Tellus
SERIES A
DYNAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Tellus (2008), 60A, 992–1000

TELLUS

Local eigenvalue analysis of CMIP3 climate model errors

By MIKYOUNG JUN[1]*, RETO KNUTTI[2] and DOUGLAS W. NYCHKA[3], [1]*Department of Statistics, 3143 TAMU, College Station, TX, USA;* [2]*Institute for Atmospheric and Climate Science, ETH Zurich, Switzerland;* [3]*National Center for Atmospheric Research, PO Box 3000, Boulder, CO, USA*

(Manuscript received 20 December 2007; in final form 24 June 2008)

ABSTRACT

Of the two dozen or so global atmosphere–ocean general circulation models (AOGCMs), many share parameterizations, components or numerical schemes, and several are developed by the same institutions. Thus it is natural to suspect that some of the AOGCMs have correlated error patterns. Here we present a local eigenvalue analysis for the AOGCM errors based on statistically quantified correlation matrices for these errors. Our statistical method enables us to assess the significance of the result based on the simulated data under the assumption that all AOGCMs are independent. The result reveals interesting local features of the dependence structure of AOGCM errors. At least for the variable and the timescale considered here, the Coupled Model Intercomparison Project phase 3 (CMIP3) model archive cannot be treated as a collection of independent models. We use multidimensional scaling to visualize the similarity of AOGCMs and all-subsets regression to provide subsets of AOGCMs that are the best approximation to the variation among the full set of models.

1. Introduction

Projections of future climate for different scenarios of anthropogenic fossil fuel emissions are based on coupled climate models representing the ocean, atmosphere, land and sea ice. Results from individual models are sometimes averaged into a multimodel mean (Meehl et al., 2007b), with the implicit assumption that some of the errors in individual model simulations will cancel if the models are independent and distributed around the true evolution of climate. Some studies use more complicated methods to assign weight to individual models (Tebaldi et al., 2005), but in most of these methods, there is an implicit assumption of independence, such that the uncertainty in the projection decreases as more models are combined (Tebaldi et al., 2005; Furrer et al., 2007). However, despite the fact that an average of models often compares more favorably to observations, it is known already from earlier model intercomparisons that models tend to have similar deficiencies (Lambert and Boer, 2001). Little analysis of the characteristics of these error patterns and how they relate across models, has been done so far.

Jun et al. (2008) present statistical methods to quantify the dependence of atmosphere–ocean general circulation models (AOGCMs) and assess the statistical significance of the dependence through simulated present day mean of surface temperature, under the assumption that all models are independent. The independence is defined in the statistical sense that if the ith AOGCM error, D_i, is decomposed into a fixed term, M_i, and a random part with mean zero, ϵ_i, that is, $D_i = M_i + \epsilon_i$, then D_i and $D_j (i \neq j)$ are defined to be independent of each other when ϵ_i and ϵ_j are independent. Their results indicate that the similarities of the model errors are obvious and statistically significant, in particular, between models that share components. How the model error in the present day climate relates to model error in projections is largely unknown at this point. But the fact that, already, the errors in the present day climatology are similar in many models, suggests that this problem needs further attention if multiple models are combined into a single prediction or projection.

In this paper, we present a new method to look at the AOGCM spatial error patterns, which extends the preliminary result in Jun et al. (2008). We perform local eigenanalysis, based on the correlation matrices for the AOGCM errors, with the goal of uncovering local features of the model dependence. This analysis is not possible using the usual empirical orthogonal function (EOF) analysis. An important feature of this method is a companion statistical model for the spatial structure of each AOGCM's error field. This model is used to determine the sampling properties of the local eigenanalysis, under the hypothesis of independence among models and serves as a check to avoid conclusions from these data that could arise by chance. Furthermore, we explore

*Corresponding author.
e-mail: mjun@stat.tamu.edu
DOI: 10.1111/j.1600-0870.2008.00356.x

and visualize similarities of AOGCMs at various subregions of interest and identify subgroups of models that can best summarize the full set of AOGCMs.

2. Data and model experiments

We consider the surface temperature from 1970 to 1999 in the latitude range from 45°S to 72°N and the full longitude range from −180° to 180°, to define model error. The 'data product' that we use are monthly averages, aggregated on a regular spatial grid by the Hadley Centre, UK MetOffice and Climate Research Unit (CRU), East Anglia, UK (HADCRU) (Jones et al., 1999; Rayner et al., 2006) and is a composite of land and ocean data sets.

The climate model experiments consist of more than 20 separate models that were run as part of the coordinated modelling effort, in support of the IPCC Fourth Assessment Report (Meehl et al., 2007a). A list of the models used here, as well as their resolution, is given in Table 1. The model output is archived in a common format and can be downloaded from the Program for Climate Model Diagnosis and Intercomparison website (PCMDI, `http://www-pcmdi.llnl.gov/`). More details on the model output used in this work is described in Jun et al. (2008). We originally have 20 models, as listed in Table 1, but model 1 was excluded from the analysis. As explained in Jun et al. (2008), this model has problems with model setup and data post-processing. Model 10 has a relatively large bias compared with other models, but we included this model in the analysis.

