Probabilistic Multimodel Regional Temperature Change Projections

ARTHUR M. GREENE, LISA GODDARD, AND UPMANU LALL

The International Research Institute for Climate and Society, The Earth Institute at Columbia University, Palisades, New York

(Manuscript received 31 May 2005, in final form 28 December 2005)

ABSTRACT

Regional temperature change projections for the twenty-first century are generated using a multimodel ensemble of atmosphere–ocean general circulation models. The models are assigned coefficients jointly, using a Bayesian linear model fitted to regional observations and simulations of the climate of the twentieth century. Probability models with varying degrees of complexity are explored, and a selection is made based on Bayesian deviance statistics, coefficient properties, and a classical cross-validation measure utilizing temporally averaged data. The model selected is shown to be superior in predictive skill to a naïve model consisting of the unweighted mean of the underlying atmosphere–ocean GCM (AOGCM) simulations, although the skill differential varies regionally. Temperature projections for the A2 and B1 scenarios from the Intergovernmental Panel on Climate Change (IPCC) Special Report on Emissions Scenarios are presented.

1. Introduction

Projections of regional climate change are of potential economic value (Katz and Murphy 1997) in that they can provide—at least in theory—a window into the future on spatial scales that are politically and operationally meaningful. Interest in such projections also arises in the context of recent progress in seasonal-to-interannual forecasting (Barnston et al. 2000; Palmer et al. 2000; Goddard et al. 2003), coupled with the realization that value may derive from considering the behavior of climate on longer time horizons as well. Historically, coupled atmosphere–ocean general circulation models (AOGCMs) have demonstrated only limited skill in the simulation of regional climates (see, e.g., Houghton et al. 2001, chapter 10). However, given the ongoing development of such models, continuing assessment of their potential utility in generating regional projections is prudent.

In this study, regional temperature projections are generated by combining simulations from 14 AOGCMs, using a Bayesian linear model. Initially, three probability model structures of differing complexity are consid-

ered, the fit to observations being assessed using both objective numerical measures and direct examination of coefficient properties. The probability model selected by this means has a hierarchical structure, in which regional sets of AOGCM coefficients are modeled as draws from a parent, or “population” distribution. The coefficients themselves are returned in the form of probability density functions (PDFs) rather than as point estimates; these PDFs are applied to AOGCM simulations for the twenty-first century to generate the projections and their associated uncertainties. The projections then also take the form of PDFs, and may thus be considered probabilistic. However, verification, in the sense of comparing categorical forecast probabilities with observed occurrence frequencies, is supplanted here by the use of Bayesian deviance statistics and the more conventional mean squared error (MSE), owing to the long time horizons considered and the consequent lack of forecast-verification cycles.

At base, the methodology consists of interposing a statistical translation layer between AOGCM simulations and the climate variable to be predicted, conditional on the observed past relationship between them. It thus bears some similarity to model calibration, as applied in the quantification of climate change uncertainty (Allen et al. 2000) and in fingerprinting studies (Allen and Tett 1999). More broadly, it may be considered interpretation of computer simulations of a complex natural process through the application of a prob-

Corresponding author address: Arthur M. Greene, The International Research Institute for Climate and Society, The Earth Institute at Columbia University, 202 Monell Building, 61 Route 9W, Palisades, NY 10964.
E-mail: amg@iri.columbia.edu

ability model. Light shed on the process by the simulations is then refracted through the prism of probabilistic interpretation. Craig et al. (2001) discuss some fundamentals of estimating uncertainty in Bayesian probability structures in this context. There is also some similarity between the methodology described here and the model output statistics (MOS) approach developed by Glahn and Lowry (1972). There are also significant differences, however, the most important being the use herein of a multimodel ensemble rather than just a single dynamic weather prediction model. On the other hand, only a single AOGCM variable, rather than a suite of such variables, is utilized here.

The data, including the AOGCMs whose simulations are employed, are described in section 2, while methodology, including probability model structure, is discussed in section 3. Model comparison is addressed in section 4 and regional temperature projections are presented in section 5. The paper concludes with a discussion and a summary, in sections 6 and 7, respectively. Additional details regarding the probability models and estimation are provided in an appendix.

2. Data

For observations, the land surface temperature dataset “CRU TS 2.0” of the Climatic Research Unit, University of East Anglia, gridded at 0.5° , was utilized (New et al. 1999, 2000). No attempt is made in this dataset to compensate for the effects of land-use changes or urbanization, and only a minority of the AOGCMs utilized (see below) incorporates such changes explicitly. However, for the “Twentieth Century Climate in Coupled Models” (20C3M) simulations employed here, all utilize time-varying trace-gas concentrations, including sulfate aerosols that vary in both space and time (Boucher and Pham 2002). Thus, to some extent the regional atmospheric effects of urbanization are implicitly included. In the analyses to be discussed, the relatively finescale observational data are aggregated into regions comprising on the order of 10^3 individual grid boxes; on this scale, effects of urbanization are likely to be small. Data aggregation also mitigates potential problems arising from the filling of missing data, while the high spatial resolution permits precise masking in the delineation of regions.

The 14 AOGCMs whose outputs are considered comprise those contributing to the Fourth Assessment Report (AR4) of the Intergovernmental Panel on Climate Change (IPCC) for which, by mid-April 2005, outputs for three experiments were available from the Program for Climate Model Diagnosis and Intercomparison data archive. These experiments are 20C3M

(Hegerl et al. 2003), and scenarios A2 and B1 as described in the IPCC Special Report on Emissions Scenarios (SRES) (Nakićenović et al. 2000). The 20C3M experiment, in which a “best effort” is made to simulate the climate of the twentieth century, was utilized in fitting the probability model to observations, while A2 and B1 were used in the generation of projections.

Guidance for the 20C3M simulations did not dictate the use of specified forcings, and those utilized do differ somewhat from model to model. All account for greenhouse gases and sulfate aerosols at a minimum, but the treatment of aerosol indirect effects, CFCs, solar variability, and other forcings varies. Simulations are utilized in the form of ensemble means, the number of ensemble members varying among AOGCMs (Table 1). This variation is taken into account in one, but not all, of the candidate probability models. For A2 and B1, fewer ensemble members were generally provided than for 20C3M. Owing to the varying dates at which the archived 20C3M runs begin and end, as well as the desire to use contiguous December–February (DJF) values, the time periods ultimately utilized for the twentieth and twenty-first century are 1902–98 and 2005–98, respectively.

