PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2007) **365**, 2103–2116
doi:10.1098/rsta.2007.2070
Published online 14 June 2007

Probabilistic climate change predictions applying Bayesian model averaging

By Seung-Ki Min*, Daniel Simonis and Andreas Hense

Meteorologisches Institut, Universität Bonn, 53121 Bonn, Germany

This study explores the sensitivity of probabilistic predictions of the twenty-first century surface air temperature (SAT) changes to different multi-model averaging methods using available simulations from the Intergovernmental Panel on Climate Change fourth assessment report. A way of observationally constrained prediction is provided by training multi-model simulations for the second half of the twentieth century with respect to long-term components. The Bayesian model averaging (BMA) produces weighted probability density functions (PDFs) and we compare two methods of estimating weighting factors: Bayes factor and expectation–maximization algorithm. It is shown that Bayesian-weighted PDFs for the global mean SAT changes are characterized by multi-modal structures from the middle of the twenty-first century onward, which are not clearly seen in arithmetic ensemble mean (AEM). This occurs because BMA tends to select a few high-skilled models and down-weight the others. Additionally, Bayesian results exhibit larger means and broader PDFs in the global mean predictions than the unweighted AEM. Multi-modality is more pronounced in the continental analysis using 30-year mean (2070–2099) SATs while there is only a little effect of Bayesian weighting on the 5–95% range. These results indicate that this approach to observationally constrained probabilistic predictions can be highly sensitive to the method of training, particularly for the later half of the twenty-first century, and that a more comprehensive approach combining different regions and/or variables is required.

Keywords: global climate change; Bayesian model averaging; probabilistic prediction; surface air temperature

1. Introduction

There have been increasing studies on regional-scale climate change detection and attribution using surface air temperatures (SATs; Karoly *et al.* 2003; Stott 2003; Zwiers & Zhang 2003; Zhang *et al.* 2006; Min & Hense 2007). They have found significant anthropogenic influence (greenhouse gases and sulphate aerosols) on the observed SAT changes over continental and even smaller spatial scales. Based on these regional assessment results, Stott *et al.* (2006) suggested producing probabilistic climate predictions weighted with some

* Author and address for correspondence: Climate Research Division, Environment Canada, 4905 Dufferin Street, Downsview, Ontario, Canada M3H 5T4 (seung-ki.min@ec.gc.ca).

One contribution of 13 to a Theme Issue 'Ensembles and probabilities: a new era in the prediction of climate change'.

measure of the model skills evaluated by observations (Allen *et al.* 2000; Stott & Kettleborough 2002). Here, the main assumption is that past observed changes attributable to anthropogenic forcing can be used as a constraint to future warming. This seems to be reasonable considering that future scenarios such as the well-known special report on emissions scenarios (SRES, Nakicenovic & Swart 2000) are based only on anthropogenic forcing factors.

Since climate predictions are inherently uncertain, the information on the uncertainty is essential to decision-makers. There have been recent efforts to develop methods for a probabilistic treatment of uncertainty in the global warming predictions. Murphy *et al.* (2004), Piani *et al.* (2005) and Stainforth *et al.* (2005) used 'perturbed physics ensembles' in which model parameters are changed within expert-defined ranges. Using distributed computing resources through the climate-prediction.net project (Allen 1999), they obtained multi-thousand simulations of an atmospheric general circulation model coupled to mixed layer ocean and estimated climate uncertainty from that ensemble. Stott *et al.* (2006) showed continental-scale temperature predictions in which weighting factors are obtained from detection/attribution results using an ensemble of atmosphere–ocean coupled climate models (AOGCMs) while the necessary uncertainty is estimated from the control run with the same model.

