Bayesian Modeling of Uncertainty in Ensembles of Climate Models

Richard L. Smith*, Claudia Tebaldi†, Doug Nychka‡and Linda O. Mearns‡

August 12, 2008

Abstract

Projections of future climate change caused by increasing greenhouse gases depend critically on numerical climate models coupling the ocean and atmosphere (GCMs). However, different models differ substantially in their projections, which raises the question of how the different models can best be combined into a probability distribution of future climate change. For this analysis, we have collected both current and future projected mean temperatures produced by nine climate models for 22 regions of the earth. We also have estimates of current mean temperatures from actual observations, together with standard errors, that can be used to calibrate the climate models. We propose a Bayesian analysis that allows us to combine the different climate models into a posterior distribution of future temperature increase, for each of the 22 regions, while allowing for the different climate models to have different variances. Two versions of the analysis are proposed, a univariate analysis in which each region is analyzed separately, and a multivariate analysis in which the 22 regions are combined into an overall statistical model. A cross-validation approach is proposed to confirm the reasonableness of our Bayesian predictive distributions. The results of this analysis allow for a quantification of the uncertainty of climate model projections as a Bayesian posterior distribution, substantially extending previous approaches to uncertainty in climate models.

Keywords: Bayesian modeling of uncertainty, Climate change, Cross-validation, Prediction

*Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27599-3260
†Climate Central, 895 Emerson Street, Palo Alto, CA 94301
‡National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307

1 Introduction

Global climate models (GCMs) are complex computer programs that simulate the physics and chemistry of the atmosphere and oceans to obtain projections of temperature and other meteorological variables under various assumptions about the composition of the atmosphere and other influences such as variations in solar energy. They have successfully explained past variations in the earth's temperature and are used to simulate future variations in climate under various assumptions about emissions of greenhouse gases and other man-made substances (such as sulfate aerosols) that are known to influence climate. These future simulations, known as projections, are an important tool in tracing the influence of human activity on the Earth System. An excellent reference source for climate models and climate science more generally are the Assessment Reports of the Intergovernmental Panel on Climate Change, the most recent of which is IPCC (2007).

All the major climate models project increases in both global and regional mean temperatures throughout the twenty-first century, under differing assumptions (or scenarios) about future trends in population growth and economic and technological development, among other factors. The consistency of these results across different climate models has greatly strengthened the belief that many climate scientists have in global warming, but there are also considerable variations among climate models, which raises the question of how different climate models can best be combined to derive climate projections with appropriate measures of uncertainty.

In this paper, we explore these issues for several datasets compiled by Giorgi and Mearns (2002), which consist of current (1961–1990) and future (2071–2100) projections of the mean temperature in 22 regions for nine climate models, which form the suite of models assessed in the IPCC 2001 report (Giorgi *et al.* (2001)). The 22 regions are depicted in Figure 1 and the nine climate models are summarized in Table 1; more details about the model calculations were given by Giorgi and Mearns (2002). Also shown — in the last column of Table 1 — is the "climate sensitivity" parameter, which is defined to be the mean warming of the whole earth, in equilibrium conditions, associated with a doubling of atmospheric carbon dioxide compared with pre-industrial conditions. As can be seen in Table 1, the nine models have quite different climate sensitivities, the lowest three being for the models MRI, CSM and PCM. As will be seen later, these models consistently produce the lowest projections for future warmings.

Also part of the data are estimates of the true temperature averages for the 22 regions for 1961–1990, based on observational data, with associated standard errors. The datasets are prepared for

two seasons of the year, DJF (December, January, February) and JJA (June, July, August) to allow some contrast between summer and winter conditions. The future projections are also prepared for two different scenarios of future emissions of greenhouse gases, the so-called SRES A2 and B2 scenarios. These scenarios, originally prepared as part of the IPCC Special Report on Emissions Scenarios (Nakićenović *et al.* (2000), IPCC (2001)), represent two (of many) possible projections of future emissions, with the A2 scenario representing faster growth and consequently higher emissions.

In this paper we address the issue of constructing a probability density function (pdf) for the mean temperature difference between the two time periods in each of the 22 regions. Our approach is Bayesian and takes account of the fact that different models have different variances that are *a priori* unknown. The approach is directly motivated by the Giorgi-Mearns (2002) "reliability ensemble average" (REA) which is reviewed in Section 3.

The method proposed here has two forms: the "univariate" approach treats each of the 22 regions as a separate variable while the "multivariate" approach treats them together, in particular by pooling information across regions in estimating the variances of the models. A version of the univariate approach has been presented previously (Tebaldi *et al.* 2005, 2004), but is extended here to allow verification of the model by cross-validation. The multivariate approach is developed for the first time in this paper.

Before going into details, we set some general context for this paper within the field of climate science research. In IPCC (2007), a report charged with assessing the state of the science on climate change, Chapter 11 is dedicated to regional projections (Christensen *et al.* (2007)). Unlike previous IPCC reports (e.g., Giorgi et al., 2001), which offered only a qualitative summary of inter-model agreement about regional mean projections, this one included discussion of two formal statistical assessments of uncertainty derived from multi-model ensembles for regional projections, one due to Tebaldi *et al.* (2005, 2004) which was a precursor to the present approach, the other due to Greene *et al.* (2006). Both approaches have been criticized for relying too much on the comparison between observed and modeled regional trends in the 20th century; for example the Tebaldi *et al.* (2005) approach sometimes produces unrealistically low estimates of uncertainty in future model projections. As shown at the end of Section 6, we now believe that to be an artifact of the method, that the new approaches presented in this paper overcome. In summary, the improved robustness of our methods and the inclusion of a cross-validation step should go a long way to resolve criticisms of earlier approaches, and will facilitate the step from methodological exercise to actual application in studies of impacts where probabilistic information is crucial to effective decision making.

The rest of the paper is organized as follows. Section 2 summarizes existing approaches to the assessment of uncertainty in climate models. Section 3 reviews the REA. Sections 4 and 5 present the details of our Bayesian approach in both its univariate and multivariate forms. Section 6 discusses the overall goodness of fit and presents a comparison between the univariate and multivariate approaches. Finally Section 7 summarizes our conclusions and suggestions for future work.

The Bayesian methods proposed in Sections 4 and 5 have been programmed in R (R Development Core Team (2008)), and are publicly available, along with the datasets, from the website `http://www.image.ucar.edu/~nychka/REA`.

2 Approaches to Uncertainty in Climate Change

Climate scientists recognize the need to take account of uncertainty in presenting projections of future climate. A report like IPCC (2007) must integrate many individual pieces of research into an overall assessment, and for this purpose they recommend assessing the likelihood of a future event using broad categories, e.g. "virtually certain" (>99% probability of occurrence), "very likely" (> 90%), "likely" (> 66%), and so on, but they emphasize that "likelihood may be based on a quantitative analysis or on an elicitation of expert views". However, individual papers within the climate science field have increasingly used a wide range of rigorous statistical approaches including both frequentist and Bayesian analyses. In this section, we summarize a few of the leasing developments.

Uncertainties in climate change projections are broadly of three types (Meehl *et al.* (2007)), (a) natural climate variability, (b) uncertainties in the responses to climate forcing factors, such as changes in atmospheric levels of greenhouse gases and sulfate aerosols, and (c) uncertainties in future emissions of greenhouse gases and other factors that could influence climate. The first two types of uncertainty are typically assessed in "detection and attribution" studies, which calibrate climate models based on their fit to existing observational data and which attempt to decompose observed changes into components associated with greenhouse gases, aerosols, solar fluctuations, and other known influences on the earth's climate, as well as internal variability, which is the inherently stochastic component of the climate system. The review paper by IDAG (2005) summarized research over several years on these topics. Further discussion is contained in chapter 9–11 of the 2007 IPCC report (Hegerl *et al.* (2007), Meehl *et al.* (2007), Christensen *et al.* (2007)). As an example of a specific paper using this approach, Allen *et al.* (2000) used several climate models

to estimate mean climate changes up to 2046, with confidence intervals that take into account both natural variability and the uncertainty in the regression coefficients. Their results showed reasonable agreement across models, but they did not attempt to combine the results of different models. See also Levine and Berliner (1999), Berliner *et al.* (2000), for a more rigorous statistical discussion of detection and attribution approaches.

Uncertainties in emissions were assessed by SRES (Nakićenović *et al.* (2000)), who developed numerous scenarios representing different assumptions about population growth and economic and technological developments. However, the SRES authors declined to assess probabilities associated with the different scenarios. Subsequent commentators such as Schneider (2001) and Webster (2003) have argued that a probabilistic assessment by experts, even if imperfect and controversial, would be invaluable in generating informed assessments of climate impacts. A full discussion and assessment of this controversy is in Parson *et al.* (2007). On the other hand, Stott and Kettleborough (2002) applied the same method as Allen *et al.* (2000) to four SRES scenarios up to 2050, and after taking uncertainties of the individual projections into account, found little discrepancy among the projections associated with different scenarios. They argued that this was because of the smaller divergence among emission scenarios in the first half of the century, and the time lag between changes in emissions and changes in climate, and one could expect greater discrepancies among scenarios after 2050.

Wigley and Raper (2001) derived probabilistic projections of future climate change by running a simplified climate model under different combinations of model parameters (including climate sensitivity) and emissions scenarios. They used subjectively determined prior distributions for the physical parameters needed to run the GCM, and (controversially) assumed that all SRES scenarios were equally likely. Their approach was Bayesian in the sense of using subjectively determined probabilities, but not in the more formal sense of calculating posterior distributions based on observational data. Other authors including Forest and co-authors (2000, 2001, 2002) and Webster *et al.* (2003) have taken an approach closer to formal Bayesian methods, combining detection and attribution methods with a subjectively-determined prior on model parameters to derive a posterior distribution for future climate changes. Forest *et al.* (2002) implicitly criticized the use of subjective-judgment priors by Wigley and Raper (2001), highlighting the need for "an objective means of quantifying uncertainty in the long-term response". More recent material is summarized in Meehl *et al.* (2007), Christensen *et al.* (2007) and Tebaldi and Knutti (2007).

Santer *et al.* (1990) appear to have been the first authors to suggest explicitly that formal

statistical methods, such as confidence intervals and hypothesis tests, should be applied to combine the results of different climate models, and their ideas have been applied in subsequent reports on climate change such as Wigley (1999), and a publicly available computer package (Hulme *et al.* (2000)) for generating and interpreting climate models. Räisänen (1997) proposed a test of significance which, in each grid cell, computes the deviation from the global mean climate change, separately for each model, and then performs a t test to determine whether the deviation in that grid box is significantly different from 0, assuming that the model responses are independently distributed about the true mean. Räisänen and Palmer (2001) used model ensembles to generate probabilistic projections that could be assessed according to various decision-theoretic criteria. However, none of these authors allowed explicitly for the different variances of different climate models.

Apart from the preceding literature on climate change, the field of ensemble-based weather forecasting has advanced extensively in recent years, and new statistical approaches have been developed in that context, especially in a series of papers by Gneiting, Raftery and co-workers (Gel *et al.* 2004, Gneiting and Raftery 2005, Gneiting *et al.* 2005, Raftery *et al.* 2005, Berrocal *et al.* 2007, Sloughter *et al.* 2007, Wilson *et al.* 2007). The central concept of their methodology is laid out in Gneiting *et al* 2005, and uses the central formula of Bayesian model averaging,

$$p(y) \quad = \quad \sum_{k=1}^{K} p(y \mid M_k) p(M_k \mid \mathbf{y}^T)$$

where, in this context, y denotes the quantity to be forecast, $M_1, ..., M_K$ denote K models — here identified with K forecasts from an ensemble — and $p(M_k \mid \mathbf{y}^T)$ denotes the posterior probability of model k given training data $\mathbf{y}^T$ (i.e. past values of the weather field). Thus their prediction equation becomes

$$p(y \mid f_1, ..., f_K) \quad = \quad \sum_{k=1}^{K} w_k g_k(y \mid f_k)$$

where w_k is the posterior probability that forecast k is best given the training data. For the densities g_k, they assume normality with a mean $a_k + b_k f_k$ and a variance σ^2, where a_k and b_k are interpreted as bias correction terms from the kth model. Based on a spatial-temporal field of past observations for each model, they are able to estimate the parameters a_k, b_k and σ^2, the weights w_k, and hence complete the probabilistic forecast based on the ensemble.

This approach is conceptually different from ours, but there are some similarities. Both approaches use weighted averages of the ensemble members, but in the Gneiting-Raftery approach

these are interpreted as posterior probabilities while we, in equation (1) and subsequently, use weights λ_i which are interpreted as inverse variances of the individual models. Also, because we typically do not have multiple replications to use as training data, we are unable to incorporate a bias correction analogous to their $a_k + b_k f_k$ formula. However in the multivariate version of our model (section 5), we are able to incorporate a bias term for each model and also for each variable being predicted.

We should also point out the recent paper by Gneiting *et al.* (2007), which has addressed more systematically the assessment of probability forecasts. Although we are not aware of this paper when the present research was being done, there are in fact a number of common elements. Gneiting *et al.* discuss the well-known use of probability integral transforms (PITs) as a means of calibrating forecasters (see e.g. Dawid 1984, Seillier-Moiseiwitsch and Dawid 1993), which we use extensively in our subsequent development of cross-validation statistics (Sections 4.1, 5.1), though they also point out disadvantages to the PIT approach. In particular, it seems clear that simply requiring forecasters to be well-calibrated, in the sense that the PIT of the forecasts closely approximates a uniform distribution, is not a sufficient requirement for a good forecasting system, and some additional requirement of "sharpness" is needed. In fact, this requirement in some form has been recognized for a long time, e.g. Murphy (1972), DeGroot and Fienberg (1983). Its principal application in the present paper is in Section 6, where we directly use the width of predictive intervals calculated under the univariate and multivariate approaches to compare the two approaches.

Summarizing, there is growing acceptance of the need for statistical, and even Bayesian, approaches to the assessment of uncertainty in climate change, but methods that rely too heavily on subjective probability assessments, especially with respect to emissions scenarios, are viewed with suspicion. Moreover, Bayesian methods have been developed for turning ensembles into probabilistic forecasts in the context of numerical weather prediction, for which there is typically far more extensive data than we are able to use in our approach. The present paper advances these methodologies by proposing a Bayesian approach to the combination of projections from different climate models, but as far as possible, using uninformative prior distributions. We do not make any attempt to place a prior distribution on emissions scenarios, instead focussing on two of the SRES scenarios to compare the results.

3 The Reliability Ensemble Average

In this section we review the approach of Giorgi and Mearns (2002), which also serves to introduce notation for our Bayesian development in Sections 4 and 5.

Suppose there are M climate models, X_j is a projection of some current climate variable generated by model j, and Y_j a projection of some future climate variable generated by model j. We also have an observation X_0 of the true current climate, with some associated measure of variability ϵ. In a typical application, X_j is the mean temperature or precipitation simulated by the jth GCM in a particular region for the period 1961–1990, X_0 is the corresponding value calculated from the observational climate record with standard error ϵ, and Y_j is either the corresponding variable calculated for 2071–2100 or the difference between the 2071–2100 and 1961–1990 values. (Giorgi and Mearns typically took the latter as their variable of interest; we generally prefer to define Y_j directly as the predicted 2071–2100 mean, but later will interpret our results in terms of projected climate change, which is analogous with Giorgi and Mearns.) We view X_j and Y_j as random variables in the sense that, as the index j ranges over all possible models, we observe a range of both current and future projections and can make inferences about their distributions.