In scenario B1, the atmospheric CO_2 concentration in the year 2100 reaches a level of 549 ppm, about twice the preindustrial level; in A2 the corresponding value is 856 ppm, somewhat more than twice the level of the 1990s. Thus, B1 represents a moderate outlook, relatively speaking, while A2 is more extreme (Nakićenović et al. 2000). Most of the AOGCMs have atmospheric components of medium resolution, ranging from 2° to 4° . A listing is provided in Table 1.

3. Methodology

a. Data preprocessing

Temperature data series are considered here in the form of regional means, the regions being those previously defined by Giorgi and Francisco (2000; Fig. 1). These regional definitions have been employed in a number of other studies, so their use here should facilitate comparison, while the data aggregation itself may be expected to improve the statistical properties of the resultant series.

Values for both simulations and observations are expressed as anomalies relative to 1902–98, the full extent of the twentieth-century data record employed. Use of anomalies amounts to the removal of individual additive model bias (Fig. 2b), while the full data period is taken for climatology in order to minimize bias resulting from differential AOGCM climate sensitivity. The

TABLE 1. Some characteristics of the models initially considered, listed alphabetically by model name. Modeling groups: National Center for Atmospheric Research (NCAR); Canadian Centre for Climate Modelling and Analysis (CCCMA); Météo-France/Centre National de Recherches Météorologiques (CNRM); Commonwealth Scientific and Industrial Research Organisation (CSIRO); Max Planck Institute for Meteorology (MPI); Goddard Institute for Space Studies (GISS); Hadley Centre for Climate Prediction and Research/Met Office (Hadley Centre); Institute for Numerical Mathematics, Russian Academy of Science (INM); U.S. Department of Commerce/NOAA/Geophysical Fluid Dynamics Laboratory (GFDL); Institut Pierre-Simon Laplace (IPSL); Center for Climate System Research (CCSR; University of Tokyo), National Institute for Environmental Studies (NIES), and Frontier Research Center for Global Change (JAMSTEC); Meteorological Research Institute of Japan (MRI). “Resolution” refers to the atmospheric model component and is given as latitude $\times$ longitude if these resolutions differ. This figure is approximate for spectral models. Here n_{ens} refers to the number of ensemble members provided for the 20C3M simulations. For SRES scenarios A2 and B1 the number is generally lower (for those AOGCMS for which $n_{ens} > 1$). See references cited for more detailed information.

Group	Country	Model	Resolution	n_{ens}	References
NCAR	United States	CCSM3	2.8°	8	Collins et al. (2006)
CCCMA	Canada	CGCM3.1 (T47)	3.8°	1	http://www.cccma.bc.ec.gc.ca/models/cgcm3.shtml
CNRM	France	CNRM-CM3	2.8°	1	Salas-Mélia et al. (2005, manuscript submitted to <i>Climate Dyn.</i>)
CSIRO	Australia	CSIRO-Mk3.0	1.9°	1	Gordon et al. (2002)
MPI	Germany	ECHAM5/MPI-OM	1.9°	3	Jungclaus et al. (2006)
GISS	United States	GISS_E_R.	4° $\times$ 5°	9	Schmidt et al. (2006); Russell et al. (1995)
Hadley Centre	United Kingdom	HadCM3	2.5° $\times$ 3.75°	2	Gordon (2000)
INM	Russia	INM-CM3.0	4° $\times$ 5°	1	Diansky and Volodin (2002)
GFDL	United States	GFDL-CM2.0	2° $\times$ 2.5°	3	Delworth et al. (2006); http://data1.gfdl.noaa.gov/CM2.X/
GFDL	United States	GFDL-CM2.1	2° $\times$ 2.5°	3	See GFDL-CM2.0
IPSL	France	IPSL-CM4	2.5° $\times$ 3.75°	1	http://dods.ipsl.jussieu.fr/omamce/
CCSR/NIES/JAMSTEC	Japan	MIROC (med-res)	2.8°	3	Hasumi and Emory (2004)
MRI	Japan	MRI-CGCM2.3.2	2.8°	5	Yukimoto et al. (2001)
NCAR	United States	PCM1	2.8°	2	Washington et al. (2000)

range of offsets among AOGCM representations of regional temperature can be considerable (Fig. 2a).

b. Some salient data characteristics

Figure 3 shows a substantial difference in interannual variance between typical high- and low-latitude regions.

FIG. 1. Regional definitions from Giorgi and Francisco (2000).

Such differences, with greater variability at higher latitudes, tend to be consistent across seasons. Comparison between the simplest of the probability models considered, in which these differences are not represented, and a more complex model in which they are (see section 4), indicates clearly that the inclusion of regional differentiation in variance significantly improves model fit.

Figure 4 shows correlation matrices for the observations and AOGCMs for two contrasting regions, NAS and EAF, for the annual mean, DJF and July–August (JJA) for 1902–98. In the extratropics during the summer months (Fig. 4c), correlations between AOGCM simulations, as well as between simulations and observations, tend to be higher than during the winter or for the annual mean, and large-scale patterns of inter-AOGCM correlation are somewhat more apparent. Seasonal variation in low-latitude correlation structure, as one might expect, is weak, while the dark marginal bands at top and left in Figs. 4d–f indicate that most AOGCMs are poorly correlated with observations in EAF. This feature is not uniform across regions, however, suggesting that predictability is likely to vary. The intercorrelations on these plots, particularly for EAF, suggest that an attempt might profitably be

FIG. 2. Observations (heavy lines) and (a) raw AOGCM simulations and (b) bias-corrected temperatures, for region CAM for the annual mean. Dotted lines are used in (b) for legibility.

FIG. 3. Example data series for high-latitude region ALA (heavy line, scale at left) and low-latitude EAF (scale at right), for the annual mean.

made to account for covariance in the probability model.

c. Probability models

Three probability model structures, designated A, B, and C, are considered.