Multi-AOGCM analyses have been carried out as well for climate change (e.g. Cubasch *et al.* 2001; Giorgi & Bi 2005). More recently, in order to consider model uncertainty systematically, Bayesian approaches have been suggested and applied (Tebaldi *et al.* 2005; Greene *et al.* 2006; Min & Hense 2006*a*). Although these approaches take Bayesian statistics as their basis, they are different in dealing with variables and methods to obtain weighting factors. Tebaldi *et al.* (2005) evaluated models with respect to the present-day climatology and the inter-model consistency in predictions. Greene *et al.* (2006) fitted a linear model to observation and model simulation, and the method of Min & Hense (2006*a*) is based on measuring a generalized distance between observation and simulation. On the other hand, Raftery *et al.* (2005) suggested the Bayesian model averaging (BMA) as a method of calibrating multi-model weather forecast ensembles. BMA produces a weighted probability density function (PDF) using the posterior probability of each participating model as a weighting factor. Raftery *et al.* (2005) showed the superiority of BMA in the probabilistic forecast as well as deterministic one in the case of mesoscale weather forecasts.

The objective of this study is to examine the effect of BMA on probabilistic predictions by comparing weighted and unweighted PDFs given a rather small number of a multi-model ensemble of opportunity very similar in method to the case of Raftery *et al.* (2005). The questions are: can we apply the BMA method in general directly to climate projections, can it be modified and do the results lead to interpretable results? To answer these questions, we will apply the BMA method to climate change predictions using a multi-AOGCM dataset from the Intergovernmental Panel on Climate Change (IPCC) fourth assessment report (AR4). As a first step, we focus on SAT changes in the twenty-first century under the SRES A1B scenario. The method is applied to global mean SATs and then extended into continental regions. To test modifications, we use two methods of estimating the weighting factors with both based on an analysis of the likelihood.

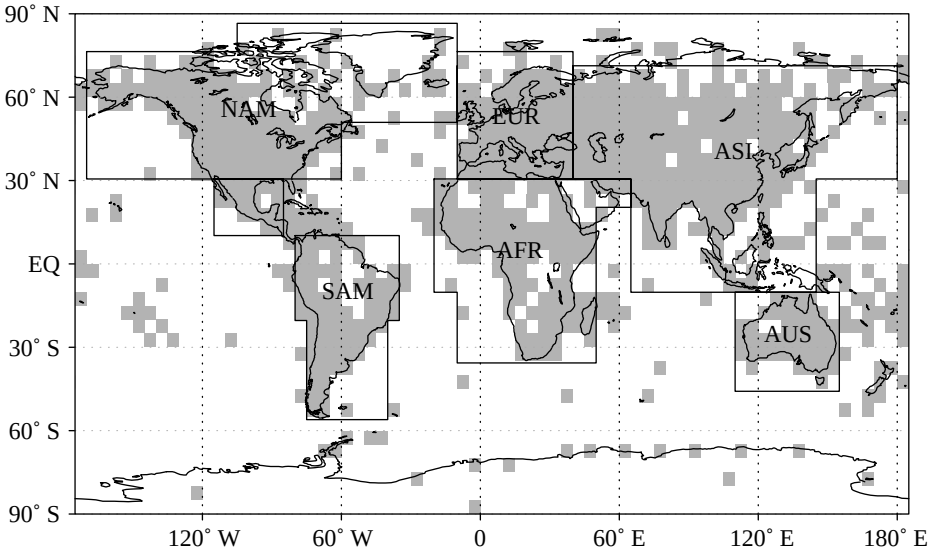

Figure 1. Spatial domain of observations (filled grids) for 1950–1999 applied to model simulations for both training and prediction periods. Continental regions are shown for NAM, ASI, SAM, AFR, AUS and EUR following Stott (2003).

2. Data

Observations of monthly SAT anomalies over land are taken from the Climate Research Unit dataset (CRUTEM2v) for the period 1950–1999 (Jones & Moberg 2003). Area-averaged SATs are calculated over global and continental regions using a temporally varying observational mask for the analysis period. Six continental regions are defined as North America (NAM), Asia (ASI), South America (SAM), Africa (AFR), Australia (AUS) and Europe (EUR) following Stott (2003) and Min & Hense (2007). Figure 1 shows the distribution of observational grids where monthly mean data exist at least once. The same time-varying spatial coverage is applied to the model simulations for the training period (1950–1999) while the constant pattern shown in figure 1 is used for the future simulations (2001–2099).

