Spatial Analysis to Quantify Numerical Model Bias and Dependence: How Many Climate Models Are There?

Mikyoung JUN, Reto KNUTTI, and Doug NYCHKA

A limited number of complex numerical models that simulate the Earth's atmosphere, ocean, and land processes are the primary tool to study how climate may change over the next century due to anthropogenic emissions of greenhouse gases. A standard assumption is that these climate models are random samples from a distribution of possible models centered around the true climate. This implies that agreement with observations and the predictive skill of climate models will improve as more models are added to an average of the models. In this article we present a statistical methodology to quantify whether climate models are indeed unbiased and whether and where model biases are correlated across models. We consider the simulated mean state and the simulated trend over the period 1970–1999 for Northern Hemisphere summer and winter temperature. The key to the statistical analysis is a spatial model for the bias of each climate model and the use of kernel smoothing to estimate the correlations of biases across different climate models. The spatial model is particularly important to determine statistical significance of the estimated correlations under the hypothesis of independent climate models. Our results suggest that most of the climate model bias patterns are indeed correlated. In particular, climate models developed by the same institution have highly correlated biases. Also, somewhat surprisingly, we find evidence that the model skills for simulating the mean climate and simulating the warming trends are not strongly related.

KEY WORDS: Cross-covariance model; Intergovernmental Panel for Climate Change; Kernel smoother; Numerical model evaluation.

1. INTRODUCTION

Recent changes in the Earth's climate [Intergovernmental Panel on Climate Change (IPCC) 2001], related to increasing anthropogenic emissions of greenhouse gases, have raised questions about the risk of future changes in the climate system. The most detailed knowledge of potential future climate change comes from coupled atmosphere ocean general circulation models (AOGCMs). An AOGCM is a large, deterministic numerical model that simulates the Earth's climate system. Besides the ocean and atmosphere, AOGCMs often include a sea ice and land surface component. AOGCMs can be used to understand the observed changes in the climate over the industrial period (Meehl et al. 2004) and to quantify the human contribution to the observed changes (Barnett et al. 2005). When these models are run under future emission scenarios from socioeconomic models for greenhouse gases, aerosols, and other radiatively active species (Nakićenović et al. 2000), they can estimate future changes in the climate system on time scales of decades to many centuries. These simulations, termed climate *projections*, form the basis for a quantitative description of how human activities will influence the future climate for a give scenario. This work describes a spatial statistical analysis on a prominent suite of AOGCM simulations.

1.1 Climate Model Uncertainty

All climate projections are necessarily uncertain (Knutti, Stocker, Joos, and Plattner 2002). The largest contribution to this uncertainty arises due to the limited understanding of all of the interactions and feedbacks in the climate system. Because of computational constraints, many geophysical processes must be simplified, and their effect is *parameterized* in terms of the large-scale variables available in the model. For example, cloud condensation processes occur on spatial scales of micrometers, yet the typical grid size of global climate models is on the order of a 100 km or more. Therefore, the effect of cloud processes within a grid cell must be represented in terms of the average temperature, humidity, vertical stability, and other parameters within that grid cell. The parameters used in these representations are often uncertain, being derived empirically from limited observations, or being tuned to give a realistic simulation of observable quantities. Structural uncertainty in climate models is introduced through the choice of processes that are explicitly represented, the specific form of parameterizations (e.g., whether cloud cover depends linearly or nonlinearly on some other quantity), but also through the choice of the grid, and the numerical schemes. Initial conditions are also not well known, particularly, in the deep ocean or the biosphere, although initial condition uncertainty is a rather small contribution to the total uncertainty on long time scales, because only the statistical description of climate (i.e., a climate mean state and its variability averaged over a decade or more) are assessed, not individual weather events.

The ability of an AOGCM to reproduce twentieth century climate, for which there are observations, is a measure of the skill of the model and provides some indication of its reliability for future projections. Climate models are evaluated on how well they simulate the current mean climate state, how they can reproduce the observed climate change over the last century, how well they simulate specific processes, and how well they agree with proxy data for very different time periods in the past (e.g., the last glacial period). Although it seems a necessary condition for a model to simulate a reasonable current mean state, it might not be sufficient to guarantee a realistic simulation of future changes. Different models might agree well for the present, yet disagree for the future climate (see, e.g., Stainforth et al. 2005), which is one aspect discussed in this article.

Currently, there are about 20 AOGCMs constructed by institutions and modeling groups throughout the world with the

Mikyoung Jun is Assistant Professor, Department of Statistics, Texas A&M University, College Station, TX 77843 (E-mail: *mjun@stat.tamu.edu*). Reto Knutti is Assistant Professor, Institute for Atmospheric and Climate Science, Swiss Federal Institute of Technology, Zurich, Switzerland (E-mail: *reto.knutti@env.ethz.ch*). Doug Nychka is Senior Scientist, Institute for Mathematics Applied to Geosciences at National Center for Atmospheric Research, Boulder, CO 80307-3000 (E-mail: *nychka@ucar.edu*). This research was supported in part by National Science Foundation grant DMS-0355474. Mikyoung Jun also acknowledges support by National Science Foundation grant ATM-0620624. The authors acknowledge the comments by the associate editor and two anonymous referees that greatly helped improve the article.

Journal of the American Statistical Association
???? 0, Vol. 0, No. 00, Applications and Case Studies
DOI 10.1198/00

complexity to produce credible simulations of current climate and future projections. Based on the uncertainty in the modeling process described earlier and different choices of parameterizations, model components, and initial conditions, these models would be expected to have different responses. The motivation to use several models for prediction is based on the experience from many applications that the combined information of many models (in many cases simply an average of several models) performs better than a single model. Examples where this has been confirmed are seasonal weather prediction (Yun, Stefanova, and Krishnamurti 2003), detection and attribution (Gillett et al. 2002), health (Thomson et al. 2006), and agriculture (Cantelaube and Terres 2005). The average of several models has been shown to agree better with observations for the current mean climate state (Lambert and Boer 2001), indicating that in some cases at least, errors in individual climate models tend to cancel when many models are averaged.

1.2 Statistical Analysis of Climate Models

There has been some recent work on combining several climate model outputs (and/or ensemble simulations) with observations into probabilistic future climate projections (e.g., Tebaldi, Smith, Nychka, and Mearns 2005; Furrer, Sain, Nychka, and Meehl 2007; Smith, Tebaldi, Nychka, and Mearns 2006). Tebaldi and Knutti (2007) have provided a more thorough review, and Smith et al. (2006) have provided more references. Most of those studies either explicitly or implicitly assume that each climate model is independent from the others and is a random sample from a distribution with the true climate as its mean. This implies that the average of a set of models converges to the true climate as more and more models are added. Although some of the biases do indeed cancel by averaging, many problems are persistent across most of the models for the following reasons. First, many models are based on the same theoretical or sometimes empirical assumptions and parameterizations. Second, all models have similar resolution, and thus cannot adequately resolve the same small-scale processes. And third, for practical reasons, the development of individual models is not independent. Models are constantly compared, and successful concepts are copied. In some cases, whole model components are transferred to reduce the effort in model development. In the most recent coordinated modeling effort, this is particularly pronounced, because several institutions have submitted more than one model or model version. In some cases only the model resolution is different, or only one component is different (e.g., the atmosphere is different, but the ocean is the same).

1.3 Outline

The goal in this work is to apply statistical models to quantify some of the biases in AOGCMS and thus support an understanding of the uncertainties in model projections. Specifically, we consider the simulated mean temperature state and the simulated temperature trend over the 1970–1999 period for Northern Hemisphere summer (JJA) and winter (DJF). For a reader outside the climate science community, it should be noted that the ensemble of 20 climate model experiments considered in this work is currently a definitive and comprehensive archive. This ensemble provides a nearly complete representation of the state of the art in climate model science and was coordinated for the IPCC Fourth Assessment Report, an international collaboration of several hundred scientists. The model experiments contributed by the National Center for Atmospheric Research (NCAR) alone cost tens of millions of dollars in computer resources and produced more than 100 terabytes of detailed model output.

We use a nonstationary spatial process model to characterize the bias in a given model to observed surface temperature or temperature trends and quantify the similarity in biases among different AOGCMs. Deriving a statistical model for each bias field is important, because it facilitates testing whether correlations among model biases are statistically significant. The fields representing the correlations between model biases are estimated by kernel smoothing and so provide a flexible method that can capture nonstationarity in cross-covariances. We provide evidence that the current biases among a sample of current AOGCMs are not independent, and thus the ensemble has a reduced effective sample size.

In Section 2 we describe the observations of global surface temperature and corresponding outputs from 20 AOGCMs used in this study (one AOGCM is omitted from most of the subsequent analysis). Sections 3 and 4 present the main results on mean state and on trend. Spatial models for the bias of each AOGCM model, as well as estimates of cross-correlation among different model biases, are presented, and these correlations are compared with simulated correlations under the assumption that AOGCM model biases are independent. Section 5 concludes with a discussion.

2. DATA

2.1 Observational Data

We use surface temperature measurements (in °C) with global coverage for comparison with the AOGCM simulations. The actual “data product” that we incorporate into the analysis are monthly averages given on a regular spatial grid and are created by the Climate Research Unit (CRU), East Anglia, and the Hadley Centre, U.K. MetOffice (Jones, New, Parker, Martin, and Rigor 1999; Rayner et al. 2006). The surface temperature data set is a composite of land and ocean data sets. Temperature over land is measured at stations, whereas temperature over the ocean is derived from sea surface temperature and marine air temperature measurements taken by ships and buoys. Individual measurements are checked for inhomogeneities, corrected for instrument biases, and averaged within each grid cells. Thus the number of measurements used differs for each grid cell. Inhomogeneities in the data arise mainly due to changes in instruments, exposure, station location (elevation, position), ship height, observation time, urbanization effects, and the method used to calculate averages. However, these effects are all well understood and taken into account in the construction of the data set (see Jones et al. 1999 for a review). The uncertainties in the temperature data set, particularly on seasonal averages for a 30-year period, are small compared with the much larger biases in the climate models compared with observations.

