Climate Projections Using Bayesian Model Averaging and Space-Time Dependence

K. Sham Bhat, Murali Haran, Adam Terando, and Klaus Keller. *

Abstract

Projections of future climatic changes are a key input to the design of climate change mitigation and adaptation strategies. Current climate change projections are deeply uncertain. This uncertainty stems from several factors, including parametric and structural uncertainties. One common approach to characterize and, if possible, reduce these uncertainties is to confront (calibrate in a broad sense) the models with historical observations. Here, we analyze the problem of combining multiple climate models using Bayesian Model Averaging (BMA) to derive future projections and quantify uncertainty estimates of spatiotemporally resolved temperature hindcasts and projections. One advantage of the BMA approach is that it allows the assessment of the predictive skill of a model using the training data, which can help identify the better models and discard poor models. Previous BMA approaches have broken important new ground, but often neglect spatial and temporal dependencies and/or impose prohibitive computational demands. Here we improve on the current state of the art by incorporating spatial and temporal dependence while using historical data to estimate model weights. We achieve computational efficiency using a kernel mixing approach for representing a space-time process. One key advantage of our new approach is that it enables us to incorporate multiple sources of uncertainty and biases, while remaining computationally tractable for large data sets. We introduce and apply our approach using BMA to an ensemble of Global Circulation Model output from the Intergovernmental Panel on Climate Change Fourth Assessment Report of surface temperature on a grid of space-time locations.

*K. Sham Bhat is a Scientist at Los Alamos National Laboratory, Statistical Sciences Division, Los Alamos, NM 87545 (E-mail: *bhat9999@lanl.gov*). Murali Haran is Associate Professor, Department of Statistics, Pennsylvania State University, University Park, PA 16802 (E-mail: *mharan@stat.psu.edu*). Adam Terando is Climate Change Research Coordinator, Department of Biology, North Carolina State University, Raleigh, NC 27695 (E-mail: *adam_terando@ncsu.edu*). Klaus Keller is Associate Professor, Department of Geosciences, Pennsylvania State University, University Park, PA 16802 (E-mail: *klaus@psu.edu*).

Keywords: Bayesian model averaging, Bayesian hierarchical modeling, Gaussian process, climate model, space-time data, climate change.

1 Introduction

Anthropogenic climate change has increasingly captured the attention of both scientists and the general public. The need to carefully project future climate change has grown in importance, as climate change projections are used to inform climate change mitigation and adaptation decisions (IPCC, 2007). Current climate projections are typically based on simulations from General Circulation Models (GCM) that approximate the complex physical, chemical, and biological processes (cf. Meehl et al., 2007). Given the complex nature of the Earth system, our incomplete knowledge about this system, our limited computational capabilities, and often sparse observational constraints, it is not surprising that current climate projections are subject to considerable uncertainty (Meehl et al., 2007). Uncertainties about the unknown input parameters for the climate model are referred to as parametric uncertainties, while uncertainties due to the form or structure of the climate model (e.g. incomplete understanding of cloud formation) are defined as structural uncertainties.

One approach to improve the characterization of the structural uncertainty is to confront an ensemble of multiple model structures with historical observations. It has been argued that because each individual GCM has strengths and weaknesses in approximating the climate system, future projections should include the results from all GCMs, as no single model can be declared as the best (Knutti et al., 2010; McAvaney et al., 2001). To obtain more accurate results and to improve our understanding of the possible range of model variability, a multi-model ensemble of GCM output is often used. Furthermore, using multi-model ensembles has improved the accuracy and consistency of model projections, in applications from weather and climate to agriculture and health (Cantelaube and Terres, 2005; Tebaldi and Knutti, 2007; Thomson et al., 2006; Vrugt et al., 2008). For a variety of resolutions of climate models, the average of a multi-model ensemble agrees more with climate observations than any single model (Gleckler et al., 2008; Knutti et al., 2010; Pierce et al., 2009). Projections are usually improved by using a weighted average of a multi-model ensemble, where the

weights are selected by the model skill (Doblas-Reyes et al., 2003; Tebaldi and Knutti, 2007; Yun et al., 2003).

Climate model simulations often simulate a large number of climate attributes (e.g. temperature, precipitation) on a fine spatial and temporal scale. Many of these attributes, especially surface temperature, are usually influenced by underlying processes of the general climate system at nearby locations, and aggregation across space or time are usually insufficient to describe these processes. Failure to account for space-time dependence may result in biased estimates of mean coefficients and non-spatial error variance components (Schabenberger and Gotway, 2005), and inefficient predictions (see Cressie, 1993, p. 16-17). Hence there is a need to account for spatial and temporal dependence. Further, there is the potential of model bias, as well as uncertainty in the historical measurements. Depending on the resolution of the locations, the output may be on the order of hundreds of thousands to millions of data points on a 3-D space-time field for each climate model, and any approach to analyze such data must account for computational challenges. Hence, there is a need for a computationally tractable approach to obtain climate projections while incorporating variability from the climate model forecasts as well as spatial and temporal dependence, bias, and sources of uncertainty for large spatial data. The primary computational bottleneck for large spatial data are matrix operations involving high-dimensional covariance matrices; some approaches to ensure tractability for large spatial data are Kronecker products, kernel mixing (Higdon, 1998), low rank matrix reduction (Cressie and Johannesson, 2008), and covariance tapering (Furrer et al., 2006).

We consider the problem of making probabilistic projections based on hindcasts and forecasts from a multi-model ensemble output and historical data, all of which are large spatial fields. Bayesian Model Averaging (cf. Hoeting et al., 1999; Leamer and Leamer, 1978) provides a formal statistical approach for combining information from multiple models by assigning weights based on model skill, defined as the ability of model hindcasts to predict historical data during a training period. Unlike most statistical approaches where analysis is conditioned on a single "best" model, BMA combines information from a class of models to obtain a probability distribution for a quantity of interest and accounts for model uncertainty. Raftery et al. (2005) use BMA for weather forecasting

models to obtain a predictive probability distribution function (pdf) of future temperature. They assess the skill of the individual weather models using hindcasts and forecasts of surface temperature from an ensemble of weather models along with temperature observations. Their approach accounts for bias in the hindcasts and forecasts, but does not account for spatial and temporal dependence. Berrocal et al. (2007) extends Raftery et al. (2005) by using variogram-based techniques to incorporate spatial and temporal dependence. However, since weather forecasting requires fast computational approaches, it is often difficult to estimate space-time dependence using more sophisticated statistical methodology within the computational constraints of weather forecasting.

Our approach expands on these previous studies by developing a computationally tractable Bayesian approach for applying BMA to hindcasts and forecasts from ensembles of climate models. Specifically, we improve upon previous studies by (i) incorporating spatial and temporal dependence using likelihood based approaches, (ii) incorporating uncertainty of the weights and bias-correction parameters using Bayesian approaches, (iii) enabling analysis of larger data sets using dimension reduction approaches. To our knowledge, this is the first BMA-based approach applied to climate projection from large spatial output from a multimodel ensemble that accounts for space-time dependence.

In Section 2, we discuss current approaches for incorporating uncertainty from multiple models. We describe the basics of the BMA approach for multiple climate models in Section 3, and introduce our computationally efficient approach for combining large space-time output from multi-model ensembles. Section 4 applies our approach to two sets of ensemble output from AR4 models. We conclude in Section 5 with a discussion and caveats.

2 Current Statistical Approaches for Combining Multiple Models for Climate Projections

In this section, we discuss current statistical approaches for deriving future climate projections by incorporating model variability from a multi-model ensemble while accounting for uncertainty from these climate models. Complex climate model output has many sources of uncertainty, including

(1) incomplete knowledge of the climate system (e.g., climate sensitivity to a doubling of CO_2; cf. Knutti and Hegerl (2008)), (2) computational limits on small-scale physical processes that can be simulated (such as cloud formation and extent, cf. Williams and Tselioudis (2007)), (3) inability to perfectly forecast a chaotic system (Lorenz, 1963; Palmer, 2001). In addition, there is uncertainty about projections of future climate change because it is not known how the most recent dominant climate forcing (i.e., anthropogenic greenhouse gas emissions) will evolve through time (Meehl et al., 2007). The most common strategy to address this uncertainty is through scenario development. For example, a set of emission scenarios were created for the Third and Fourth IPCC Assessment Reports (also known as the Special Report on Emission Scenarios or SRES) based on different technological and economic assumptions (Nakićenović et al., 2001). Notably, these are pure scenario exercises; that is, no likelihood was attached to any particular scenario. See Section 4 for more discussion on these scenarios.

Much research has been done on deriving probability distributions of future temperature projections (see Allen et al., 2000; Webster et al., 2003; Wigley and Raper, 2001). Some of this research implicitly uses Bayesian methodology by using subjective priors on model parameters in their analysis (see Forest et al., 2002; Webster et al., 2003). Räisänen and Palmer (2001) used multi-model ensembles to derive probability distributions of future projections, but did not directly account for the variability or bias of individual models.

As described by Knutti et al. (2010), there are two general approaches for the problem of combining output from multi-model ensembles. One is that the different models should not be combined or compared and that any ensemble averages or variability measures should treat all the models equally. The second approach is that some models are "better" than others, and the "better" models should receive more weight when the models are combined. In general, the second approach where model weights are determined by model skill performs better than equally weighting all models (e.g. Doblas-Reyes et al., 2003; Tebaldi and Knutti, 2007; Yun et al., 2003). However, there is no consensus for a single approach for determining these skill-based model weights using quantitative methods. We briefly describe one such approach, the Reliability Ensemble Average (REA) approach (Giorgi and Mearns, 2003), which serves as a basic mathematical framework for much of

the statistical methodology in analyzing multi-model ensembles. The REA approach improves on previous attempts to characterize bias and uncertainty in combining multiple models and serves as a framework that continues to be expanded by others. It gives increased weight to models which have smaller bias and better agreement with the other models and the observed climate. While the REA estimator is robust under certain conditions (Nychka and Tebaldi, 2003), this approach lacks an explicit probabilistic foundation. Tebaldi et al. (2005) extended the REA approach with a formal statistical framework using Bayesian methods by allowing the modeled historical data, hindcasts, and forecasts from different models to share common mean and variance parameters. They applied their approach to regionally averaged surface temperature data for 22 regions and nine Regional Climate Models (RCM), and assumed that hindcasts and forecasts from different regions are independent. Smith et al. (2009) extends this approach in a multivariate framework, relaxing the assumption of independence between the regions by modeling the hindcasts and forecasts from different regions to share common mean and variance parameters. Tebaldi and Sansó (2009) proposes an approach with a similar flavor but several substantially novel features to model a bivariate response of surface temperature and precipitation, using globally and decadally averaged ensembles from 18 GCMs over six decades in the past and nine future decades. Berliner and Kim (2008) develop an approach where climate models themselves are assumed to be drawn from a superensemble of possible models. This framework results in noisy hindcasts and forecasts while the physical observations are assumed to be known.