If we assume $\mathrm{Var}(Y_j) = \sigma^2/\lambda_j$, with σ^2 unknown but λ_j (for the moment) assumed known, then a suitable ensemble estimate of the future climate state is

$$\tilde{Y} = \frac{\sum_{j=1}^{M} \lambda_j Y_j}{\sum_{j=1}^{M} \lambda_j}. \tag{1}$$

Routine calculations show that an unbiased estimator of the variance of $\tilde{Y}$ is

$$\tilde{\delta}_Y^2 = \frac{\sum_{j=1}^{M} \lambda_j (Y_j - \tilde{Y})^2}{(M-1)\sum_{i=1}^{M} \lambda_j}, \tag{2}$$

so $\tilde{\delta}_Y$ may be interpreted as a standard error.

Giorgi and Mearns called λ_j the "reliability" of model i and formula (1) the "reliability ensemble estimator" or REA. Their presentation of (2) omitted the factor $M-1$ in the denominator.

To estimate the reliabilities, Giorgi and Mearns proposed

$$\lambda_i = (\lambda_{B,i}^m \lambda_{D,i}^n)^{1/mn} \tag{3}$$

where

$$\lambda_{B,i} = \min\left(1, \frac{\epsilon}{|X_i - X_0|}\right), \quad \lambda_{D,i} = \min\left(1, \frac{\epsilon}{|Y_i - \tilde{Y}|}\right), \tag{4}$$

where $|X_i - X_0|$ is the "bias" of model i, $|Y_i - \tilde{Y}|$ the "convergence" of model i, and the parameters m and n control the relative importance given to these two quantities (Giorgi and Mearns suggested

$m = n = 1$). The justification for introducing ϵ is, loosely, to avoid giving a model too much credit when, purely by chance, either the bias or the convergence is much smaller than the natural variability ϵ.

Giorgi and Mearns proposed an iterative procedure to find a set of weights λ_i satisfying the relations (1)–(4). In most cases, stability is achieved with a few iterations.

Although this procedure appears to lack formal statistical justification, Nychka and Tebaldi (2003) showed that it can be interpreted as a robust estimator, choosing $\tilde{Y}$ to minimize a sum of the form $\sum C_i |Y_i - \tilde{Y}|^{1-1/n}$ for suitable weights C_i. In the case $n = 1$, this reduces to a weighted median.

By using the data directly to assess uncertainty, but avoiding the assumption that all climate models have the same variability, the Giorgi-Mearns approach potentially improves on previous attempts to assess uncertainty in climate models. Nevertheless it has several seemingly *ad hoc* features, in particular its treatment of bias and convergence. From a Bayesian viewpoint, we would prefer to express uncertainty via a posterior density than simply a point estimate and standard error.

4 Univariate Model

The first version of our analysis is univariate in the sense that it treats each of the model output variables X_i and Y_i as univariate random variables. In practice we will apply this model separately to each of the 22 regions. In Section 5, this will be extended to a multivariate analysis, in which we combine the 22 regions into a single overall model.

A version of the univariate analysis has been presented previously (Tebaldi *et al.* 2005), but there are several modifications in the present approach, which we discuss after outlining the basic method.

As in Section 3, we assume X_0 is the current observed mean temperature, X_j is the current modeled mean temperature for a particular region for model $j = 1, ..., M$, and Y_j is the future modeled mean temperature for model $j = 1, ..., M$. In the following, $N[\mu, \sigma^2]$ will denote the normal distribution with mean μ and variance σ^2; $U[a, b]$ the uniform distribution on the interval $[a, b]$; and $G[a, b]$ the gamma distribution whose density is proportional to $x^{a-1}e^{-bx}$. With these definitions we assume

$$X_0 \sim N[\mu, \lambda_0^{-1}], \quad (\lambda_0 \text{ known}) \tag{5}$$

$$X_j \sim N[\mu, \lambda_j^{-1}], \tag{6}$$

$$Y_j \mid X_j \sim N[\nu + \beta(X_j - \mu), (\theta\lambda_j)^{-1}], \tag{7}$$

where parameters $\mu,\ \nu,\ \beta,\ \theta,\ \lambda_j$ have prior distributions

$$\mu, \nu, \beta \sim U(-\infty, \infty), \tag{8}$$

$$\theta \sim G[a, b], \tag{9}$$

$$\lambda_1, ..., \lambda_M \sim G[a_\lambda, b_\lambda], \tag{10}$$

$$a_\lambda, b_\lambda \sim G[a^*, b^*]. \tag{11}$$

Here the hyperparameters $a,\ b,\ a^*,\ b^*$ are chosen so that each of $\theta,\ a_\lambda,\ b_\lambda$ has a proper but diffuse prior. In practice we set $a = b = a^* = b^* = 0.01$.

We discuss briefly the rationale for these assumptions. The λ_js represent reliabilities for the M models and have the same interpretation as in Section 3. The parameter θ (typically, between 0 and 1) represents a differential between the reliabilities of current and future model projections. We could not estimate a statistical model in which the reliabilities of future observations were completely arbitrary, but we can estimate a posterior distribution for θ under the assumption that the variances are as in (6) and (7). We choose λ_0 so that $\lambda_0^{-1/2}$ is the estimated standard deviation of $X_0 - \mu$. Since this estimate is based on plentiful observational data, there is no loss of model accuracy by treating it as known.

The regression parameter β is a convenient way of introducing correlation between X_i and Y_i. If $\beta = 0$, this is equivalent to assuming X_i and Y_i are independent. Under this assumption, the weighted average (1) is directly justifiable as an estimator for ν, assuming λ_j are known. Alternatively, as already noted in Section 3, Giorgi and Mearns actually defined Y_j to be the *difference* between future and present climate under model j. That is equivalent to assuming $\beta = 1$ in the present notation. We take the view that the correlation between X_i and Y_i is best treated as unknown and arbitrary, which is equivalently represented by (7) with arbitrary unknown β. Tebaldi *et al.* (2005) discussed further the role of this parameter and made comparisons with cases when β was fixed at 0 or 1.

The parameters μ and ν are means (respectively, for X_j and for $Y_j - \beta(X_j - \mu)$) that are assumed to be the same for all models. Our approach therefore makes no explicit allowance for model bias — in other words, we are assuming that any deviations between model projections and the corresponding true climate values can be characterized by the variance terms in (6) and

(7). However in the absence of either (informative) prior knowledge about the performance of each model, or replications of either X_j or Y_j for a given model, such bias terms would not be identifiable. In Section 5, we do include bias terms as part of our development of a multivariate model.

The main difference between the present model and the one given in Tebaldi *et al.* (2005) is in the prior distribution for $\lambda_1, ..., \lambda_M$. Here we assume they are $G(a_\lambda, b_\lambda)$ with a_λ, b_λ having a hyperprior distribution of their own, whereas Tebaldi *et al.* (2005) simply assumed $a_\lambda = b_\lambda = 0.01$. Exactly why this apparently small change to the prior distribution makes a critical difference in the model interpretation will be explained in Section 4.1.

Under the model (5)–(11), the joint density of θ, μ, ν, β, a_λ, b_λ, X_0 and $(\lambda_j,\ X_j,\ Y_i,\ j = 1, ..., M)$ is proportional to

$$\theta^{a+M/2-1} e^{-b\theta} e^{-\frac{1}{2}\lambda_0 (X_0-\mu)^2} a_\lambda^{a^*-1} e^{-b^* a_\lambda} b_\lambda^{a^*-1} e^{-b^* b_\lambda} \cdot \\ \cdot \prod_{j=1}^{M} \left[\frac{b_\lambda^{a_\lambda} \lambda_j^{a_\lambda} e^{-b_\lambda \lambda_j}}{\Gamma(a_\lambda)} \cdot e^{-\frac{1}{2}\lambda_j (X_j-\mu)^2 - \frac{1}{2}\theta\lambda_j \{Y_j - \nu - \beta(X_j-\mu)\}^2} \right]. \tag{12}$$

Define

$$\tilde{\mu} = \frac{\lambda_0 X_0 + \sum \lambda_j X_j - \theta\beta \sum \lambda_j (Y_j - \nu - \beta X_j)}{\lambda_0 + \sum \lambda_j + \theta\beta^2 \sum \lambda_j}, \tag{13}$$

$$\tilde{\nu} = \frac{\sum \lambda_j \{Y_j - \beta(X_j - \mu)\}}{\sum \lambda_j}, \tag{14}$$

$$\tilde{\beta} = \frac{\sum \lambda_j (Y_j - \nu)(X_j - \mu)}{\sum \lambda_j (X_j - \mu)^2}. \tag{15}$$

In a Monte Carlo sampling scheme, all the parameters in (12), with the exception of a_λ and b_λ, may be updated through Gibbs sampling steps, as follows:

$$\mu \mid \text{rest} \sim N\left[\tilde{\mu}, \frac{1}{\lambda_0 + \sum \lambda_j + \theta\beta^2 \sum \lambda_j}\right], \tag{16}$$

$$\nu \mid \text{rest} \sim N\left[\tilde{\nu}, \frac{1}{\theta \sum \lambda_j}\right], \tag{17}$$

$$\beta \mid \text{rest} \sim N\left[\tilde{\beta}, \frac{1}{\theta \sum \lambda_j (X_j - \mu)^2}\right], \tag{18}$$

$$\lambda_j \mid \text{rest} \sim G\left[a + 1, b + \frac{1}{2}(X_j - \mu)^2 + \frac{\theta}{2}\{Y_j - \nu - \beta(X_j - \mu)\}^2\right], \tag{19}$$

$$\theta \mid \text{rest} \sim G\left[a + \frac{M}{2}, b + \frac{1}{2}\sum \lambda_j \{Y_j - \nu - \beta(X_j - \mu)\}^2\right]. \tag{20}$$

For the parameters a_λ, b_λ, the following Metropolis updating step is proposed instead:

1. Generate U_1, U_2, U_3, independent uniform on $(0, 1)$.

2. Define new trial values $a'_\lambda = a_\lambda e^{\delta(U_1-1/2)}$, $b'_\lambda = b_\lambda e^{\delta(U_2-1/2)}$. The value of δ (step length) is arbitrary but $\delta = 1$ seems to work well in practice, and is therefore used here.

3. Compute

$$\begin{aligned}\ell_1 &= Ma_\lambda \log b_\lambda - M\log\Gamma(a_\lambda) + a_\lambda \sum \log\lambda_j - b_\lambda \sum \lambda_j + a^* \log(a_\lambda b_\lambda) - b^*(a_\lambda + b_\lambda),\\ \ell_2 &= Ma'_\lambda \log b'_\lambda - M\log\Gamma(a'_\lambda) + a'_\lambda \sum \log\lambda_j - b'_\lambda \sum \lambda_j + a^* \log(a'_\lambda b'_\lambda) - b^*(a'_\lambda + b'_\lambda).\end{aligned}$$

   This computes the log likelihood for both (a_λ, b_λ) and (a'_λ, b'_λ), allowing for the prior density and including a Jacobian term to allow for the fact that the updating is on a logarithmic scale.

4. If $\log U_3 < \ell_2 - \ell_1$ then we accept the new (a_λ, b_λ), otherwise keep the present values for the current iteration, as in a standard Metropolis accept-reject step.

This process is iterated many times to generate a random sample from the joint posterior distribution. In the case where a_λ, b_λ are treated as fixed, the Metropolis steps for these two parameters are omitted and in this case the method is a pure Gibbs sampler, as in Tebaldi *et al.* (2005). For the version presented here, an R program (REA.GM.r) to perform the sampling is available for download from `http://www.image.ucar.edu/~nychka/REA`.

4.1 Cross-Validation in the Univariate Model

A difficulty with this kind of Bayesian analysis is how to validate the statistical assumptions. Of course, direct validation based on future climate is impossible. However the following alternative viewpoint is feasible: if we think of the given climate models as a random sample from the universe of possible climate models, we can ask ourselves how well the statistical approach would do in predicting the response of a new climate model. This leads to a cross-validation approach. In effect, this makes an assumption of exchangability among the available climate models.

In more detail, suppose someone gave us a new climate model for which the projected current and future temperature means were $X^\dagger$ and $Y^\dagger$. Conditionally on the hyperparameters μ, ν, β, θ, a_λ and b_λ, the distribution of $Y^\dagger - X^\dagger$ is derived from (i) $\lambda^\dagger \sim G[a_\lambda, b_\lambda]$, (ii) $Y^\dagger - X^\dagger \mid \lambda^\dagger \sim N[\nu - \mu, \{(\beta-1)^2 + \theta^{-1}\}/\lambda^\dagger]$. By mixing this conditional predictive distribution over the posterior distribution of $(\mu, \nu, \beta, \theta, a_\lambda, b_\lambda)$, we obtain a full posterior predictive distribution.

This suggests a cross-validatory approach in which each climate model j in turn is dropped from the analysis, a predictive distribution for $Y^\dagger - X^\dagger$ calculated from the remaining eight climate

models, and this is applied to the observed value of the dropped model $Y_j - X_j$. In practice we apply a probability integral transformation to convert this value to a standard uniform U_j, and then assess the goodness of fit using standard tests such as Kolmogorov-Smirnov. Details are as follows:

1. For each $j \in \{1, ..., M\}$, rerun the REA.GM procedure without model j.

2. The hyperparameter values in the nth row of the REA.GM output, say $a_\lambda^{(n)}, b_\lambda^{(n)}, \nu^{(n)}, \mu^{(n)}, \beta^{(n)}, \theta^{(n)}$, correspond to one draw from the posterior distribution. Therefore, draw a random $\lambda_{j,n} \sim G[a_\lambda^{(n)}, b_\lambda^{(n)}]$ and calculate

$$U_j^{(n)} = \Phi\left\{ \frac{Y_j - X_j - \nu^{(n)} + \mu^{(n)}}{\sqrt{\{(\beta_x^{(n)} - 1)^2 + {\theta^{(n)}}^{-1}\}(\lambda_{j,n})^{-1}}} \right\}.$$

3. Let U_j be the mean value of $U_j^{(n)}$ over all n draws from the posterior distribution. This is therefore an estimate of the predictive distribution function, evaluated at the true $Y_j - X_j$. If the model is working correctly, U_j should have a uniform distribution on $(0, 1)$.

4. Recompute steps 1–3 for each region, so we have a set of test statistics $U_{ij},\ i = 1, ..., R,\ j = 1, ..., M$.

5. Plot the U_{ij}'s to look for systematic discrepancies, and apply standard tests of fit, such as Kolmogorov-Smirnov, for a formal test that the predictive distribution is consistent with the data.

This procedure is encoded in the `REA.CV` function, also available from
`http://www.image.ucar.edu/~nychka/REA`.

Note that it is essential, for this procedure, that the values of a_λ and b_λ define a realistic posterior distribution for the λ_j's. This is a critical difference from the earlier approach of Tebaldi *et al.* (2005), where a_λ and b_λ were simply defined in such a way as to produce an uninformative prior distribution (the paper actually took $a_\lambda = b_\lambda = 0.01$). Within that approach, no cross-validation appears to be possible.