The AOGCM errors are defined for the climatological mean state. We focus on Boreal winter (DJF) and summer (JJA) mean surface temperature, averaged over 30 yr (1970–1999). Accordingly, the sample error for each AOGCM at each grid cell is the difference between the model 30-year mean and the observed 30-year mean, based on the HADCRU data product. For this multimodel data set, the number of ensemble runs is limited, and so, one would expect uncertainty in the error estimates,

Table 1. The names of modelling groups, country, IPCC I.D. and resolutions of the 20 IPCC model outputs used in the study

Sl. no.	Group	Country	IPCC I.D.	Resolution
1	Beijing Climate Center	China	BCC-CM1	192 × 96
2	Canadian Center for Climate Modelling & Analysis	Canada	CGCM3.1	96 × 48
3	Météo-France/ Centre National de Recherches Météorologiques	France	CNRM-CM3	128 × 64
4	CSIRO Atmospheric Research	Australia	CSIRO-Mk3.0	192 × 96
5	US Dept. of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.0	144 × 90
6	US Dept. of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	USA	GFDL-CM2.1	144 × 90
7	NASA/Goddard Institute for Space Studies	USA	GISS-AOM	90 × 60
8	NASA/Goddard Institute for Space Studies	USA	GISS-EH	72 × 46
9	NASA/Goddard Institute for Space Studies	USA	GISS-ER	72 × 46
10	LASG/Institute of Atmospheric Physics	China	FGOALS-g1.0	128 × 60
11	Institute for Numerical Mathematics	Russia	INM-CM3.0	72 × 45
12	Institut Pierre Simon Laplace	France	IPSL-CM4	96 × 72
13	Center for Climate System Research, National Institute of Environmental Studies, and Frontier Research Center for Global Change	Japan	MIROC3.2 (medres)	128 × 64
14	Meteorological Institute of the University of Bonn, Meteorological Research Institute of KMA, and Model and Data group	Germany/ Korea	ECHO-G	96 × 48
15	Max Planck Institute for Meteorology	Germany	ECHAM5/MPI-OM	192 × 96
16	Meteorological Research Institute	Japan	MRI-CGCM2.3.2	128 × 64
17	National Center for Atmospheric Research	USA	CCSM3	256 × 128
18	National Center for Atmospheric Research	USA	PCM	128 × 64
19	Hadley Centre for Climate Prediction and Research/ Met Office	UK	UKMO-HadCM3	95 × 73
20	Hadley Centre for Climate Prediction and Research/ Met Office	UK	UKMO-HadGEM1	192 × 145

The resolution of the observation is 72 × 36 (5° × 5°).

simply due to the inherent variability of the model integrations. Thus, it is important to distinguish between the sample error that is estimated by each of these statistics and the theoretical model error obtained from a large ensemble where the internal variability has been eliminated by averaging.

3. Statistical methods

The statistical methodology to quantify the correlations for pairs of model errors is based on the results in Jun et al. (2008). With a single response from each model, effectively a single observation, it is not feasible to estimate the correlations among model errors for each gridpoint. The key idea in Jun et al. (2008) is to estimate the correlations between pairs of AOGCM by a local spatial weighting, using a non-negative kernel function. This spatial aggregation overcomes the problem of a single observation but will have less spatial resolution than the model grid.

3.1. Local covariances for the model errors

Here, we outline the local covariance estimator. Let $X(\mathbf{s}, t)$ denote the observations and $Y_i(\mathbf{s}, t)$ the ith model output (DJF or JJA) at spatial grid location $\mathbf{s}$ and year t ($t = 1, \cdots, 30$). We only consider the difference of observation and model data or the model error $D_i(\mathbf{s}, t) = X(\mathbf{s}, t) - Y_i(\mathbf{s}, t)$, and let $D_i(\mathbf{s})$ be its average over time. Finally, let $\sigma_{ij}(\mathbf{s}) = \mathrm{Cov}\{\mathrm{D_i}(\mathbf{s}), D_j(\mathbf{s})\}$ be the covariance between the error statistics—the theoretical target of our estimator.

The kernel estimator for $\sigma_{ij}(\mathbf{s})$, has the form,

$$\hat{\sigma}_{ij}(\mathbf{s}) = \sum_{k=1}^{N} K\left(\frac{|\mathbf{s}, \mathbf{s}_k|}{h}\right) \tilde{D}_{ij}(\mathbf{s}_k) \Bigg/ \sum_{k=1}^{N} K\left(\frac{|\mathbf{s}, \mathbf{s}_k|}{h}\right), \qquad (1)$$

for non-negative kernel function K and bandwidth h. Here, $\tilde{D}_{ij}(\mathbf{s})$ denotes $\tilde{D}_i(\mathbf{s}) \cdot \tilde{D}_j(\mathbf{s})$ and $\tilde{D}_i(\mathbf{s})$ is the filtered $D_i(\mathbf{s})$. We filter $D_i(\mathbf{s})$ by subtracting the estimated linear regression term mentioned in Section 3.3. For two spatial locations $\mathbf{s}_1$ and $\mathbf{s}_2$, $|\mathbf{s}_1, \mathbf{s}_2|$ denotes the great circle distance between the two locations. In our case, we assume a Gaussian kernel function, $K(u) = \exp(-u^2)$, and so, the spatial weighting of the cross-products decreases with distance, the grid cells beyond a distance of $2h$ getting less than 5% of the total weight. The denominator in (1) is simply to normalize the kernel weights to sum to 1.

