point variances and that by spatial variances indicate that increasing ALFA degrades predictions of  $T$  ( $\Delta\text{cost}$  of 4.88 and 4.28, respectively). For the JJA ALFA = 0.1 experiment, the primary difference from observations (National Centers for Environmental Prediction (NCEP) reanalysis data in this case) is that the model has a  $4^{\circ}$ – $5^{\circ}$ C cold bias in the upper tropical troposphere (near the tropopause) around 100 mbar and a  $3^{\circ}$ – $4^{\circ}$ C warm bias in the stratosphere around 30 mbar in the tropics. A  $1^{\circ}$ – $2^{\circ}$ C warm bias also exists at 70 mbar around  $60^{\circ}$ N as well as a  $1^{\circ}$ – $2^{\circ}$ C warm bias in the midtropical troposphere (Figure 6a). Increasing ALFA to 0.3 cools the tropical midtroposphere and lower stratosphere around 70 mbar. The cooling in the lower stratosphere makes worse the cold bias in the lower stratosphere but reduces some of the warm bias in the upper stratosphere (Figure 6b). The latitude-height cross section of variance (Figure 6c) shows enhanced tropical variability at 70 mbar up to the top of the model as compared to the troposphere. The cooling near the tropical tropopause (near 100 mbar), which brings the model further