Many of these approaches are computationally efficient due to heavy aggregation. However by using global or regional space-time aggregation, they lose much of the data richness. Further, they do not distinguish between regions or times, and do not explicitly model spatial or temporal dependence. In addition, these models require many assumptions about independence among variance parameters to ensure identifiability; in the most general case without any independence assumptions, there are just two fewer parameters than data points. While these approaches may be extended to incorporate spatial and temporal dependence, computational tractability and identifiability would still be difficult to resolve for the large spatial fields that are the interest of this paper. Here we use a different approach, Bayesian Model Averaging (BMA), to combine information from

multi-model ensembles and address these issues.

3 A BMA Approach for Space-Time Climate Projection

In this section, we begin by describing the motivation and the basic ideas of our Bayesian Model Averaging (BMA) approach for climate model ensemble output following the ideas of Raftery et al. (2005) and Berrocal et al. (2007). We then discuss our computationally tractable Bayesian approach for incorporating space-time dependence for large data sets using a Bayesian approach.

3.1 Bayesian Model Averaging Basics

Statistical analyses usually assume that there is a single "best" model, often selected from a class of several possible models. Although it is typically not the case that the selected model is always the best, the analysis ignores model uncertainty and assumes that the data are obtained from the selected model. This results in overconfident inference and predictions, especially when substantially different scientific results may be obtained from alternative models (see Hoeting et al., 1999). For a more detailed treatment of model uncertainty, see Draper (1995), Kass and Raftery (1995), and Chatfield (1995). Bayesian Model Averaging reduces the potential overconfidence by conditioning not on a single model, but on a class of models. Specifically, if we are interested in a projection of a quantity Z from training data Z^h and models $M_1 \cdots M_K$, the probability distribution of Z may be derived using the law of total probability as follows,

$$p(Z) = \sum_{k=1}^{K} p(Z \mid M_k) p(M_k \mid Z^h),$$

where $p(Z \mid M_k)$ is the projection of Z conditioned on choosing the model M_k and $p(M_k \mid Z^h)$ is the posterior probability that M_k is the best model given the training data. The latter can also be interpreted as the skill of model M_k with respect to the training data. For computational ease, it might be helpful to retain only models whose skill exceeds a particular threshold when K is large (Madigan and Raftery, 1994; Volinsky et al., 1997). The BMA framework has both strong theoretical properties and performs well for both simulated and real data applications (Raftery and Zheng, 2003).

The BMA approach may also be extended to non-statistical models, such as climate models, providing a statistical framework to quantify model uncertainty. Again, the fundamental concept is that while there is a "best" model among this class of models, it is not easy to identify such a model. Raftery et al. (2005) introduced the use of BMA in the context of weather forecasting to obtain probability density functions (pdfs) for projections of future weather characteristics such as surface temperature. The BMA pdf is a weighted average of the pdf of the forecast of a particular attribute from the individual models of the ensemble. The measurements at different locations are assumed to be independent, and spatial error correlations are not considered.

Consider the problem of projecting future surface temperature. Let $Z(\mathbf{s}, t)$ be the projection of surface temperature at a location $\mathbf{s}$ in space and a time t in the future. Let $Z^h(\mathbf{s}, t)$ be the historical surface temperature at a location $\mathbf{s}$ in space and a time t in the past. Similarly, let $Y_k^f(\mathbf{s}, t)$ and $Y_k^h(\mathbf{s}, t)$ be the forecast or hindcast from model M_k, depending on t. For a particular location $\mathbf{s}$ and time t in the future (for notational convenience we drop the space-time indices), $p(Z) = \sum_{k=1}^{K} w_k g_k(Z|Y_k^f)$, where $g_k(Z|Y_k^f)$ is the projection of surface temperature given that the model M_k is the best model, and thus $w_k = p(M_k|Z)$ is the BMA weight given to model M_k.

The BMA weights w_k quantify the relative skill of the model by comparing hindcasts with historical data. Following Raftery et al. (2005), for a particular location $\mathbf{s}$ and time t in the past, $Z^h \mid Y_k^h \sim N(a_k + b_k Y_k^h, \sigma_k^2)$, where a_k and b_k are bias correction terms and σ_k^2 is the BMA variance associated with model M_k. The bias correction terms may be estimated using the regression of Z^h on Y_k^h for all past space-time locations. The BMA weights and parameters are obtained by using maximum likelihood methods (Fisher, 1922) using historical data and hindcasts (Z^h and Y^h). The log-likelihood for estimating these parameters is as follows:

$$\ell(a_1 \cdots a_K, b_1 \cdots b_K, \sigma_1^2, \cdots \sigma_K^2) = \sum_{\mathbf{s},t} \sum_{k=1}^{K} w_k \sigma_k^{-1} \exp\left(-\frac{1}{2\sigma_k^2}(Z^h(\mathbf{s}, t) - a_k - b_k Y_k^h(\mathbf{s}, t))^2\right).$$

Raftery et al. (2005) also assumed that $\sigma_k^2 = \sigma^2$ for all models, although they note that this assumption may be relaxed. When simple optimization algorithms fail to converge to a global maximum for this likelihood, one may use other approaches such as an expectation-maximization (EM) algorithm (Dellaert, 2002; Dempster et al., 1977; Hartley, 1958) or a differential evolution

algorithm (Ardia, 2007; Storn and Price, 1997). Using the parameter estimates, the BMA predictive mean and variance can be obtained, and a pdf for the projection of surface temperature can be derived at each future space-time location. The BMA predictive mean and variance, as derived in Raftery et al. (2005), are as follows,

$$\mathrm{E}(Z(\mathbf{s},t) \mid Y_1^f(\mathbf{s},t) \cdots Y_k^f(\mathbf{s},t)) = \sum_{k=1}^{K} w_k(a_k + b_k Y_k^f(\mathbf{s},t)), \text{and}$$

$$\mathrm{Var}(Z(\mathbf{s},t) \mid Y_1^f(\mathbf{s},t) \cdots Y_k^f(\mathbf{s},t)) = \sum_{k=1}^{K} w_k \left((a_k + b_k Y_k^f(\mathbf{s},t)) - \sum_{i=1}^{K} w_i(a_i + b_i Y_i^f(\mathbf{s},t)) + \sigma_k^2 \right)^2.$$

3.2 Bayesian Model Averaging for Large Spatial Data

We now describe our approach to combine multi-model ensembles of AR4 GCM models to obtain probabilistic projections of surface temperature on a high resolution global scale while accounting for spatial and temporal dependence in a computationally efficient manner. In Berrocal et al. (2007) and Berrocal et al. (2008), the BMA approach in Raftery et al. (2005) was expanded to incorporate spatial dependence in the context of weather forecasting. However, due to the need for very fast computational results in the field of weather forecasting, Berrocal et al. (2007) use empirical variograms to estimate spatial parameters. Since our requirements for speedy results are substantially less, we expand on their spatial BMA approach to incorporate more sophisticated statistical techniques such as likelihood methods and Gaussian processes to estimate spatial and temporal parameters and then use a Bayesian approach to infer the BMA weights, bias, and non-spatial variance terms. Due to the large amount of data from hindcasts, forecasts, and historical data, we use dimension reduction techniques such as kernel mixing and matrix identities to ensure computational tractability as described in this section.

Following Berrocal et al. (2007), we model the space-time field of historical data for a single variable below as follows,

$$f(\mathbf{Z} \mid \mathbf{Y}_1^h \cdots \mathbf{Y}_k^h) = \sum_{k=1}^{K} w_k g_k(\mathbf{Z} \mid \mathbf{Y}_k^h),$$

where $\mathbf{Z}$ is the vector of historical data, $\mathbf{Y}_k^h$ are the hindcasts for model M_k, $w_k = p(M_k \mid \mathbf{Z})$ is the BMA weight, which is the probability that M_k is the best model given the historical data, and

$g_k(\mathbf{Z} \mid \mathbf{Y}_k^h)$ is the conditional distribution of $\mathbf{Z}$ given that M_k is the "best" model. We then model $g_k(\mathbf{Z} \mid \mathbf{Y}_k^h)$ in Equation (1), as

$$g_k(\mathbf{Z} \mid \mathbf{Y}_k^h) \sim a_k \mathbf{1} + b_k \mathbf{Y}_k^h + \boldsymbol{\delta}_k + \boldsymbol{\epsilon}_k, \tag{1}$$

where a_k and b_k are correction terms for additive and multiplicative bias, $\boldsymbol{\delta}_k$ represents space-time dependence and $\boldsymbol{\epsilon}_k$ represents non-spatial error for model M_k. We model $\boldsymbol{\epsilon}_k$ as iid error terms, i.e., $\boldsymbol{\epsilon}_k \sim N(\mathbf{0}, \psi_k \mathbf{I})$, with $\psi_k > 0$, while we model $\boldsymbol{\delta}_k$ as a zero mean linear Gaussian process with a squared exponential covariance function. We provide descriptions of important parameters in Table 1.