We consider the 30-year interval of 1970–1999, because observations tend to be more complete and of better quality toward the end of the observational period, and because there is a strong signal for temperature increase during that period. A 30-year period is commonly used to define a climatologic mean state, which does not focus on specific weather situations. Due to the lack of observations in high latitudes, we consider only the spatial regions of the latitude range 45°S to 72°N, with the full longitude range from −180° to 180°. We then have only very few missing observations, and we naively impute these by taking averages of spatially neighboring cells (eight neighboring cells if all are available). The method of imputation has very little impact on the results of our analysis, because there are only 10 out of 1,656 grid cells with missing data.

For the analysis of the climatologic mean state, we focus on Boreal winter and Boreal summer mean surface temperature; that is, we average the monthly temperatures over December to February (DJF) and June to August (JJA) and over 30 years. For the trends, we calculate least squares linear trends at each grid point, separately for DJF and JJA.

2.2 AOGCM Output

In a recent coordinated modeling effort in support of the IPCC Fourth Assessment Report, many AOGCMs were used to simulate the climate over the twentieth century. Such experiments are driven by observed changes in radiatively active species at the top of the atmosphere and do not explicitly include any observed meteorological observations. This has the goal of simulating the anthropogenic influence on climate change observed so far. The model runs were then continued into the future following several possible emission scenarios, to quantify expected future climate change. Although the projections are not analyzed in this work, they form a cornerstone for IPCC reports on the future of Earth's climate. For a statistician, climate is an expectation or a long-term average of weather events. Each AOGCM simulation produces a time series of weather that is then averaged to estimate the climate state. If a model were run many times with just slightly different initial conditions, then the climate of the AOGCM could be estimated to high accuracy. In practice, only a few realizations of the AOGCM are run, and thus there is some uncertainty in the actual climate simulated. This sampling error, also known as internal variability, contributes an additional random component to a statistical model for the AOGCM output. However, its contribution is small for a 30-year mean. A list of the models used here as well as their resolution is given in Table 1. The "data" produced by all of the models are archived in a common format and can be downloaded from the Program for Climate Model Diagnosis and Intercomparison website (*http://www-pcmdi.llnl.gov/*). In contrast to the observations, there are no missing data in the climate model output fields. The models' resolution, complexity, and completeness in terms of the processes that they include vary substantially across the models. A first comparison of each model to observations showed that model 1 (BCC–CM1) could not be used in our

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

	Group	Country	IPCC ID	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	U.S. Department of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	U.S.	GFDL–CM2.0	144 × 90
6	U.S. Department of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory	U.S.	GFDL–CM2.1	144 × 90
7	NASA/Goddard Institute for Space Studies	U.S.	GISS–AOM	90 × 60
8	NASA/Goddard Institute for Space Studies	U.S.	GISS–EH	72 × 46
9	NASA/Goddard Institute for Space Studies	U.S.	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	U.S.	CCSM3	256 × 128
18	National Center for Atmospheric Research	U.S.	PCM	128 × 64
19	Hadley Centre for Climate Prediction and Research/ Met Office	U.K.	UKMO–HadCM3	95 × 73
20	Hadley Centre for Climate Prediction and Research/ Met Office	U.K.	UKMO–HadGEM1	192 × 145

NOTE: The resolution of the observation is 72 × 36 (5 × 5 degrees).

analysis. Apart from large biases and large discrepancy in variograms compared with observations, the lack of documentation and problems in reproducing the numerical simulation suggest that there were issues with the model output possibly caused not by biases in the physics of the model but by the model setup or the data postprocessing. Model 10 (FGOALS-g1.0) also has larger biases (particularly in the high latitudes) than the other models (excluding model 1), but those were confirmed to be actually produced by biases in the ocean model. This model 10 is included in the analysis. If the statistical methodology presented here works as expected, it must differentiate models with larger biases, like model 10, from the others.

Several models have been developed by a single organization. For example, two models were developed by NOAA GFDL (models 5 and 6), two by NCAR (17 and 18), three by NASA GISS (7, 8, and 9), and two by the U.K. MetOffice (19 and 20). Those models often share components or parameterizations of subgrid-scale processes; therefore, we suspect that these models should have similar biases compared to observations.

2.3 Registering Model Output and Observations to a Common Grid

We quantify the model biases by comparing the AOGCM data with observations. Specifically, we need the difference between observations and model output. Unfortunately, the model output and observations are on several different grids. Because the observations have the coarsest grid (see Table 1), we use bilinear interpolation of the model output to the observational grid. One reason for using bilinear interpolation is that because the grid resolutions of model output and observations are not nested, it is not obvious how we should aggregate model output to match the observational grids without serious statistical modeling. (See Banerjee, Carlin, and Gelfand 2004 for a hierarchical Bayesian modeling approach for nonnested block-level realignment.) Another reason is that, as reported by Shao, Stein, and Ching (2006), bilinear interpolation seems to work better than naive interpolation in aggregating numerical model outputs in general.

Instead of taking differences between observations and model output, we could jointly model the observations and model output. However, the difference fields tend to have a much simpler covariance structure than the observations or model output themselves (see Jun and Stein 2004 for details). Furthermore, the differences are much closer to Gaussian than the observations and model output themselves. Therefore, we develop statistical models for the differences rather than a joint model for the observations and model output.

2.4 An Example of AOGCM Results for Mean Temperature

Figure 1 shows the differences of observations and model output for DJF and JJA climatologic averages. Examples are given for two models with very similar bias patterns (models 5 and 6, especially for DJF), one model with poor agreement [large bias pattern (model 10)], and a model with reasonably good agreement [small amplitude of the bias pattern (model 17)]. Regional biases can be large for both DJF and JJA averages for many models. Although the DJF difference of model 10 shows distinct patterns compared with the others, overall many models have similar problems in regions with steep topography (e.g., Himalayas and Andes regions), regions of ocean deep water formation (e.g., the North Atlantic) or upwelling (e.g., west of South America), and high-latitude areas, where snow or ice cover influences the climate. This is not surprising, because all models cannot properly resolve steep mountains or ocean convection due to their limited resolution. No single model performs consistently better than all of the other models in all spatial regions and for both DJF and JJA.

The problems in simulating a high-altitude and high-latitude climate in most models are illustrated in Figure 2. The left column shows the difference between observations and the multimodel mean (i.e., the average of the 19 models) for each season. Note that although the magnitudes of the differences between observations and the multimodel mean are slightly less than the magnitude of the differences in the individual models (Fig. 1), the problems of the multimodel mean over the high-altitude and high-latitude areas are still present. This is a first sign that the assumption of model averages converging to the true climate is not fully justified. If all models have similar biases in the same regions, then adding more models with similar biases to an average will not eliminate those biases. The right column shows the root mean squared (RMS) errors of the 19 models (i.e., the RMS of the bias patterns of all models, averaged across all models). It shows that the regions in which model biases have large spread (high RMS error) tend to be the same as those in which the multimodel mean deviates from observations.

3. ANALYSIS ON MEAN STATE

3.1 Statistical Models for the Climate Model Biases

In this section we build explicit statistical models to quantify the model biases on the mean state. Let $X(\mathbf{s}, t)$ denote the observations and $Y_i(\mathbf{s}, t)$ denote the ith model output (DJF or JJA averages) at spatial grid location $\mathbf{s}$ and year t ($t = 1, \ldots, 30$). As mentioned earlier, we model the difference of observation and model data, or the model bias $D_i(\mathbf{s}, t) = X(\mathbf{s}, t) - Y_i(\mathbf{s}, t)$. The process D_i varies over space and time, and we decompose it as $D_i(\mathbf{s}, t) = b_i(\mathbf{s}) + u_i(\mathbf{s}, t)$. Here b_i is a purely spatial field with possibly nonstationary covariance structure and represents the bias of the ith model with respect to observed climate. The residual, u_i, has mean 0 and is assumed to be independent of b_i. This term u_i includes contributions from the measurement error and year-to-year variation of climate model outputs. We are interested mainly in modeling b_i, especially the cross-covariance structure of b_i and b_j for $i \neq j$. Most of the information for modeling b_i comes from the average of D_i over 30 years, that is, $\bar{D}_i(\mathbf{s}) = \sum_{t=1}^{30} D_i(\mathbf{s}, t)/30$, because the noise component of weather in a 30-year seasonal average is small.

One may wonder whether $D_i(\mathbf{s}, t)$ should contain a spatial field that represents the bias due to the observational errors. However, as mentioned in Section 2.1, the climate scientists have fairly strong confidence in the quality of their observational data compared with the climate model biases. Therefore, we assume that the effect of observational errors to D_i is negligible. If the observational errors do turn out to be important, then they would induce correlations among the climate model biases.