Now, let $\hat{\Sigma}(\mathbf{s})$ be the estimated covariance matrix for all pairs of models, using the kernel estimator for each entry given in (1). It can be shown that $\hat{\Sigma}(\mathbf{s})$ is a positive definite matrix, with the interpretation that it is a local estimate of the covariance, at location $\mathbf{s}$ for the model errors. This covariance can be further analysed using an eigenfunction decomposition, to quantify the amount of variability and dependence among the models. In particular, we focus on the fraction of variance, which is explained by the leading eigenfunctions of this matrix. This statistic gives an idea of the effective degrees of freedom, which are explained by the models. Note that if the bandwidth is made very large, this estimate will be the same for all locations and will reproduce an EOF analysis, applied to the different models and grid locations. A very small bandwidth will result in a covariance matrix estimate that has rank one and is just the outer product of the model errors at location $\mathbf{s}$.

3.2. Comparison with empirical orthogonal functions

Given the role of the bandwidth, it is clear that the local eigenanalysis provides different information from the usual EOF analysis. An EOF analysis would quantify the overall dependence of the AOGCM error fields by expanding the spatial error fields in terms of orthogonal spatial fields (i.e. the EOFs) and the companion singular values are related to the variance explained by each orthogonal component. The orthogonal components are global in extent, and it is often difficult to discern how isolated spatial features are explained by a global set of EOFs. In contrast to EOF analysis, the local covariance method can detect heterogenous variability among model errors at different grid locations. One of the characteristics of this procedure is that for the locations that are far apart, the leading eigenvectors can be different. Therefore, even if the same amount of variability is explained, they can be based on different linear combinations of the models.

3.3. Reference distribution under independent models

Even if the model errors are independent, the spatial averaging from the kernel and the selection of the largest eigenvalues will suggest some dependence among the models. In fact, Section 4 shows that even under the independence assumption across models, the effective degrees of freedom in 19 independent models is smaller than 19 (see Fig. 1). Thus, the eigenanalysis of $\hat{\Sigma}$ should be interpreted with care, and we approach this problem by a Monte Carlo simulation of a 'reference distribution' to guide the statistical interpretation of the data analysis. Jun et al. (2008) develop statistical models that describe the spatial dependence and structure of the error for each model. Briefly, these spatial models include a linear regression term that adjusts the model for systematic longitude, altitude and land/ocean effects and a correlated, non-stationary random component that accounts for additional spatial dependence. Under the the assumption that the model errors are independent, one can simulate synthetic model error fields and quantify the distribution of the percentage variation explained by the eigenvalues. This reference distribution, based on the hypothesis of independent model errors, allows one to determine whether the actual error fields have some evidence for dependence. Throughout the paper, the reference distribution is generated from 1000 synthetic model errors (1000 independent simulations, based on the spatial model described above).

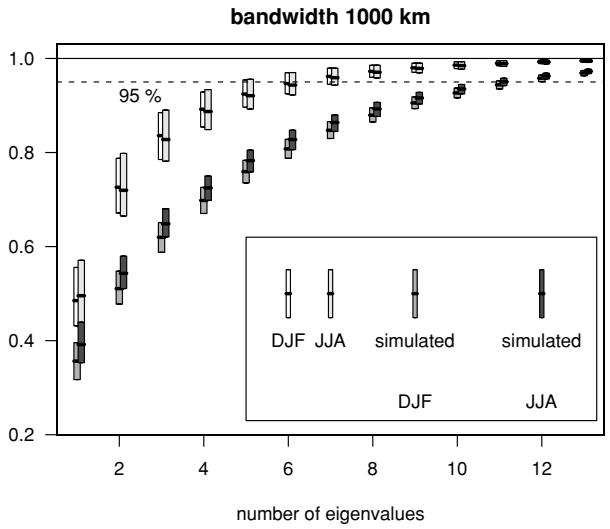

Fig. 1. The quartiles and medians of the amount of variation (across the entire domain), explained with respect to the number of eigenvalues from local eigenanalysis. The bandwidth used is 1000 km and both of the results from the data and the reference distribution are displayed for both seasons.