Since the size of the data poses substantial computational hurdles, we model $\boldsymbol{\delta}_k$ using a kernel mixing approach. We briefly describe the basics of the kernel mixing approach (Higdon, 1998), which is especially useful for modeling space-time dependence. Here, we use the idea that a continuous process z can be created by convolving a continuous white noise process ω with a convolution kernel k. However, we approximate the continuous white noise process ω using a finite sum approximation ω defined on a lattice $\mathbf{u}_1, \ldots, \mathbf{u}_J$ that spans the region of interest. $\mathbf{u}_j$ is denoted as a "knot location". Then

$$z(\mathbf{s}) = \sum_{j=1}^{J} k(\mathbf{s} - \mathbf{u}_j)\omega(\mathbf{u}_j),$$

where $\omega(\mathbf{u}_j)$ is the value of the white noise process at location $\mathbf{u}_j$. We use a slightly modified version of the following kernel, which corresponds to a squared exponential covariance function,

$$k(\mathbf{u}) = \kappa \exp\left\{-\frac{\|\mathbf{u}\|^2}{\phi}\right\}.$$

Stein (1999) cautions against the use of this kernel due to its smoothness. However, based on our exploratory analysis and the fact that we may assume that the underlying climate model output is a smooth process, this kernel appears to be a reasonable assumption here.

We apply kernel mixing to a large data set with N space-time locations in order to obtain substantial computational gains by reducing the dimension of the matrices to be inverted from $N \times N$ to $J \times J$. Our set of knots are $((\mathbf{u}_1, v_1), \ldots, (\mathbf{u}_J, v_J))^T$, where $(\mathbf{u}_j, v_j)$ is the location (latitude and longitude) and the time of the jth knot location. The $\omega(\mathbf{u}_j, v_j)$'s are the white noise

process at the jth knot and are normally distributed with zero mean and variance 1. These knots define a lattice over the entire region of interest. We note that we do not actually estimate these knot processes, and instead use the marginal covariance after integrating out the $\omega(\mathbf{u}_j, v_j)$'s. The marginal covariance matrix has the same dimensions as the original covariance, $N \times N$. However due to the special matrix structure of the marginal covariance matrix, we can apply the Sherman-Morrison-Woodbury identity so inversions involve matrices of dimension of $J \times J$ instead. This results in dramatic savings in computational effort. We discuss the selection of the number and location of the knots in the context of the ocean tracer case study in Section 4.2.

Using kernel mixing, we model the spatial component, $\boldsymbol{\delta}_k \mid \phi_{s_k}, \phi_{t_k}, \kappa_k \sim N(\mathbf{0}, K_k K_k^T)$. The kernel function K_k is described as follows,

$$K_k(i,j) \mid \phi_{s_k}, \phi_{t_k}, \kappa_k = \sqrt{\kappa_k} \exp\left(-\frac{\|\mathbf{s}_i - \mathbf{u}_j\|^2}{\phi_{s_k}^2} - \frac{|t_i - v_j|^2}{\phi_{t_k}^2} \right)$$

where $\kappa_k, \phi_{s_k}, \phi_{t_k} > 0$. The kernel matrix K_k has dimensions $N \times J$.

We have now specified the entire model, and in theory we can infer the BMA weights, bias-correcting, spatial, and non-spatial parameters above in a fully Bayesian approach, after specifying prior distributions. However, such an approach may impose prohibitive computational requirements and result in identifiability issues. To mitigate these challenges, we first estimate the spatial parameters $\kappa_k, \phi_{s_k}, \phi_{t_k}$ and the non-spatial BMA variance ψ_k and bias-correcting parameters a_k and b_k separately using maximum likelihood on the conditional distribution $g_k(\mathbf{Z} \mid \mathbf{Y}_k^h)$. In other words, we estimate these parameters using only the historical data and the output from model M_k. To compute MLEs, we need to invert an $N \times N$ matrix for each evaluation of the likelihood, which is computationally expensive since N is quite large. By using kernel mixing, we obtain a covariance matrix with a specific structure, $\Sigma = \psi_k \mathbf{I} + K_k K_k^T$. By rewriting the inverse of this matrix and using the Sherman-Morrison-Woodbury identity (see Appendix A), the matrix inversions are on a reduced $J \times J$ matrix, where we choose J=252 for our dataset (see Section 4.2). This approach has the additional advantage of being highly parallelizable, which further improves computational efficiency.

After obtaining point estimates for the spatial parameters, we use a fully Bayesian approach

for sample based inference on the w_k (BMA weights), the ψ_k (BMA variances), and the a_k and b_k (bias-correcting parameters), resulting in estimating $4K$ parameters. Using Markov Chain Monte Carlo (MCMC), we can estimate the posterior distribution for these parameters after specifying prior distributions. Here we re-estimate the non-spatial variances and the bias-correcting parameters, and use the maximum likelihood estimates already obtained to select priors for these parameters (prior selection are discussed in Section 4.2, and are provided in Table 1.). Having obtained sample-based inference for the parameters above, we obtain probabilistic projections and uncertainty estimates at future space-time locations via the standard kriging methods using the bias-corrected forecasts and the BMA weights. For each sample of the posterior distribution of the parameters, we compute the predictive distribution, which is a weighted average of multivariate normals. and sample from that predictive distribution. We thus obtain a sample-based probability distribution of the projected surface temperatures for a set of space-time locations. For computational expediency, one may use the posterior mean of the parameters to obtain the pdf for future surface temperature.

4 Application of BMA Approach to AR4 GCM Hindcasts and Forecasts

In this section, we apply our computationally efficient BMA approach to low resolution spatial output from an ensemble of three different AR4 GCMs described below to project surface temperature in the future. We also apply our approach to a multi-model ensemble of 20 different GCMs at a finer spatial grid scale.

4.1 Climate Data Description

We first apply our approach to combine the output of three AR4 GCMs (specifically, ECHO, ECHAM5, and GISS (see Meehl et al., 2007)). Major climate forcings such as anthropogenic greenhouse gas emissions, anthropogenic and volcanic aerosols, and solar luminosity changes are included in these simulations. We use a bilinear interpolation (cf. Tribbia, 1997) to obtain output from these GCMs to a resolution of 10° by 10° grid boxes, yielding 648 spatial grid cells, and we use

the centroid of the grid cell as the spatial location. Forecasts from these models are computed under three different emissions scenarios: A1B, A2, and B1 (Meehl et al., 2007), from 2004 until 2100, at the same global 10° by 10° resolution. Scenario A1B assumes strong economic growth, a globalized economy with converging income levels between nations, a global population of 9 million in 2050 but stable or decreasing afterwards, and reliance on both fossil-fuels and non-fossil energy sources. Scenario A2 assumes a less globalized and more localized economy, an increasing global population, and fewer technological changes. Scenario B1 is similar to Scenario A1B, however, it assumes that more aggressive action is taken to conserve energy through efficiency and reduced fossil fuel use over time (see Nakićenović et al., 2001, for more information on these scenarios). We use historical data from the University of East Anglia Climate Research Unit HadCRUT3 dataset (Brohan et al., 2006), for surface temperature between 1880 and 2000 that are interpolated to the hindcast grid locations. However, historical data are missing for many locations in the HadCRUT3 dataset, and for computational ease we limit our study to locations where both hindcasts and historical data are available. Many of these missing locations are at very high or low latitudes (above 80°N or below 70°S) or before 1900. Approximately 31% of possible spatial locations have no data for any time point, and were discarded, resulting in N=28,058 space-time locations. All hindcasts, forecasts, and historical data are in terms of anomalies, which are deviations from data during a reference time period. Plots of the surface temperature anomalies forecasts in 2100 for the three GCMs are shown in Figure 1.

We also apply our BMA approach to an ensemble of twenty AR4 GCM hindcasts and forecasts and historical data on a much finer scale. The observed historical data, obtained from the NASA Goddard Institute for Space Studies (GISS, Hansen et al. (2010)), are gridded to 2° by 2° cells. We use a bilinear interpolation to obtain AR4 GCM output at the locations of GISS observational dataset. We believe this simple interpolation method is appropriate given the scale of aggregated observations and the method in which the GCM output are produced (i.e. the finite difference method, Tribbia, 1997). And while a few of the GCMs are at a coarser resolution than the grid of observations, there is no reason to assume that this would result in spurious results indicating higher model fidelity to the observations than is actually the case. Again, we limit our study to

locations where both hindcasts and historical data are available as many of these missing locations are at very high or low latitudes or before 1900. Approximately 16% of possible space-time locations were discarded, resulting in N=1,225,332 space-time locations. For this data set, both the GCM hindcasts and GISS historical data are referenced annually between 1900 and 2000. The average annual surface temperature anomaly between the years 1900 and 2000 is shown in Figure 2. There appears to be a 0.7° global temperature increase over the last 100 years and substantial variability among the models (Figure 2).

4.2 Implementation Details of BMA Approach

We next discuss some of the details of the application of our BMA approach to the AR4 GCM ensemble hindcasts and historical data. We account for the curvature of the earth by using a geodesic distance formula to determine the distance between locations (see Banerjee, 2005). The kernel mixing approach greatly reduces the computational challenges since the matrix computations depend on the number of knot locations selected rather than on the number of spatial locations. To ensure that the knots for kernel mixing span the entire space-time field, we chose 7 equally spaced latitude locations, 9 longitude locations, and 4 different times. We selected each combination of these values as a knot, for a total of J=252 knots. Knots in space were selected with the help of the heuristic described in Short et al. (2007), and the median of the spatial range parameter was greater than the smallest knot-to-knot distance (of 63 knots) in space. Analysis of a subset of this data using twice as many temporal knots (and thus an increase in the total number of knots from J= 252 to 504) gave similar inferential results. Furthermore, for two randomly selected GCMs, the Bayesian Information Criterion (BIC) in fitting the conditional distribution $g_k(\mathbf{Z} \mid \mathbf{Y}_k^h)$ to the GCM output using J=504 knots was actually slightly greater than using J=252 knots.

We specify a Dirichlet$(0.05 \cdots 0.05)$ prior for the BMA weights, which induces a discrete uniform distribution for all valid selections of BMA weights. However, the Dirichlet prior requires that all weights are positive, and this prior puts near zero mass on a set of weights where very poor models receive infinitesmal weights. We use a wide inverse gamma prior of $\psi_k \sim IG(2, 0.6)$ for the non-spatial variance terms. Flat priors were used for the bias-correcting terms a_k and b_k. For some

parameters which were estimated in the first stage, priors were guided by the point estimates obtained from likelihood methods.