For model A we write

$$Y_{ik} \sim N(\mu_{ik}, \sigma^2), \quad (1)$$

while for models B and C we have

$$Y_{ik} \sim N(\mu_{ik}, \sigma_k^2), \quad (2)$$

where Y_{ik} represents the observed temperature for year i at region k , and in each case, the rhs describes the distributional attributes of Y . Thus, for model A, regional differences in error variance are ignored, while in B and C they are represented.

All models utilize the same functional core:

$$\mu_{ik} = \beta_{0k} + \sum_j \beta_{jk} X_{ijk}, \quad (3)$$

where μ_{ik} is the expectation of Y_{ik} , as given in Eqs. (1) or (2), β_{0k} is a regionally dependent constant term, β_{jk} is the coefficient for AOGCM j at region k , and X_{ijk} is the temperature simulated at region k by AOGCM j for year i . Equations (1) and (2) describe the stochastic node Y , while Eq. (3) encodes the fundamental model structure; that is, the expected temperatures are given by a linear combination of AOGCM simulations for each region k and time i , plus a regional offset.

In addition, for models A and B we have

$$\beta_{jk} \sim N(\theta_{jk}, \tau_{jk}^2), \quad (4)$$

while for model C the corresponding relation is

$$\beta_{jk} \sim \text{MVN}(\theta_{(j)}, \Sigma). \quad (5)$$

Equation (4) states that for structures A and B the individual AOGCM coefficients β_{jk} are modeled as having independent normal distributions, with θ_{jk} and τ_{jk}^2 the associated means and variances. For structure C, on the other hand [Eq. (5)], the β_{jk} are modeled as having some *common* dependence, in the form of a multivariate normal *population* distribution with vector mean $\theta_{(j)}$ and covariance matrix Σ . [The parenthetical subscript is appended as an indication that the distribution is multivariate on j , i.e., over AOGCMs; it is not an index in the sense of the jk appended to θ in Eq. (4)]. For clarification, say the 308 values of β_{jk} occupy a 14×22 matrix $\mathbf{B}$, indexed by row j and column k , that is, AOGCMs in the rows, regions in the columns. Then $\theta_{(j)}$ is a vector with 14 elements, Σ is a 14×14 matrix, and each column of $\mathbf{B}$, consisting of the 14 β_j for region k , represents a draw from the distribution $\text{MVN}(\theta_{(j)}, \Sigma)$. A model structure of this kind, in which low-level parameters are presumed to be members of a population that is in turn described by a set of *hyperparameters*, is referred to as multilevel, or hierarchical. Note that the θ_{jk} and τ_{jk}^2 in Eq. (4) are not intended to represent hyperparameters; in the estimation process these means and variances are assigned fixed priors. Models A and B thus have no hierarchical structure.

Consideration of Eqs. (1), (2), (4), and (5) indicates that a comparison of models A and B tests the utility of

FIG. 4. Correlation matrices for (top) NAS and (bottom) EAF for (left) annual mean, (center) DJF, and (right) JJA, for 1902–98.

modeling the data as having regionally differentiated error variance. Similarly, a comparison between B and C assesses the utility of adopting the hierarchical model, with its structured covariance representation.

We note that the β_{jk} in each region are not strictly weights, since they are not required to have positive values that sum to unity. On the one hand, this means that a linear combination of AOGCM simulations using the β_{jk} is not constrained to lie within the envelope of those simulations. However, this also enables the linear model to account for scale bias in the simulations. Probability models utilizing true weights can certainly be devised, and represent a potentially viable alternative to those presented herein. However, a choice between the two types of model structure cannot be made on the basis of fit alone, since such a choice rests on unverifiable beliefs about the joint future behavior of the observations and AOGCMs. Further discussion of this matter is offered in section 6.

4. Probability model comparison

Model parameters were estimated using Markov chain Monte Carlo (MCMC) sampling, as implemented in the software package “Bugs” (Spiegelhalter et al. 1996). Specifics are provided in the appendix.

a. Assessing model fit

Preliminary comparison among models is made using Bayesian deviance statistics. These are computed from the sampling distributions of model parameters (see appendix) and are labeled, $\bar{D}$, $\hat{D}$, DIC, and pD (Spiegelhalter et al. 2002; Gelman et al. 2003); $\bar{D}$ and $\hat{D}$ are related measures of lack of fit, expressed as a negative factor multiplied by log-likelihoods of the data, given the model and its parameters. For $\bar{D}$, the log-likelihood is an average taken over values returned by the MCMC sampling, while for $\hat{D}$ it is a point estimate computed on the parameter means. Here $pD(=\bar{D} - \hat{D})$ is a measure of the effective number of free parameters in the model, and $DIC(=\bar{D} + pD)$, the deviance information criterion, provides an assessment of model predictive skill. The DIC functions like other “information criteria” (Kuha 2004), in that a penalty, based on the number of model parameters, is assessed. The smaller the value of the DIC, the better the estimated predictive capacity of the model.

Table 2 shows deviance statistics for the three candidate models, as well as a “simple” model (SM), consisting of the unweighted mean of the underlying AOGCM simulations. (The simple model is discussed more fully in section 4c, where it plays a role in cross-validation.)

Significance of the differences among the values shown in Table 2 was examined by repeating the MCMC estimation 3 times for each model, in each case starting from different initial values. These tests indicate a high degree of stability, with values of all deviance statistics remaining invariant to within 0.1, the precision with which they are returned by the sampling routine in Bugs.

Several patterns are evident in Table 2. First, the DIC decreases consistently in going from model A to B and then to C, the reduction being greatest for the first of these transitions. Thus, modeling interregional variance produces the larger improvement in model fit, while the inclusion of covariance structure yields an additional, but lesser, gain. Second, deviance is largest for DJF, followed by JJA, and then the annual mean. This may be a consequence of the unequal distribution of land between Northern and Southern Hemispheres, so that DJF, in essence, represents winter. The conclusion would be that interannual fluctuations in winter land surface temperatures are noisier than their summer counterparts. However, the greater DIC might also signal poorer AOGCM representation of DJF, compared with the other seasons. Third, pD consistently increases in going from model A to model B, but *decreases* sharply from B to C, even though the number of nominal parameters increases at each stage of model refinement. This reduction in pD may be seen as a result of the partial pooling of information that occurs when the multilevel structure is introduced. If all regions behaved identically, for example, then the large array of regional parameters would be redundant, since the parent distribution alone would suffice to describe the entire population. On the other hand, if there were no common behavior among regions, all the regional parameters would be “effective” in the model. The reduction in pD in going from model B to C is thus an indication that at least some common structure, of the form described by Eq. (5), must exist among regions.