3.4. Bandwidth selection

The method to estimate correlation through kernel smoothing is sensitive to the bandwidth. Depending on the bandwidth used, the results may vary quite substantially. Although there are more statistically rigorous methods to choose an appropriate bandwidth, we choose to use 1000 km in this study due to the typical spatial variations in a climatological mean temperature. However, as a safeguard, we also investigate the sensitivity of our results to this choice. With the bandwidth 100 km, the largest eigenvalue explains more than 95% of the variation for both the data and the reference distribution, whereas with the bandwidth 5000 km, we require at least 11 or more largest eigenvalues to explain the same amount of the variation in the data. Again, the EOF analysis is an extreme case of the local analysis, with the bandwidth being extremely large.

4. Main result

In this section, we demonstrate the results of local eigenanalysis and compare those with the results from EOF analysis. For both analyses, the dependence of model errors are tested through the reference distribution as described in Section 3.3.

4.1. Dependence of model errors and its local feature

In the local eigenanalysis, one obtains a 19 × 19 correlation matrix at each pixel as explained in Section 3.1. The number of 'dominant' eigenvalues of this matrix should give an idea on how many 'independent' AOGCMs we really have. Although we have 19 eigenvalues in total, we suspect, only a few of them should have relatively large magnitude, since many AOGCMs have highly correlated errors, as demonstrated in Jun et al. (2008). We compare the number of eigenvalues that explain a certain amount of variation in the actual model errors from the data with a reference distribution under the hypothesis of independence. Figure 1 gives the numbers of eigenvalues needed to explain 95% of the variation, for both seasons, in the data and in the reference distribution. We obtain these numbers for each gridpoint; so, the figure shows the median and quartiles of these numbers across the entire spatial domain (summarized as boxplots). A bandwidth of 1000 km is used for these local estimates. Approximately six eigenvalues for the climate model output are required to explain 95% of variation in the data, but almost 12 eigenvalues are required for the reference distribution. This does not imply that there are only six 'independent' AOGCMs that explain almost all the variation in the data, rather that there are six linear combinations of 19 AOGCMs that explain 95% of the variation in the data. Note that the form of these linear combinations vary pixel by pixel, and so, the models with large loadings in one location may be different from that at a location that is widely separated. Figure 2 gives the analogous result from the EOF analysis. For the EOF analysis, we do not get results in each pixel; so, the result from the data is summarized as points. For the reference distribution, however, the result is summarized as boxplots, since we have 1000 Monte Carlo cases. The overall results are similar for both seasons. It is interesting to note that at most 85% of variation is explained by the first six eigenvalues. As in local eigenanalysis, the effective degrees of freedom in the data is much smaller than that in the reference distribution,

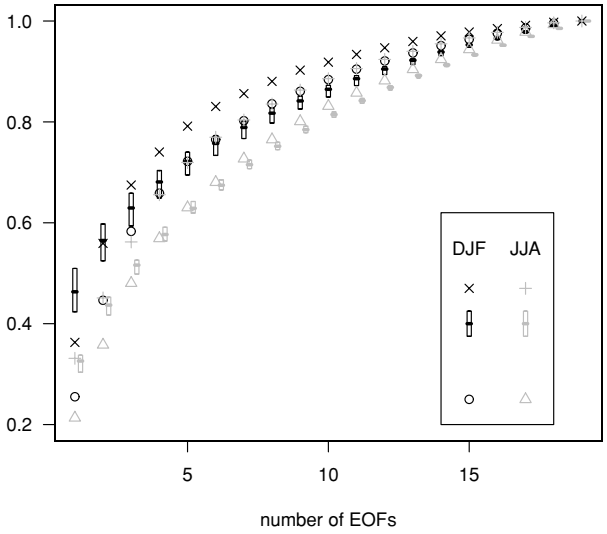

Fig. 2. Similar to Fig. 1, but the result of the EOF analysis is displayed instead of local eigenanalysis. The symbols × and + are from the data, and the boxplots are from the reference distribution. The symbols ○ and △ display R^2 for the best regressors given in Table 2, for comparison.

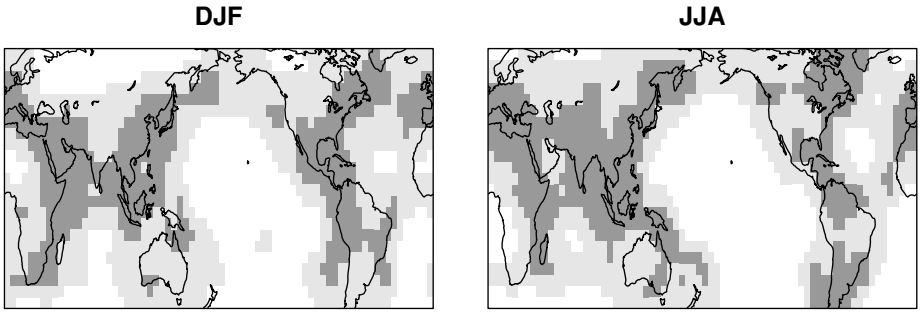

Fig. 3. The amount of variation explained by the five largest eigenvalues for the data. The pixels with white are for values above 0.95, those with light grey are for values between 0.9 and 0.95 and those with dark grey are for values below 0.9.

although for DJF, the first two EOFs explain a smaller amount of the variation in the data compared with the reference distribution. This may be due to the fact that model 10 is very different from the rest of the models, and the effective degrees of freedom in the data therefore should be at least 2.