MCMC was run for 200,000 iterations, allowing for a burn-in of 10,000 samples. To ensure convergence of our MCMC based estimates in the second stage, we obtained Monte Carlo standard errors for the posterior estimates of parameters computed by consistent batch means (Flegal et al., 2008; Jones et al., 2006). The posterior mean estimates of these parameters had MCMC standard errors below 10^{-3}. We implemented the computer code in R (Ihaka and Gentleman, 1996) using a 3.0 GHz Intel Xeon on a Dell PowerEdge server with 32GB of RAM. We computed MLE in the first stage using the differential evolution algorithm (Ardia, 2007; Storn and Price, 1997) for each of the 20 GCMs in parallel. For the ensemble of twenty AR4 GCMs, computing the MLE of spatial and temporal covariance parameters and the 200,000 samples using MCMC required approximately 5 hours and about 80 hours of computer time, respectively. We obtained the MLE for six parameters for each GCM, and we obtained the posterior distribution of the model weight, additive and multiplicative biases, and non-spatial variance, resulting in a total of 80 parameters. For our data set of N=1,225,332 space-time locations, computation of the MLE and posterior distribution would have required many years without our methods to enhance computational tractability.

We used a cross-validation approach in order to verify that our BMA approach is appropriate for this data. We performed cross-validation for spatial modeling by holding out three different 20° by 20° spatial regions for all years, resulting in more than 1400 holdout locations for each of the regions. The latitude/longitude coordinates of centroids of the three spatial regions held out are (40,-40), (30,130),(-50,-120). The coverage probabilities using a 90% credible region are 0.91, 0.87, and 0.81 respectively. We also hold out three ten year time periods to verify the average yearly predictions. The time periods held out are 1910-1920, 1950-1960, and 1980-1990, and the coverage probabilities were 0.89, 0.84, and 0.92, respectively (Table 3). The cross-validation results seem to suggest that our model does a reasonable job of fitting the data well.

To justify the need for spatial modeling, we compare our approach to a non-spatial BMA approach. In particular, we compare the fit of the conditional distribution $g_k(\mathbf{Z} \mid \mathbf{Y}_k^h)$ to the kth GCM output using models with and without space-time effects. We discovered that the Bayesian

Information Criterion was smaller for the model including space-time effects than the alternate model without space-time effects for all 20 GCMs. Furthermore, for many GCMs, the BIC value was substantially smaller for the model including space-time effects, justifying the need for spatial modeling. Our ability to compare the fit of alternative models to the GCM output highlights an advantage of using a hierarchical approach. In addition, when space-time effects were ignored, we obtained wider intervals for the BMA weights, such that it was impossible to say with statistical confidence that any model received more weight than any other model.

4.3 Results

In this section we first present and discuss the results obtained from applying our BMA approach to the hindcasts and forecasts from the ensemble of three AR4 GCMs and the HadCRUT3 historical data. We obtain sample-based projections and 90% credible regions for the 10° by 10° grid of locations from 2004-2100 for all three scenarios above. For Scenario A1B, we found a statistically significant positive surface temperature anomaly for areas outside the Southern Ocean (i.e. areas above 50°S, see Figure 3), as there are positive anomalies for both the lower and upper bound of the 90% credible region. For Scenario A2, we found a significant positive surface temperature anomaly for areas outside the Southern Ocean and eastern Russia. For Scenario B1, we found a significant positive surface temperature anomaly in 2100 for areas outside the Southern Ocean and eastern Russia.

We then computed probabilistic projections and uncertainty estimates for a globally averaged annual surface temperature anomaly. To obtain an estimate for the globally averaged annual surface temperature, we take an average of surface temperature at all spatial locations in that year, weighted by the cosine of the latitude to account for different areas of the cells. The average annual surface temperature anomaly in 2100 is between 2 and 3.5°C, 2.8 and 4.2°C, and 1.4 and 2.6°C under scenarios A1B, A2, and B1 respectively (see Figure 4(a)-(c)). As expected, average annual surface temperatures increase much more slowly under scenario B1 than under scenario A1B or A2. (see Figure 4(d)).

We also apply our BMA approach to combine information of the hindcasts and forecasts for

20 AR4 GCMs and the GISS historical data at a much finer scale for a single scenario A1B. The posterior model weights and their 90% credible regions are shown in Table 2. The BMA derived model weights vary across the considered GCMs, ranging from approximately two to eight percent of the total weight (see Table 2) with no obvious clustering. We found a significant positive surface temperature anomaly for areas outside the Southern Ocean and the North Atlantic after combining the information from the multi-model ensemble of GCMs (see Figure 5). The average annual surface temperature anomaly in 2100 under scenario A1B after combining information from these 20 GCMs is between 2.0 and 3.8°C (see Figures 6 and 7).

5 Discussion

We develop an improved approach to combine large spatial output from a multi-model ensemble of GCMs using Bayesian Model Averaging, while accounting for spatial and temporal dependence in a computationally tractable manner. By using an ensemble from a class of AR4 GCMs rather than a single GCM, we have incorporated structural model uncertainty in our approach and obtained a better quantification of uncertainty for an important climate projection problem. Our approach improves on previous work by using a likelihood approach to estimate spatial and temporal parameters and a Bayesian approach to estimate BMA weights, non-spatial variances, and bias-correcting parameters in a computationally tractable manner. Our approach, in theory, not only provides improved projections and a more complete picture of uncertainty, but also allows for the comparison of the skill of the climate models in the ensemble.

The derived model weights (Table 2) decay relatively smoothly from the highest to the lowest weight and the highest ranking model receives less than four times the weight as the lowest ranking model. This suggests that, at least for the considered observations, all the models contribute considerably (but unevenly) to the projection. This ranking is, of course, not a substitute for a careful and mechanistically motivated analysis of the physical realism of the different model; more detailed analyses (e.g., as synthesized by Randall et al. (2007)) are needed for this.

One possible concern is that the BMA weights of all models are required to be positive due to

the Dirichlet prior on the BMA weights. Further, a Dirichlet prior discourages very low weights even for very poor models, which result in even poor models receiving non-zero weights, making it hard to completely rule out the influence of such models on projections. Hence, we may be unable to determine a poor model with statistical certainty, and we may need to resort to likelihood approaches to determine whether there are models that can be discarded. One scientific caveat that we note is that for some climate models, the hindcasts are on a coarser scale than the historical data, and regridding to a finer scale may ignore small scale effects.

In this work, we have combined the information for only a single variable from multimodel ensembles of the same resolution, i.e., coupled atmosphere-ocean GCMs. Climate model output today often comes from multivariate spatial fields, and there is a need to expand our framework to allow for analysis of large multivariate spatial fields from multi-model ensembles with more models.

In addition to the GCM output analyzed in this study, one could also add information from Earth System Models of Intermediate Complexity (EMIC). EMICs use a coarser spatial resolution than the GCMs used in this analysis. While EMICs are generally less accurate at small scales the GCMs analyzed here, they are may be sufficient for making useful projections at coarser scales, and these projections can be obtained at a much reduced computational demand. Hence, it may be useful to analyze an expanded ensemble of that includes additional EMIC projections, to test whether adding these EMIC runs improves the hindcast skill and/or the out-of-sample cross-validation errors. Another issue that is neglected so far is the quantification of structural forecast uncertainty, which results in the inaccurate assumption that forecasts in 2100 or beyond are equally uncertain as a forecast in 2050.

6 Acknowledgements

This work was partially supported by the National Science Foundation and from the US Geological Survey. Any opinions, findings, and conclusions expressed in this work are those of the authors alone, and do not necessarily reflect the views of the NSF and USGS. The authors also thank Nathan Urban, Veronica Berrocal, Kary Myers, Jim Gattiker, Dave Higdon, and Matt Pratola for

helpful insights.

Appendix A: Matrix Identities

The Sherman-Morrison-Woodbury identity states that the inverse of a matrix of the form $A+UCV$, where A is of dimension $N \times N$, U is dimension $N \times J$, V is dimension $J \times N$, and C is dimension $J \times J$ can be expressed as follows,

$$(A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

The determinant of a matrix $A+UCV$ can be expressed as the follows,

$$|A+UCV| = |C^{-1} + VA^{-1}U| \times |C| \times |A|$$

using the matrix determinant lemma (Harville, 2008). This identity reduces matrix inversions and determinant computations to dimension J rather than N (cf. Golub and Van Loan, 1996, p. 50).

The matrix form $(\psi I_N + KK^T)$ comes up regularly in our computations, for which we obtain the inverse and determinant (using Sylvester's Theorem, see Golub and Van Loan (1996)) below in equation (2), which only require computations of matrices on dimension $J \times J$:

$$\begin{aligned}(\psi I_N + K(I_J)^{-1}K^T)^{-1} &= \frac{I_N}{\psi} - \frac{K}{\psi}\left(I_J - \frac{K^TK}{\psi}\right)^{-1}\frac{K^T}{\psi}, \text{and} \\ |\psi I_N + K(I_J)^{-1}K^T| &= \left|I_J - \frac{K^TK}{\psi}\right| \cdot \psi^N.\end{aligned} \tag{2}$$

To compute the likelihoods in this paper, we compute the Cholesky decomposition of the matrix $\left(I_J - \frac{K^TK}{\psi}\right)$ rather than the inverse directly, which also reduces the computational cost of the determinant (Golub and Van Loan, 1996).

References

Allen, M., Stott, P., Mitchell, J., Schnur, R., and Delworth, T. (2000). Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature*, 407(6804):617–620.

Ardia, D. (2007). The DEoptim package: Differential Evolution Optimization. *R Foundation for Statistical Computing.*

Banerjee, S. (2005). On Geodetic Distance Computations in Spatial Modeling. *Biometrics*, 61(2):617–625.

Berliner, L. and Kim, Y. (2008). Bayesian Design and Analysis for Superensemble-based Climate Forecasting. *Journal of Climate*, 21:1891–1910.