4.2 Example

Some results from applying the univariate analysis just described are summarized in Figures 2 and 3. In Figure 2 six regions (Southern AUstralia, the AMaZons, Central AMerica, GReenLand, Western

AFrica and South ASia) are chosen to exemplify the quality of the posterior distributions of future temperature change derived through our analysis. For reference, the 9 individual GCM projections are shown as circles along the x-axis. The black curves pertain to the posterior distributions estimated by the region-specific analysis of the univariate model presented above, for winter (DJF) projected temperature change, under the A2 scenario. (The red curves, to be discussed later, are based on the corresponding multivariate analysis.) As far as can be assessed, the PDFs are smooth envelopes of most of the individual projections. However, in some of the regions (SAU, GRL and SAS in this example), individual GCM values may behave as outliers, as a result of the statistical assumptions by which the estimate of each GCM's reliability parameter, λ_j, bears a direct relation to that GCM's degree of agreement with the rest of the ensemble's projections.

Figure 3 is a graphical representation of the cross-validation exercise, that we perform for all four combinations of seasons and scenarios in our dataset. Each of the image plots represents a matrix of U_{ij} values, for the 22 regions (along the horizontal dimension) and the 9 models (along the vertical dimension). In general, the models with low climate sensitivity in Table 1 also produce low values of the test statistic (blue end of the color scale) which those with high sensitivity produce high values of the test statistic (red end of the color scale), but this effect is not universal, e.g. MRI which has the lowest climate sensitivity is not at the blue end of the scale. It is obvious that the U_{ij} statistics are not independent from region to region, but the intent of the cross-validation statistics is that within each row, the values of $U_{i1}, ..., U_{i9}$ are approximately independent draws from a uniform distribution on (0,1). In Section 6, we consider formal goodness-of-fit statistics.

5 Multivariate Model

A disadvantage of the approach so far is that each of the 22 regions is treated as an entirely separate data analysis. The data available for any one region consist solely of the nine climate model projections X_j and Y_j, plus a single observational value X_0, and the analysis is open to the objection that it is trying to produce rather complicated inferences based on a very limited set of data. In this section, we propose an extension of the method in which the data for all climate models and regions are treated within a single statistical model. The hope is that we will be able to estimate some of the variance parameters more precisely and hence not have such diffuse predictive distributions.

We assume we have current and future climate model projections, X_{ij} and Y_{ij}, which, in addition

to representing different models $j = 1, ..., M$, also represent different variables $i = 1, ..., R$. In the subsequent discussion, X_{ij} and Y_{ij} represent the current and future projection of model j for the temperature average over region i. We also assume that X_{i0} is the current observed mean temperature in region i, $i = 1, ..., R$, which is an estimate of the true current mean temperature with known standard deviation $\lambda_{0i}^{-1/2}$.

Note that in the current application, the index i is associated with the ith region — the model is "multivariate" in the sense that the projections of temperature over different regions are treated as a multivariate response. We could also consider using the same approach for an analysis that is multivariate in the sense of jointly modeling different meteorological variables, but that has not been attempted in the present application.

The assumed model in this case is of the following form:

$$X_{i0} \sim N[\mu_0 + \zeta_i, \lambda_{0i}^{-1}], \tag{21}$$

$$X_{ij} \sim N[\mu_0 + \zeta_i + \alpha_j, (\eta_{ij}\phi_i\lambda_j)^{-1}], \tag{22}$$

$$Y_{ij} \mid X_{ij} \sim N[\nu_0 + \zeta_i' + \alpha_j' + \beta_i(X_{ij} - \mu_0 - \zeta_i - \alpha_j), (\eta_{ij}\theta_i\lambda_j)^{-1}]. \tag{23}$$

With the exception of λ_{0i} (which is again treated as a known constant) these random variables depend on unknown parameters whose joint prior densities are assumed to be as follows:

$$\mu_0, \nu_0, \zeta_i, \zeta_i', \beta_i, \beta_0 \sim U(-\infty, \infty), \tag{24}$$

$$\theta_i, \phi_i, \psi_0, \theta_0, c, a_\lambda, b_\lambda \sim G[a, b], \tag{25}$$

$$\lambda_j \mid a_\lambda, b_\lambda \sim G[a_\lambda, b_\lambda], \tag{26}$$

$$\eta_{ij} \mid c \sim G[c, c], \tag{27}$$

$$\alpha_j \mid \psi_0 \sim N[0, \psi_0^{-1}], \tag{28}$$

$$\alpha_j' \mid \alpha_j, \beta_0, \theta_0, \psi_0 \sim N[\beta_0\alpha_j, (\theta_0\psi_0)^{-1}], \tag{29}$$

all mutually independent unless explicitly indicated otherwise.

The following discussion is intended to illuminate our reasons for making these specific assumptions. The general philosophy behind our approach is to make the statistical model as general as possible, subject to being identifiable and estimable, as judged by our ability to construct predictive distributions. Or to turn Einstein's famous quote on its head, we believe the model should be "as complicated as possible, but not more so."

Regarding the mean terms in (21)–(23), we interpret μ_0 and ν_0 as global mean values, and the parameters ζ_i and ζ_i' as region-specific departures from the global mean for the present and future

time periods. The parameters α_j and α'_j represent global biases associated with a particular climate model: we have already seen that the different climate models have different climate sensitivities and are therefore expected to differ systematically in some of their projections (models with large climate sensitivities tend to project more warming than those with small climate sensitivities). Note, however, the different structure for the prior distributions of (α_j, α'_j) as compared with those for (ζ_i, ζ'_i). In the case of ζ_i and ζ'_i, we take the view that the prior should be as uninformative as possible and therefore take a uniform prior density over $(-\infty, \infty)$. This also reflects the fact that the different regions are physically very different from each other and there is no reason to adopt a statistical model that assumes that the warming in a polar region such as Alaska is in any way correlated with the warming in equatorial regions. For the climate model parameters α_j and α'_j, however, we are adopting the same viewpoint as Section 4, whereby the different climate models in our survey are treated as a random sample from a supposedly infinite population of climate models, whose parameters are linked through hyperparameters $\psi_0, \beta_0, \theta_0$.

Another way of thinking about this distinction is in terms of the well-known statistical phenomenon of shrinkage. Our model shows a tendency to shrink the values of α_j and α'_j towards a common mean, which is natural if we think of these as samples from a population of climate models. However, for the region effects, there is a much less compelling reason to do any shrinkage. The models consistently project more warming for polar regions such as Alaska; we have every reason to believe this is a true physical effect (e.g. it's supported by current data on the melting of the polar ice caps), and there is no reason to shrink our projections towards a common mean. In preliminary studies, we have experimented extensively with variations on these assumptions; assuming a hyperprior for (ζ_i, ζ'_i) does indeed produce shrinkage (sometimes projecting polar warming of 2–3 K less than we get with a uniform prior) but we do not find this physically meaningful. On the other hand, the hyperprior assumption for α_j and α'_j makes it possible to construct a predictive distribution for a new climate model, which is the basis of our proposed cross-validation technique in Section 5.1.

Now let us turn to the variance assumptions in (21)–(23). Consider first the special case $\eta_{ij} \equiv 1$, which can also be achieved by letting $c \to \infty$ in (27). In this case, the variance of climate model j in region i factorizes as $(\phi_i \lambda_j)^{-1}$. We call this the factorization assumption. In contrast with the model of Section 4, where there were $22 \times 9 = 198$ separate variance parameters to estimate, in this model there are only $22 + 9 = 31$ such parameters. Therefore, if this assumption is correct, we should be able to estimate the individual variance parameters much more precisely, resulting in

tighter posterior distributions for the quantities we ultimately want to estimate.

However, there is a clear disadvantage to this assumption, in that it assumes the same pattern of variation among climate models holds across all regions. For example, if climate model A has twice the variance of climate model B in one region, it will have twice the variance in every region. In preliminary discussions of these analyses, climate modelers have repeatedly expressed skepticism that such a simple assumption is correct.

Therefore, we introduce η_{ij} as a multiplicative interaction term: the value of η_{ij} for a specific region $\times$ climate model combination reflects the extent to which the variance for that region $\times$ climate model differs from what would hold under the factorization assumption. We assume a common gamma prior for all the η_{ij}, and there is no loss of generality in assuming this prior distribution has mean 1, so we make the gamma shape and scale parameters a common value c. We can think of c as a tuning parameter; the limiting cases $c \to \infty$ and $c \to 0$ correspond, respectively, to the factorization model and to the case where the region $\times$ climate model variances are completely unconstrained, which is in effect the assumption of Section 4. Our hope is that by taking some intermediate value of c, we will be able to improve on Section 4 without making the unrealistic factorization assumption.

In preliminary analyses, we have experimented with different approaches to the parameter c, for example, simply fixing c to be some common-sense value (such as 0.1, 1 or 10) while finding reasonable consistency across analyses with different values of c. However, treating c as a hyper-parameter with its own prior distribution, given by (25), seems to be the most general and flexible approach. In the results to be reported later, we generally find the median posterior value of c to be greater than 10, confirming that the factorization assumption is not too far from reality (and at the same time, that the present model likely is an improvement on that of Section 4), but still allowing that there may be some region $\times$ climate model combinations where the variance is very different from the factorization assumption.

We have chosen to give more attention to the assumptions in the equations (21)—(29) than to the actual analysis, which is similar to Section 4. In particular, we use Gibbs sampling to update most of the unknown parameters but a Metropolis update for a_λ, b_λ and c. The method is available as an R program (REAMV.GM.r) from `http://www.image.ucar.edu/~nychka/REA`. Details of the updating steps are in the Appendix (Section 10).

5.1 Cross validation in the Multivariate Model

As with the univariate model, we can calculate cross-validation statistics by dropping one climate model at a time, constructing predictive distributions for the dropped climate model based on the other eight climate models, using this predictive distribution via a probability integral transformation to convert the actual data from the dropped climate model to a uniform distribution on $[0, 1]$, and then performing goodness of fit tests.

In this case, the variable that we use for cross-validation is $Y_{ij} - X_{ij}$, the projected increase in region i for model j. Note that

$$Y_{ij} - X_{ij} \mid \text{rest} \sim N\left[\nu_0 - \mu_0 + \zeta_i' - \zeta_i + \alpha_j' - \alpha_j, \frac{1}{\eta_{ij}\lambda_j}\left\{\frac{(\beta_i - 1)^2}{\phi_i} + \frac{1}{\theta_i}\right\}\right].$$

For the jth-model cross validation, we run the Gibbs/Metropolis simulation described in Section 10 for N iterations leaving out climate model j. For every set of parameters saved as the nth iteration, we generate corresponding values of $\lambda_j^{(n)}$, $\alpha_j^{(n)}$, $\alpha_j'^{(n)}$ and $\eta_{ij}^{(n)}$ as

$$\begin{aligned}
\lambda_j^{(n)} &\sim G\left[a_\lambda^{(n)}, b_\lambda^{(n)}\right] \\
\alpha_j^{(n)} &\sim N\left[0, \frac{1}{\psi_0^{(n)}}\right] \\
\alpha_j'^{(n)} &\sim N\left[\beta_0^{(n)}\alpha_j^{(n)}, \frac{1}{\psi_0^{(n)}\theta_0^{(n)}}\right], \\
\eta_{ij}^{(n)} &\sim G\left[c^{(n)}, c^{(n)}\right].
\end{aligned}$$

From these values we compute the statistic

$$U_{ij} = \frac{1}{N}\sum_{n=1}^{N} \Phi\left[\frac{Y_{ij} - X_{ij} - (\nu_0^{(n)} - \mu_0^{(n)}) - (\zeta_i'^{(n)} - \zeta_i^{(n)}) - (\alpha_j'^{(n)} - \alpha_j^{(n)})}{\sqrt{(\lambda_j^{(n)}\eta_{ij}^{(n)})^{-1}\left\{(\phi_i^{(n)})^{-1}(\beta_i^{(n)} - 1)^2 + (\theta_i^{(n)})^{-1}\right\}}}\right]. \tag{30}$$

As with the univariate analysis, we then perform various goodness of fit tests on the statistics U_{ij}. If the model is a good fit, the results within each row should be consistent with independent draws from the uniform distribution on $[0, 1]$.

5.2 Results

The solid red lines in Figure 2 are PDFs of posterior densities for DJF temperature change under scenario A2 derived through the multivariate model just described, for the six regions chosen as examples. As indicated by the six pairs of curves in Figure 2, the comparison with the univariate

model shows substantial agreement of the two posterior estimates. These results are representative of all 22 regions.

With regard to the other parameters of interest in the model, we show in Figures 4 and 5 the posterior distribution of the hyperparameters c, a_λ and b_λ. As previously discussed, the tuning parameter c reflects the degree of interaction among the variances in different climate models and regions. For all four datasets analyzed, the median values of c are larger than ten, suggesting a significant interaction effect (and highlighting the advantage of the multivariate model).

The convergence of the MCMC algorithm to its underlying stationary distribution was tested through standard diagnostics available in CODA, which is available as a downloadable package within R (R Development Core Team (2008)). All the individual components of the Markov chain for the univariate model pass the convergence tests. We show in Figures 6 and 7 the traces of three easily interpretable and relevant parameters (temperature change, future and current temperature) for two regions, ALA and NAS, representative of the entire set of 22 regions. The stationarity in mean and variance of the time series is evident by eye and confirmed by CODA.

For some of the individual model parameters in the multivariate approach, the traces of the sampled values show non-stationary behavior and significant auto- and cross-correlation. The high auto-correlation (within a single parameter chain) was addressed by running the MCMC simulation for a large number of iterations (125000), and saving only one out of every 100 samples, after discarding the first 25000. The non-stationary behavior and large cross-correlation across parameters is attributable to the structure of the statistical model, where some parameters are tightly coupled, but never affects the interpretable quantities of interest, which result from aggregating the individual parameters (e.g. $\Delta T_i \equiv \nu_0 - \mu_0 + \zeta_i' - \zeta_i$). For all quantities of interest, traces appear stationary and the diagnostic tests confirm it. In the right columns of Figures 6 and 7 we show traces of the quantities from the multivariate model corresponding to the parameters of the univariate model in the left-columns.

We have also run the cross-validation exercise for the multivariate model. Figure 8 is a graphical representation of the U-statistics values determined for the four sets of estimates (DJF and JJA under SRES A2, DJF and JJA under SRES B2). Here as in Figure 3 a good model fit would generate values across each of the 22 horizontal bands in every panel not significantly different from a random draw of nine variates from a uniform distribution on $(0, 1)$. The results again seem to show a pattern consistent with what would be expected from the climate sensitivities — in fact, this pattern is more consistent than the one in Figure 3 (for example, with model MRI, which has

the lowest climate sensitivity, the U_{ij} statistics are consistently small).

6 Comparisons of univariate and multivariate approaches

In this section, we make some direct comparisons of the univariate and multivariate approaches, focussing on three issues, (a) goodness of fit, (b) width of the posterior densities, (c) robustness.