As model data, we take 21 AOGCMs of IPCC AR4 which provide simulations under the SRES A1B scenario (model description and data are available from the Coupled Model Intercomparison Project phase 3, CMIP3 archive). According to the implemented external forcing, the models are divided into two groups: MME_ALL (natural plus anthropogenic forcing, 10 models) and MME_ANTH (anthropogenic-only forcing, 11 models). Here, we treat ALL and ANTHRO members as a common ensemble, considering that, for the second half of the twentieth century, both ALL and ANTHRO signals are detectable in the observed SAT changes with similar amplitudes (Min & Hense 2007). This approach would not be possible if the complete twentieth century would be taken for the model-data assessment. For the list of analysed models, we refer the reader to figure 2. The ensemble mean of each model is used as an input variable

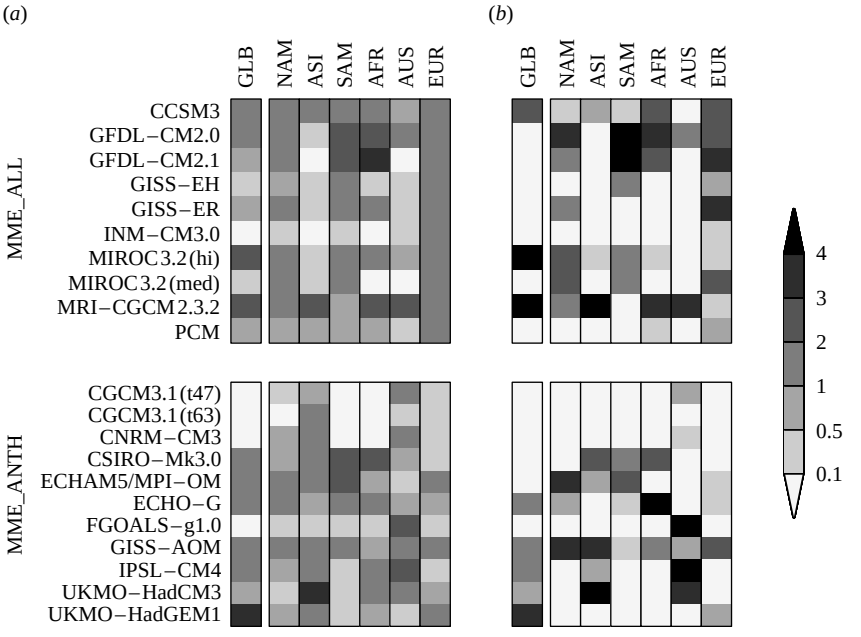

Figure 2. Distributions of normalized weighting factors for global (GLB) and six continental (figure 1) mean SATs simulated by 21 coupled climate models, (*a*) BF and (*b*) EM. Ten models include both natural and anthropogenic forcing (MME_ALL) while 11 models take anthropogenic-only forcing (MME_ANTH). Note that weighting factors are normalized (divided by mean value of $1/21 = 0.048$). See text for details.

for skill comparisons among models. For models with a single member, we regard the single realization as ensemble mean. To estimate the internal variability of area-averaged SATs, pre-industrial control runs of the 21 models (MME_PI) are used as in the previous studies (Min & Hense 2006*b*, 2007). There are 80 independent samples of 100-year long SATs available. Therefore, this ensemble of opportunity addresses mainly the modelling (epistemic) uncertainty and its effect upon the projections.

This study is based on the assumption that models which simulate better consistency with the observed change will provide more reliable predictions of future climate changes. Dealing with large-scale SAT changes, we also assume that some of the models are plausible representations of the real world and that one can ignore the interactions with and the behaviours of other variables. Additionally, we consider only long-term components of model responses to given external forcing (mainly the anthropogenic ones for 1950–1999). To accomplish the latter, projections on Legendre polynomials in time are used as a low-pass filtering method following Min & Hense (2006*b*). Legendre coefficients from the first to fourth degree (LP1–LP4) are computed for the 50-year SAT changes of observations and model simulations. Zero degree coefficients (LP0) which represent time averages are omitted here to avoid any effect from selecting different reference periods and to reduce the influence of the difference between the ALL and ANTHRO forcing runs which are basically visible in LP0

(Min & Hense 2007). This corresponds to using SAT anomalies relative to 1950–1999 for both observations and model simulations. Model data are interpolated to the observational grid of 5°×5° before analysis.