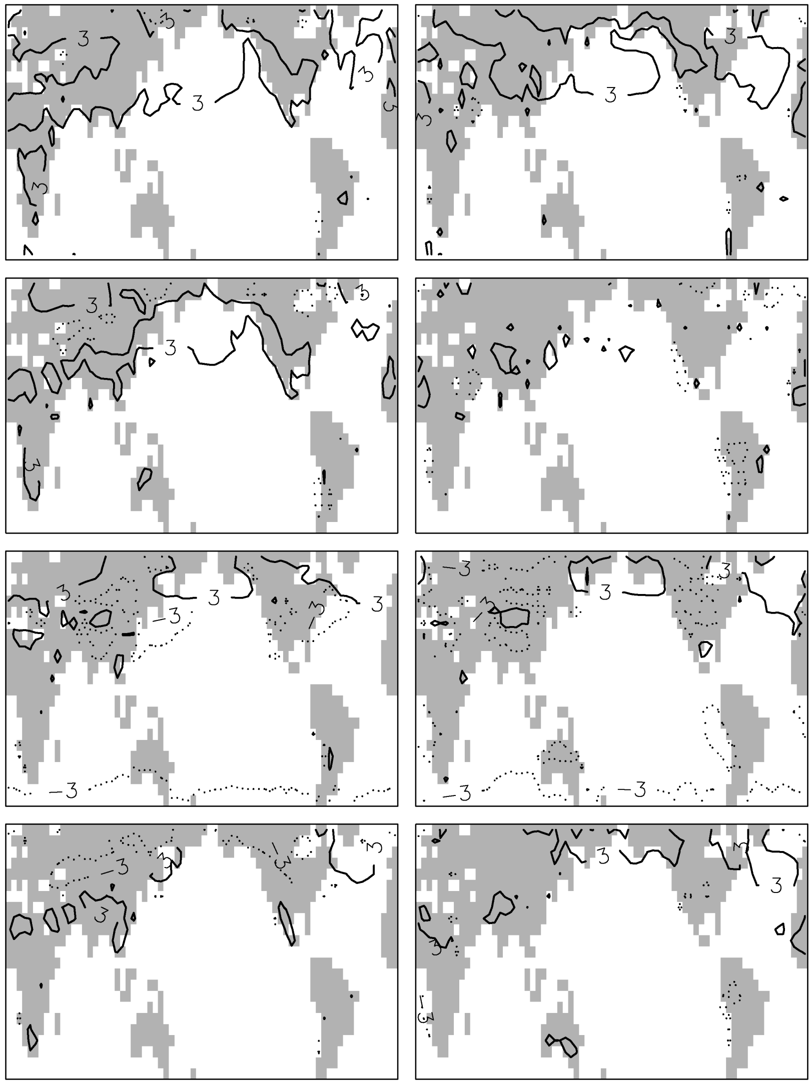

Figure 1. Differences between observation and model outputs for models 5, 6, 10, and 17 (from top to bottom). The left column is for DJF averages, and the right column is for JJA averages. Contour levels are -3 (dashed) and $+3$ (solid).

We model b_i as a Gaussian random field with a mean structure depending on certain covariates. The patterns in Figures 1 and 2 suggest that we need to include the latitude and altitude in the mean term. We also find that the ocean/land indicator has large significant effects on the differences of observation and model outputs. Thus for $i = 2, \ldots, 20$, we let

$$b_i = \mu_{0i} + \mu_{1i} L(\mathbf{s}) + \mu_{2i} \mathbf{1}_{(\mathbf{s} \in \text{land})} + \mu_{3i} A(\mathbf{s}) + a_i(\mathbf{s}), \qquad (1)$$

where L denotes the latitude and A denotes the altitude (over the ocean, $A = 0$). Here every term except for a_i is a fixed effect. The term a_i is stochastic and is modeled as a Gaussian process with mean 0.

In modeling the covariance structure of a_i, we need to have a covariance model that satisfies at least two conditions. First, the covariance model should be valid on a sphere. Second, it should have nonstationary structure; even the simplest possible model should have at least different covariance over the land and over the ocean, because the differences over the land have higher covariances than those over the ocean.

Jun and Stein (2007) gave a class of flexible space-time covariance models valid on a sphere that are nonstationary in space. We use a spatial version of this model class to model the covariance structure of a_i. Following Jun and Stein (2007), for d_i's that are constants ($i = 2, \ldots, 20$), we model a_i as

$$a_i(\mathbf{s}) = \eta_i \frac{\partial}{\partial L} Z_i(\mathbf{s}) + d_i Z_i(\mathbf{s}). \qquad (2)$$

The differential operator in the first term of (2) allows for possible nonstationarity depending on latitude. Now Z_i is a nonstationary process defined on a sphere, and with $\delta_i > 0$, we write it as

$$Z_i(\mathbf{s}) = \left(\delta_i \mathbf{1}_{(\mathbf{s} \in \text{land})} + 1\right) \widetilde{Z}_i(\mathbf{s}), \qquad (3)$$

where $\widetilde{Z}_i$ is a Gaussian process with mean 0 and covariance

$$\operatorname{cov}\{\widetilde{Z}_i(\mathbf{s}_1), \widetilde{Z}_i(\mathbf{s}_2)\} = \alpha_i \mathcal{M}_{\nu_i+1}(\beta_i{}^{-1} d). \qquad (4)$$

Here $\alpha_i, \beta_i, \nu_i > 0$; d is the chordal distance between $\mathbf{s}_1$ and $\mathbf{s}_2$; $\mathcal{M}$ is the Matérn class; $\mathcal{M}_\nu(x) = x^\nu \mathcal{K}_\nu(x)$; and $\mathcal{K}$ is a modified Bessel function (Stein 1999). This covariance model is valid on a sphere, because the Matérn class is valid on $\mathbb{R}^3$, and through the chordal distance d, we get a valid covariance model on a sphere (Yaglom 1987). Due to the differential operator in (2),

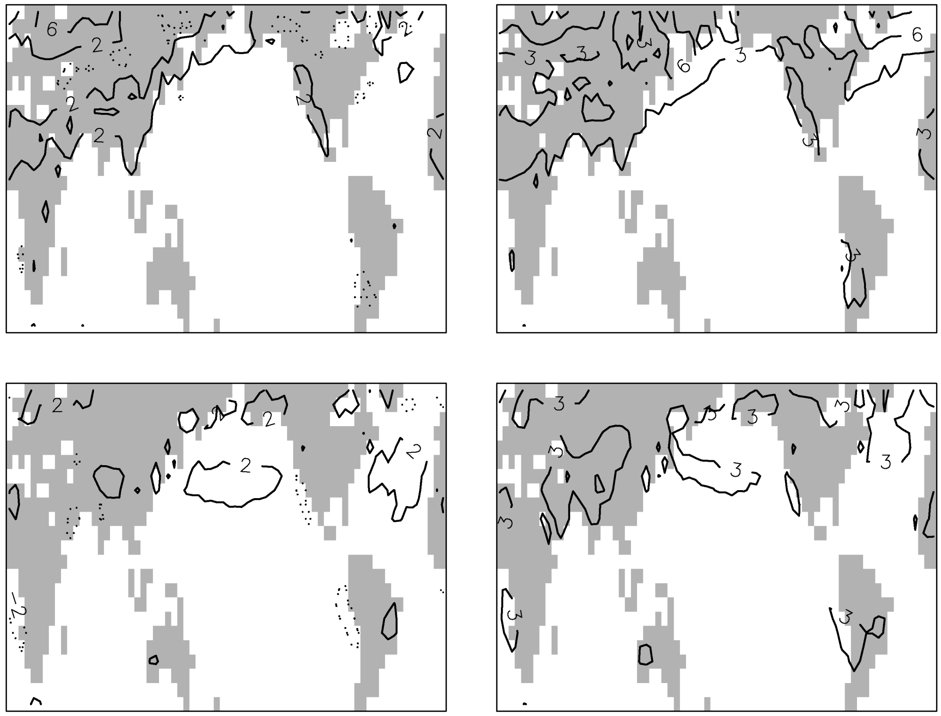

Figure 2. The difference between observation and the average of 19 models (left column) and the RMS errors (right column). The top panel is for DJF averages, and the bottom column is for JJA averages. The solid lines represent positive levels, and the dotted lines represent negative levels.

the smoothness parameter of a Matérn covariance function in (4) should be >1 ($\nu_i + 1 > 0$). In (3), the term δ_i gives higher covariance over the land than over the ocean for the process a_i and the amount of this inflation is allowed to vary across AOGCMs.

The variance can be expected to depend much more on latitude than the correlation. Thus we tried the few variations of (3), such as $Z_i(\mathbf{s}) = (\delta_i \mathbf{1}_{(\mathbf{s}\in\text{land})} + \psi_i |L(\mathbf{s})| + 1)\widetilde{Z}_i(\mathbf{s})$ and $Z_i(\mathbf{s}) = (\delta_i \mathbf{1}_{(\mathbf{s}\in\text{land})} + \psi_i L(\mathbf{s}) \mathbf{1}_{\{L(\mathbf{s})>0\}} + 1)\widetilde{Z}_i(\mathbf{s})$, for $\psi_i > 0$; however, we found no significant increase in log-likelihood values.

3.2 Results for Model Experiments

To estimate the covariance parameters, we use restricted maximum likelihood estimation (REML) and then obtain regression parameters in the mean function through generalized least squares (Stein 1999). We find that constraining $\eta_i = 0$ in (2) gives a likelihood comparable to $\eta_i > 0$ for all 19 models. This may not mean that there is no nonstationarity depending on latitude, but it may suggest that we need to allow η_i to vary over latitude or some other appropriate variables. For parsimony, we set $\eta_i = 0$ for $i = 2, \ldots, 20$. We also set $d_i = 1$ to avoid the identifiability problem in the marginal variance.