The practical significance of the difference in deviance statistics between models A and B can be seen in Fig. 5, which shows observations and fitted model distributions for a high- and a low-latitude region (ALA and SEA, respectively; cf. Fig. 3), for the annual mean. In the case of model A (top row), error variance may assume only a single value, the σ^2 of Eq. (1); the corresponding standard deviation is estimated as 0.44°C . For region ALA (Fig. 5a) this would appear to be an underestimate, while for SEA (Fig. 5b), the opposite is true. This inflexibility characterizes neither B nor C (middle and bottom rows, respectively), and explains the relatively high DIC of model A. For comparison,

TABLE 2. Deviance statistics for models A, B, and C, and the “Simple” model SM.

Model	$\bar{D}$	$\hat{D}$	DIC	pD
		Annual		
A	2532	2217	2847	315.3
B	1768	1429	2108	339.4
C	1735	1505	1964	229.8
SM	2076	2037	2115	39.1
		DJF		
A	6268	5956	6581	312.6
B	4900	4562	5237	337.5
C	4841	4623	5058	217.5
SM	5055	5029	5081	26.3
		JJA		
A	2728	2413	3043	314.9
B	2241	1903	2580	338.9
C	2211	1976	2447	235.7
SM	2622	2594	2650	27.8

standard deviations for model B are 0.83° and 0.19°C for ALA and SEA, respectively, and for model C, 0.81° and 0.18°C , respectively.

A difference in fit between models B and C is more difficult to discern by inspection of Fig. 5, so we turn to the properties of the β_{jk} . Figure 6a indicates that values are shifted to the right, while dispersion over regions and AOGCMs is reduced, in going from B to C, while Fig. 6b shows an increase in the precision with which the β_{jk} are estimated by model C, with considerably less mass in regions of high uncertainty. The result will be more precisely estimated projections for this model.

The model estimate of Σ is shown in correlation form in Fig. 7, for the annual mean. (Corresponding matrices for DJF and JJA are similar in character but with somewhat lower off-axis values.) This figure, which represents the population covariance structure [see Eq. (5)], shows no strong patterns. The values are also low, suggesting that there is little common structure in inter-model covariance. However it is the abstraction, in the form of $\theta_{(j)}$ and Σ , of whatever common structure exists among the β_{jk} that is responsible for the reduction in the effective number of free parameters, and the corresponding reduction in DIC, for model C (Table 2). This reduction, as well as the improvement in the statistics of the β_{jk} (Fig. 6), suggests that the weak structure shown in Fig. 7 belies a significant improvement in model characteristics in moving from B to C.

b. A note on AOGCM skill

The parent vector $\theta_{(j)}$ describes the population distribution of multivariate means of the β_{jk} . For

FIG. 5. Observations and fitted values for the three candidate models for ALA and SEA for the annual mean. Observations are represented by the heavy black line; fitted values are shown as the mean (red line), 25th and 75th quantiles (green lines), and 5th and 95th quantiles (blue lines). (top) Model A, (center) model B, and (bottom) model C.

AOGCM p , θ_p will be close to the mean-over-regions of its assigned coefficient, i.e., $\theta_p \sim (1/K) \sum_{k=1}^K \beta_{pk}$, where $K = 22$ is the number of regions. Element θ_p thus represents the globally averaged contribution of AOGCM p to regional model fit. It is not an unbiased measure of individual model skill, however, since the values of $\theta_{(j)}$ are conditional on the presence of all the other covariates in the model. The number of ensemble members associated with a given AOGCM (Table 1) also influences its efficacy as a predictor. With these caveats, standardized values of the elements of θ_j are shown in Fig. 8. Many, but not all, of the θ_j are small, relative to their standard errors, and there are no significant negative values, indicating relative independence of predictors at this level. There is also some consistency across seasons, although the degree is somewhat lower for DJF. Such consistency as does exist indicates a degree of stability in relative AOGCM performance.

Consideration of the population distribution of coefficients raises the possibility of fitting a reduced probability model, perhaps retaining only those predictors represented by the statistically significant elements of $\theta_{(j)}$. This would amount to the application of a strictly global metric in the selection of predictors, but would not necessarily result in the choice of a best predictor subset for all regions, for which a measure requiring some balance between global and regional skill would seem preferable. In fact, such a compromise is implicit in model C, since the β_{jk} in any given region are drawn from the single population distribution, but are also conditioned locally.

c. Cross-validation

While the DIC may be viewed as a measure of predictive skill, it has been computed here on data having considerable variance at high frequencies. The projec-

FIG. 6. Distributions over regions and AOGCMs of (a) the means and (b) the variances of the 308 β_{jk} , for models B and C. Values shown are for the mean annual temperature.

tions, on the other hand, are expressed as temporal averages. It thus seems sensible to consider, in addition to the DIC, a verification metric based on time-mean statistics. To this end, a classical cross-validation measure was utilized, in which nine periods, each of decade length, were withheld in succession (from all regions at once), and the three probability models fitted in turn to the remaining data. The withheld intervals were then hindcast and the resulting values compared, in the mean squared error sense, with observations. High-frequency variability was suppressed by computing error values, not on individual years within the withheld data decades, but instead on the decadal means. Thus, for the twentieth century each region provides nine values, beginning with the 1909–18 decade and ending with 1989–98, and there are a total of 198 values contributing to the MSE for the 22 regions. Seasons are examined separately, while SM serves as the “null model” with which the contending probability models are compared.