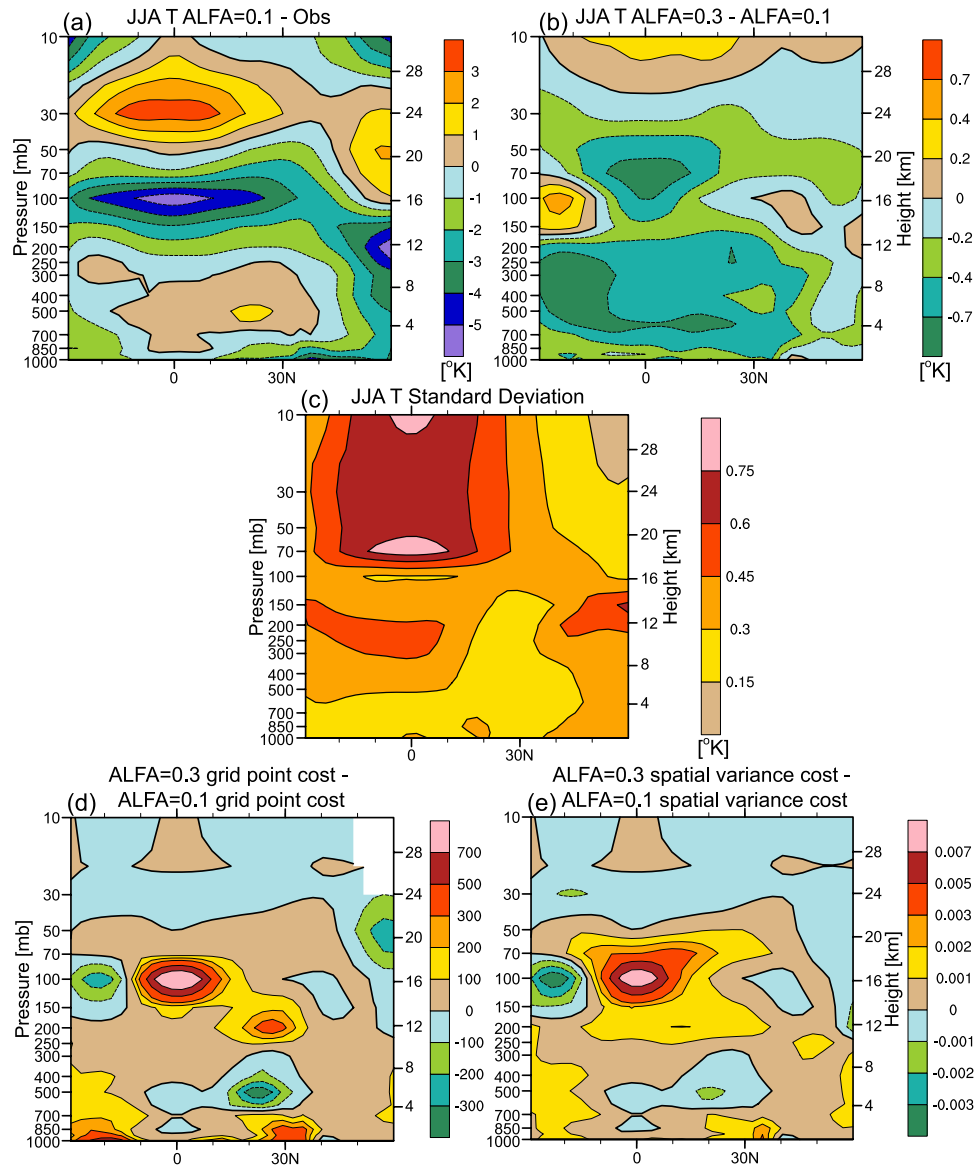


**Figure 5.** June through August mean surface sensible heat flux SHFLX. (a) Difference between CCM3.10 with ALFA = 0.1 and observations of SHFLX (obtained from NCEP), (b) difference in SHFLX with ALFA = 0.3 and ALFA = 0.1, (c) standard deviation of June–August mean SHFLX within the NCAR Community Climate System Model (CCSM2.0), (d) spatial distribution of change in cost function normalized by grid point variances when changing ALFA from 0.1 to 0.3, and (e) spatial distribution of change in cost function normalized by spatial variances when changing ALFA from 0.1 to 0.3.

away from observations and is in a location of small variance, dominates the grid point variance perspective that ALFA = 0.3 is worse than ALFA = 0.1. This degradation is tempered a bit by improvements in the tropopause and portions of the upper atmosphere (Figure 6d). The cost function normalized by spatial variances also emphasizes the stronger cold bias near the tropical tropopause but emphasizes slightly differently the heterogeneous mixture of areas of improvement and degradation (Figure 6e).

[20] Whereas the cost function normalized by spatial variances should be proportional to the area average of the

anomalies present within Figure 6e, the cost function based on EOFs has a less intuitive weighting system. To evaluate which aspects of the ALFA = 0.3 experiment improved from the EOF perspective, we have divided the primary anomaly field (Figure 6b) into a sum of EOFs that increase the cost value in going from ALFA = 0.1 to ALFA = 0.3 (Figure 7a) and a sum of EOFs that decrease the cost value (Figure 7b). The most significant aspect of these reconstructions is that the set of EOFs associated with a decrease in cost function values (model improvements) represent smaller amplitude changes than those features that are associated with an increase in cost



**Figure 6.** June through August mean, zonal mean, and height cross sections of air temperature  $T$ . (a) Difference between CCM3.10 with ALFA = 0.1 and observations of  $T$  (obtained from NCEP), (b) difference in  $T$  with ALFA = 0.3 and ALFA = 0.1, (c) standard deviation of June–August mean, zonal mean, and  $T$  within the NCAR Community Climate System Model (CCSM2.0), (d) spatial distribution of change in cost function normalized by grid point variances when changing ALFA from 0.1 to 0.3, and (e) spatial distribution of change in cost function normalized by spatial variances when changing ALFA from 0.1 to 0.3.

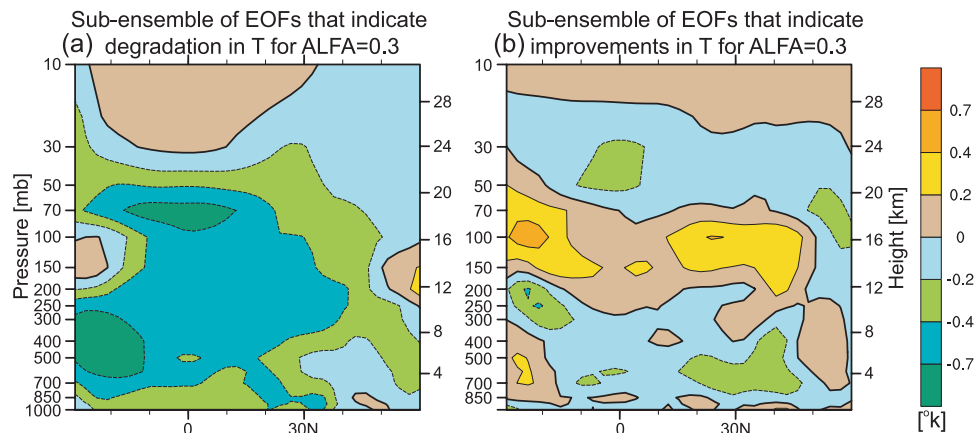
function values. This shows why the EOF-based measure of model performance may at times seem counterintuitive, as we know in this case that the smaller amplitude changes dominate in significance over the larger amplitude changes. This may in part be related to the fact that the anomalies presented in Figure 7b counter more precisely the model biases present in the ALFA = 0.1 experiment (Figure 6a). In contrast, the anomalies presented in Figure 7a represent a mixture of areas of improvement and degradation.