The fitted values of the parameters in (1) for DJF and JJA averages are given in Tables 2 and 3. The unit of spatial distance is km. Based on these estimates, we compare the estimated fixed part and the random part in (1). Figures 3 and 4 compare $\bar{D}_i$ (for model i, $i = 2, \ldots, 20$), the fixed part of the difference, and the random part as in (1) for each season. For most of the models, the random part is close to mean 0 relative to $\bar{D}_i$ and the magnitudes of the fixed part and random part are similar. The covariance parameters are β_i, the spatial range; ν_i, the smoothness parameter for the Matérn class; and δ_i, the inflation parameter for covariances over land. The parameter α_i is related to the sill in the sense that the variance of the process $\widetilde{Z}_i$ in (3) over the ocean is $\alpha_i\, 2^{\nu_i} \Gamma(\nu_i + 1)$. Note that the ν_i values in Tables 2 and 3 are not the same as ν_i in (4). Because $\eta_i = 0$, we do not have to have the smoothness in (4) >1, as explained before, and so we report the actual smoothness parameter values of the Matérn class, ν_i. Overall, the smoothness of the bias processes is around .5, which corresponds to the exponential covariance class. Although δ_i seems small relative to α_i, we inflate the covariance over the land by $(1 + \delta_i)^2$ times the covariance over the ocean, so it is a significant change in the covariance over the land. Finally, we note that the estimates for DJF and JJA are fairly similar.

3.3 Correlations Between the Model Biases

One of our main interests is how the biases of each model outputs are correlated with each other. To build a joint statistical model for the b_i's defined in (1), we particularly need models for $\sigma_{ij}(\mathbf{s}) = \operatorname{cov}\{a_i(\mathbf{s}), a_j(\mathbf{s})\}$ ($i, j = 2, \ldots, 20$). This is different from the covariance model for a_i that we discussed in Section 3.1, because $\sigma_{ij}(\mathbf{s})$ is a *cross-covariance*. Usual spatial covariance functions, such as Matérn class, may not be appropriate for modeling σ_{ij} because, for instance, it can take negative values.

Our idea for estimating $\sigma_{ij}(\mathbf{s})$ is to apply a kernel smoother to the sample statistics $\tilde{\bar{D}}_{ij}(\mathbf{s}) = \tilde{\bar{D}}_i(\mathbf{s})\tilde{\bar{D}}_j(\mathbf{s})$. Here $\tilde{\bar{D}}_i(\mathbf{s})$ equals $\bar{D}_i(\mathbf{s})$ but with the estimated mean values using the parameter estimates in Tables 2 and 3 subtracted. The $\tilde{\bar{D}}_i$'s are assumed to be mean 0, and thus $\tilde{\bar{D}}_{ij}$ is related to the cross-covariances of

Table 2. MLE estimates for DJF averages

Model	μ_{0i}	μ_{1i}	μ_{2i}	μ_{3i}	β_i	ν_i	α_i	δ_i
2	−.0447	.0071	1.4629	−.0006	1,305.4203	.4588	2.1599	.6477
3	.8954	−.0108	.5618	.0005	1,423.8011	.5795	3.6890	.5125
4	.8728	−.0075	.4995	.0010	1,975.6505	.4026	2.5803	.6713
5	1.6158	.0233	.6556	.0006	2,941.1116	.5336	4.8570	.4117
6	.6472	.0095	.6749	.0010	1,381.9561	.5683	2.7177	.6395
7	.1093	−.1630	.2392	.0008	5,575.8392	.3684	7.0323	.2296
8	−.4683	.0358	−.1747	0	2,063.9520	.4942	3.8876	.5642
9	.1034	.0208	.8656	.0001	1,423.0634	.5466	3.4714	.6810
10	−.1884	.1879	−.4706	−.0018	3,989.8403	.6182	31.8067	.0255
11	.9452	−.0359	.3916	−.0012	1,955.9095	.5428	5.8097	.3058
12	1.4438	−.0540	.2757	−.0009	5,202.5915	.5095	7.5262	.4763
13	1.1964	−.0381	.3456	−.0008	1,713.7778	.5538	3.3057	.5221
14	.4183	.0068	.3443	−.0010	822.8218	.6095	1.7618	.7915
15	−.5040	−.0329	.9223	.0011	1,117.3721	.4769	1.9867	.6211
16	.6465	.0231	.4969	−.0008	1,177.6320	.5385	1.9344	.4993
17	.2020	−.0130	.3952	.0004	1,021.6539	.6523	2.3250	.5074
18	1.3609	.0878	−.0802	−.0008	2,199.3238	.6468	1.0881	.0541
19	−.4049	−.0259	2.2031	.0008	2,646.4107	.1849	1.0559	1.0476
20	.9534	−.0062	2.0699	.0004	2,245.2868	.4316	3.2668	.8793

biases of model i and j. To be precise, for each AOGCM pair i and j, we assume that

$$\tilde{\tilde{D}}_{ij}(\mathbf{s}) = \sigma_{ij}(\mathbf{s}) + \epsilon_{ij}(\mathbf{s}), \tag{5}$$

where $\epsilon_{ij}(\mathbf{s})$ is a spatial process with mean 0. Then consider a kernel estimator for $\sigma_{ij}(\mathbf{s})$,

$$\hat{\sigma}_{ij}(\mathbf{s}) = \sum_{k=1}^{1,656} K\left(\frac{|\mathbf{s}, \mathbf{s}_k|}{h}\right) \tilde{\tilde{D}}_{ij}(\mathbf{s}_k) \cdot \left[\sum_{k=1}^{1,656} K\left(\frac{|\mathbf{s}, \mathbf{s}_k|}{h}\right)\right]^{-1}$$

for nonnegative kernel function K and bandwidth h. 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. Now let $\mathbf{\Sigma}(\mathbf{s}) = (\sigma_{ij}(\mathbf{s}))_{i,j=2,\ldots,20}$, and denote its kernel estimate as $\hat{\mathbf{\Sigma}}(\mathbf{s}) = (\hat{\sigma}_{ij}(\mathbf{s}))_{i,j=2,\ldots,20}$. For each $\mathbf{s}$, $\hat{\mathbf{\Sigma}}(\mathbf{s})$ is nonnegative definite; for $\tilde{\tilde{D}}(\mathbf{s}) = (\tilde{\tilde{D}}_2(\mathbf{s}), \ldots, \tilde{\tilde{D}}_{20}(\mathbf{s}))^T$ and for any nonzero $\mathbf{x} = (x_1, \ldots, x_{19})^T \in \mathbb{R}^{19}$,

$$\begin{aligned}
&\mathbf{x}^T \hat{\mathbf{\Sigma}}(\mathbf{s})\mathbf{x} \\
&\quad = \sum_{k=1}^{1,656} K\left(\frac{|\mathbf{s}, \mathbf{s}_k|}{h}\right) \{\mathbf{x}^T \tilde{\tilde{D}}(\mathbf{s}_k)\}^2 \cdot \left[\sum_{k=1}^{1,656} K\left(\frac{|\mathbf{s}, \mathbf{s}_k|}{h}\right)\right]^{-1} \\
&\quad \geq 0.
\end{aligned}$$

But $\hat{\mathbf{\Sigma}}(\mathbf{s})$ may not be positive definite. Because we have a fixed number of $\tilde{\tilde{D}}_i$'s, and thus the dimension of $\mathbf{\Sigma}(\cdot)$ is fixed, we

Table 3. MLE estimates for JJA averages

Model	μ_{0i}	μ_{1i}	μ_{2i}	μ_{3i}	β_i	ν_i	α_i	δ_i
2	−.0865	.0434	.2780	−.0002	1,759.548	.4376	1.4964	.6016
3	.6538	.0190	.1892	.0007	1,759.293	.5332	1.9254	.5999
4	1.0446	.0609	−1.0807	.0012	2,650.400	.2908	1.5231	.4944
5	1.1957	.0271	.1957	.0010	3,356.275	.4483	2.7490	.3808
6	.3627	.0110	−.0166	.0012	1,323.982	.4647	1.5883	.5082
7	.3979	−.0205	−.5429	.0010	2,645.100	.4027	2.5659	.4344
8	−.3541	−.0087	−.4084	.0003	1,487.233	.5427	3.2541	.4623
9	.0462	.0222	.3924	.0004	2,046.440	.5440	4.8100	.2944
10	−.9498	.0422	−.1926	−.0016	2,852.056	.4997	5.7854	.6555
11	1.3586	−.0664	.2870	−.0013	2,503.948	.5109	3.4285	.4081
12	1.4088	.0156	−.3069	−.0005	4,560.329	.5438	6.2419	.3750
13	.8719	−.0067	.2550	−.0008	2,707.467	.4745	1.9458	.8710
14	.6508	.0319	−.1562	−.0006	2,474.549	.3860	1.7212	.6869
15	−.4343	.0330	−.3780	.0016	2,101.735	.3405	1.3852	.5711
16	.7512	.0072	.8560	−.0007	1,652.461	.4671	1.2025	.9530
17	.4178	.0419	.2092	.0004	1,681.918	.5115	1.9174	.4088
18	1.6444	−.0038	.4404	−.0006	2,823.030	.5164	3.5510	.3124
19	−.2534	.0671	−.5807	.0013	2,606.165	.2826	1.8235	.3605
20	.6320	.0559	−.7369	.0012	1,830.724	.3471	1.7890	.3003