Results appear in Table 3, where at least two patterns are evident. First, there is steady improvement (i.e., MSE ratios increase) in going from model A to B and then to C, for all seasons. However, the larger difference now occurs in going from B to C, rather than from A to B (cf. Table 2). This can be seen as a direct result of the decadal averaging, which attenuates the regional differences in temporal variance by whose representation models A and B are distinguished. Second, the greatest improvement over SM occurs in the case of the annual mean, followed by JJA and DJF, for all three models. The pattern in this case is similar to that ex-

FIG. 7. Posterior distribution for Σ , for the annual mean.

hibited by the DIC (Table 2). If the MSE ratios are treated as F statistics, models A and B would be essentially indistinguishable, B and C would differ with p values ranging from 0.18 to 0.29, and A and C with p values of 0.13 to 0.23, depending on season. Thus, the three models are not well discriminated by this metric.

Deviance statistics were also computed for SM (Table 2). In terms of DIC, SM appears to consistently outperform model A, and for DJF, model B as well. Much of the improvement derives from reduction in pD (SM has many fewer actual parameters than any of the candidate models), although $\bar{D}$ for SM still lies somewhere between that of models A and B, for all seasons.

Regional variability in MSE ratio, in this case for model C, is shown in Fig. 9. (Plots for models A and B exhibit very similar patterns, with values tending to be slightly smaller, as might be expected from the values in Table 3.) Model C consistently outperforms SM, with ratios exceeding unity in all but two cases. Variability in performance advantage is greater for DJF and JJA than for the annual mean, with about half the regional values falling below the 0.05 significance level in each of the outlying seasons. This is still considerably better than would be expected by chance alone.

To summarize, in the context of an MSE measure based on decadal means, all models outperform SM, while deviance statistics place SM somewhere between models A and B. The difference in ranking is evidently a consequence of the averaging used in computing MSE, and is thus very likely a result of the omission, in model A, of any representation of regional differences in error variance. As was demonstrated earlier (Figs. 5a and 5b), this omission leads to suboptimal fit in many

FIG. 8. Elements of θ_0 for (a) the annual mean, (b) DJF, and (c) JJA. Values are standardized. The x-axis labels correspond to the AOGCMs listed in Table 1 (some names are abbreviated). Standard deviations do not differ greatly among elements; plots of the unstandardized values therefore appear similar (but not identical). For reference, largest unstandardized values (not shown) and corresponding AOGCMs are annual mean, 0.16 (GISS); DJF, 0.11 (MRI); JJA, 0.18 (CCSM3).

regions. The weak discrimination among models exhibited by the MSE metric suggests that cross-validated decadal means are similarly close to observations from model to model, and that other properties should be considered as well in model selection.

The two metrics considered are consistent, however, in that they indicate a progression in goodness of fit in going from A to B to C, the last of these providing the greatest advantage over SM. Further, the improvement of coefficient properties in going from B to C (Fig. 6) indicates that projections made using the latter model will have smaller associated uncertainties. From a conceptual point of view, the incorporation of both global and regional information through the use of a hierarchical structure, as well as the modeling of covariance, would seem intuitively sensible choices. Taken together, these considerations point to C as the model of choice among the three candidates, and this will be the model utilized in the generation of projections. Further discussion of potential model structures is provided in section 6.

5. Regional temperature projections

a. General characteristics

Using the coefficient PDFs produced by model C, projections were generated for all regions for scenarios A2 and B1 (Figs. 10, 11, and 12). These projections reflect the underlying differences in anthropogenic forcing described in section 2, with A2 showing greater warming in every region. All seasonal variants exhibit enhanced high-latitude warming, this being more extreme for scenario A2. In light of past climate change simulations (Houghton et al. 2001), these results are not surprising.

b. Distributional attributes

Projected temperature changes derive from distributions generated by the MCMC estimation, each initially comprising 5000 samples for each year. Final distributions were obtained by averaging each of the 5000 sequences over the 2079–98 interval. Time averaging in this manner narrows the distributions only slightly, and is consistent with that applied to the individual AOGCM simulations with which the distributions are compared. Temperature changes are referenced in each case to 1979–98 means; in the case of model C this is a weighted mean computed using the β_{jk} .

Figure 13 shows a number of representative distributions, as well as corresponding values for the underlying AOGCM simulations (SM would correspond to the unweighted mean of these simulations). Figures 13a–c contrast two regions, AMZ and NEU, for which model C exhibits differing degrees of skill compared with SM (Fig. 9): for AMZ, MSE ratios are consistently large across seasons, while for NEU they are small. Figure 13a indicates that warming for AMZ in the annual mean, as projected by model C, is only about half that of SM, with model C's distribution centered near the AOGCM showing the least warming for this region. By contrast, for NEU (Fig.

TABLE 3. MSE ratios computed on independent data (10-yr means) over all regions. Values shown are the ratios $\text{MSE}_{\text{SM}}/\text{MSE}_X$, where X represents the model being compared with the “Simple” model SM.

Season	Model A	Model B	Model C
Annual	4.26	4.31	4.73
DJF	2.43	2.51	2.72
JJA	2.91	2.99	3.41

FIG. 9. Regional MSE ratios (MSE_{SM}/MSE_C), computed on nine values per region, one for each of the nine withheld decades of data. Ratios for (a) annual mean, (b) DJF, and (c) JJA. Solid line is plotted at the nominal 0.05 significance level; dotted line is plotted at the value 1.

13b) the median projected warming is not far from that of SM.

From the differences in central tendency in Figs. 13a and 13b, one might be tempted to conclude that regions where model C has little advantage over SM are those where the two models *agree*. However, Fig. 13c, for NEU in JJA, shows that this is not necessarily the case, since agreement is weaker here, but performance advantage for model C remains low. This situation reflects the fact that a similar MSE does not necessarily imply similarity in fit, but only a similar degree of closeness to the observations, in the cross-validation.

The larger uncertainty surrounding the projections for NEU, relative to the spread in individual AOGCM means, reflects greater uncertainty in the estimates of the β_{jk} , and ultimately, differences in both the signal-to-noise ratio of the underlying data and the ability of the probability model to account for variance in the nonrandom component of the record. The deviance statistics (Table 2) are considerably higher for DJF than for either JJA or the annual mean; this is reflected in the spread of mid- and high-latitude projections for Northern Hemisphere regions, as shown in Figs. 13d–f. A comparison of Figs. 13d and 13e (see x-axis range) shows that uncertainty is substantially greater for the DJF season in this high northern latitude region; comparison of Figs. 13c and 13f shows that a similar differential applies in region NEU in going from JJA to DJF. A comparison of Figs. 13a and 13d, meanwhile, shows that tighter distributions characterize the low-latitude regions.