Figure 3 displays the percentage of variation explained by the first five eigenvalues across the entire spatial domain. One of the interesting findings here is that the variation explained by the five largest eigenvalues in the data is much larger over the ocean than over the land. The fact that the spatial scale of variations over land is generally smaller than over ocean, may contribute to the signal. In addition, there are more physical processes that affect climate over the land than over the ocean, for example, topography, soil properties, plant types, surface roughness. Therefore, climate models tend to be less similar over land than over ocean. Another interesting point is that the amount of variation explained is somewhat smaller along the coast than over the land. This may be partly due to the fact that we have an indicator for the land and the ocean in the regression model described in Section 3.3, but the results with or without this term do not change much. Another possibility is that since the data are on grids, there are grid cells after the interpolation that contain partly land and partly ocean. The land–ocean mask also differs slightly between models. Interpolation from different grids is therefore a problem, even at finer resolution. The fact that this effect is particularly strong in southeast Asia, the Caribbean, the Mediterranean and the Canadian Arctic, all areas where the land–ocean mask is very complex, is consistent with that explanation. As expected, the reference distribution did not show any notable patterns across the entire domain. Moreover, the amount of variation explained is significantly less than that of the data.

To provide a more formal statistical test on the dependence of the model errors, we generate the distributions (or histograms) of (1) number of grid cells where more than 85% of the variation is explained and (2) the maximum amount of variation explained across the entire spatial domain, based on the reference distribution (1000 synthetic model errors). Figure 4 gives these histograms for both (1) and (2). The actual number of pixels with more than 85% variation explained in the data, are

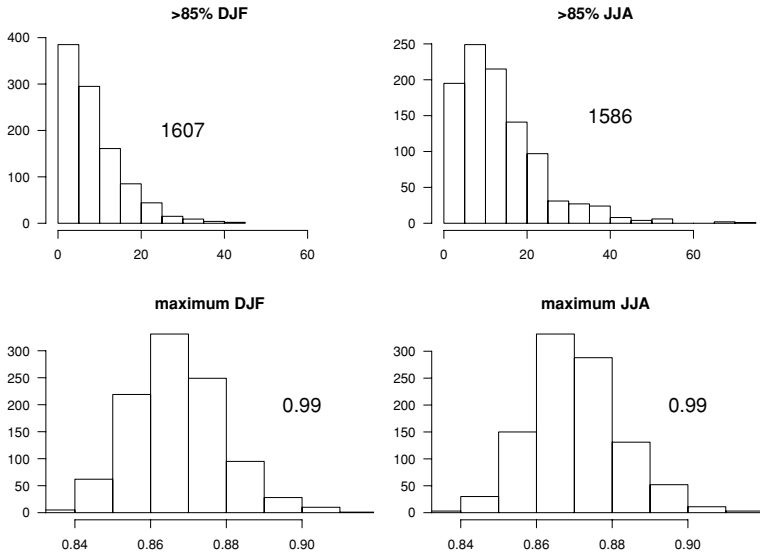

Fig. 4. Top: histograms of the number of pixels with 85% or more variation explained by the five largest eigenvalues from the reference distribution. The numbers in the figures are the corresponding values from the data. Bottom: similar to the top figures except that the results are for the maximum amount of variation explained across the entire spatial domain.

1607 and 1586 for DJF and JJA, respectively. These numbers are far above the range of the histograms. Moreover, the maximum amount of variation in the data for both seasons is 0.99, which is again far greater than the maximum value in the range of the histograms for both seasons. These two comparisons give us strong confidence that the model errors are, indeed, highly correlated throughout the entire spatial domain.

4.2. Selection of representative subgroups of models

So far, we have demonstrated, in several ways, that the effective degrees of freedom in the entire model set are much smaller than the actual number of models. We have given thorough statistical testing to support the claim. We now raise the following question: how many models are needed to explain most of the variation in the data? We apply two statistical techniques to answer this question, classical multidimensional scaling (Young and Householder, 1938; Torgerson, 1952; Cox and Cox, 2001) and all-subsets regression.

4.2.1. Classical multidimensional scaling (CMS). The CMS technique is a graphical technique to visualize relative distances among data points. Based on a distance metric, the technique assigns each data point to a location in a low-dimensional space, commonly either two- or three-dimensional space. The 'similar' data points should be located close to each other, and the degree of dissimilarity is displayed as distances between the points. For the main algorithm and more details of CMS, see Cox and Cox (2001). We apply the CMS technique to our AOGCM errors, using the correlation matrices as the similarity measure. Although CMS is a simple visualization technique, it conveys some interesting findings, regarding the correlations among AOGCM errors, that are consistent with our results in the previous sections. The CMS technique is applied for both EOF analysis and local eigenanalysis cases.