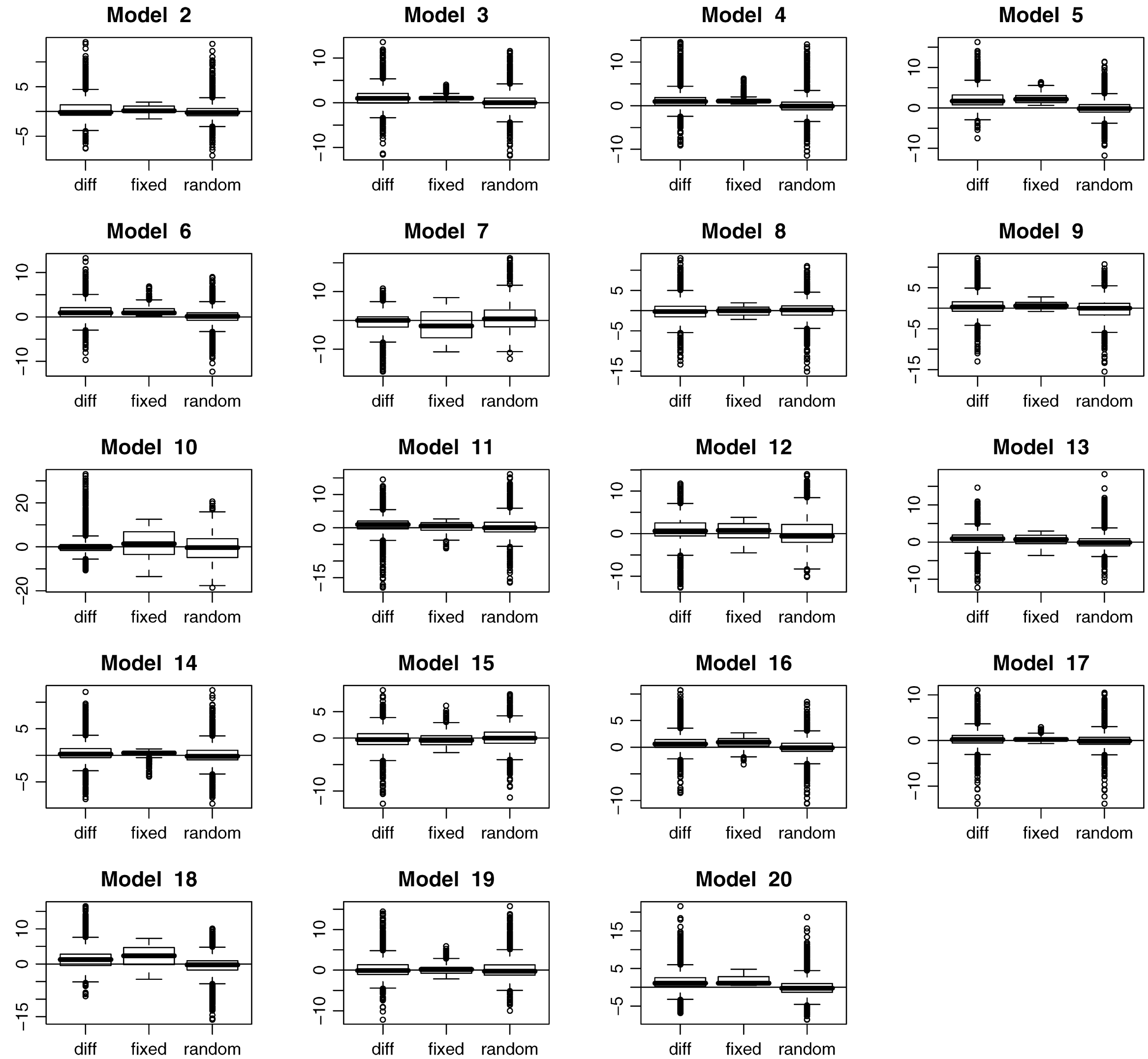

Figure 3. Comparison of the difference between the observation and ith model output ($\bar{D}_i$ for model i) and the estimated fixed part and random part in (1) (DJF).

should assess the consistency of the kernel estimate $\hat{\boldsymbol{\Sigma}}(\cdot)$. One technical issue is that the error ϵ_{ij} in (5) is spatially correlated. Altman (1990) discussed the consistency of kernel smoothers with correlated errors in the one-dimensional case. Because we expect the spatial correlation of ϵ_{ij} to die out after a certain spatial distance lag, we may expect similar consistency result on $\hat{\boldsymbol{\Sigma}}$ as the bandwidth goes to 0.

In our analysis, we use a Gaussian kernel for K with bandwidth $h = 800$ km. The choice of the bandwidth here is based on typical spatial variations in a climatologic mean temperature field. For every $\mathbf{s}$ and for all data cases, we found $\hat{\boldsymbol{\Sigma}}(\mathbf{s})$ to be positive definite.

We are interested in the correlation matrix $\mathbf{R}(\mathbf{s}) = (r_{ij}(\mathbf{s}))_{i,j=2,\ldots,20}$, with $r_{ij}(\mathbf{s}) = \hat{\sigma}_{ij}(\mathbf{s})/\sqrt{\hat{\sigma}_{ii}(\mathbf{s})\hat{\sigma}_{jj}(\mathbf{s})}$. The matrix $\mathbf{R}(\cdot)$ is useful for answering several questions. First, we can quantify the correlation between biases of pairs of models at a fixed spatial location. Second, using 1 minus the correlation values as a distance measure, we can classify models into subgroups with highly correlated biases. Third, we can identify spatial subregions in which certain pairs of models have higher correlated biases.

As an example, Figure 5 displays r_{5j} over the whole spatial domain for DJF averages, the correlation between model 5 bias and the other model biases. They show that models 5 and 6 have particularly highly correlated biases over the whole domain. The models 5 and 6 are both developed by the NOAA GFDL group and use different numerical schemes but the same physical core. This result confirms the hypothesis that models developed by the same group have highly correlated biases and thus cannot be assumed to be independent. Similarly, other model pairs developed by the same organization (e.g., models 8 and 9, or models 19 and 20) have noticeably higher correlated biases than other pairs of models, independent of the season (not shown). Those types of figures also can indicate the regions in which a pair of models has highly correlated biases.

Figure 6 gives a summary of correlation values for each pair of model biases. Each symbol denotes a correlation between biases of model i and j averaged over the whole spatial domain. Note first that most of the correlation values are positive and rather high. Pairs of models, some developed by same group of people, show very high correlation values that are consistent across seasons. Note that correlations between the model 10

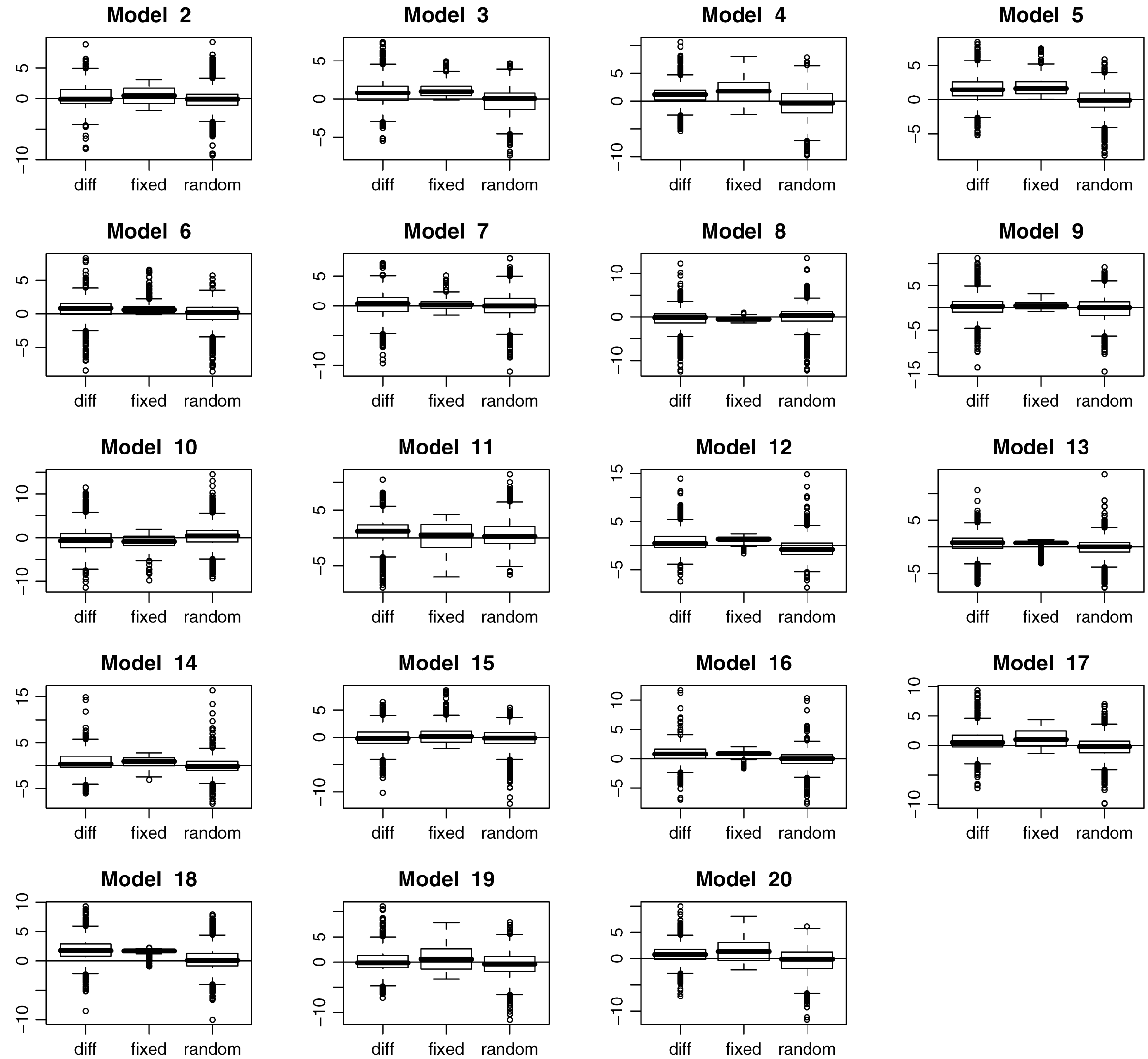

Figure 4. Comparison of the difference between the observation and ith model output ($\bar{D}_i$ for model i), and the estimated fixed part and random part in (1) (JJA).

bias and the other model biases are small and especially for DJF season, some of the model biases have negative correlation with model 10 bias (in crosses). This is another sign of the unusual behavior of model 10.