Goodness of fit is assessed using the cross-validatory statistics discussed in Sections 4.1 and 5.1. If the statistical model is correct, these statistics U_{ij} should be consistent with a uniform distribution on $(0, 1)$. It is evident from Figures 3 and 8 that these are not independent from region to region (for example, a model with low climate sensitivity tends to produce low values of U_{ij} across all regions). However, the values of $U_{ij},\ j = 1, ..., 9$ should be approximately independent for each region i. This hypothesis is assessed using four common goodness of fit statistics: Kolmogorov-Smirnov (henceforth, K-S), Cramér-von Mises (C-vM), Anderson-Darling (A-D) and a correlation test (Cor) in which the test statistic is 1 minus the correlation coefficient of the ordered $\{U_{ij},\ j = 1, .., 9\}$ with the values 0.1, 0.2,...,0.9. The latter is analogous to the Shapiro-Wilk test often used with normally distributed data; we subtract the correlation coefficient from 1 so that (as with the other tests) small values correspond to a very close fit between the empirical and theoretical distribution functions.

For each of the four tests and each of the season/scenario combinations, Table 2 computes the number of regions (out of 22) on which the univariate model resulted in a smaller (better) test statistic than the multivariate model. Overall, this happened in about one-third of the possible cases, implying a clear though not overwhelming superiority for the multivariate model.

We can also perform formal goodness of fit tests by simulation. For each i, 50000 simulated independent samples of $U_{ij},\ j = 1, .., 9$, were drawn from the uniform distribution, and the same test statistics calculated. These were used to calculate empirical p-values. Table 3 shows the number of regions (out of 22) in which this procedure led to a rejection of the null hypothesis of uniformity, for each of the univariate and multivariate procedures, for all four goodness of fit tests, and for each season/scenario combination. A two-sided .05-level test was used. This analysis showed more rejections for the multivariate analysis than for the univariate analysis. However, with only one exception, all the rejections occurred in the *lower* tail of the test statistic, implying that the agreement between the empirical and theoretical distributions was better than would be obtained by random sampling.

This conclusion was unexpected, but we are inclined not to over-interpret it. The calculation of p-values assumes independence across different climate models for each region, but it is evident that such an assumption cannot be literally correct. (For example, the U_{ij} statistics for climate model j depend on parameter estimates computed from the other eight models.) Therefore, our p-values can only be regarded as approximate. We regard Table 3 as confirming the overall fit of our statistical model, in either its univariate or multivariate manifestations.

We next turn to the question of whether the multivariate approach leads to tighter predictive densities than the univariate approach — to the extent that "shrinkage" reduces the variance of posterior densities, we would expect this to be the case. An obvious tool for comparison is the inter-quartile range, defined as the difference between the 75th and 25th percentiles of the empirical predictive distribution obtained from the MCMC output. Analogously to the IQR but giving more emphasis to the tails, we also consider test statistics that we call I15R (difference between the 85th and 15th percentiles) and I5R (difference between the 95th and 5th percentiles). We prefer robust statistics such as these to moment-based measures of scale such as standard deviation because the latter are more likely to be influenced by a few outliers in the MCMC sample.

The results of this comparison are summarized in Table 4. In each cell, we compute the mean ratio of IQR, I15R or I5R in the predictive distribution obtained from the univariate method (numerator) that of the multivariate method (denominator). We also show (in parentheses) the number of regions (out of 22) that the multivariate method resulted in a smaller IQR, I15R or I5R than the univariate method. Consistently, the multivariate method performed better in more than half the regions and the average ratio of scale parameters was greater than 1, indicating the multivariate method should be preferred.

However, examination of results for individual regions (not tabulated) shows a less clear-cut picture. The ratios of IQR, I15R and I5R for individual regions show a wide variability, with many values less than 1, and in three regions (SAU, SAH, SEA) the univariate method always beats the multivariate method when assessed by IQR, I15R and I5R for both seasons and both emissions scenarios. In contrast, for nine regions (AMZ, WNA, CNA, ENA, NEU, EAS, CAS, TIB, NAS) the multivariate method is always better.

In summary, the overall comparison favors the multivariate method as producing tighter predictive distributions, but this result is not uniform over all regions, so the comparison is not completely clear-cut.

Finally, we compare the univariate and multivariate approaches from a robustness viewpoint,

focussing on one specific region (WAF in JJA/A2), though we believe the discussion applies generally in cases where (as here) several climate models produced nearly identical predictions in the original data. We artificially perturbed one of those models by adding $0.5, 1, 1.5, 2$ K to the value of its future projection, and performing the univariate and multivariate versions of the statistical analysis over both the original and the perturbed datasets (a total of 10 analyses). We also applied this procedure to the original version of the univariate analysis as proposed by Tebaldi *et al.* (2005). The results, in Figure 9, show that the Tebaldi *et al.* (2005) model displays a high degree of sensitivity of the posterior density to the perturbations. This sensitivity is much reduced for the present paper's univariate approach, and reduced still further for the multivariate approach. Based on that comparison, we believe that there are robustness advantages to either of the approaches of the present paper, compared with that of Tebaldi *et al.* (2005).

7 Summary and Conclusions

In this paper, we have presented two approaches (univariate and multivariate) to the calculation of posterior distributions for future climate change based on an ensemble of GCMs.

A feature of our approach is the use of cross-validation statistics to develop goodness-of-fit tests. This feature was missing from the approach of Tebaldi *et al.* (2005), and we view that as a significant advantage of the present method. Calculations of test statistics based on the cross-validations generally confirm that the univariate and multivariate approaches both lead to adequate fits, with the multivariate model showing a slightly better fit. Comparisons of the predictive distributions themselves, as assessed through the robust scale measures IQR, I15R and I5R, also show slight superiority of the multivariate model, though there is substantial variability from region to region and in some regions the univariate approach leads to tighter predictive distributions.

In Figure 10, we use a color-coded map to summarize the actual predictive distributions (in terms of the mean and several quantiles for the projected temperature change) for each combination of region, season and scenario, using our multivariate approach. As examples of the interpretation of these tables, consider the results for DJF under the A2 scenario. The largest projected increases in mean temperature change are for the three northernmost regions (Alaska, Greenland, North Asia) with posterior means of 7.0, 6.9 and 6.4 K respectively, compared with means in the range 2.9–5.0 K for the other 19 regions. We also calculate 95% posterior intervals for ALA, GRL and NAS are respectively (5.0,9.2), (5.6,8.3), (4.6,8.1). When compared with the corresponding intervals for the

other regions, these results seem to confirm rather decisively that these three regions will warm substantially more than the overall global average, a conclusion that cannot be drawn from the posterior means alone. These are the same regions as were identified for significant warmings in the previous studies by Giorgi *et al.* (2001) and Tebaldi *et al.* (2005), but the present study provides superior calculations of the posterior distribution and hence more precise summaries of uncertainty. The corresponding results for JJA do not appear to show nearly such a strong polar region effect, confirming that this phenomenon of strong polar warming is primarily a northern hemisphere winter phenomenon. The results for the B2 scenario are qualitatively similar, but generally show less warming over all regions, as would be expected from the fact that B2 represents a smaller increase in greenhouse gas emissions compared with A2. Note, however, that for all region/season/scenario combinations, the 95% posterior interval excludes 0, implying a clear warming effect over all regions.

There are of course some limitations to what these procedures can achieve. Although the different climate modeling groups are independent in the sense that they consist of disjoint groups of people, each developing their own computer code, all the GCMs are based on similar physical assumptions and if there were systematic errors affecting future projections in all the GCMs, our procedures could not detect that. On the other hand, another argument sometimes raised by so-called climate skeptics is that disagreements among existing GCMs are sufficient reason to doubt the correctness of any of their conclusions. The methods presented in this paper provide some counter to that argument, because we have shown that by making reasonable statistical assumptions, we can calculate a posterior density that captures the variability among all the models, but that still results in posterior-predictive intervals that are narrow enough to draw meaningful conclusions about probabilities of future climate change.

Future work will apply these methods to a wider range of climate models, including models from the Fourth Assessment Report of IPCC (2007) and to regional climate models. There is a large archive of Fourth Assessment model output available through the Program for Climate Model Diagnosis and Intercomparison (http://www-pcmdi.llnl.gov) and results from these model runs will be presented in future papers. At this time, the website http://www.rcpm.ucar.edu provides regional analyses upon user's specification of latitude/longitude boundaries using the modified version of the univariate method that is described in Tebaldi *et al.* (2004), with large θ. These regional results are based on the latest suite of models/scenarios runs from the PCMDI archive. Work is in progress on implementing the method described in the current paper.

8 Acknowledgements

This research was partially supported by the NCAR Weather and Climate Impacts Assessment Science Program, which is funded by the National Science Foundation. Additional support was provided by the National Science Foundation (grants DMS-0084375 to Smith and DMS-0355474 to the Geophysical Statistics Project at NCAR) and NOAA (grant NA05OAR4310020 to Smith).

9 References

Allen, M.R., Stott, P.A., Mitchell, J.F.B., Schnur, R. and Delworth, T.L. (2000), Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **407**, 617–620.

Berliner, L.M., Levine, R.A and Shea, D.J. (2000), Bayesian climate change assessment. *Journal of Climate* **13**, 3805–3820.

Berrocal, V.J., Raftery, A.E. and Gneiting, T. (2007), Combining spatial statistical and ensemble information in probabilistic ensemble forecasting. *Monthly Weather Review* **135**, 1386–1402.

Christensen, J.H., Hewitson, B., Busuioc, A., Chen, A., Gao, X., Held, I., Jones, R., Kolli, R.K., Kwon, W.-T., Laprise, R., Rueda, V.M., Mearns, L., Menéndez, C.G., Räisänen, J., Rinke, A., Sarr, A. and Whetton, P. (2007), Regional Climate Projections. Chapter 11 of IPCC (2007), pp. 847–940.

Dawid, A.P. (1984), Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, **147**, 278–292.

DeGroot, M.H. and Fienberg, S.E. (1983), THe comparison and evaluation of forecasters. *The Statistician* **32**, 12–22.

Forest, C.E., Allen, M.R., Stone, P.H. and Sokolov, A.P. (2000), Constraining uncertainties in climate models using climate change detection techniques. *Geophysical Research Letters* **27**, 569–572.

Forest, C.E., Allen, M.R., Sokolov, A.P. and Stone, P.H. (2001), Constraining climate model properties using optimal fingerprint detection methods. *Climate Dynamics* **18**, 277–295.

Forest, C.E., Stone, P.H., Sokolov, A.P., Allen, M.R. and Webster, M.D. (2002), Quantifying uncertainties in climate system properties with the use of recent climate observations. *Science* **295**, 113–117.

Gel, Y., Raftery, A.E. and Gneiting, T. (2004), Calibrated probabilistic mesoscale weather

forecasting: The geostatistical output perturbation (GOP) method (with discussion and rejoinder). *Journal of the American Statistical Association* **99**, 575–590.

Giorgi, F., Hewitson, B., Christensen, J., Hulme, M., Von Storch, H., Whetton, P., Jones, R., Mearns, L. and Fu, C. (2001), Regional Climate Information — Evaluation and Projections. Chapter 10 of IPCC (2001), 583–638.

Giorgi, F. and Mearns, L.O. (2002), Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the "Reliability Ensemble Averaging" (REA) method. *J. Climate* **15**, 1141–1158.

Gneiting, T., Balabdaoui, F. and Raftery, A.E. (2007), Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society, Series B* **69**, 243–268.

Gneiting, T. and Raftery, A.E. (2005), Weather forecasting with ensemble methods. *Science* **310**, 248–249.

Gneiting, T., Raftery, A.E., Westfeld, Anton H. III and Goldman, T. (2005), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review* **133**, 1098–1118.

Greene, A.M., Goddard, L. and Lall, U. (2006), Probabilistic multimodel regional temperature change projections. *J. Climate* **19**, 4326–4346.

Hegerl, G.C., Zwiers, F.W., Braconnot, P., Gillett, N.P., Luo, Y, Marengo Orsini, J.A., Nicholls, N., Penner, J.E. and Stott, P.A. (2007), Understanding and Attributing Climate Change. Chapter 9 of IPCC (2007), 663–745.

Hulme, M., Wigley, T.M.L., Barrow, E.M., Saper, S.C.B., Centella, A., Smith, S. and Chipanshi, A.C. (2000), *Using a Climate Scenario Generator for Vulnerability and Adaptation Assessment: MAGICC and SCENGEN Version 2.4 Workbook.* Climatic Research Unit, Norwich, U.K., 52pp.

IDAG (2005), Detecting and attributing external influences on the climate system: A review of recent advances. The International Ad Hoc Detection and Attribution Group, *Journal of Climate* **18**, 1291–1314.

IPCC (2001), *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press, Cambridge and New York, 881pp.

IPCC (2007), *Climate Change 2007 - The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the IPCC.* Cambridge University Press. Available online (with other IPCC Assessment Reports) from http://www.ipcc.ch/ipccreports/assessments-

reports.htm.

Levine, R.A. and Berliner, L.M. (1999), Statistical principles for climate change studies. *Journal of Climate* **12**, 564–574.

Meehl, G.A., Stocker, T.F., Collins, W.D., Friedlingstein, P., Gaye, A.T., Gregory, J.M., Kitoh, A., Knutti, R., Murphy, J.M., Noda, A., Raper, S.C.B., Watterson, I.G., Weaver, A.J. and Zhao, Z.-C. (2007), Global Climate Projections. Chapter 10 of IPCC (2007), pp. 747–845.

Murphy, A.H. (1972), Scalar and vector partitions of the probability score: part 1. Two-state situation. *Journal of Applied Meteorology* **11**, 273–282.

Nakićenović, N., Alcamo, J., Davis, G., de Vries, B., Fenhann, J., Gaffin, S., Gregory, K., Grübler, A., Jung, T.Y., Kram, T., La Rovere, E.L., Michaelis, L., Mori, S., Morita, T., Pepper, W., Pitcher, H., Price, L., Raihi, K., Roehrl, A., Rogner, H.-H., Sankovski, A., Schlesinger, M., Shukla, P., Smith, S., Swart, R., van Rooijen, S., Victor, N. and Dadi, Z. (2000), *IPCC Special Report on Emissions Scenarios*, Cambridge University Press, Cambridge and New York, 599 pp.

Nychka, D. and Tebaldi, C. (2003), Comments on "Calculation of average, uncertainty range and reliability of regional climate changes from AOGCM simulations via the 'Reliability Ensemble Averaging' (REA) method". *Journal of Climate* **16**, 883–884.

Parson, E.A., Burkett, V.R., Fisher-Vanden, K., Keith, D.W., Mearns, L.O., Pitcher, H.M., Rosenzweig, C.E. and Webster, M.D. (2007). *Global Change Scenarios: Their Development and Use.* US Climate Change Science Program Synthesis and Assessment Product 2.1b. Department of Energy, Office of Biological and Environmental Research, Washington, DC., USA 106 pp.

R Development Core Team (2008), *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
`http://www.R-project.org`.

Raftery, A.E., Gneiting, T., Balabdaoui, F. and Polakowski, M. (2005), Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review* **133**, 1155–1174.

Räisänen, J. (1997), Objective comparison of patterns of CO_2 induced climate change in coupled GCM experiments. *Climate Dynamics* **13**, 197–211.

Räisänen, J. and Palmer, T.N. (2001), A probability and decision-model analysis of a multimodel ensemble of climate change simulations. *Journal of Climate* **14**, 3212–3226.