Berrocal, V., Raftery, A., and Gneiting, T. (2007). Combining Spatial Statistical and Ensemble information in probabilistic weather forecasts. *Monthly Weather Review*, 135:1386–1402.

Berrocal, V., Raftery, A., and Gneiting, T. (2008). Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *Annals*, 2(4):1170–1193.

Brohan, P., Kennedy, J., Harris, I., Tett, S., and Jones, P. (2006). Uncertainty estimates in regional and global observed temperature changes: a new dataset from 1850. *J. Geophys. Res*, 111(D12).

Cantelaube, P. and Terres, J. (2005). Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A*, 57(3):476–487.

Chatfield, C. (1995). Model Uncertainty, Data Mining and Statistical inference. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 158(3):419–466.

Cressie, N. and Johannesson, G. (2008). Fixed Rank Kriging for Very Large Spatial Data Sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Cressie, N. A. (1993). *Statistics for Spatial Data*. John Wiley & Sons, New York, 2nd. edition.

Dellaert, F. (2002). The expectation maximization algorithm. *Georgia Institute of Technology, Technical Report Number GIT-GVU-02-20*.

Dempster, A., Laird, N., Rubin, D., et al. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Doblas-Reyes, F., Pavan, V., and Stephenson, D. (2003). The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Climate Dynamics*, 21(5):501–514.

Draper, D. (1995). Assessment and Propagation of Model Uncertainty. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):45–97.

Fisher, R. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222:309–368.

Flegal, J., Haran, M., and Jones, G. (2008). Markov Chain Monte Carlo: Can We Trust the Third Significant Figure? *Statist. Sci*, 23(2):250–260.

Forest, C., Stone, P., Sokolov, A., Allen, M., and Webster, M. (2002). Quantifying Uncertainties in Climate System Properties with the use of Recent Climate Observations. *Science*, 295(5552):113–117.

Furrer, R., Genton, M., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 5:502–523.

Giorgi, F. and Mearns, L. (2003). Probability of regional climate change based on the Reliability Ensemble Averaging (REA) method. *Geophysical Research Letters*, 30(12):1629.

Gleckler, P., Taylor, K., and Doutriaux, C. (2008). Performance metrics for climate models. *Journal of Geophysical Research*, 113(D6):D06104.

Golub, G. and Van Loan, C. (1996). *Matrix Computations*. Johns Hopkins University Press.

Hansen, J., Ruedy, R., Sato, M., and Lo, K. (2010). Global surface temperature change. *NASA Goddard Institute for Space Studies. New York. Disponível em: http://data. giss. nasa. gov/gistemp/paper/gistemp2010_draft0803. pdf.*

Hartley, H. (1958). Maximum likelihood estimation from incomplete data. *Biometrics*, 14(2):174–194.

Harville, D. (2008). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag New York Inc.

Higdon, D. (1998). A process-convolution approach to modelling temperatures in the North Atlantic Ocean (Disc: p191-192). *Environmental and Ecological Statistics*, 5:173–190.

Hoeting, J., Madigan, D., Raftery, A., and Volinsky, C. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314.

IPCC (2007). Climate change 2007: the physical science basis.

Jones, G. L., Haran, M., Caffo, B. S., and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association*, 101:1537–1547.

Kass, R. and Raftery, A. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430).

Knutti, R., Cermak, J., Furrer, R., Tebaldi, C., and Meehl, G. (2010). Challenges in combining projections from multiple climate models. *Journal of Climate*.

Knutti, R. and Hegerl, G. (2008). The equilibrium sensitivity of the Earth's temperature to radiation changes. *Nature Geoscience*, 1(11):735–743.

Leamer, E. and Leamer, E. (1978). *Specification searches: ad hoc inference with nonexperimental data*. Wiley.

Lorenz, E. (1963). Deterministic non-periodic flow. *Atmos. Sci*, 20:130–141.

Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

McAvaney, B., Covey, C., Joussaume, S., Kattsov, V., Kitoh, A., Ogana, W., Pitman, A., Weaver, A., Wood, R., Zhao, Z., et al. (2001). Model evaluation. *Climate Change 2001: The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, pages 471–523.

Meehl, G., Stocker, T., Collins, W., Friedlingstein, P., Gaye, A., Gregory, J., Kitoh, A., Knutti, R., Murphy, J., Noda, A., Raper, S., Watterson, I., Weaver, A., and Zhao, Z.-C. (2007). Climate Change 2007: The Physical Science Basis. *Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*.

Nakićenović, N., Alcamo, J., Davis, G., De Vries, B., Fenhann, J., Gaffin, S., Gregory, K., Gr "ubler, A., Jung, T., Kram, T., et al. (2001). *IPCC Special Report on Emissions Scenarios (SRES)*. Cambridge: Cambridge University Press.

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.

Palmer, T. (2001). A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quarterly Journal of the Royal Meteorological Society*, 127(572):279–304.

Pierce, D., Barnett, T., Santer, B., and Gleckler, P. (2009). Selecting Global Climate Models for Regional Climate Change Studies. *Proceedings of the National Academy of Sciences*, 106(21):8441.

Raftery, A., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174.

Raftery, A. and Zheng, Y. (2003). Discussion. *Journal of the American Statistical Association*, 98(464):931–938.

Räisänen, J. and Palmer, T. (2001). A Probability and Decision-Model Analysis of a Multimodel Ensemble of Climate Change Simulations. *Journal of Climate*, 14:3212–3226.

Randall, D. A., Wood, R. A., S. Bony, R. C., Fichefet, T., J. Fyfe, V., Kattsov, Pitman, A., Shukla, J., Srinivasan, J., Stouffer, R. J., Sumi, A., and Taylor, K. E. (2007). *Climate models and their evaluation, in Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by*

S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K. B. Averyt, M. Tignor and H. L. Miller. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Schabenberger, O. and Gotway, C. (2005). *Statistical Methods For Spatial Data Analysis*. CRC Press.

Short, M. B., Higdon, D. M., and Kronberg, P. P. (2007). Estimation of Faraday Rotation Measures of the Near Galactic Sky Using Gaussian Process Models. *Bayesian Analysis*, 2(4):665–680.

Smith, R., Tebaldi, C., Nychka, D., and Mearns, L. (2009). Bayesian Modeling of Uncertainty in Ensembles of Climate Models. *Journal of the American Statistical Association*, 104(485):97–116.

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

Storn, R. and Price, K. (1997). Differential Evolution–A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization*, 11(4):341–359.

Tebaldi, C. and Knutti, R. (2007). The use of the multi-model ensemble in probabilistic climate projections. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 365(1857):2053.

Tebaldi, C. and Sansó, B. (2009). Joint Projections of Temperature and Precipitation Change from Multiple Climate Models: A Hierarchical Bayesian Approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):83–106.

Tebaldi, C., Smith, R., Nychka, D., and Mearns, L. (2005). Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. *Journal of Climate*, 18(10):1524–1540.

Thomson, M., Doblas-Reyes, F., Mason, S., Hagedorn, R., Connor, S., Phindela, T., Morse, A., and Palmer, T. (2006). Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, 439(7076):576–579.

Tribbia, J. (1997). Weather prediction. In *Economic Value of Weather and Climate Forecasts. R.W. Katz and A.H. Murphy*, pages 1–12. Cambridge University Press.

Volinsky, C., Madigan, D., Raftery, A., and Kronmal, R. (1997). Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 46(4):433–448.

Vrugt, J., Diks, C., and Clark, M. (2008). Ensemble bayesian model averaging using markov chain monte carlo sampling. *Environmental fluid mechanics*, 8(5):579–595.

Webster, M., Forest, C., Reilly, J., Babiker, M., Kicklighter, D., Mayer, M., Prinn, R., Sarofim, M., Sokolov, A., Stone, P., et al. (2003). Uncertainty analysis of climate change and policy response. *Climatic Change*, 61(3):295–320.

Wigley, T. and Raper, S. (2001). Interpretation of high projections for global-mean warming. *Science*, 293(5529):451.

Williams, K. and Tselioudis, G. (2007). GCM intercomparison of global cloud regimes: Present-day evaluation and climate change response. *Climate Dynamics*, 29(2):231–250.

Yun, W., Stefanova, L., and Krishnamurti, T. (2003). Improvement of the multimodel superensemble technique for seasonal forecasts. *Journal of Climate*, 16(22):3834–3840.

Parameter	Description	Prior Distribution (if applicable)
$\mathbf{s}$	Spatial location (Latitude (°N) and Longitude(°E))	
t	Time (Year)	
$Z(\mathbf{s},t)$	Projection of surface temperature	
$Z^h(\mathbf{s},t)$	Observation of surface temperature anomaly	
$Y_k^h(\mathbf{s},t)$	Hindcast value for model M_k	
$Y_k^f(\mathbf{s},t)$	Forecast value for model M_k	
w_k	BMA weight given to model M_k	Dirichlet$(0.05 \cdots 0.05)$
a_k	additive bias for model M_k	flat prior
b_k	multiplicative bias for model M_k	flat prior
σ_k	BMA variance for model M_k	IG around MLE
κ_k	Spatial variance for model M_k	computed using ML
ψ_k	Observational error variance for model M_k	IG(2,0.6)
ϕ_{s_k}	Spatial range for model M_k	computed using ML
ϕ_{t_k}	Temporal range for model M_k	computed using ML

Table 1: Description and prior distributions of statistical model parameters. Many of the space-time parameters were estimated using maximum likelihood and thus have no prior distribution.