Monte Carlo simulation is used to determine the distribution of correlation fields when two models have independent bias fields. Based on the spatial model for biases [as in (2) and Tables 2 and 3], we simulate bias fields that are independent between models. From these simulated values, we calculate the correlation values in the same way as for the model and observation differences. We repeat this procedure 1,000 times, and calculate correlation values averaged across space for each time. These correlation values are comparable to the dots (or crosses) in Figure 6. For reference, we calculate the 99th and 1st percentiles of the correlation for all model pairs [$171 = \binom{19}{2}$ total pairs]. The maximum values of the 99th percentile and minimum for the 1st percentile from 1,000 simulations are the dotted lines in Figure 6. Another reference that is suitable for multiple comparison is the distribution of the maximum entry from the average correlation matrix. The 99th percentile and 1st percentile (dashed) and median (combination of dashed and dotted) of this maximum is included in Figure 6. Many of the symbols are above the dotted and even quite a number of the symbols are above the dashed lines. This supports our conclusion that the model biases are indeed correlated.

3.4 Verification of Our Methodology

We are able to test our methodology on initial condition ensemble runs that were simulated by the same models. Initial condition ensemble runs are created by running the same AOGCM but only changing the initial conditions slightly. This results in model outputs that have a different realization of the noise component (i.e., different weather) but very similar climatologic mean states. If the statistical model were accurate, then we would expect high correlation values for the ensembles from the same model and smaller correlation values from pairs of different models. Using four ensemble runs of model 2 and two ensembles each of models 5 and 6, and assuming that these eight ensemble members are from eight different models, we apply our methodology to calculate correlations among their biases. As in Section 3.3, before calculating the correlations, we

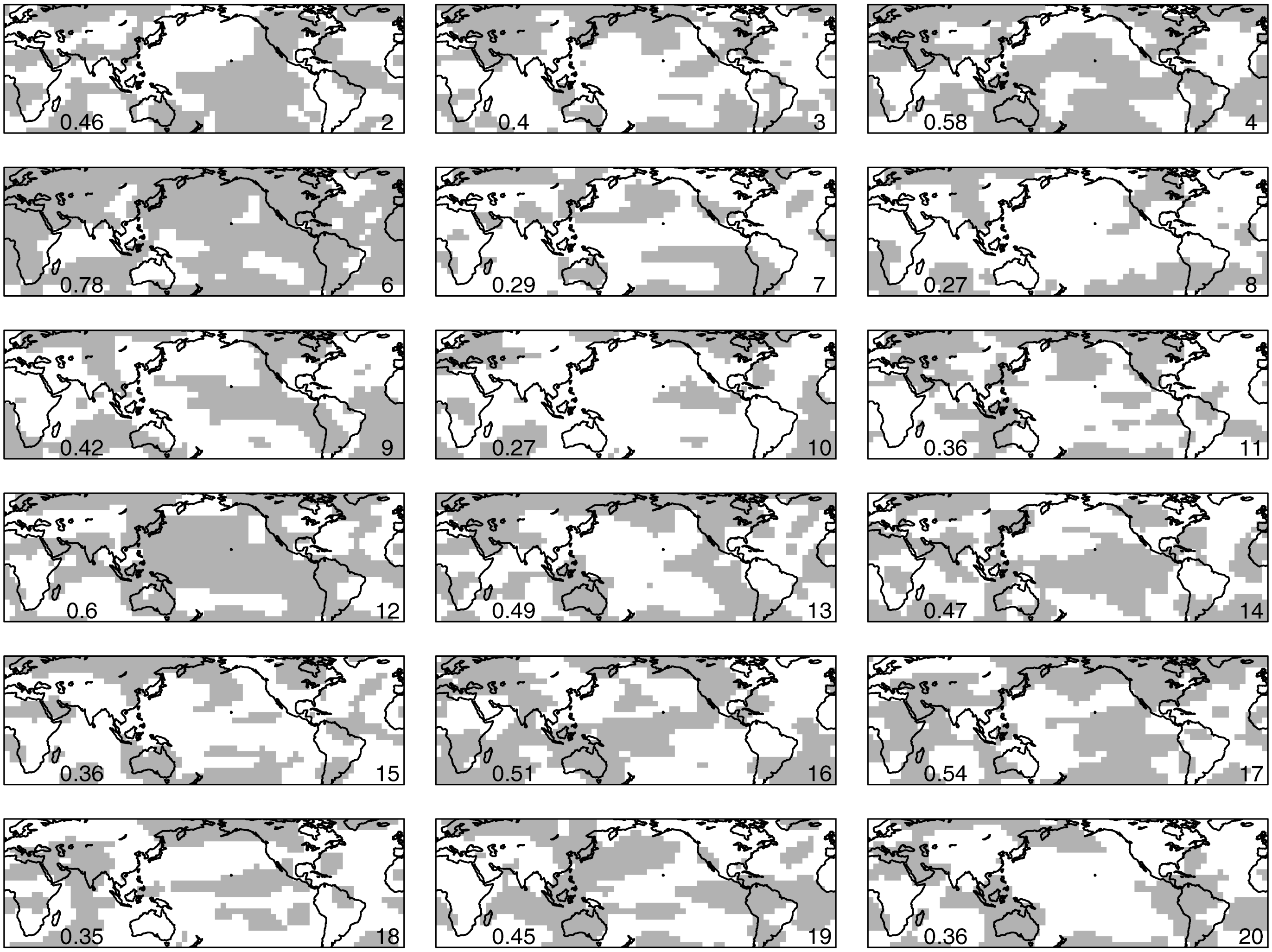

Figure 5. The correlation r_{5j} for $j = 2, \ldots, 20$ for DJF averages. The value of j is shown at the bottom right corner of each picture. The grid points with $r_{5j} > .6$ are in gray, and the rest are in white. The average value of r_{5j} over the spatial domain is given in the lower Indian ocean area.

subtract the mean field using the estimated parameter values in Tables 2 and 3.

The results are consistent with our expected outcome that climate values for each ensemble runs are similar and so correlations among biases of ensembles from the same model are all $>.97$. In addition, correlations among biases of ensembles from different models are similar to the values obtained from the original model runs.

4. ANALYSIS ON THE TREND

To study the climate model biases of the trend, we use seasonal averages of surface temperature data for each of the 30 years over the whole spatial domain. We first examine the relationship between the biases on the mean state and the biases on trend. One reason for developing climate models is to quantify possible climate change for the future. Thus accurate prediction of the trend is important, and comparing the simulated warming trend over the last decades with the observed trend is a way to assess the dynamic response of the model.

For practical reasons, models are still evaluated mostly on their mean state. Relatively accurate data sets are available for many variables, and the mean state requires only short equilibrium simulations, in many cases with an atmospheric model only. But evaluation on the trend requires a more expensive run of the full model from a preindustrial state (before the year 1900) up to the present day. A common assumption is that climate models that simulate the present-day mean climate state well will also be accurate in simulating the trend (Tebaldi et al. 2005). We test this hypothesis by comparing simple statistics from the biases. Then we apply the method of calculating correlations between model biases on the trend, as we did earlier for the biases on the mean state.

4.1 Estimated Spatial Trends in Temperature

To define the biases on the trend, we determine the slope of the temperature trend at each grid point for both observations and model data. At each grid point, we regress the seasonal averages on time, that is,

$$Z(\mathbf{s}, t) = \gamma_0(\mathbf{s}) + \gamma_1(\mathbf{s})(t - 15.5),$$

where $Z(\mathbf{s}, t)$ is the seasonal average of observations or climate model output at year t for the location $\mathbf{s}$ ($t = 1, \ldots, 30$). Instead

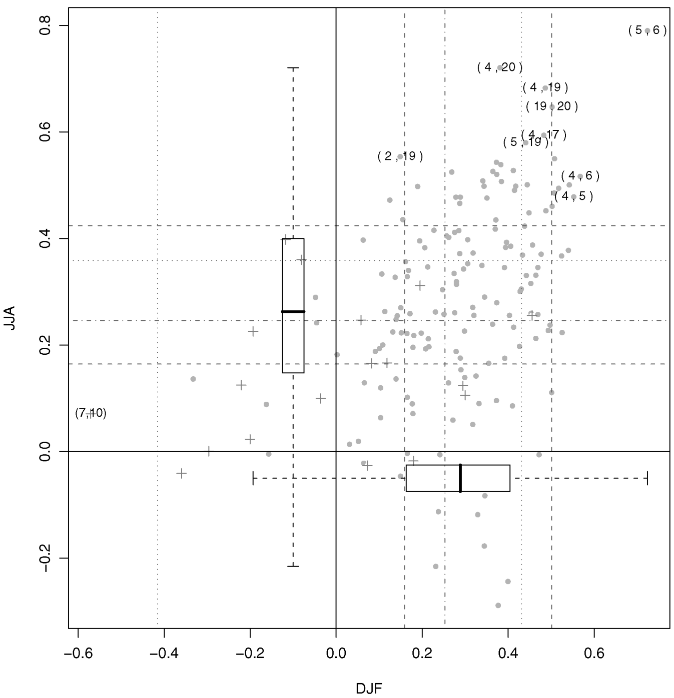

Figure 6. Average of r_{ij}'s over the whole spatial domain displayed for each pair of models as points (dots or crosses) for both seasons. When either i or j is 10, crosses are used; otherwise, dots are used. The distribution of the points for each season is given in boxplots. Refer to Section 3.3 for how the values for lines are obtained. Pairs (i, j) are displayed if $r_{ij} > .55$ for at least one season.

of regressing on t, we regress on $t - 15.5$, to have γ_0 and γ_1 independent of each other ($t = 1, \ldots, 30$, so that the average of all t values is 15.5).