Figure 5 gives the result of the CMS technique applied to the EOF results, for each season. For DJF, except for models 7, 8, 9 and 10, most of the models are clustered together, and we cannot find any notable patterns regarding the dependence of model errors. The results for JJA also do not reveal any notable

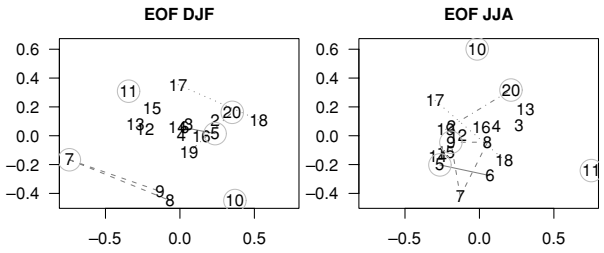

Fig. 5. The result of the Classical Multidimensional Scaling from the EOF analysis. The models connected by lines are those developed by the same institutions. The models with a circle are among the best five regressors given in Table 2.

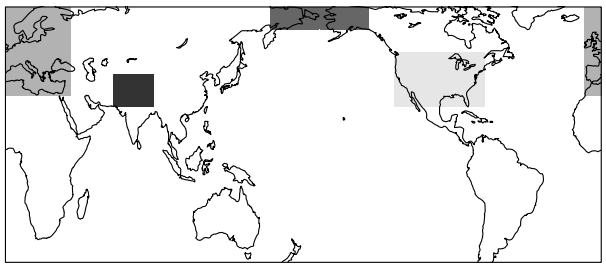

Fig. 6. The four subregions (US, Europe, Himalayas region, sea ice area) used for the Classical Multidimensional Scaling from the local eigenanalysis (see Fig. 7).

patterns. For the local eigenanalysis, we pick four subregions as in Fig. 6: US, Europe, sea ice region over the Pacific and Himalayas area. Unlike the previous EOF analysis case, Fig. 7 shows many interesting patterns. Throughout all four regions, the GFDL models (numbers 5 and 6) stay relatively close to each other. The GISS models (numbers 7, 8 and 9) are separated from the rest of the models in quite a few regions/seasons, and they stay fairly close to each other. In terms of the similarity between models developed by the same groups, we find many interesting patterns. In particular, these patterns are clear in the Himalayas regions and the sea ice area, where most of the models have poor performance. In the sea ice area, GFDL, GISS, NCAR or UKMO models stay very close to their own groups and suggest that the model errors between models developed by the same groups are highly correlated over this region. We also tried a similar analysis over the tropical Pacific (Nino3.4 region), but there is no significant pattern among most of the models. This may be because we do not take time information into our analysis (we take the climatological mean state), that is, the amplitude and frequency spectrum of ENSO are not considered. Finally, it is interesting to note that the result of the CMS technique is not the same for each season, in both EOF analysis and local eigenanalysis.

4.2.2. All-subsets regression. Although the results in Section 4.2.1 give some ideas of which models are 'close', either over the entire region on average or in some local regions, it is not clear how to pick out specific subsets of models that explain most of the variation in the data. To study this in more details, we use the all-subsets regression technique. The main idea is to divide all 19 models into two subgroups and regress one subgroup on the other. If the model subgroup is of size n ($1 \leq n \leq 18$), then there are $\binom{19}{n}$ different possible choices of subgroups. When we regress $19 - n$ model errors onto n model errors, we choose the best n models that have the minimum error sum of squares (SSE). It can be shown that the principal component analysis (or the EOF analysis) has an equivalent optimization criteria as the multiple linear regression approach and, in particular, the ratio between the sum of the leading eigenvalues and the sum of the all eigenvalues are equivalent to the ratio between the regression sum of squares (SSR) and the total sum of squares

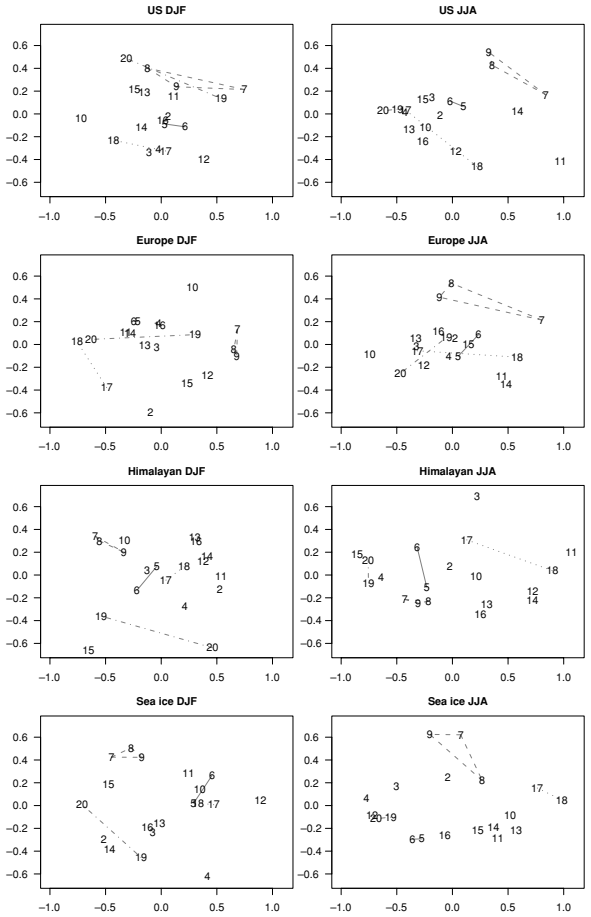

Fig. 7. Similar to Fig. 5 except that this result is from the local eigenanalysis over the subregions given in Fig. 6.