Model Number	Model	% of Weight	Lower Bound	Upper Bound
1	gfdl_cm2_1	7.82	5.46	10.20
2	gfdl_cm2_0	7.04	4.46	9.01
3	ncar_ccsm3_0	6.86	5.73	8.03
4	cccma_cgcm3_1	6.64	4.55	9.89
5	iap_fgoals1_0_g	6.61	5.48	7.74
6	ncar_pcm1	5.90	3.91	7.56
7	ipsl_cm4	5.52	3.61	7.63
8	miub_echo_g	5.38	4.18	7.13
9	csiro_mk3_0	5.36	4.59	6.18
10	cccma_cgcm3_1_t63	5.32	4.30	6.63
11	cnrm_cm3	4.84	3.78	5.90
12	mpi_echam5	4.51	3.21	5.68
13	bccr_bcm2_0	4.42	3.20	5.44
14	csiro_mk3_5	4.32	1.88	5.51
15	miroc3_2_medres	4.16	2.48	5.64
16	ukmo_hadcm3	3.83	1.76	5.35
17	ukmo_hadgem1	3.28	0.81	6.78
18	ingv_echam4	3.10	2.01	3.99
19	inmcm3_0	2.75	1.06	4.39
20	mri_cgcm2_3_2a	2.31	0.45	4.00

Table 2: Posterior model weights for the twenty AR4 models derived by hindcast skill.

Region/Time Period	Latitude	Longitude	Years	Coverage Probability
1	30°N-50°N	30°W-50°W	all	0.9133
2	20°N-40°N	120°E-140°E	all	0.8688
3	40°S-60°S	110°W-130°W	all	0.8104
1	all	all	1910-1920	0.8945
2	all	all	1945-1955	0.8358
3	all	all	1980-1990	0.9176

Table 3: Coverage probabilities of cross-validation regions.

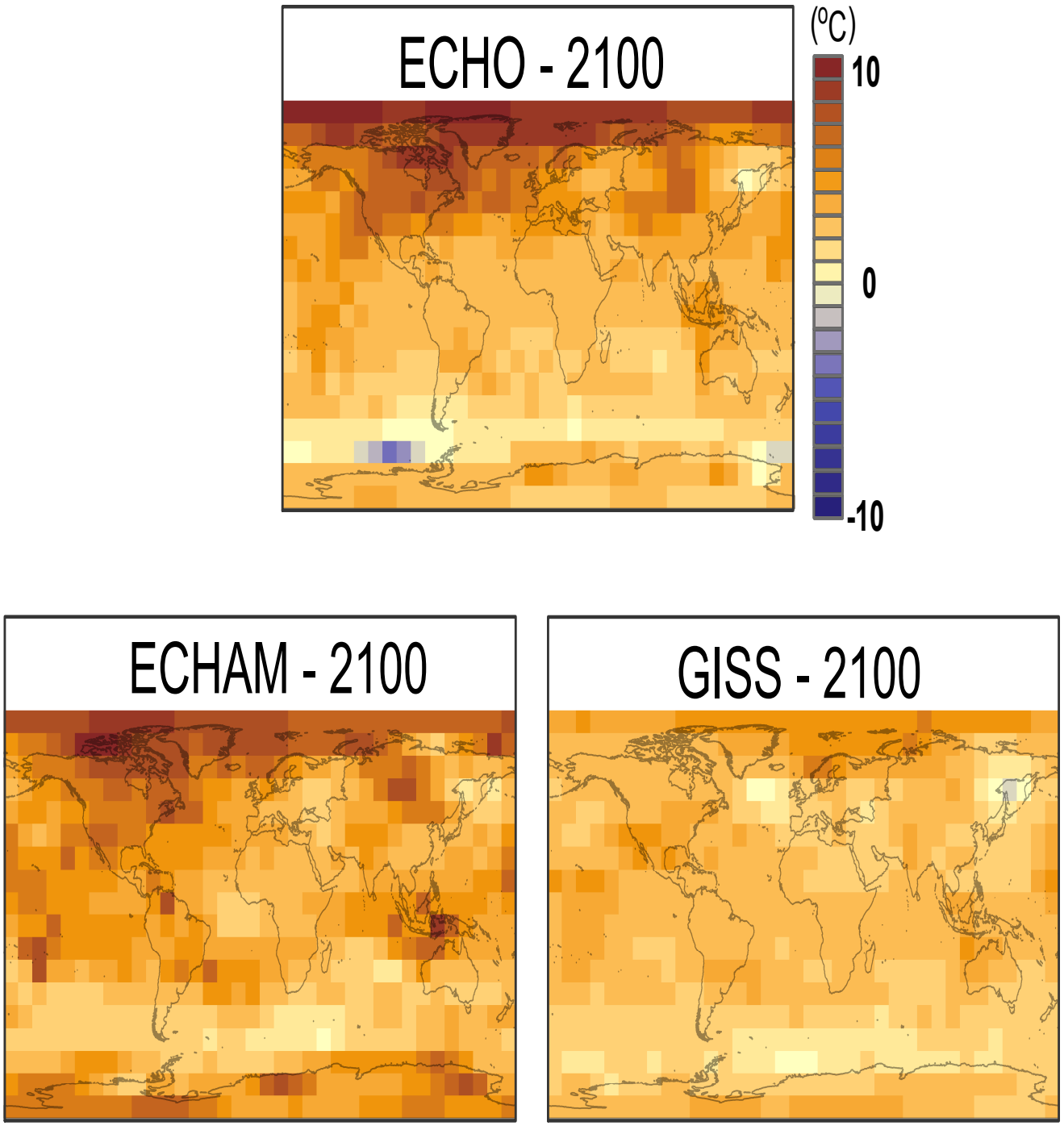

Figure 1: Surface plots of projections of surface temperature anomalies for grid locations in 2100 for scenario A1B (top left). Model output of surface temperature anomalies for 2100 for GCMs ECHO, ECHAM, and GISS, clockwise.

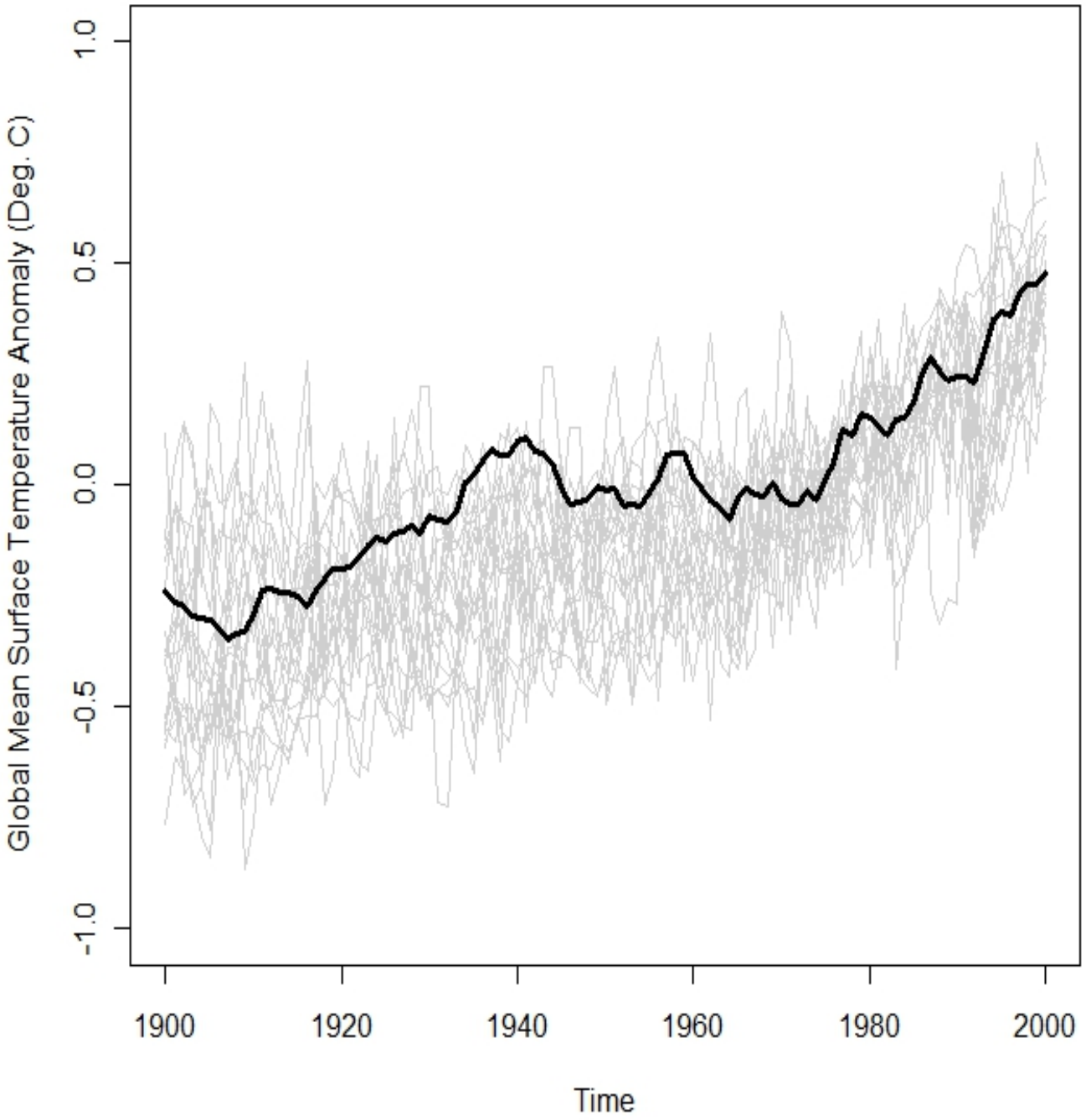

Figure 2: Annual global mean surface temperature anomaly of 20 GCM hindcasts and GISS observations (in black) from 1900 to 2000. There appears to be a small increase in surface temperature between 1900 and 2000.