The left column of Figure 7 shows the slope values $\gamma_1(\cdot)$ for observations and some model outputs. The surface of slopes for observations are rougher than those for the model output. For both observation and model outputs, many grid cells do not have significant slopes. Also, some grid cells have negative slopes, indicating a temperature decrease over the 30-year period.

We define the bias of the trend as the difference between the slope in the observations and the slope in the model data. An alternative would be to use the ratio of the two slopes, but for many grid cells, the very small slopes would make this ratio unstable.

Now we compare the biases from the trend and the biases from the mean state. Figure 8 shows scatterplots of several combinations of the four values: DJF RMS mean, JJA RMS mean, DJF RMS trend, and JJA RMS trend. DJF RMS mean is calculated as the RMS of the bias in the mean state for DJF averages (as in Fig. 1), separately for each model, and JJA RMS mean is the same as DJF RMS mean but for Boreal summer. DJF RMS trend is the RMS of the trend biases (observation slope − model slope) for DJF. The DJF RMS mean and JJA RMS mean are highly correlated (correlation value = 0.73); that is, models that agree well with observations for one season tend to also agree well with those for the other season. Model 10 has very large RMS mean values for both seasons. On the other hand, DJF and JJA RMS mean are only weakly correlated with trend RMS for the corresponding seasons. Although model 10 has exceptionally large DJF and JJA RMS means, it does not have the largest trend RMS. The results call into question to some degree the common assumption that a model, that does well in simulating the climate mean state can be trusted to project changes into the future (the latter being temperature increase over time, i.e., a trend). This assumption is either explicit (e.g., Tebaldi et al. 2005) or implicit (e.g., IPCC 2001) in many studies.

4.2 Correlation of the Trend Biases

To quantify the correlation between the model biases on the linear warming trend, we apply the same analysis described in Section 3.3 to the biases of the trend. Our goal is to test the relationship between pairs of models with highly correlated biases on the mean state and on the trend. Figure 9 shows the correlations among the biases of trend for both seasons. The pairs of models with high correlations are not the same as the pairs of models from the analysis on the mean state (Fig. 6). But what is more surprising is that for almost all model pairs, the correlation level is consistently high [Fig. 9(a)]. Several factors could cause such a result. The observations are obviously only one possible realization and contain internal variability. Whereas the noise contribution is small in a 30-year climatologic mean, linear trends at a single grid point can be influenced substantially by noise, that is, the internal unforced variability in the climate system. Part of that result also could be caused by some local biases in the observations, particularly in regions with poor data coverage. Those obviously would not be picked up by any climate model, so all models would differ from the observations in a similar way. Another possible explanation is that many models do not include all of the radiative forcing components. For example, the cooling effect of volcanic eruptions is not included in some models, causing them to overestimate temperature in most regions shortly after a volcanic eruption.

We are particularly concerned about the rough surface of slopes from the observation. To quantify the effect of this on the correlation values, we recreated Figure 9(a) but smoothed the observation and model slopes with a Gaussian kernel and bandwidth of 1,600 km before the correlation analysis, to remove some of the high-frequency signals. The results are not sensitive to the particular choice of kernels or to the bandwidth. With the smoothed data, there are more grid points with significant positive slopes. Furthermore, Figure 9(b) shows that the average correlation level has decreased significantly, whereas the maximum correlation has remained at a similar level. The pairs of models with highly correlated biases have not changed much for the original data and the smoothed data. An interesting point here is that a comparison of Figures 6 and 9 reveals no obvious correspondence in the correlation across models for the biases of the mean state and the biases of the trend. Recall that in Figure 8, DJF RMS for trend and JJA RMS for trend have negative correlation (top right). However, in Figure 9, the correlations from both seasons seem to have high correlation; pairs of models with high correlated biases on trend for DJF averages tend to have high correlated biases on trend for JJA averages as well.

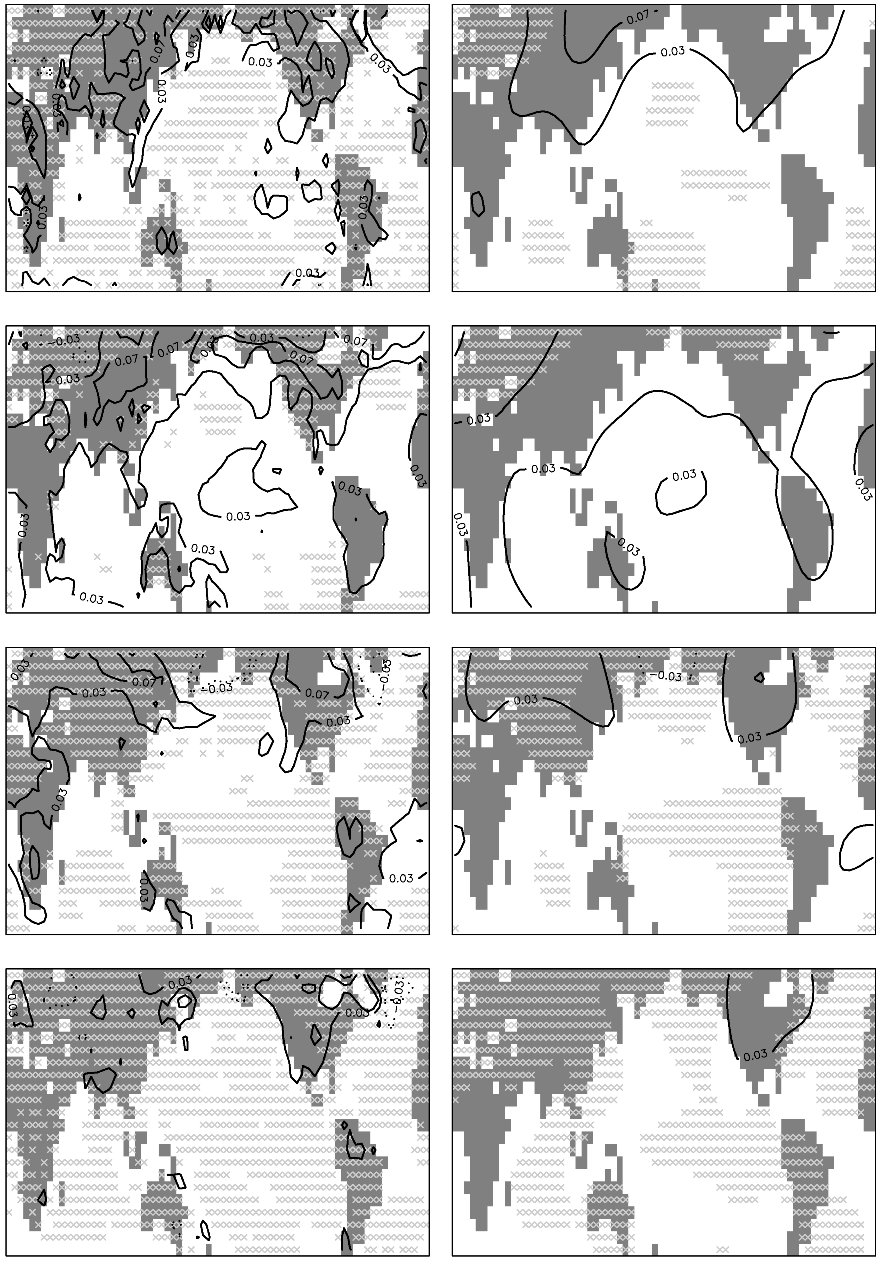

Figure 7. Plot of $\gamma_1(\mathbf{s})$ for observation and models 2, 10, and 18 outputs (top to bottom) from DJF averages. The left column shows values from the original data, and the right column shows the smoothed data (with bandwidth 1,600 km). The grid points with crosses are where p values of the regression are $>.1$.

5. CONCLUSIONS AND DISCUSSION

We have presented the results of quantification of AOGCM model biases and their dependence across different models. Based on our analysis, many AOGCM models (especially those developed by the same parent organizations) have highly correlated biases, and thus the effective number of "independent" AOGCM models is much lower than the actual number of models. This lets us form subgroups of models that share "common" features and to find a better strategy for combining the information from different model outputs rather than taking a naive average. We also have demonstrated that the performance of AOGCM models on the mean temperature state has little relationship with its performance in reproducing the observed spatial temperature trend. This conflicts with a standard assumption used to interpret different AOGCM projections of future climate. Our results suggest the need for better model validation procedures that are multivariate.

The reason why we fit the $a_i(\mathbf{s})$'s separately instead of modeling them jointly is because in building a joint multivariate model for $\mathbf{a}(\mathbf{s}) = (a_2(\mathbf{s}), \ldots, a_{20}(\mathbf{s}))$, we need to specify the cross-covariance structure between $a_i(\mathbf{s})$ and $a_j(\mathbf{s})$, $i \neq j$. This is a challenging problem, and we are not aware of flexible cross-covariance models that would be suitable for modeling the $\mathbf{a}(\mathbf{s})$'s. We contend that using a limited and inflexible cross-covariance model for modeling the $a_i(\mathbf{s})$'s jointly would lead to less satisfactory estimates for the r_{ij}'s compared with our results.