Santer, B.D., Wigley, T.M.L., Schlesinger, M.E. and Mitchell, J.F.B. (1990), *Developing Climate Scenarios from Equilibrium GCM Results*. Report No. 47, Max Planck Institut für Meteorologie, Hamburg, 29pp.

Schneider, S.H. (2001), What is 'dangerous' climate change? *Nature* **411**, 17–19.

Seillier-Moiseiwitsch, F. and Dawid, A.P. (1993), On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association* **88**, 355–359.

Sloughter, J.M., Raftery, A.E., Gneiting, T. and Fraley, C. (2007), Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review* **135**, 3209–3220.

Stott, P.A. and Kettleborough, J.A. (2002), Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature* **416**, 723–726 (Erratum: **417**, 205).

Tebaldi, C. and R. Knutti (2007), The use of the multimodel ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A*, **365**, Number 1857, pp. 2053–2075. Part of a Theme Issue on "Ensembles and probabilities: a new era in the prediction of climate change", edited by Matthew Collins and Sylvia Knight.

Tebaldi, C., Mearns, L.O., Nychka, D. and Smith, R.L. (2004), Regional probabilities of precipitation change: A Bayesian analysis of multimodel simulations. *Geophysical Research Letters* **31**, L24213, doi:10.1029/2004GL021276, 2004.

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 multi-model ensembles. *Journal of Climate* **18**, 1524–1540 (corrigendum p. 3405).

Webster, M. (2003), Communicating climate change uncertainty to policy-makers and the public: An editorial comment. *Climate Change* **61**, 1–8.

Webster, M., Forest, C., Reilly, J., Babiker, M., Kicklighter, D., Mayer, M., Prinn, R., Sarofim, M., Sokolov, A., Stone, P and Wang, C. (2003), uncertaintly analysis of climate change and policy response. *Climate Change* **61**, 295–320.

Wigley, T.M.L. (1999), *The Science of Climate Change: Global and U.S. Perspectives.* Pew Center on Global Climate Change, Arlington, VA, USA, 48pp.

Wigley, T.M.L. and Raper, S.C.B. (2001), Interpretations of high projections for global-mean warming. *Science* **293**, 451–454.

Wilson, L.J., Beauregard, S., Raftery, A.E. and Verret, R. (2007), Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Monthly Weather Review* **135**, 1364–1385.

10 Appendix: Derivation of Gibbs-Metropolis updating equations for the multivariate model

We assume the statistical model defined by (21)—(29). Omitting unnecessary constants, the joint density of all the parameters and random variables is

$$
\begin{aligned}
&(ca_\lambda b_\lambda)^{a-1}e^{-b(c+a_\lambda+b_\lambda)}\cdot\left[\prod_{i=0}^{R}\theta_i^{a-1}e^{-b\theta_i}\right]\cdot\left[\prod_{i=1}^{R}\phi_i^{a-1}e^{-b\phi_i}\right]\cdot\left[\prod_{j=1}^{M}\lambda_j^{a_\lambda-1}e^{-b_\lambda\lambda_j}\frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)}\right]\cdot\left[\psi_0^{a-1}e^{-b\psi_0}\right]\cdot\\
&\cdot\left[\prod_{i=1}^{R}\prod_{j=1}^{M}\eta_{ij}^{c-1}e^{-c\eta_{ij}}\frac{c^c}{\Gamma(c)}\right]\cdot\left[\prod_{j=1}^{M}\sqrt{\psi_0}e^{-\frac{1}{2}\psi_0\alpha_j^2}\right]\cdot\left[\prod_{j=1}^{M}\sqrt{\theta_0\psi_0}e^{-\frac{1}{2}\theta_0\psi_0(\alpha'_j-\beta_0\alpha_j)^2}\right]\cdot\\
&\cdot\left[\prod_{i=1}^{R}e^{-\frac{1}{2}\lambda_{0i}(X_{i0}-\mu_0-\zeta_i)^2}\right]\cdot\left[\prod_{i=1}^{R}\prod_{j=1}^{M}\sqrt{\eta_{ij}\phi_i\lambda_j}e^{-\frac{1}{2}\eta_{ij}\phi_i\lambda_j(X_{ij}-\mu_0-\zeta_i-\alpha_j)^2}\right]\cdot\\
&\cdot\left[\prod_{i=1}^{R}\prod_{j=1}^{M}\sqrt{\eta_{ij}\theta_i\lambda_j}e^{-\frac{1}{2}\eta_{ij}\theta_i\lambda_j\left\{Y_{ij}-\nu_0-\zeta'_i-\alpha'_j-\beta_i(X_{ij}-\mu_0-\zeta_i-\alpha_j)\right\}^2}\right]. \qquad (31)
\end{aligned}
$$

Define

$$
\tilde{\mu}_0 = \frac{\sum_i\lambda_{0i}(X_{i0}-\zeta_i)+\sum_i\phi_i\sum_j\eta_{ij}\lambda_j(X_{ij}-\zeta_i-\alpha_j)-\sum_i\beta_i\theta_i\sum_j\eta_{ij}\lambda_j\{Y_{ij}-\nu_0-\zeta'_i-\alpha'_j-\beta_i(X_{ij}-\zeta_i-\alpha_j)\}}{\sum_i\{\lambda_{0i}+(\phi_i+\beta_i^2\theta_i)\sum_j\eta_{ij}\lambda_j\}}, \qquad (32)
$$

$$
\tilde{\nu}_0 = \frac{\sum_i\theta_i\sum_j\eta_{ij}\lambda_j\{Y_{ij}-\zeta'_i-\alpha'_j-\beta_i(X_{ij}-\mu_0-\zeta_i-\alpha_j)\}}{\sum_i\theta_i\{\sum_j\eta_{ij}\lambda_j\}}, \qquad (33)
$$

$$
\tilde{\zeta}_i = \frac{\lambda_{0i}(X_{i0}-\mu_0)+\phi_i\sum_j\eta_{ij}\lambda_j(X_{ij}-\mu_0-\alpha_j)-\beta_i\theta_i\sum_j\eta_{ij}\lambda_j\{Y_{ij}-\nu_0-\zeta'_i-\alpha'_j-\beta_i(X_{ij}-\mu_0-\alpha_j)\}}{\lambda_{0i}+(\phi_i+\beta_i^2\theta_i)\sum_j\eta_{ij}\lambda_j}, \qquad (34)
$$

$$
\tilde{\zeta}'_i = \frac{\sum_j\eta_{ij}\lambda_j\{Y_{ij}-\nu_0-\alpha'_j-\beta_i(X_{ij}-\mu_0-\zeta_i-\alpha_j)\}}{\sum_j\eta_{ij}\lambda_j}, \qquad (35)
$$

$$
\tilde{\beta}_0 = \frac{\sum_j\alpha'_j\alpha_j}{\sum_j\alpha_j^2}, \qquad (36)
$$

$$
\tilde{\beta}_i = \frac{\sum_j\eta_{ij}\lambda_j(Y_{ij}-\nu_0-\zeta'_i-\alpha'_j)(X_{ij}-\mu_0-\zeta_i-\alpha_j)}{\sum_j\eta_{ij}\lambda_j(X_{ij}-\mu_0-\zeta_i-\alpha_j)^2}, \quad (i\neq 0), \qquad (37)
$$

$$
\tilde{\alpha}_j = \frac{\beta_0\theta_0\psi_0\alpha'_j+\lambda_j\sum_i\eta_{ij}\phi_i(X_{ij}-\mu_0-\zeta_i)-\lambda_j\sum_i\eta_{ij}\theta_i\beta_i\{Y_{ij}-\nu_0-\zeta'_i-\alpha'_j-\beta_i(X_{ij}-\mu_0-\zeta_i)\}}{\psi_0+\beta_0^2\theta_0\psi_0+\lambda_j\sum_i\eta_{ij}\phi_i+\lambda_j\sum_i\eta_{ij}\theta_i\beta_i^2}, \qquad (38)
$$

$$
\tilde{\alpha}'_j = \frac{\beta_0\theta_0\psi_0\alpha_j+\lambda_j\sum_i\eta_{ij}\theta_i\{Y_{ij}-\nu_0-\zeta'_i-\beta_i(X_{ij}-\mu_0-\zeta_i-\alpha_j)\}}{\theta_0\psi_0+\lambda_j\sum_i\eta_{ij}\theta_i}. \qquad (39)
$$

The conditional distributions required for the Gibbs sampler are as follows:

$$\mu_0 \mid \text{rest} \sim N\left[\tilde{\mu}_0, \frac{1}{\sum_i\{\lambda_{0i} + (\phi_i + \beta_i^2\theta_i)\sum_j \eta_{ij}\lambda_j\}}\right], \tag{40}$$

$$\nu_0 \mid \text{rest} \sim N\left[\tilde{\nu}_0, \frac{1}{\sum_i\{\theta_i \sum_j \eta_{ij}\lambda_j\}}\right], \tag{41}$$

$$\zeta_i \mid \text{rest} \sim N\left[\tilde{\zeta}_i, \frac{1}{\lambda_{0i} + (\phi_i + \beta_i^2\theta_i)\sum_j \eta_{ij}\lambda_j}\right], \tag{42}$$

$$\zeta_i' \mid \text{rest} \sim N\left[\tilde{\zeta}_i', \frac{1}{\theta_i \sum_j \eta_{ij}\lambda_j}\right], \tag{43}$$

$$\beta_0 \mid \text{rest} \sim N\left[\tilde{\beta}_0, \frac{1}{\theta_0\psi_0 \sum_j \alpha_j^2}\right], \tag{44}$$

$$\beta_i \mid \text{rest} \sim N\left[\tilde{\beta}_i, \frac{1}{\theta_i \sum_j \eta_{ij}\lambda_j(X_{ij} - \mu_0 - \zeta_i - \alpha_j)^2}\right], \quad (i \neq 0) \tag{45}$$

$$\alpha_j \mid \text{rest} \sim N\left[\tilde{\alpha}_j, \frac{1}{\psi_0 + \beta_0^2\theta_0\psi_0 + \lambda_j \sum_i \eta_{ij}\phi_i + \lambda_j \sum_i \eta_{ij}\theta_i\beta_i^2}\right], \tag{46}$$

$$\alpha_j' \mid \text{rest} \sim N\left[\tilde{\alpha}_j', \frac{1}{\theta_0\psi_0 + \lambda_j \sum_i \eta_{ij}\theta_i}, \right], \tag{47}$$

$$\theta_0 \mid \text{rest} \sim G\left[a + \frac{M}{2}, b + \frac{1}{2}\psi_0 \sum_j (\alpha_j' - \beta_0\alpha_j)^2\right], \tag{48}$$

$$\theta_i \mid \text{rest} \sim G\left[a + \frac{M}{2}, b + \frac{1}{2}\sum_j \eta_{ij}\lambda_j\{Y_{ij} - \nu_0 - \zeta_i' - \alpha_j' - \beta_i(X_{ij} - \mu_0 - \zeta_i - \alpha_j)\}^2\right], \quad (i \neq 0) \tag{49}$$

$$\phi_i \mid \text{rest} \sim G\left[a + \frac{M}{2}, b + \frac{1}{2}\sum_j \eta_{ij}\lambda_j(X_{ij} - \mu_0 - \zeta_i - \alpha_j)^2\right], \tag{50}$$

$$\lambda_j \mid \text{rest} \sim G\left[a_\lambda + R, b_\lambda + \frac{1}{2}\sum_i \eta_{ij}\phi_i(X_{ij} - \mu_0 - \zeta_i - \alpha_j)^2 + \frac{1}{2}\sum_i \eta_{ij}\theta_i\{Y_{ij} - \nu_0 - \zeta_i' - \alpha_j' - \beta_i(X_{ij} - \mu_0 - \zeta_i - \alpha_j)\}^2\right], \tag{51}$$

$$\psi_0 \mid \text{rest} \sim G\left[a + M, b + \frac{1}{2}\sum_j \alpha_j^2 + \frac{1}{2}\theta_0 \sum_j (\alpha_j' - \beta_0\alpha_j)^2\right], \tag{52}$$

$$\eta_{ij} \mid \text{rest} \sim G\left[c + 1, c + \frac{1}{2}\phi_i\lambda_j(X_{ij} - \mu_0 - \zeta_i - \alpha_j)^2 + \frac{1}{2}\theta_i\lambda_j\{Y_{ij} - \nu_0 - \zeta_i' - \alpha_j' - \beta_i(X_{ij} - \mu_0 - \zeta_i - \alpha_j)\}^2\right]. \tag{53}$$

Were a_λ, b_λ and c fixed, as in the univariate analysis, the iteration (40)–(53) could be repeated many times to generate a random sample from the joint posterior distribution. Having added a layer by making the three parameters random variates, two Metropolis steps are added to the

iteration (40)–(53), as follows.

For the sampling of a_λ and b_λ jointly, define U_1, U_2 two independent random variables distributed uniformly over the interval $(0,1)$, and the two candidate values $a'_\lambda = a_\lambda e^{\delta(U_1-\frac{1}{2})}$ and $b'_\lambda = b_\lambda e^{\delta(u_2-\frac{1}{2})}$, where δ is an arbitrary increment, chosen as $\delta = 1$ in our implementation. We then compute

$$\ell_1 = Ma_\lambda \log b_\lambda - M \log \Gamma(a_\lambda) + a_\lambda \sum_j \log \lambda_j - b_\lambda \sum_j \lambda_j + a \log(a_\lambda b_\lambda) - b(a_\lambda + b_\lambda), \tag{54}$$

$$\ell_2 = Ma'_\lambda \log b'_\lambda - M \log \Gamma(a'_\lambda) + a'_\lambda \sum_j \log \lambda_j - b'_\lambda \sum_j \lambda_j + a \log(a'_\lambda b'_\lambda) - b(a'_\lambda + b'_\lambda). \tag{55}$$

In (54) and (55) we are computing the log likelihoods of (a_λ, b_λ) and (a'_λ, b'_λ), allowing for the prior densities and including a Jacobian term, allowing for the fact that the updating is taking place on a logarithmic scale. Then, within each iteration of the Gibbs/Metropolis simulation, the proposed values (a'_λ, b'_λ) are accepted with probability $e^{\ell_2-\ell_1}$ if $\ell_2 < \ell_1$, or 1 if $\ell_2 \geq \ell_1$.

Similarly, the updating of c takes place by proposing $c' = ce^{\delta(U_3-\frac{1}{2})}$, where U_3 is a draw from a uniform distribution on $(0,1)$, and computing

$$\ell_1 = MRc \log c - M \log \Gamma(c) + c \sum_i \sum_j \log \eta_{ij} - c \sum_i \sum_j \eta_{ij} + a \log c - bc \tag{56}$$

$$\ell_2 = MRc' \log c' - MR \log \Gamma(c') + c' \sum_i \sum_j \log \eta_{ij} - c' \sum_i \sum_j \eta_{ij} + a \log c' - bc'. \tag{57}$$

Then, within each iteration of the Gibbs/Metropolis simulation, the proposed value c' is accepted with probability $e^{\ell_2-\ell_1}$ if $\ell_2 < \ell_1$, or 1 if $\ell_2 \geq \ell_1$.

The iteration is repeated many times to generate a Monte Carlo sample from the posterior distribution.

Figures

Figure 1. The 22 regions used in this study.