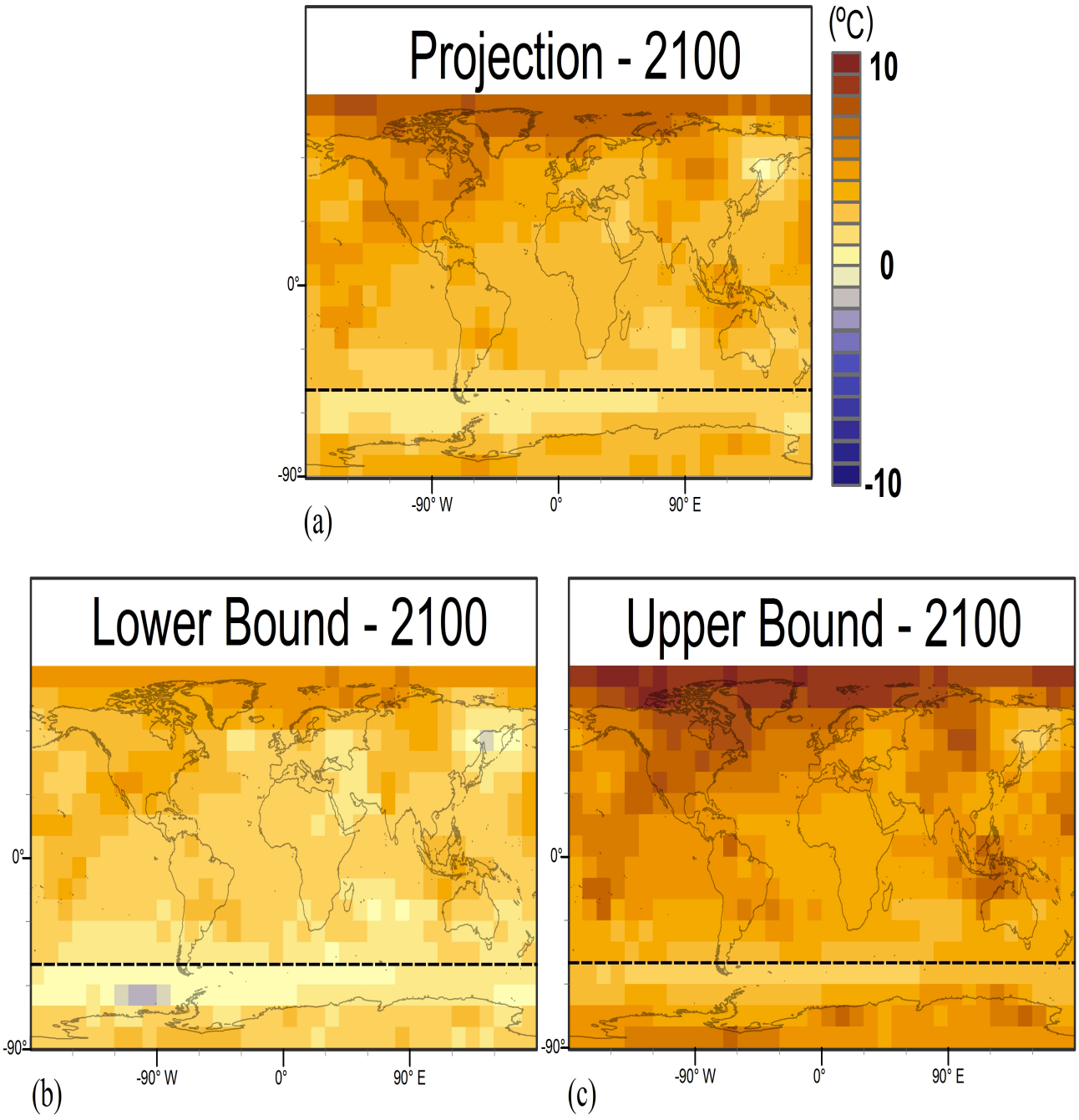

Figure 3: Top Left: Projections of surface temperature anomalies in 2100 for Scenario A1B obtained by BMA and accounting for space-time dependence for three GCM ensemble in Figure 1. Top right: 90% lower bound, Bottom left: upper bound. There appears to be a significant increase in surface temperature everywhere but the Southern Ocean.

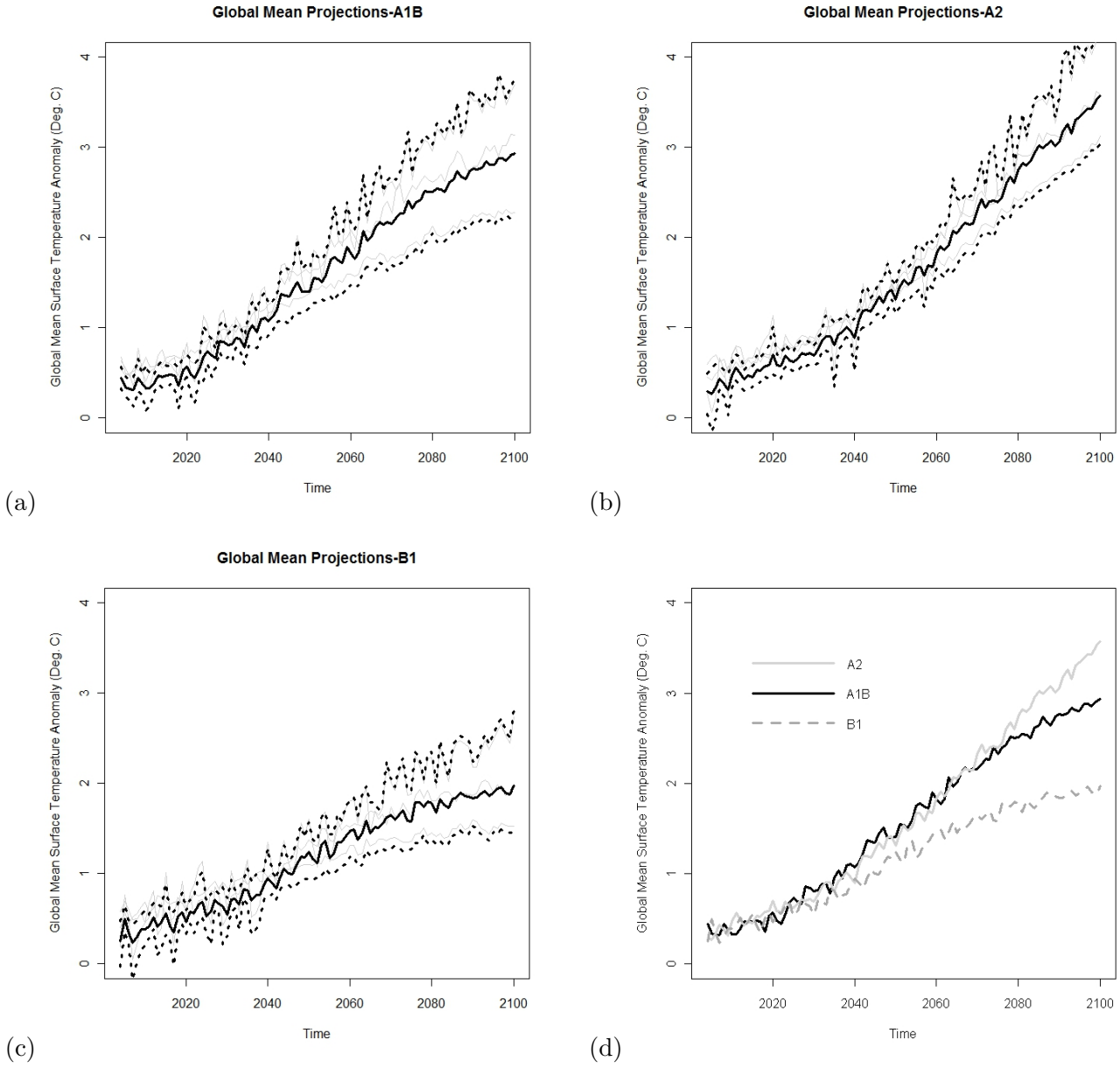

Figure 4: Projections (solid black line) and uncertainty bounds (dotted black lines) of average annual surface temperature anomaly between 2001 and 2100 under scenarios A1B, A2, and B1 for Figures (a), (b), (c). Other (gray) lines show average annual surface temperature anomaly of GCM forecasts. Combining the information of the models allows us to conclude that there will be a significant increase in the average annual surface temperature by 2100, under all three scenarios, however, the increase in temperature is smallest under scenario B1, as seen in (d).