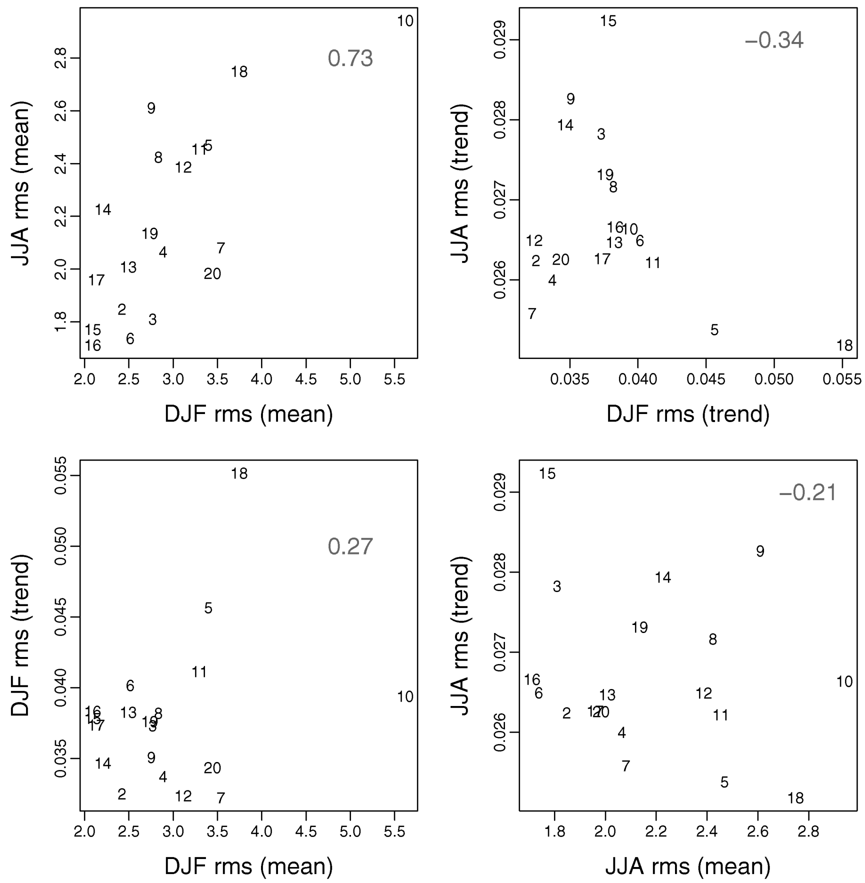

Figure 8. Comparison of biases the mean state and the trend for each season. The biases are summarized over the whole spatial region as rms errors. Each number denotes the model number; the number in gray is the correlation between the two rms errors.

Eventually we are interested in building joint statistical models to combine all of the climate models with observations. Our results demonstrate that the statistical approaches of Tebaldi et al. (2005), Furrer et al. (2007), and Smith et al. (2006) may need to be extended due to the biasedness of the climate models and, more importantly, the dependence among biases from different AOGCMs. Achieving this requires flexible cross-covariance models that are valid on a sphere. Another challenge in this task is the spatial analysis of large data sets. Dealing with a large number of global processes and modeling them jointly

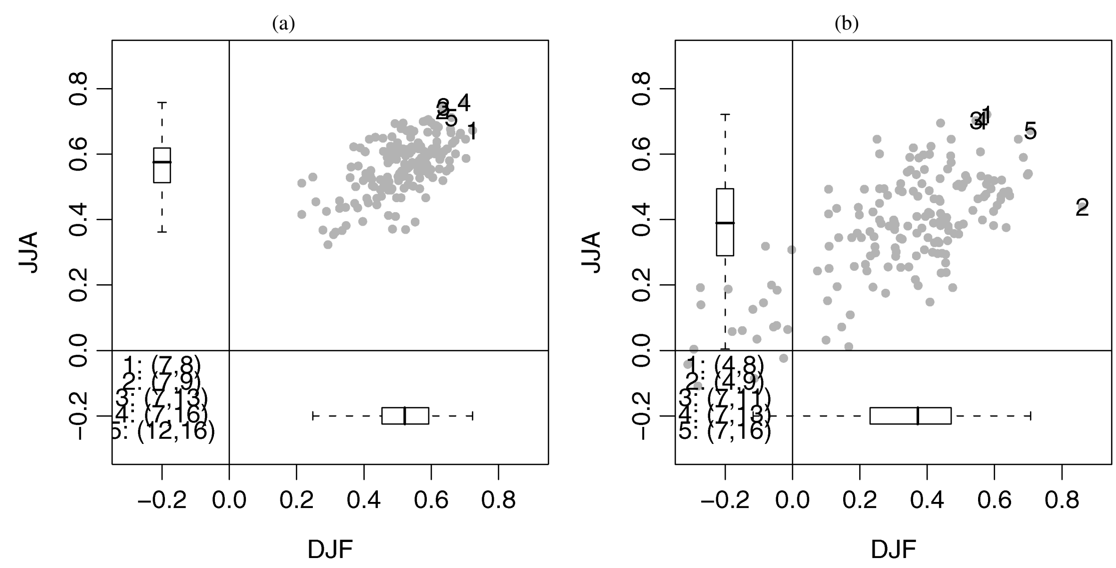

Figure 9. Similar figures as in Figure 6 for the biases of the trend from the original data (a) and from the smoothed data (b), with Gaussian kernel, bandwidth 1,600 km. Pair numbers displayed are pairs of model biases with correlation >.7 (a) or .71 (b) for at least one season.

requires significant computing resources and efficient computational methods.

Our correlation estimates are based on the maximum likelihood error estimates given in Tables 2 and 3, and the necessary uncertainty about these estimates has not been discussed in the article. To study the uncertainty of the estimates, it would be natural to consider a Bayesian hierarchical model framework.

[Received December 2006. Revised May 2007.]

REFERENCES

Altman, N. (1990), "Kernel Smoothing of Data With Correlated Errors," *Journal of the American Statistical Association*, 85, 749–759.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004), *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman & Hall/CRC.

Barnett, T., **???** (2005), "Detecting and Attributing External Influences on the Climate System: A Review of Recent Advances," *Journal of Climate*, 18, 1291–1314.

Cantelaube, P., and Terres, J.-M. (2005), "Seasonal Weather Forecasts for Crop Yield Modelling in Europe," *Tellus A*, 57A, 476–487.

Furrer, R., Sain, S. R., Nychka, D. W., and Meehl, G. A. (2007), "Multivariate Bayesian Analysis of Atmosphere-Ocean General Circulation Models," *Environmental and Ecological Statistics*, **??**, **??–??**.

Gillett, N. P., Zwiers, F. W., Weaver, A. J., Hegerl, G. C., Allen, M. R., and Stott, P. A. (2002), "Detecting Anthropogenic Influence With a Multi-Model Ensemble," *Journal of Geophysical Research*, 29.

Intergovernmental Panel on Climate Change (2001), *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, eds. J. T. Houghton, et al., Cambridge, U.K.: Cambridge University Press, pp. **??–??**.

Jones, P., New, M., Parker, D., Martin, S., and Rigor, I. (1999), "Surface Air Temperature and Its Variations Over the Last 150 Years," *Reviews of Geophysics*, 37, 173–199.

Jun, M., and Stein, M. L. (2004), "Statistical Comparison of Observed and CMAQ Modeled Daily Sulfate Levels," *Atmospheric Environment*, 38, 4427–4436.

——— (2007), "An Approach to Producing Space-Time Covariance Functions on Spheres," *Technometrics*, **??**, **??–??**.

Knutti, R., Stocker, T. F., Joos, F., and Plattner, G.-K. (2002), "Constraints on Radiative Forcing and Future Climate Change From Observations and Climate Model Ensembles," *Nature*, 416, 719–723.

Lambert, S. J., and Boer, G. J. (2001), "CMIP1 Evaluation and Intercomparison of Coupled Climate Models," *Climate Dynamics*, 17, 83–106.

Meehl, G., Washington, W., Ammann, C., Arblaster, J., Wigley, T., and Tebaldi, C. (2004), "Combinations of Natural and Anthropogenic Forcings in Twentieth-Century Climate," *Journal of Climate*, 17, 3721–3727.

Nakićenović, **??.**, **???** (2000), *Special Report on Emission Scenarios*, *Intergovernmental Panel on Climate Change*, Cambridge, U.K.: Cambridge University Press, 599 pp.

Rayner, N., Brohan, P., Parker, D., Folland, C., Kennedy, J., Vanicek, M., Ansell, T., and Tett, S. (2006), "Improved Analyses of Changes and Uncertainties in Marine Temperature Measured in situ Since the Mid-Nineteenth Century: The HadSST2 Dataset," *Journal of Climate*, 19, 446–469.

Shao, X., Stein, M., and Ching, J. (2006), "Statistical Comparisons of Methods for Interpolation the Output of a Numerical Air Quality Model," *Journal of Statistical Planning and Inference*, **??**, **??–??**.

Smith, R. L., Tebaldi, C., Nychka, D., and Mearns, L. (2006), "Bayesian Modeling of Uncertainty in Ensembles of Climate Models," unpublished manuscript, **???**.

Stainforth, D. A., **???** (2005), "Uncertainty in Predictions of the Climate Response to Rising Levels of Greenhouse Gases," *Nature*, 433, 403–406.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag.

Tebaldi, C., and Knutti, R. (2007), "The Use of the Multi-Model Ensemble in Probabilistic Climate Projections," *Phil. Trans. Royal Society*, **??**, **??–??**.

Tebaldi, C., Smith, R. L., Nychka, D., and Mearns, L. O. (2005), "Quantifying Uncertainty in Projections of Regional Climate Change: A Bayesian Approach to the Analysis of Multimodel Ensembles," *Journal of Climate*, 18, 1524–1540.

Thomson, M. C., Doblas-Reyes, F. J., Mason, S. J., Hagedorn, R., Connor, S. J., Phindela, T., Morse, A. P., and Palmer, T. N. (2006), "Malaria Early Warnings Based on Seasonal Climate Forecasts From Multi-Model Ensembles," *Nature*, 439, 576–579.

Yaglom, A. M. (1987), *Correlation Theory of Stationary and Related Random Functions*, Vol. I, New York: Springer-Verlag.

Yun, W. T., Stefanova, L., and Krishnamurti, T. N. (2003), "Improvement of the Multimodel Supersensemble Technique for Seasonal Forecasts," *Journal of Climate*, 16, 3834–3840.