[21] One may question how robust the results are for cost measures of  $T$  given how the result depends on the relative weighting assigned to areas of improvement or degradation. The weight measures assigned to specific spatial structures are defined by the eigenvalues (equation (2)) and EOFs

derived from interannual variations of a 290-year integration of CCSM2.0. As a check, we have also recalculated the cost function measures of model predictions of  $T$  within the ALFA = 0.1 and ALFA = 0.3 experiments using eigenvalues and EOFs derived from time series of 3-year seasonal averages instead of 1-year seasonal averages. We find that the changes in model performance going from ALFA = 0.1 to ALFA = 0.3 based on 3-year seasonal mean variability are nearly identical to the results described above.

## 6. Summary

[22] An EOF-based multivariate measure of climate model performance is presented so that better use is made



**Figure 7.** Subensembles of EOFs that reconstruct the change in  $T$  (experiments ALFA = 0.3 minus ALFA = 0.1). (a) Sum of all EOFs that contribute to a degradation of the EOF-based cost function. (b) Sum of all EOFs that contribute to an improvement in the EOF-based cost function.

of pattern correlations and spatial differences in natural variability. This approach uses EOFs to recast modeled-observational differences in terms of a limited number of spatial structures and weighted inversely by the amount of variance each pattern accounts for within a time series of interannual variability. As modeled-observational differences can be quite large, it is necessary to break the domain ( $30^{\circ}\text{S}$  to  $60^{\circ}\text{N}$ ) into  $30^{\circ}$  latitude regional bands, which helps keep down the number of EOFs that are necessary to represent modeled-observational differences. The EOF truncation number is determined by a criterion based on the rate at which additional EOFs improve the spatial correlations between grid point modeled-observational differences and EOF reconstructed modeled-observational differences. On average, 73 EOFs are used to represent model data differences. This number is a significant reduction in the 1450 potential degrees of freedom represented by the approximate number of grid points in each  $30^{\circ}$  latitude band. The “cost function” measure of model performance for each region, season, and field has been renormalized by the range in cost values that occurs from internal variability. The EOF-based measure of model performance has been found to be relatively insensitive to 25% changes in the number of EOFs used to represent modeled-observational differences.

[23] The idea of using EOFs to evaluate model performance is motivated by the need to make better use of knowledge of pattern correlations and spatial differences in natural variability so that (1) the composite measure of model skill is not biased toward any single variable and (2) it maximizes the ability to detect changes in model predictions (increases the signal-to-noise ratio) without placing too much weight on regions where internal variability is small. These points were illustrated by comparison of the EOF-based measure of model performance against measures based on normalization by grid point variances and normalization by spatial variances. In sum, the EOF-based measure of model performance weighs more strongly modeled-observational differences that occur over regions where these differences are large and natural variability is well defined. These characteristics

permit a fairer comparison of the statistical significance of modeled-observational differences that occur between multiple fields.

[24] **Acknowledgments.** Financial support for this research was provided by the G. Unger Vetlesen Foundation. Correspondence and requests should be addressed to C. Jackson.