Uncertainty in projected temperature change also increases with the degree of warming, that is, distance from the data distribution on which the model is built. This is to be expected, since not only do the individual AOGCMs

tend to drift apart with the increase in forcing, owing to their differing sensitivities, but uncertainties in coefficient values translate into larger and larger errors as they are multiplied by larger and larger departures in the underlying predictors. This can be seen in the reduced spread of projected distributions for the B1 scenario (Fig. 13g versus Fig. 13d and Fig. 13h versus Fig. 13b).

In Fig. 13i, which shows the EAF annual mean for B1, the projection is displaced toward lower values relative to SM, similar to what is seen in Figs. 13a and 13c. This pattern is characteristic of a number of low-latitude regions, including AMZ (but not CAM), EAF, WAF, SAS, and SEA, and can be explained by the relationship between observations and AOGCM simulations during the twentieth century. Figure 14 shows time series for two regions, one (EAF) for which projected temperatures are lower than those of SM, and one for which they are approximately aligned (ALA; cf. Figs. 13g and 13i). For region EAF (Fig. 14a), the simulations are almost all cooler than observations during the early part of the century, but similar or even a little warmer near the end. For ALA, on the other hand, the long-term trends are more alike. The model reproduces these characteristics in projecting an overall trend for EAF that is less than that of the underlying simulations, and one that more closely matches them in ALA. The modeling exercise thus involves an underlying *stationarity* assumption, with respect to the relationship between simulated and observed temperatures.

6. Discussion

The probability model selected is the most complex of the three candidates. The question then naturally arises of whether a model of *sufficient* complexity has

FIG. 10. Projected temperature changes for 2079–98, relative to 1979–98, for (a) scenario B1 and (b) scenario A2, for the annual mean. Shown in each region are the median temperature change and below it, in smaller type, the associated uncertainty range. The two values representing the latter are the 5th and 95th percentiles of the temperature change probability distribution.

DJF - SRES B1

DJF - SRES A2

FIG. 11. Same as in Fig. 10, but projected temperature change for DJF.

FIG. 12. Same as in Fig. 10, but projected temperature change for JJA.

FIG. 13. Distributions of projected temperature change (solid lines) and corresponding values for the individual AOGCMs (filled circles along the x axis), for selected regions. Top two rows derive from SRES scenario A2, and bottom row from B1.

been utilized, to which a qualified “yes” can be offered in response. A number of model structures of greater complexity were in fact explored, but such models did not produce appreciable reductions in DIC, and often the modeled hyperparameters had little or no statistical significance. Although we do not claim that this exer-

cise was exhaustive, it was ultimately decided that for the purposes of the present work, which retains an exploratory character, model C represented a reasonable stopping point.

The choice between coefficients and true weights (for which all values $w_i > 0$, and $\sum_i w_i = 1$) was alluded to

FIG. 14. Observations (heavy solid lines) and simulations for EAF and ALA for the annual mean.

in section 3. Such a choice will depend in part on the preferences of the researcher, who may have a predisposition, for example, against seeing projected temperature changes wander away from the AOGCM “consensus,” even though some justification may exist for the use of coefficients. This situation is one for which the Bayesian methodology is well suited, since in theory both structures could be incorporated into a hybrid model, in which the degree of prior belief in weights, as opposed to coefficients, was encoded. This would doubtless add additional layers of complexity, but might be deemed acceptable in view of improvements in either fit or perceived structural suitability.

Uncertainty in the temperature projections is comparable to that among AOGCMs, as can be seen from Fig. 13. So, although one may argue that the projected mean temperature changes represent an improvement over SM, the same cannot be said for precision in estimation. However, it should be remembered that the scenarios themselves embody considerable uncertainty regarding the evolution of human society on the planet. There are many conceivable pathways along which energy production, land-use change, the biosphere, emissions, and ultimately, atmospheric concentrations of radiatively active trace gases might evolve, and uncertainties here lead to equally large changes in projected temperatures (cf., e.g., Figs. 13d and 13g).

7. Summary

Regional temperature projections for the twenty-first century were generated by fitting a hierarchical linear Bayesian model to an observational dataset, using as predictors a multimodel ensemble consisting of 14 AOGCMs. In effect, the linear model is used to calibrate the ensemble against the observational dataset; the calibrated ensemble is then used to generate the projections, using simulations based on the A2 and B1 emissions scenarios from the IPCC SRES.

Probability model configurations of varying degrees of complexity were entertained. It was found that a large improvement in fit resulted from allowing modeled error variance to differ from region to region, while a smaller but still significant advance was achieved with the introduction of a multilevel structure, in which regional groups of coefficients are modeled as deriving from a population distribution with structured covariance. Final model selection involved both Bayesian deviance statistics and a cross-validated mean squared error measure, as well as direct examination of coefficient properties and certain conceptual considerations.

Temperature projections made with the Bayesian

model are not always centered on those made with a naïve model (SM) consisting of the unweighted mean of the 14 underlying AOGCM simulations. Displacements, typically toward smaller temperature increases in low-latitude regions, were seen to be due to differences between simulated and observed long-term behavior of temperature during the twentieth century, that is, the training period. A stationarity assumption underlying the modeling strategy employed was thus brought to light.

Uncertainty in the projections is comparable to the spread among the underlying simulations, and thus cannot be said to represent an improvement in precision over SM. However, such uncertainty is comparable to that attributable to differences between the two scenarios utilized in making these projections (which themselves do not represent the most extreme cases considered in the SRES). Thus, while reduction of projection error remains a desirable goal, this must be considered in the context of other contributing sources of uncertainty.

Acknowledgments. The authors are grateful to Tony Barnston, Simon Mason, and Andy Robertson, all of whom provided helpful criticism and discussion, and to two reviewers whose comments measurably improved the manuscript. The authors also acknowledge the international modeling groups that have provided their data for analysis: the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving the model data, the JSC/CLIVAR Working Group on Coupled Modelling (WGCM) and their Coupled Model Intercomparison Project (CMIP) and Climate Simulation Panel for organizing the model data analysis activity, and the IPCC WG1 TSU for technical support. The IPCC Data Archive at Lawrence Livermore National Laboratory is supported by the Office of Science, U.S. Department of Energy. The authors are supported by a cooperative agreement from the National Oceanic and Atmospheric Administration (NOAA) NA07GP0213. In addition, this research was funded by NSF and DOE as a Climate Model Evaluation Project (CMEP), Grant NSF ATM04-29299, under the U.S. CLIVAR Program (<http://www.usclivar.org/index.html>).