Figure 2. Posterior densities for mean temperature change in six regions, DJF season, A2 scenario. Black curve: univariate approach. Red curve: multivariate approach. The black dots on the bottom represent the individual GCM projections on which the analysis is based.

Figure 3. Color-coded U_{ij} statistics for the univariate approach.

Figure 4. Posterior densities for c under the multivariate approach, together with Metropolis acceptance probabilities, medians and IQRs.

Figure 5. Posterior densities for a_λ and b_λ under the multivariate approach, together with Metropolis acceptance probabilities, medians and IQRs.

Figure 6: Trace plots of samples from individual chains of the MCMC algorithms. Three parameters from the univariate model (left-hand side panels) are compared to the corresponding parameters from the multivariate model (right-hand side panels). From top to bottom, left to right: temperature change ($\Delta T \equiv \nu - \mu$), future ($\nu$) and current temperature (μ) from the univariate model; temperature change ($\Delta T \equiv \nu_0 + \zeta' - \mu_0 - \zeta$), future ($\nu_0 + \zeta'$) and current temperature ($\mu_0 + \zeta$) from the multivariate model. Values are in K. Temperature change is estimated under scenario A2, in DJF, for region ALA (Alaska).

Figure 7: Same as Figure 6, for region NAS (Northern Asia).

Figure 8. Color-coded U_{ij} statistics for the multivariate approach.

Figure 9: Sensitivity plots for Western Africa (WAF), season JJA, scenario A2. One model projection (red dots) shifted in 0.5-K increments to assess the sensitivity of predictive density to changes in original data. The original data value is given by the leftmost red dot and the associated predictive density is the solid curve; remaining curves on each plot represent the new predictive density after each shift. Top to bottom: method of Tebaldi *et al.* (2005); this paper's univariate approach; and this paper's multivariate approach. The top panel shows that the Tebaldi *et al.* (2005) model is extremely sensitive to changes in the relative position of the 9 GCM values. However the corresponding plots for the current univariate version of the model (second panel) and the multivariate version (third panel) show that both models' performance is largely insensitive to changes in a single data point, even for increasingly large perturbations. Additionally, a comparison of the two sets of PDFs for the current models indicates that the multivariate model's estimates are relatively more robust to the perturbations, shifting less to the right with the movement of the red mark than the set of PDFs from the univariate model.

Figure 10 Predictive distribution for mean temperature change under the multivariate model. Color-coded projections for each region represent (L–R) the 5% and 25% quantiles, the mean, the 75% and 95% quantiles of the predictive distribution for two seasons (DJF, JJA) and two emission scenarios (A2, B2).

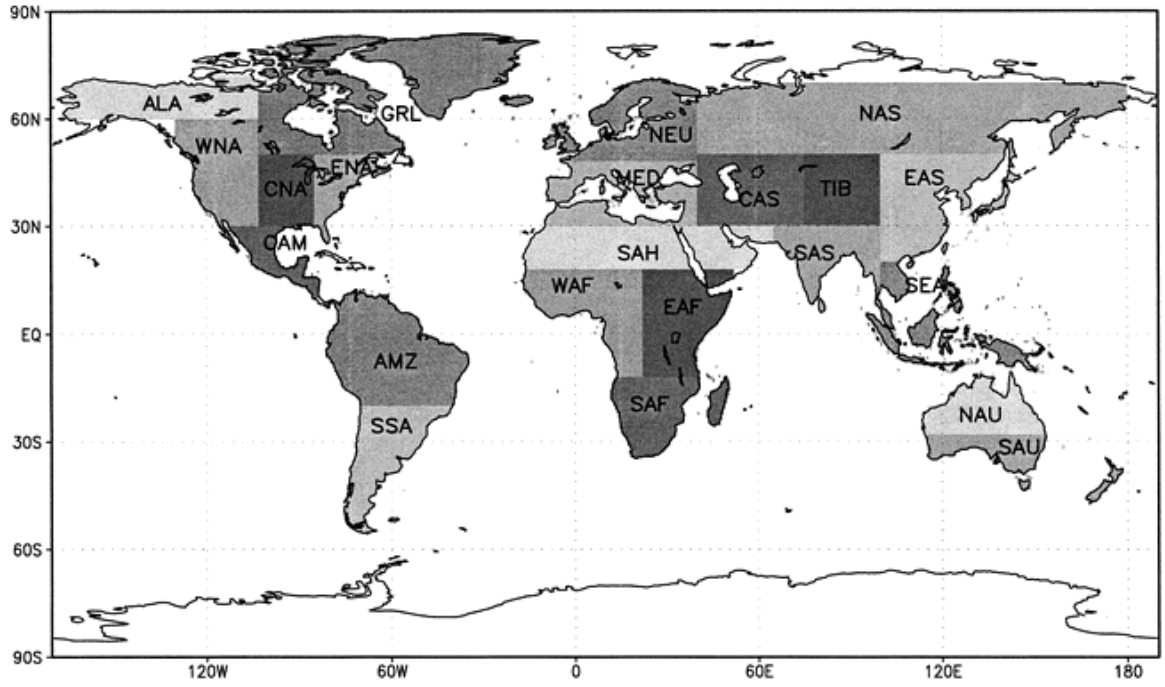
90N
60N
30N
EQ
30S
60S
90S
120W
60W
0
60E
120E
180
ALA
GRL
WNA
ENA
CNA
CAM
AMZ
SSA
NEU
NAS
MED
CAS
TIB
EAS
SAH
SAS
WAF
EAF
SEA
SAF
NAU
SAU

Figure 1:

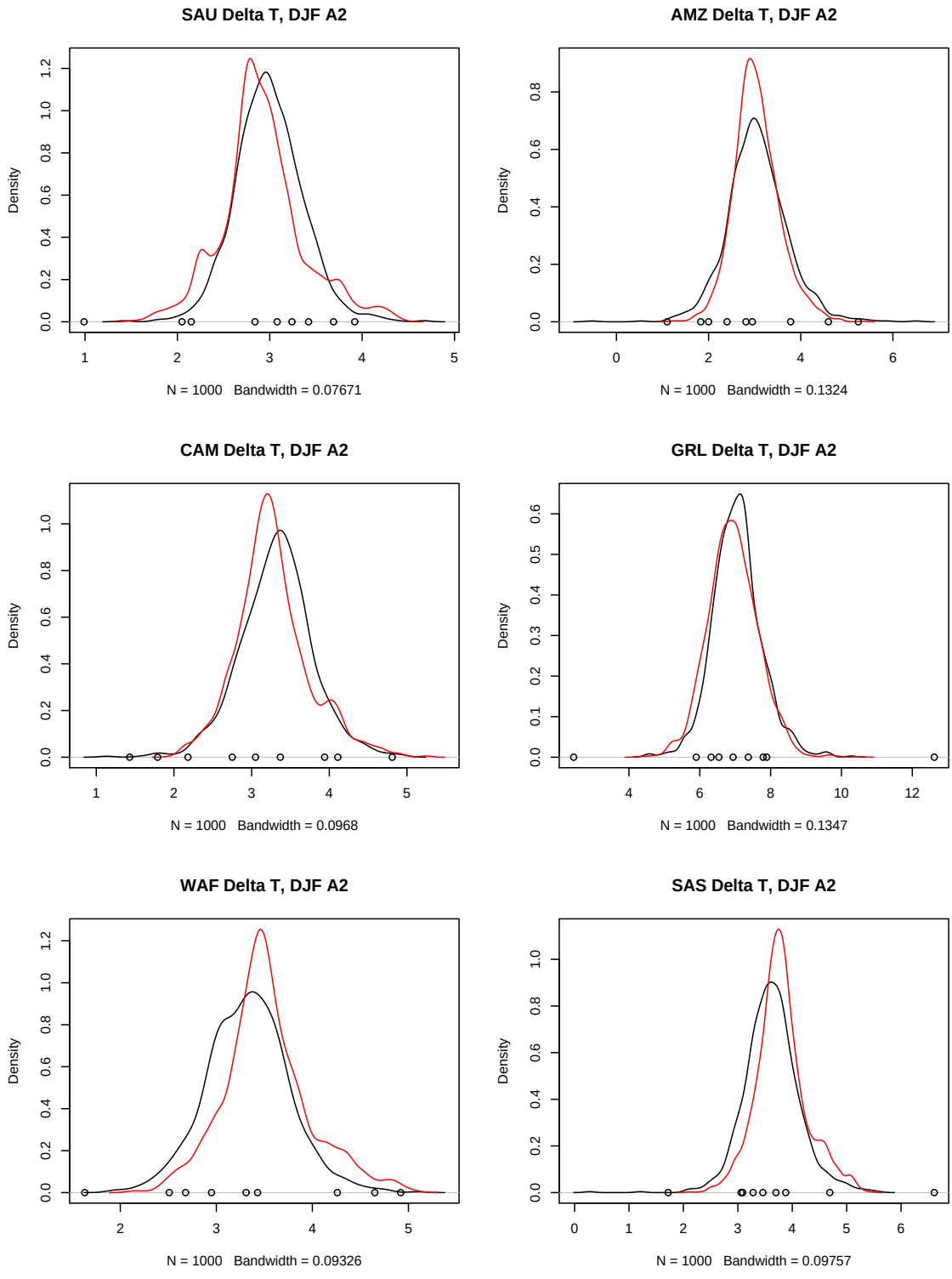
SAU Delta T, DJF A2
Density
N = 1000 Bandwidth = 0.07671
AMZ Delta T, DJF A2
Density
N = 1000 Bandwidth = 0.1324
CAM Delta T, DJF A2
Density
N = 1000 Bandwidth = 0.0968
GRL Delta T, DJF A2
Density
N = 1000 Bandwidth = 0.1347
WAF Delta T, DJF A2
Density
N = 1000 Bandwidth = 0.09326
SAS Delta T, DJF A2
Density
N = 1000 Bandwidth = 0.09757

Figure 2:

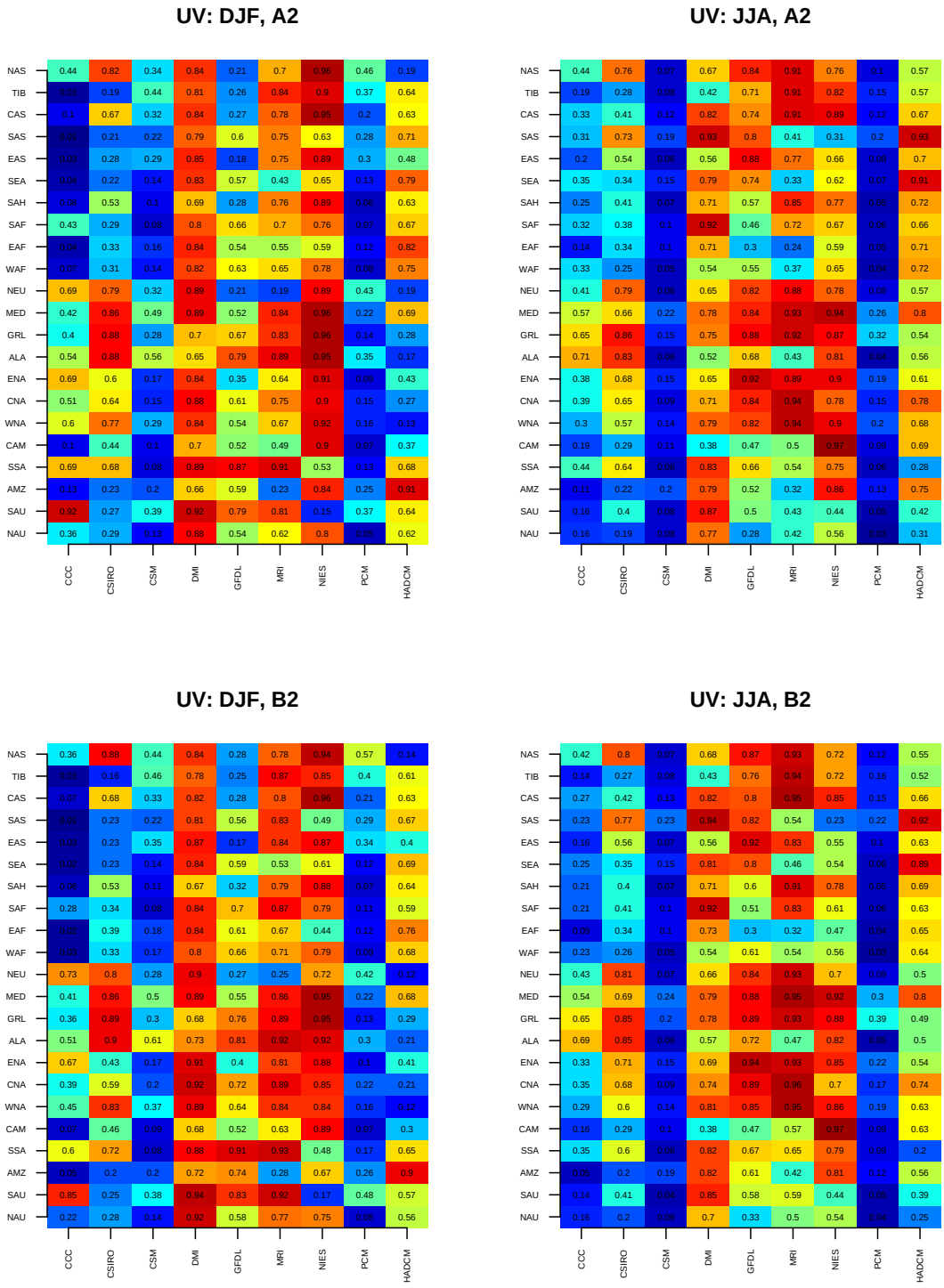
UV: DJF, A2
UV: JJA, A2
UV: DJF, B2
UV: JJA, B2
NAS
TIB
CAS
SAS
EAS
SEA
SAH
SAF
EAF
WAF
NEU
MED
GRL
ALA
ENA
CNA
WNA
CAM
SSA
AMZ
SAU
NAU
CCC
CSIRO
CSM
DMI
GFDL
MRI
NIES
PCM
HADCM

Figure 3:

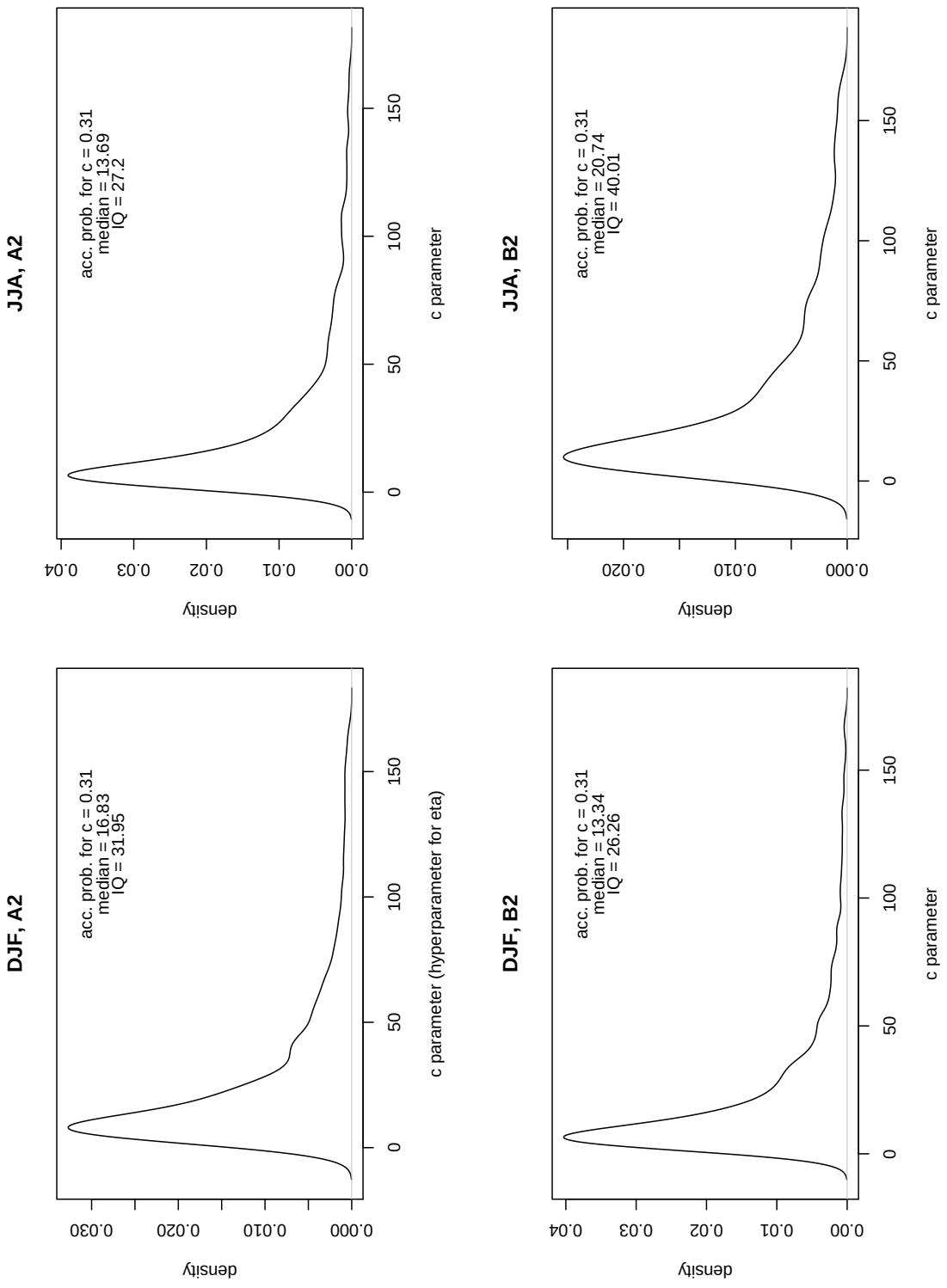
JJA, A2
acc. prob. for c = 0.31
median = 13.69
IQ = 27.2
density
c parameter
JJA, B2
acc. prob. for c = 0.31
median = 20.74
IQ = 40.01
density
c parameter
DJF, A2
acc. prob. for c = 0.31
median = 16.83
IQ = 31.95
density
c parameter (hyperparameter for eta)
DJF, B2
acc. prob. for c = 0.31
median = 13.34
IQ = 26.26
density
c parameter

Figure 4:

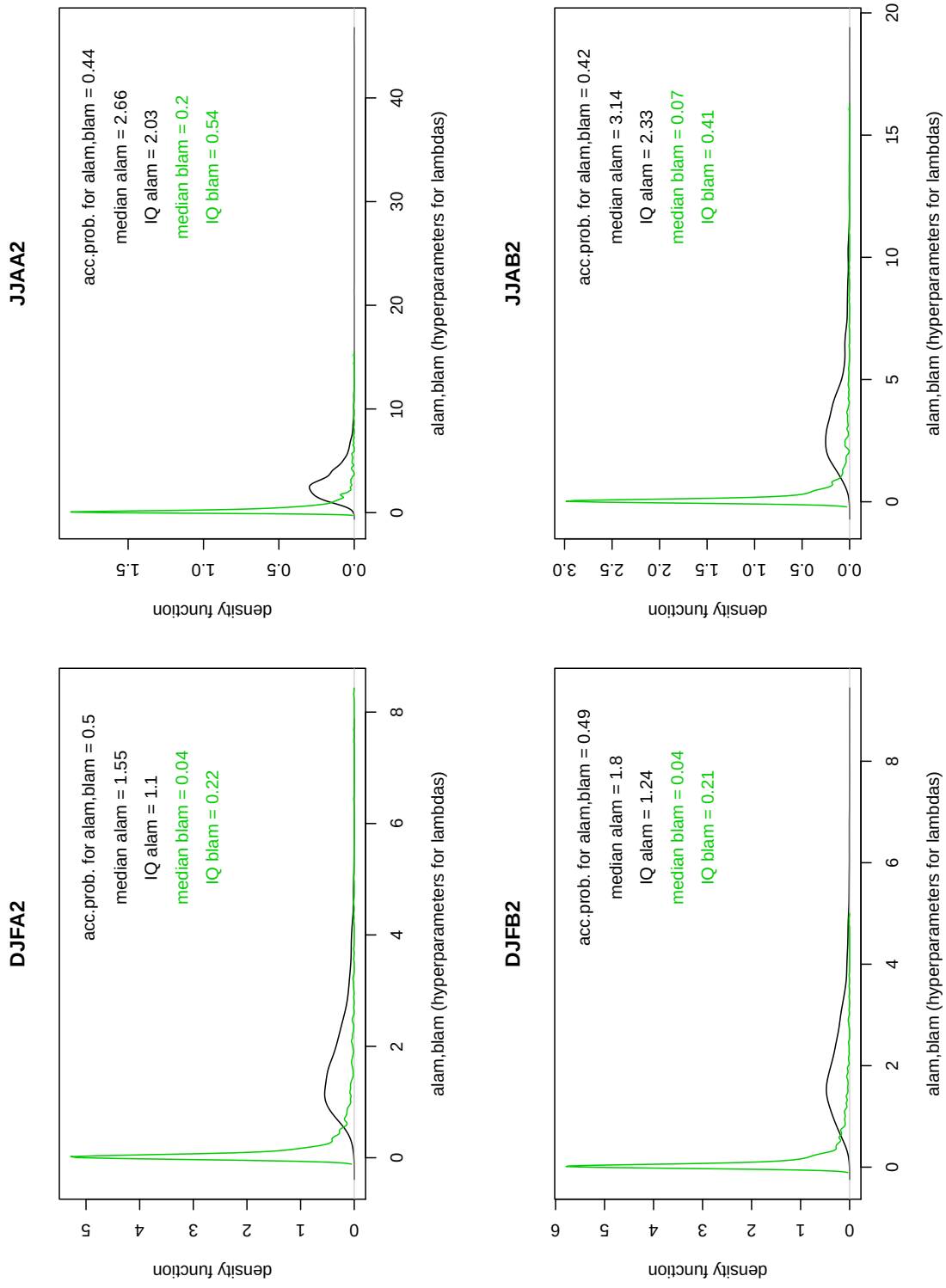
DJFA2
acc.prob. for alam,blam = 0.5
median alam = 1.55
IQ alam = 1.1
median blam = 0.04
IQ blam = 0.22
density function
alam,blam (hyperparameters for lambdas)
JJAA2
acc.prob. for alam,blam = 0.44
median alam = 2.66
IQ alam = 2.03
median blam = 0.2
IQ blam = 0.54
density function
alam,blam (hyperparameters for lambdas)
DJFB2
acc.prob. for alam,blam = 0.49
median alam = 1.8
IQ alam = 1.24
median blam = 0.04
IQ blam = 0.21
density function
alam,blam (hyperparameters for lambdas)
JJAB2
acc.prob. for alam,blam = 0.42
median alam = 3.14
IQ alam = 2.33
median blam = 0.07
IQ blam = 0.41
density function
alam,blam (hyperparameters for lambdas)

Figure 5:

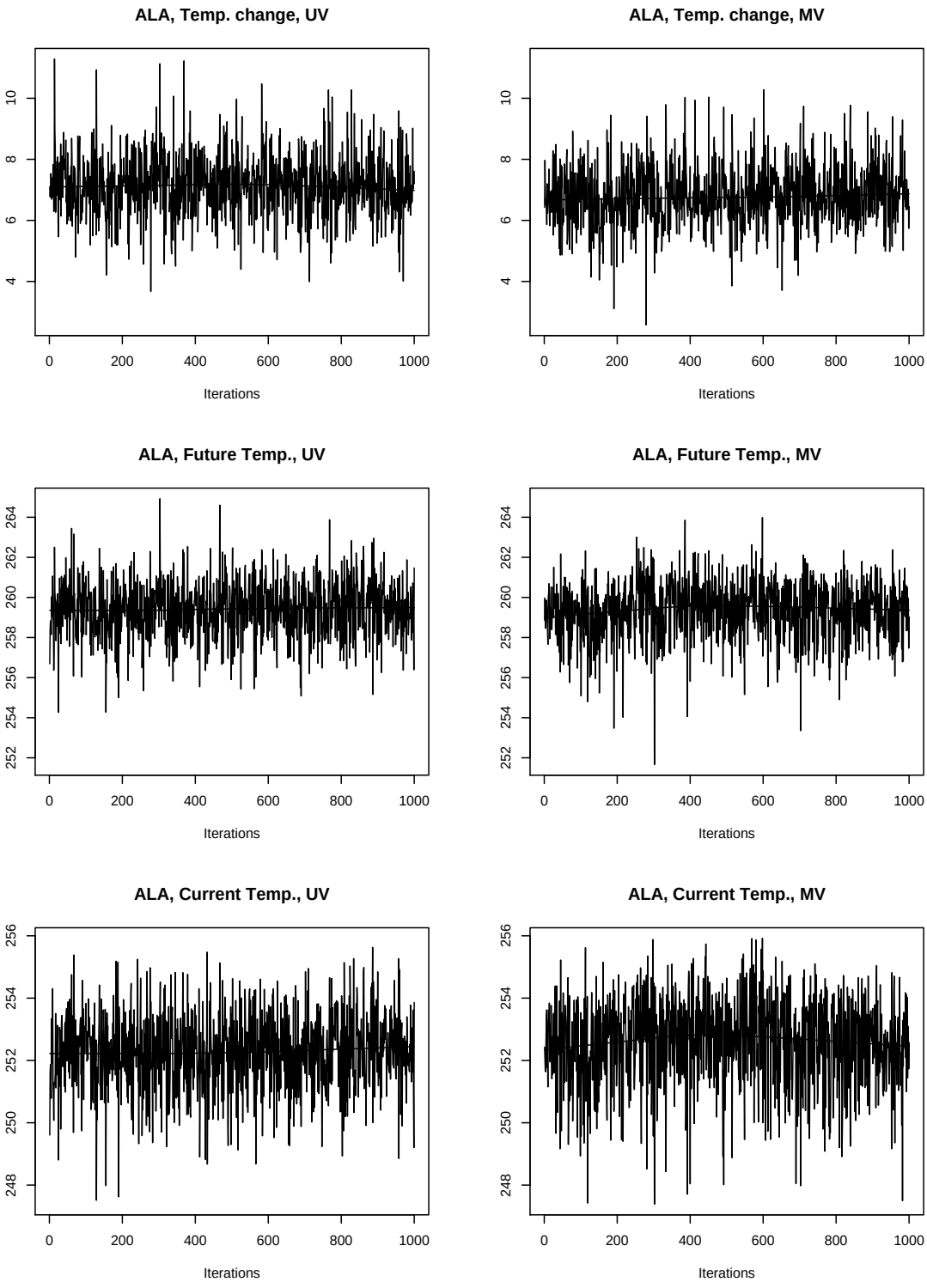
ALA, Temp. change, UV
ALA, Temp. change, MV
ALA, Future Temp., UV
ALA, Future Temp., MV
ALA, Current Temp., UV
ALA, Current Temp., MV
Iterations
0
200
400
600
800
1000

Figure 6:

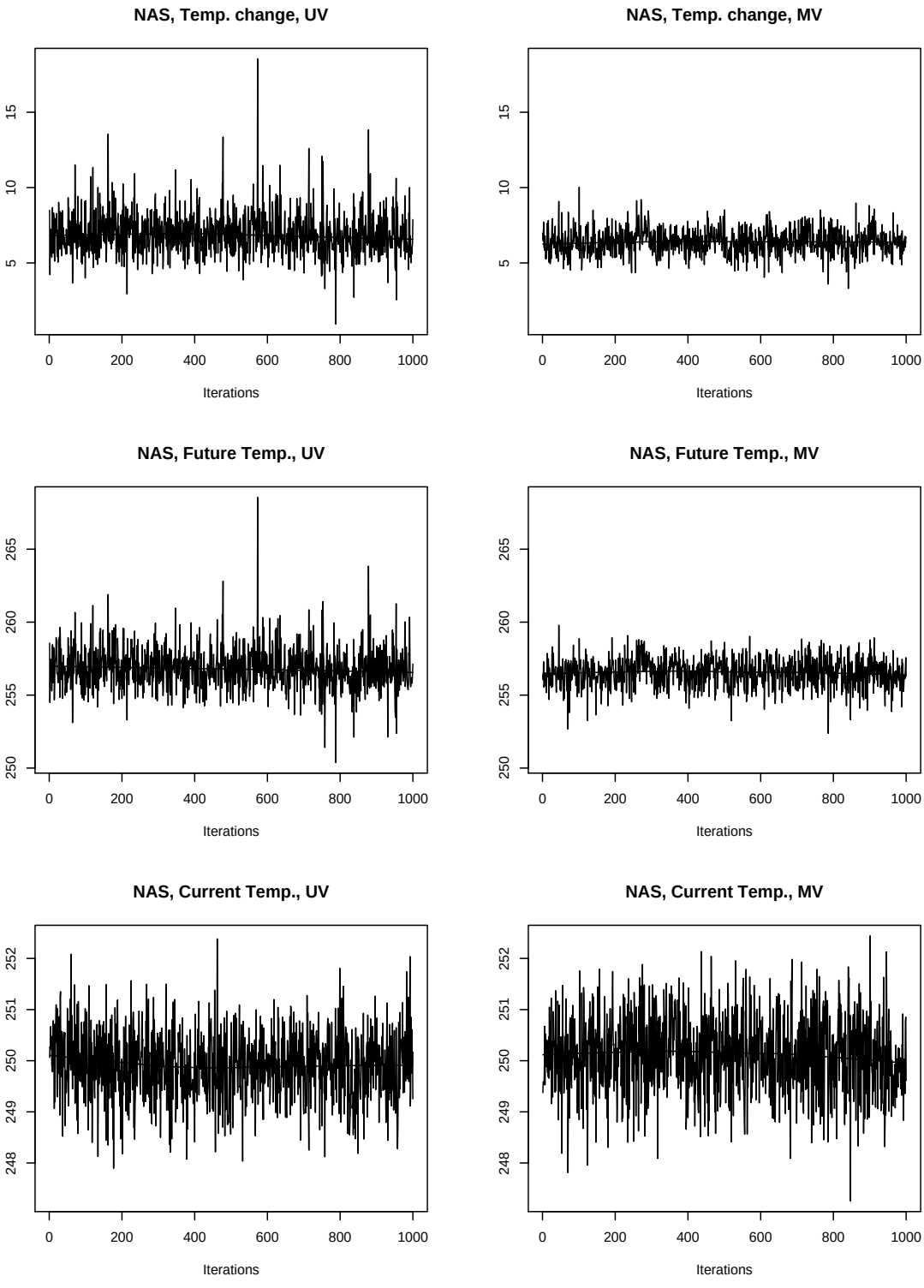
NAS, Temp. change, UV
NAS, Temp. change, MV
NAS, Future Temp., UV
NAS, Future Temp., MV
NAS, Current Temp., UV
NAS, Current Temp., MV
Iterations
0
200
400
600
800
1000
5
10
15
250
255
260
265
248
249
250
251
252