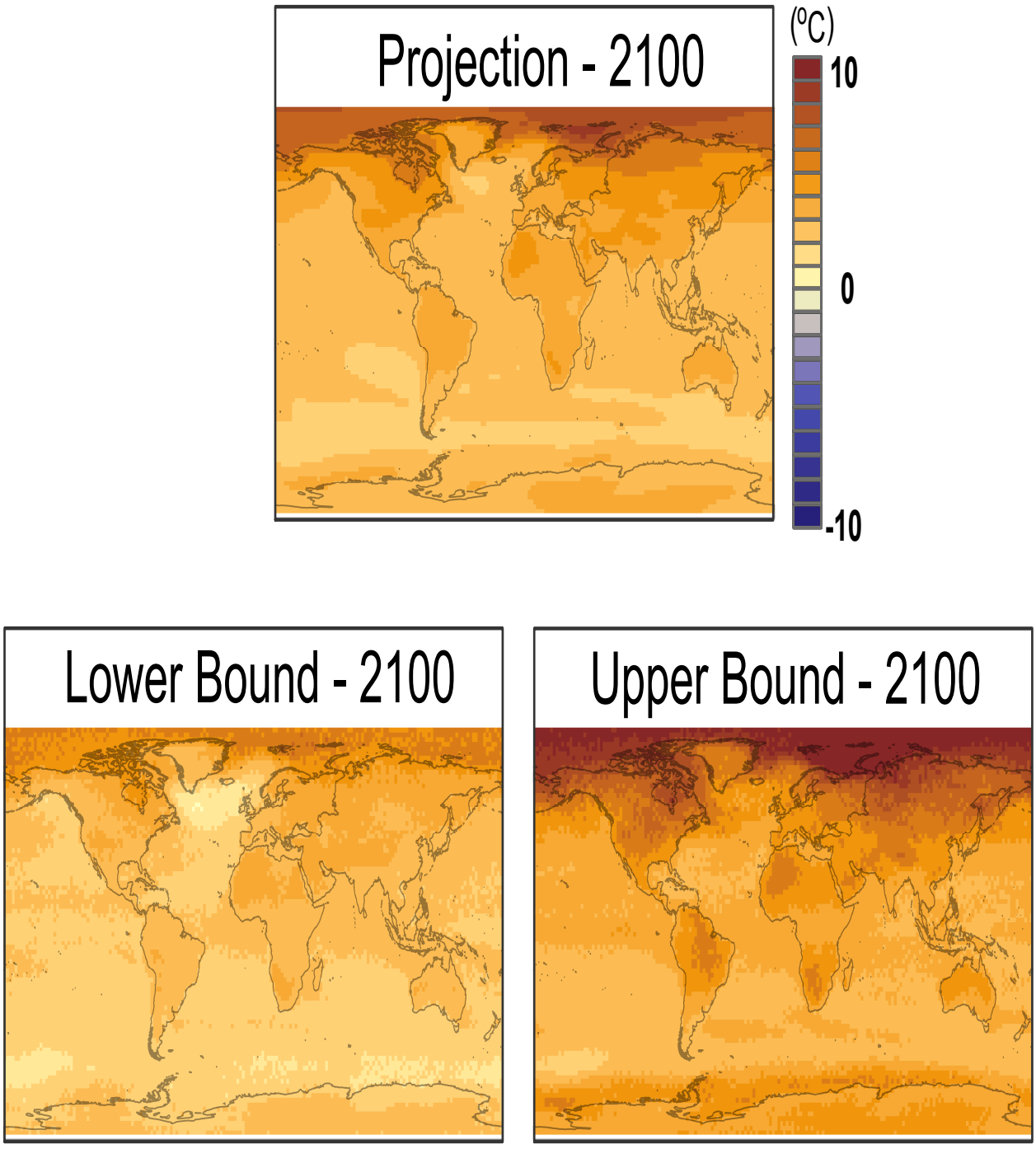

Figure 5: Top Left: Projections of surface temperature anomalies in 2100 for Scenario A1B obtained by BMA using twenty GCM ensemble while accounting for space-time dependence. Top right: 90% lower bound, Bottom left: upper bound. There appears to be a significant increase in surface temperature everywhere but the Southern Ocean and the North Atlantic.

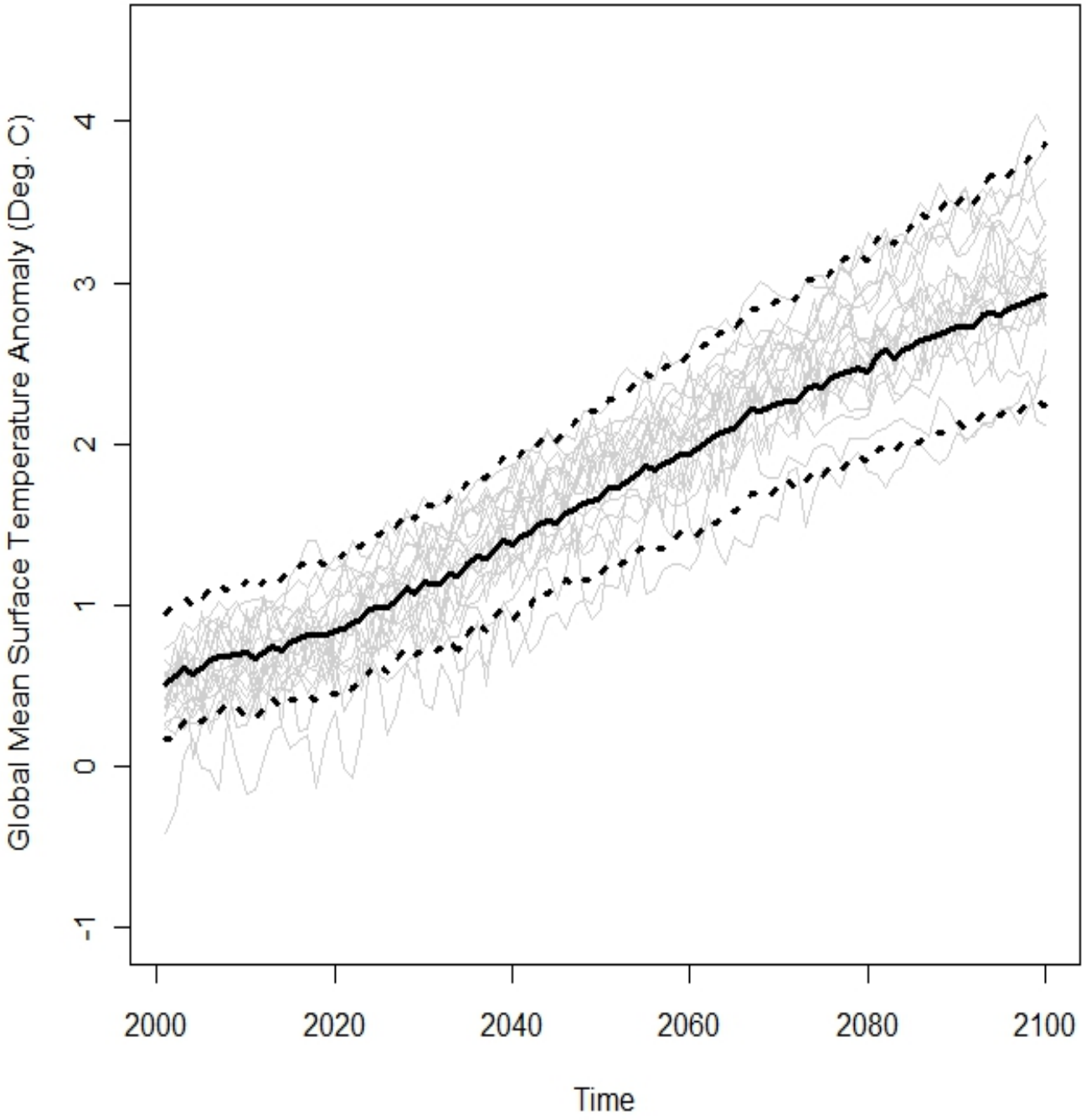

Figure 6: Projections (solid black line) and 90% credible regions (dotted black line) of annual global mean surface temperature anomaly between 2001 and 2100 under the A1B scenario. Other lines show annual global mean surface temperature anomaly of 20 different GCM forecasts.

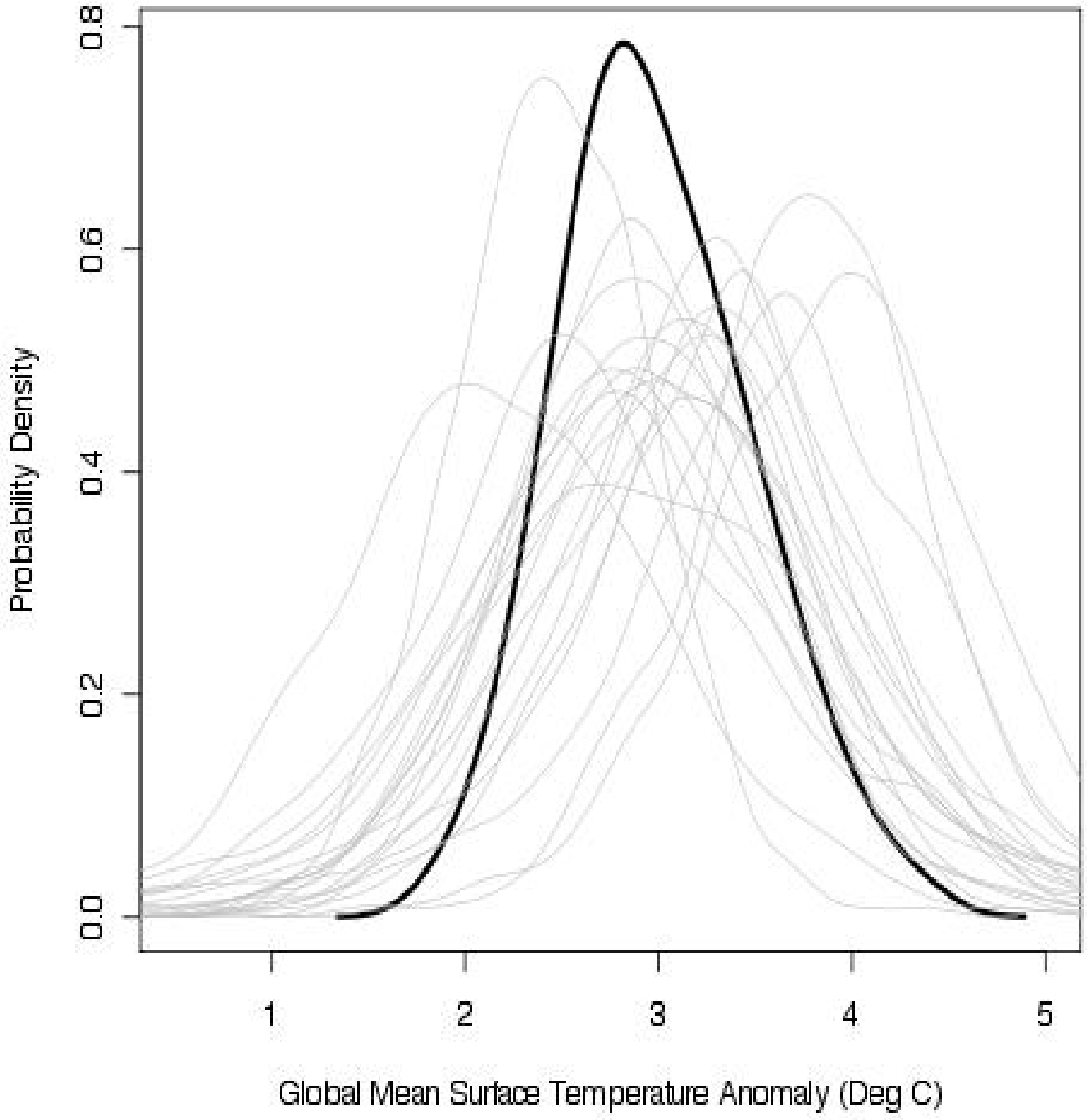

Figure 7: BMA predictive pdf for global mean surface temperature anomaly in 2100 (solid black line) and the twenty components for each GCM forecast (gray lines) under the A1B scenario.