APPENDIX

Model and Estimation Details

It is common practice to represent Bayesian probability models in the form of directed acyclic graphs (DAGs), networks in which the random variables, including observations and model parameters, as well as

FIG. A1. DAG for model C. Variables are as described in section 3, with the exception that the *precision*, τ_k , is utilized here instead of the variance ($\tau_k = 1/\sigma_k^2$). Similarly, $\mathbf{T}$, the precision matrix, is utilized in place of Σ ; Y_{ik} is enclosed in a rectangular box to designate its status as observational. All other nodes with the exception of X_{ijk} , the AOGCM simulations, are stochastic, and represent model parameters. The large rectangles, known as “plates,” denote multiplicity, and enclose the nodes to which they apply. Text notations in the corners of the plates identify the corresponding indices; arrows denote logical dependence.

predictors, are represented as nodes, connected by arrows that designate logical dependence (Lauritzen and Spiegelhalter 1988). This dependence may be either stochastic or deterministic, corresponding to the similarity and equals signs, respectively, that were utilized in the model description in section 3. A DAG for model C is provided as Fig. A1.

One group of model elements not described in detail in section 3 (and omitted from the DAG as well) are the *priors* utilized in the model. In the Bayesian formalism, estimation proceeds on the basis of the conditional identity

$$P(B|A) \propto P(A|B)P(B), \quad (\text{A1})$$

in which A represents the observations, or data, while B represents the model, or hypothesis. The term $P(B)$ on the rhs is the probability of the model (including its parameters) in the absence of any observational evidence. It thus represents *prior* belief or understanding, and is so named; $P(A|B)$ is the *likelihood* of the observations, given the model, and embodies information

carried in the data. The term on the lhs, the probability of the model, but now conditional on the observations, is the *posterior* (distribution of B), the resultant of the modeling exercise. Estimation using such a model requires that priors for the various parameters be specified; these may have an appreciable effect on the posterior, depending on the model, the observations and the prior itself (Gelman et al. 2003).

For model C, the priors specified are for the most part “noninformative”; that is, they are diffuse, or broad, distributions. With such specifications, the posterior derives principally from the likelihood term. Priors for model C are specified as follows:

$$\beta_{0k} \sim N(\mu, \beta_0, \tau, \beta_0), \quad (\text{A2})$$

$$\mu, \beta_0 \sim N(0., 0.0001), \quad (\text{A3})$$

$$\tau, \beta_0 \sim \text{Gamma}(0.001, 0.001), \quad (\text{A4})$$

$$\mathbf{T} \sim W(\mathbf{R}[\cdot, \cdot], p), \quad (\text{A5})$$

$$\mathbf{R} = \begin{cases} 1/n_j, & i = j \\ 0, & i \neq j \end{cases}. \quad (\text{A6})$$

We adopt here a slightly different notation than that of section 3 in order to conform to typical Bayesian practice. In particular, distributions are specified in terms of precision, τ , rather than variance ($\tau = 1/\sigma^2$); the same is true for the precision matrix, $\mathbf{T}$ ($\propto \Sigma^{-1}$). Thus, $\mu.\beta_0$ is given a prior distribution with precision 10^{-4} , corresponding to a variance of 10^4 . Here β_{0k} is modeled as multilevel, as is β_{jk} , but in practice this makes little difference, since both the observations and AOGCM simulations are expressed as anomalies, and the β_{0k} , as well as $\mu.\beta_0$, are very close to zero. The prior precision $\tau.\beta_0$ is specified as $\text{Gamma}(10^{-3}, 10^{-3})$, with a variance of $10^{-3}/10^{-6} = 10^3$.

Here $\mathbf{T}$, the precision “parent,” is modeled as Wishart, a matrix distribution that corresponds to the multivariate normal structure assumed for $\theta_{(j)}$. The Wishart distribution must itself be given a prior, in the form of a scale matrix $\mathbf{R}$ having degrees of freedom p . Here, $\mathbf{R}$ is specified as diagonal, with $p = 14$, the rank of $\mathbf{R}$. The diagonal elements of $\mathbf{R}$, which function as prior order-of-magnitude estimates of the variances of the θ_j , are set equal to $1/n_j$, where n_j is the number of ensemble members associated with AOGCM j . This scaling represents an attempt to account for the differing number of ensemble members provided by each AOGCM, the weights on the diagonal of $\mathbf{R}$ expressing a prior expectation that AOGCMs having larger ensemble sizes will exhibit lower error variances. The unit value in the numerator was chosen after some experimentation, based on examination of the posterior covariance matrix (Fig. 7). If the value is made too large [say $\mathbf{R}(i, i) = 10/n_j$], the posterior matrix becomes almost featureless, indicating that the prior structure is effectively being ignored. On the other hand, if too small a value is used [say $\mathbf{R}(i, i) = 0.1/n_j$], values in the posterior covariance tend to become extreme, suggesting overfitting. An intermediate value of unity was therefore chosen. Appearance of the fitted series (Fig. 5) changes little as the diagonal scaling of $\mathbf{R}$ is manipulated, although β_{jk} values do shift somewhat. This suggests that projections are not overly sensitive to this scaling, as long as values remain within a reasonable range.

As noted, parameter estimation is carried out via MCMC sampling. This is not essential to the Bayesian methodology, but rather a computational strategy that permits estimation of complex distributions that might be difficult or even impossible to address analytically (Spiegelhalter et al. 1996; Gilks et al. 1996). In essence, MCMC amounts to drawing samples from the posterior distribution without having to compute it directly. For the results shown, a single chain was run for 500 cycles as “burn-in,” followed by 5000 cycles through the

model parameters. The burn-in allows the chain to “forget” its initial state; once this has occurred, it effectively generates samples from the posterior distribution. All distributions shown are computed on the 5000 samples returned by the MCMC process. Various convergence diagnostics, including tests for stationarity of the sampling distribution over the sampling period, indicate that the burn-in was of adequate length, and that the sampling was sufficiently extended for the final distributions to provide reasonably good parameter estimates.