(SST), that is, R^2 (see Jong and Kotz, 1999 for details). Here, the SSR is the same as SST–SSE and the SST is the sum of squared deviations of the 19 model errors from their own means. Therefore, minimizing the SSE is the same as maximizing R^2 in this analysis.

We report the top three selected best regressors sets with corresponding R^2 in Table 2. For DJF, the best 1 regressor is model 10, which is surprising since it actually has the largest model error among all 19 models. This may be due to the fact that with the largest model error, model 10 is fairly independent to all other models and the SSE would be large if model 10 is not selected as a regressor. For a relatively small number of regressors, not all models from the same institutions are selected. This is what we expect since given number of models, the method should choose the models that can span most of the range of the model error space. The choices of models are not the same for both seasons. The R^2 values are compared with the amount of variation explained by the EOFs in Fig. 2. We expect these R^2 values to be smaller than the amount of variation explained by the EOFs for a given number of regressors (or number of EOFs), since the space spanned with a fixed number of regressors in multiple regression is more restricted than the space that can be spanned by the same number of EOFs. For both EOF analysis and the multiple regression, more variation is explained by a given number of EOFs (or regressors) for DJF than JJA. This can be explained by the following argument. Over land, models tend to have smaller biases and tend to be more similar in the winter season than in the summer season because the winter climate is more dominated by the large-scale atmospheric circulation, whereas more small-scale processes (e.g. soil moisture, land surface processes, thunderstorms) are important in the summer season. Because the land fraction is larger in the Northern Hemisphere, models tend to be more similar overall in DJF than in JJA, and the variance explained by a small number of models is therefore larger in DJF.

4.3. Statistical permutation test

A different perspective on this problem is to consider whether the observed climate and the AOGCM results belong to the same statistical population, and thus, whether the observations and the AOGCM outputs are exchangeable. We can test this assumption

Table 2. The result of all-subsets regression described in Section 4.2.2 for both seasons

	DJF		JJA	
Size of subgroup	Subgroup members	R^2	Subgroup members	R^2
1	10	0.255	5	0.213
	5	0.24	9	0.201
	6	0.232	19	0.199
5	5 7 10 11 20	0.722	5 9 10 11 20	0.63
	7 10 12 18 19	0.714	5 8 10 11 20	0.626
	7 10 12 18 20	0.71	9 10 11 14 19	0.622
10	6 7 8 10 11 12 14 18 19 20	0.884	6 7 9 10 11 12 13 14 18 19	0.831
	2 6 7 8 10 11 12 18 19 20	0.883	4 7 9 10 11 12 13 14 18 19	0.831
	2 6 7 8 10 11 12 14 18 20	0.882	6 7 9 10 11 12 13 18 19 20	0.83

Table 3. The number of eigenvalues to explain 95% of the variation

Omitted case	Model output and observations	Reference distribution
10	4	12
1, 7	6	12
12	7	14
8, 9	8	12
11, 18	8	14
20	9	15
19	9	16
14, 15, 17	10	14
2, 3, 4, 6, 16	10	15
5, 13	11	14

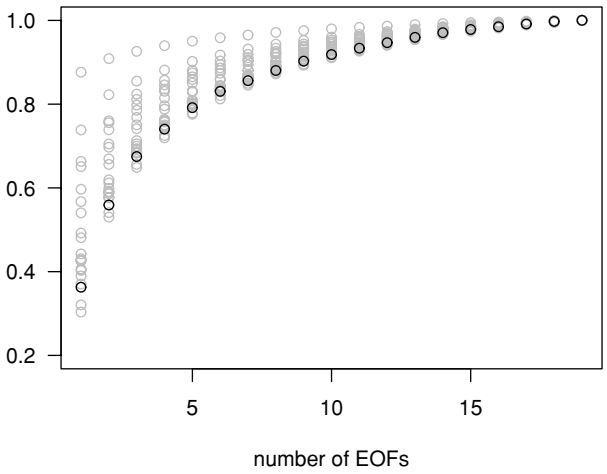

Fig. 8. EOF analysis for cases 1 to 20 in Section 4.3. Black circle is for case 1 (real observations treated as reality) and the grey circles are for cases 2 to 20 (one of the models treated as reality).