## References

- Allen, M. R., and S. F. B. Tett (1999), Checking for model consistency in optimal fingerprinting, *Clim. Dyn.*, *15*, 419–434.
- Barkstrom, B. R., E. Harrison, G. Smith, R. Green, J. Kibler, R. Cess, and ERAB Science Team (1989), Earth Radiation Budget Experiment (ERBE) archival and April 1985 results, *Bull. Am. Meteorol. Soc.*, *70*, 1254–1262.
- Boer, G. J., and S. J. Lambert (2001), Second-order space-time climate difference statistics, *Clim. Dyn.*, *17*, 213–218.
- Gates, W. L., et al. (1999), An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I), *Bull. Am. Meteorol. Soc.*, *80*(1), 29–56.
- Gibson, J. K., P. Kallberg, S. Uppala, A. Hernandez, A. Nomura, and E. Serrano (1997), ERA description, *ECMWF Re-Analysis Proj. Rep. Ser., Rep. 1*, 72 pp., Eur. Cent. for Medium-Range Weather Forecasts, Reading, England.
- Hegerl, G. C., and M. R. Allen (2002), Origins of model-data discrepancies in optimal fingerprinting, *J. Clim.*, *15*, 1348–1356.
- Hegerl, G. C., P. A. Stott, M. R. Allen, J. F. B. Mitchell, S. F. B. Tett, and U. Cubasch (2000), Optimal detection and attribution of climate change: Sensitivity of results to climate model differences, *Clim. Dyn.*, *16*, 737–754.
- Jackson, C. S., M. K. Sen, and P. L. Stoffa (2004), An efficient stochastic Bayesian approach to optimal parameter and uncertainty estimation for climate model predictions, *J. Clim.*, *17*, 2828–2841.
- Kalnay, E., et al. (1996), The NCEP/NCAR 40-year Reanalysis Project, *Bull. Am. Meteorol. Soc.*, *77*, 437–471.
- Kistler, R., et al. (2001), The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation, *Bull. Am. Meteorol. Soc.*, *82*, 247–267.
- Legates, D. R., and C. J. Willmott (1990), Mean seasonal and spatial variability in global surface air temperature, *Theor. Appl. Climatol.*, *41*, 11–21.
- Murphy, A. H. (1988), Skill scores based on the mean square error and their relationships to the correlation coefficient, *Mon. Weather Rev.*, *116*, 2417–2424.
- Murphy, A. H., and E. S. Epstein (1989), Skill scores and correlation coefficients in model verification, *Mon. Weather Rev.*, *117*, 572–581.
- North, G. R., and M. J. Stevens (1998), Detecting climate signals in the surface temperature record, *J. Clim.*, *11*, 563–577.
- North, G. R., and Q. Wu (2001), Detecting climate signals using space-time EOFs, *J. Clim.*, *14*, 1839–1863.

- Rossow, W. B., L. C. Garder, P. J. Lu, and A. W. Walker (1991), International Satellite Cloud Climatology Project (ISCCP) documentation of cloud data, *WMO/TD 266*, 76 pp. + appendices, World Meteorol. Organ., Geneva.
- Taylor, K. E. (2001), Summarizing multiple aspects of model performance in a single diagram, *J. Geophys. Res.*, *106*, 7183–7192.
- Trenberth, K. E. (1984), Signal versus noise in the Southern Oscillation, *Mon. Weather Rev.*, *112*, 326–332.
- Watterson, I. G. (1996), Non-dimensional measures of climate model performance, *Int. J. Climatol.*, *16*, 379–391.
- Watterson, I. G., and M. R. Dix (1999), A comparison of present and doubled CO<sub>2</sub> climate and feedbacks simulated by three general circulation models, *J. Geophys. Res.*, *104*, 1943–1956.
- Williamson, D. L. (1995), Skill scores from the AMIP simulations, in *Proceedings of the First International AMIP Scientific Conference*, *WMO TD-732*, edited by W. L. Gates, pp. 253–258, World Meteorol. Organ., Geneva.
- Xie, P., and P. Arkin (1996), Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions, *J. Clim.*, *9*, 840–858.
- Xie, P., and P. Arkin (1997), Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs, *Bull. Am. Meteorol. 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*(3), 407–446.

---

C. S. Jackson, Q. Mu, and P. L. Stoffa, Institute for Geophysics, John A. and Katherine G. Jackson School of Geosciences, University of Texas at Austin, 4412 Spicewood Spring Road, Building 600, Austin, TX 78759-8500, USA. (charles@ig.utexas.edu)