REFERENCES

- Allen, M. R., and S. F. B. Tett, 1999: Checking for model consistency in optimal fingerprinting. *Climate Dyn.*, **15**, 419–434.
- , P. A. S. John, F. B. Mitchell, R. Schnur, and T. L. Delworth, 2000: Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, **407**, 617–620.
- Barnston, A. G., Y. He, and D. A. Unger, 2000: A forecast product that maximizes utility for state-of-the-art seasonal climate prediction. *Bull. Amer. Meteor. Soc.*, **81**, 1271–1279.
- Boucher, O., and M. Pham, 2002: History of sulfate aerosol radiative forcings. *Geophys. Res. Lett.*, **29**, 1308, doi:10.1029/2001GL014048.
- Collins, W. D., and Coauthors, 2006: The community climate system model: CCSM3. *J. Climate*, **19**, 2122–2143.
- Craig, P. S., M. Goldstein, J. C. Rougier, and A. H. Seheult, 2001: Bayesian forecasting for complex systems using computer simulators. *J. Amer. Stat. Assoc.*, **96**, 717–729.
- Delworth, T. L., and Coauthors, 2006: GFDL’s CM2 global coupled climate models. Part I: Formulation and simulation characteristics. *J. Climate*, **19**, 643–674.
- Diansky, N. A., and E. M. Volodin, 2002: Simulation of present-day climate with a coupled atmosphere–ocean general circulation model. *Izv. Atmos. Oceanic Phys.*, **38**, 732–747.
- Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, Eds., 2003: *Bayesian Data Analysis*. 2d ed. Chapman and Hall/CRC, 696 pp.
- Gilks, W., S. Richardson, and D. Spiegelhalter, Eds., 1996: *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 512 pp.
- Giorgi, F., and R. Francisco, 2000: Evaluating uncertainties in the prediction of regional climate change. *Geophys. Res. Lett.*, **27**, 1295–1298.
- Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211.
- Goddard, L., A. Barnston, and S. Mason, 2003: Evaluation of the IRI’s “net assessment” seasonal climate forecasts: 1997–2001. *Bull. Amer. Meteor. Soc.*, **84**, 1761–1781.
- Gordon, C., 2000: The simulation of sst, sea ice extents and ocean heat transports in a version of the Hadley Centre coupled model without flux adjustments. *Climate Dyn.*, **16**, 147–168.
- Gordon, H., and Coauthors, 2002: The CSIRO Mk3 climate system model. CSIRO Atmospheric Research Tech. Paper 60, 130 pp.

- Hasumi, H., and S. Emory, 2004: K-1 coupled model (MIROC) description. Center for Climate System Research Tech. Rep. 1, University of Tokyo, 34 pp. [Available online at <http://www.ccsr.u-tokyo.ac.jp/kyosei/hasumi/MIROC/tech-repo.pdf>.]
- Hegerl, G., G. Meehl, C. Covey, M. Latif, B. McAvaney, and R. Stouffer, cited 2003: 20C3M: CMIP collecting data from 20th century coupled model simulations. [Available online at <http://www.clivar.org/publications/exchanges/ex26/supplement/>.]
- Houghton, J. T., Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden, X. Dai, K. Maskell, and C. A. Johnson, Eds., 2001: *Climate Change 2001: The Scientific Basis*. Cambridge University Press, 881 pp.
- Jungclauss, J. H., and Coauthors, 2006: Ocean circulation and tropical variability in the coupled model ECHAM5/MPI-OM. *J. Climate*, **19**, 3952–3972.
- Katz, R. W., and A. H. Murphy, Eds., 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 238 pp.
- Kuha, J., 2004: AIC and BIC: Comparisons of assumptions and performance. *Soc. Meth. Res.*, **33**, 188–229.
- Lauritzen, S., and D. Spiegelhalter, 1988: Local computations with probabilities on graphical structures and their application to expert systems. *J. Roy. Stat. Soc. B*, **50**, 157–224.
- Nakićenović, N., and Coauthors, 2000: IPCC special report on emissions scenarios. Intergovernmental Panel on Climate Change Tech. Rep., 570 pp.
- New, M., M. Hulme, and P. Jones, 1999: Representing twentieth-century space–time climate variability. Part I: Development of a 1961–90 mean monthly terrestrial climatology. *J. Climate*, **12**, 829–856.
- , —, and —, 2000: Representing twentieth-century space–time climate variability. Part II: Development of 1901–96 monthly grids of terrestrial surface climate. *J. Climate*, **13**, 2217–2238.
- Palmer, T., Č. Branković, and D. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quart. J. Roy. Meteor. Soc.*, **126**, 2013–2033.
- Russell, G. L., J. R. Miller, and D. Rind, 1995: A coupled atmosphere–ocean model for transient climate change studies. *Atmos.–Ocean*, **33**, 683–730.
- Schmidt, G. A., and Coauthors, 2006: Present-day atmospheric simulations using GISS ModelE: Comparison to in situ, satellite, and reanalysis data. *J. Climate*, **19**, 153–192.
- Spiegelhalter, D., A. Thomas, N. Best, and W. Gilks, 1996: BUGS 0.5: Bayesian inference using Gibbs sampling manual (version ii). Medical Research Council Biostatistics Unit, 59 pp. [Available online at <http://www.mrc-bsu.cam.ac.uk/bugs/documentation/Download/manual0.5.pdf>.]
- , N. Best, B. Carlin, and A. van der Linde, 2002: Bayesian measures of model complexity and fit (with discussion). *J. Roy. Stat. Soc. B*, **64**, 583–639.
- Washington, W. M., and Coauthors, 2000: Parallel climate model (PCM) control and transient simulations. *Climate Dyn.*, **16**, 755–774.
- Yukimoto, S., and Coauthors, 2001: The new Meteorological Research Institute coupled GCM (MRI-CGCM2): Model climate and variability. *Pap. Meteor. Geophys.*, **51**, 47–88.