using a permutation test, that is, by withholding a model output as the 'reality', we perform the local eigenanalysis and the EOF analysis on the differences between the withheld 'reality' and the rest of the model outputs and the observations. Let case 1 refer to the situation that we withhold the observations as the reality and case n ($n = 2, \cdots, 20$) refers to the case that we withhold model n as the reality. Table 3 gives the number of eigenvalues from the local eigenvalue analysis that explain 95% of the variation. Notice that except for cases 7 and 10, the number of eigenvalues required to explain 95% of the variation for cases 2 to 10 are bigger than that of case 1. Note that model 10 (for case 10) is the one that has a very large bias compared with others, and thus, it makes sense that when model 10 is considered to be the truth, all the rest of the models turn out to be highly correlated. The result in Table 3 suggests that the assumption that the model results belong to the same population as the observation is violated. Figure 8 is a similar figure to Fig. 2, without the reference distribution and R^2. The black circles are from case 1 and the grey circles are from cases 2 to 10. Figure 8 shows that the model-to-model differences tend to have fewer degrees of freedom than the model to observation differences. Thus, we argue that the bias of a model relative to the remaining models has a different character than the bias of a model relative to the observations. This difference is not surprising and suggests that intercomparison of the models can give limited information about the bias with respect to the observations.

5. Discussion

We present the results of a local eigenvalue analysis using the correlation values of pairs of AOGCM errors at each gridpoint. These results are compared with those from the usual EOF analysis, and we demonstrate that we find interesting local features regarding the dependence of model errors, which cannot be obtained from the usual EOF analysis. The result suggests that the data set has fewer degrees of freedom than the number of AOGCMs would suggest, overall. One of the local features discovered is that there are more degrees of freedom over the land than over the ocean. There is also a 'coastal effect', meaning that the degrees of freedom are even bigger along the coast. Moreover, models developed by the same groups tend to have highly correlated model errors in areas such as the Himalayas or the sea ice region. We also provide subsets of models that best explain the full set and have identified subsets of 10 models, which explain more than 80% of the variation in the bias.

The analysis performed in this paper can easily be applied to variables other than surface temperature. Moreover, instead of looking at the climatological mean, we could explore model errors in the trend. However, to assess the statistical significance of the dependence of model errors of the trend, we would need to build a spatial-temporal model that can simulate the synthetic model errors in the trend.

6. Acknowledgments

This research has been supported in part by the National Science Foundation DMS-0355474 and ATM-0620624. The authors acknowledge the modelling groups for making their simulations available for analysis, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the CMIP3 model output and the World Climate Research Programme (WCRP)'s Working Group on Coupled Modelling (WGCM) for organizing the model data analysis activity. The WCRP CMIP3 multimodel data set is supported by the Office of Science, U.S. Department of Energy. Finally, the authors thank two reviewers whose comments substantially improved the manuscript and for the suggestion of including a permutation test.

References

Cox, T. F. and Cox, M. A. A. 2001. *Multidimensional Scaling* 2nd Edition. Chapman & Hall/CRC Boca Raton, London, New York, Washington, D.C.

Furrer, R., Sain, S. R., Nychka, D. W. and Meehl, G. A. 2007. Spatial patterns of probabilistic temperature change projections from a multivariate Bayesian analysis. *Geophys. Res. Lett.* **34**, (doi:10.1029/2006GL027754).

Jones, P. D., New, M., Parker, D. E., Martin, S. and Rigor, I. G. 1999. Surface air temperature and its variations over the last 150 years. *Rev. Geophys.* **37**, 173–199.

Jong, J.-C. and Kotz, S. 1999. On a relation between principal components and regression analysis. *Am. Stat.* **53**, 349–351.

Jun, M., Knutti, R. and Nychka, D. W. 2008. Spatial analysis to quantify numerical model bias and dependence: how many climate models are there? *J. Am. Stat. Assoc.* In press.

Lambert, S. J. and Boer, G. J. 2001. CMIP1 evaluation and intercomparison of coupled climate models. *Clim. Dyn.* **17**, 83–106.

Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B. and co-authors. 2007a. The WCRP CMIP3 multimodel dataset: a new era in climate change research. *Bull. Ame. Meteorol. Soc.* **88**, 1383–1394.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T. and co-authors. 2007b. Global climate projections. In: *Climate Change 2007: The Physical Science Basis*. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, (eds. S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, and co-editors). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Rayner, N. A., Brohan, P., Parker, D. E., Folland, C. K., Kennedy, J. J., and co-authors. 2006. Improved analyses of changes and uncertainties in marine temperature measured in situ since the mid-nineteenth century: the HadSST2 dataset. *J. Clim.* **19**, 446–469.

Tebaldi, C., Smith, R. L., Nychka, D. W. and Mearns, L. O. 2005. Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. *J. Clim.* **18**, 1524–1540.

Torgerson, W. S. 1952. Multidimensional scaling, 1: theory and methods. *Psychometrika* **17**, 401–419.

Young, G. and Householder, A. S. 1938. Discussion of a set of points in terms of their mutual distances. *Psychometrika* **3**, 19–22.