Figure 7:

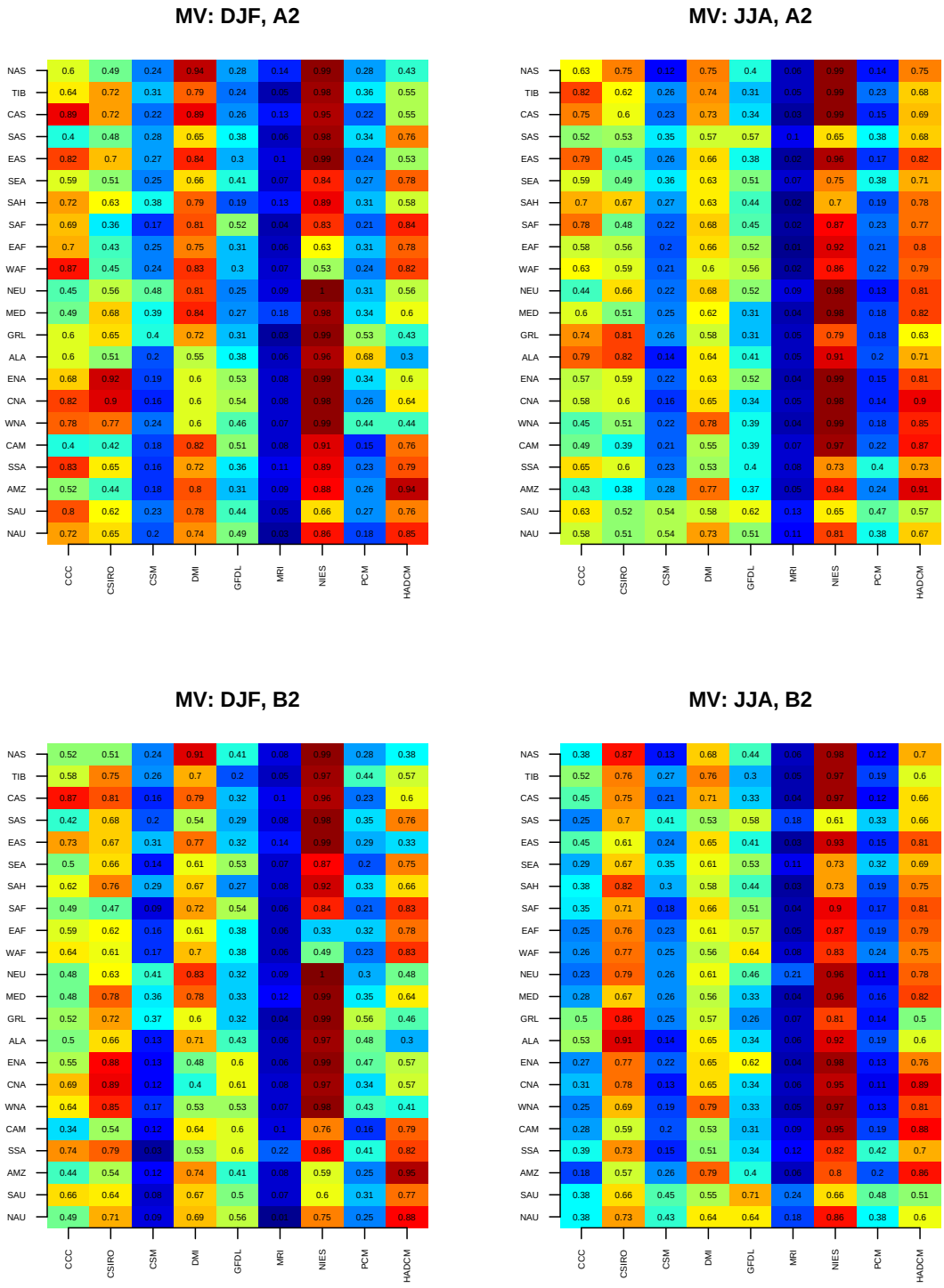
MV: DJF, A2
MV: JJA, A2
MV: DJF, B2
MV: JJA, B2
NAS
TIB
CAS
SAS
EAS
SEA
SAH
SAF
EAF
WAF
NEU
MED
GRL
ALA
ENA
CNA
WNA
CAM
SSA
AMZ
SAU
NAU
CCC
CSIRO
CSM
DMI
GFDL
MRI
NIES
PCM
HADCM

Figure 8:

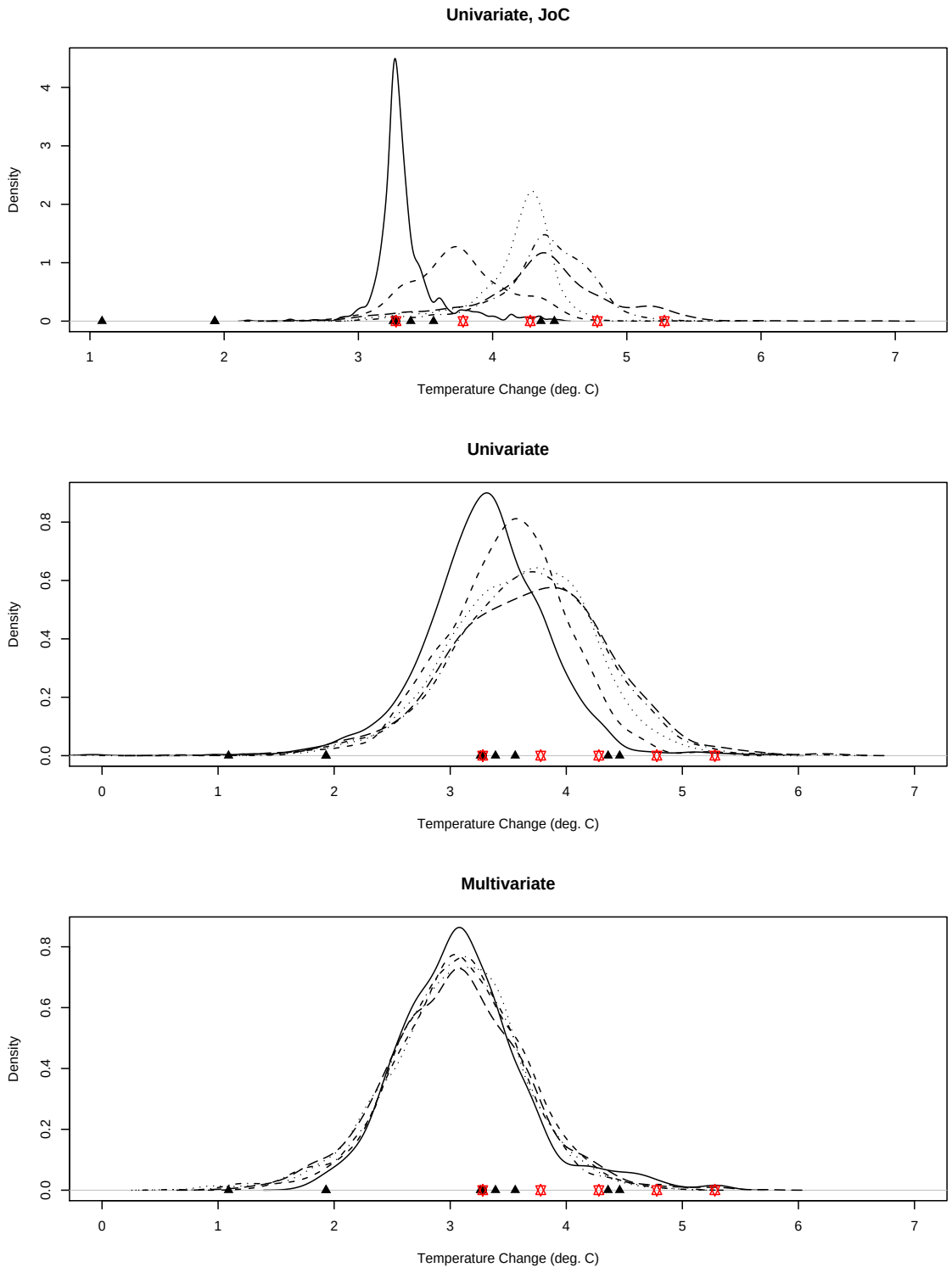
Univariate, JoC
Density
Temperature Change (deg. C)
Univariate
Density
Temperature Change (deg. C)
Multivariate
Density
Temperature Change (deg. C)

Figure 9:

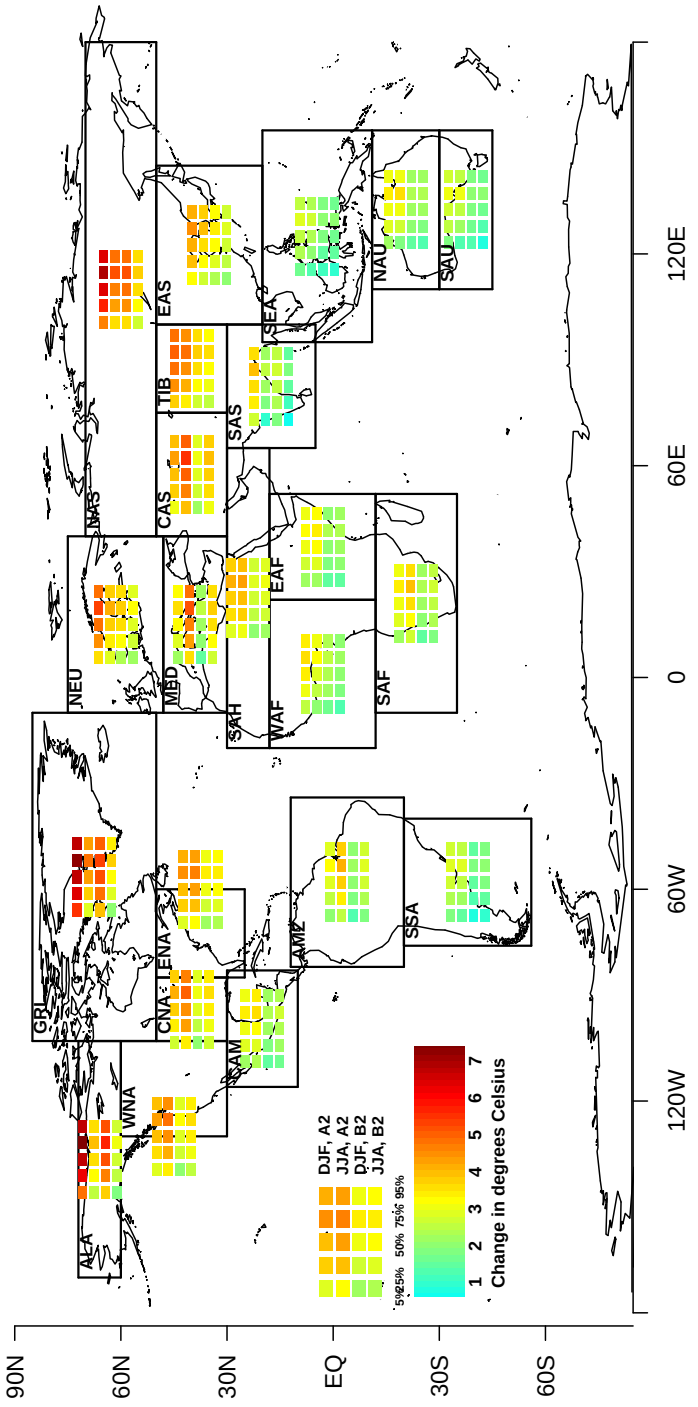
90N
60N
30N
EQ
30S
60S
120W
60W
0
60E
120E
GRL
ALA
WNA
CNA
ENA
CAM
AMZ
SSA
NEU
MED
NAS
CAS
TIB
EAS
SAS
SEA
NAU
SAU
EAF
SAH
WAF
SAF
DJF, A2
JJA, A2
DJF, B2
JJA, B2
5%
25%
50%
75%
95%
1 2 3 4 5 6 7
Change in degrees Celsius

Figure 10:

Tables

Model	Full Name	Sensitivity
CCC	Canadian Centre for Climate	3.59
CSIRO	Commonwealth Scientific and Industrial Research Organisation (Australia)	3.50
CSM	Climate System Model (NCAR, USA)	2.29
DMI	Max Planck Institute for Meteorology (Germany)	3.11
GFDL	Geophysical Fluid Dynamics Laboratory (USA)	2.87
MRI	Meteorological Research Institute (Japan)	1.25
NIES	National Institute for Environmental Studies (Japan)	4.53
PCM	Parallel Climate Model (several institutions in USA)	2.35
HADCM	Hadley Centre Coupled Model (U.K. Meteorological Office)	3.38

Table 1: Climate models used in this study, and their climate sensitivities in K.

Test	K-S	C-vM	A-D	Cor.
DJFA2	7	5	8	9
DJFB2	7	6	7	13
JJAA2	6	6	7	10
JJAB2	4	4	5	6

Table 2: Results of four goodness-of-fit statistics calculated from the U_{ij} values. For each statistic and each season/scenario combination, tabulated is the number of regions (out of 22) in which the univariate procedure of Section 4 produced a smaller (better) test statistic than the multivariate procedure of Section 5.

Test	K-S UNIV	K-S MULTIV	C-vM UNIV	C-vM MULTIV	A-D UNIV	A-D MULTIV	Cor. UNIV	Cor. MULTIV
DJFA2	1	2	3	2	3	3	0	3
DJFB2	2	3	1	3	3	3	1	1
JJAA2	0	3	0	4	1	3	0	2^*
JJAB2	1	3	2	4	2	5	0	1

Table 3: Formal tests applied to the goodness-of-fit statistics. In each case, tabulated are the number of rejections of the test, at level .05 in a two-sided test, over 22 regions. All rejections are in *lower* tail of test statistic except for one in the box marked*.

	IQR	I15R	I5R
DJFA2	1.11 (13)	1.09 (12)	1.12 (15)
DJFB2	1.04 (13)	1.04 (14)	1.05 (12)
JJAA2	1.05 (13)	1.04 (14)	1.00 (14)
JJAB2	1.10 (15)	1.08 (16)	1.08 (14)

Table 4: Comparisons of the IQR, I15R and I5R scale statistics applied to the posterior distributions of projected temperature changes for the univariate and multivariate approaches. The main entry in each cell of the table is the ratio of statistics calculated for the univariate (numerator) and multivariate (denominator) approaches, a value > 1 implying that the multivariate approach resulted in a tighter posterior distribution overall. Also shown in parentheses is the number of times (out of 22 regions) that the multivariate approach led to a smaller scale statistic than the univariate approach.