3. Bayesian model averaging

In ensemble forecasting, it is customary to take the arithmetic ensemble mean (AEM) as a prediction quantity and in most cases AEM already provides a better skill than any of the ensemble members alone. However, this approach gives no information about any kind of uncertainty contained in the predictions. BMA can be a powerful tool because it produces a complete PDF as a forecast and provides a quantification of the uncertainties. If one has a high enough number of ensemble members forming a sample of the climate model population, BMA will give a realistic estimate of the modelling uncertainty of the climate system. However, one has to admit that this assumption probably does not hold due to an undersampling of the 'climate model space': even available models cannot be regarded as independent because they share components or are from the same institution (Allen & Stainforth 2002). Accordingly we focus on examining the sensitivity of multi-model probabilistic predictions to different weighting methods. In addition, BMA delivers a way of model selection by weighting each ensemble member according to a measure of models' performance in the training period. Weighted probabilistic predictions of the twenty-first century climate change are obtained on the basis of the model evaluation results for the twentieth century.

The theory of BMA is comprehensively described in Hoeting *et al.* (1999). Given forecast from K models f_k, $k=1, \ldots, K$, and the training data y^{T}, the weighted forecast PDF for predictand y is obtained by

$$p(y|f_1, \ldots, f_K, y^{\mathrm{T}}) = \sum_{k=1}^{K} w_k g_k(y|f_k, \sigma^2), \tag{3.1}$$

where w_k is weight for each model and g_k is a normal PDF with mean f_k and variance σ^2 which is statistically denoted by $y|f_k \sim N(f_k, \sigma^2)$. The weights w_k are estimated from evaluating the models in view of y^{T} for which we use two different methods based on Bayesian statistics. Then BMA predictive mean is just the conditional expectation which is defined as weighted multi-model averaging:

$$\tilde{y} \equiv E[y|f_1, \ldots, f_K] = \sum_{k=1}^{K} w_k f_k. \tag{3.2}$$

If the weighting factors are all equal, the BMA mean becomes identical to AEM which is simply

$$\bar{y} = \frac{1}{K} \sum_{k=1}^{K} f_k. \tag{3.3}$$

As climate variables underlie long-term variations, they are correlated not only in space but also in time. Therefore, a time-series of SAT anomalies has to be treated as the realization of a multivariate normally distributed random variable.

Table 1. Standard deviations of annual and 30-year mean SATs for global and continental regions, estimated from the pre-industrial control simulations with multi-AOGCMs (MME_PI). (Refer to figure 1 for spatial domains.)

	s.d.	
region for area averages	annual mean	30-year mean
GLB	0.17	0.10
NAM	0.37	0.18
ASI	0.25	0.13
SAM	0.20	0.08
AFR	0.23	0.08
AUS	0.31	0.11
EUR	0.44	0.22

Dealing with 50-year annual time-series, a dimension reduction is required. As explained above, we apply Legendre expansions in time restricting from LP1 to LP4 by which the linear trend and the long-term variations are only considered. Now observation and model dataset are analysed on a reduced temporal space (dimension $q=4$) which are hereafter denoted by $\boldsymbol{d}$ and $\boldsymbol{\mu}_k$, respectively.

The variance σ^2 in equation (3.1) is estimated from MME_PI simulations. From 80 independent 100-year long samples, we calculate variances of global and continental-scale averaged SATs. Table 1 shows the standard deviations. The estimated standard deviation of the global mean annual SAT is 0.17 K. The corresponding continental values are larger, ranging from 0.20 to 0.44 K. NAM and EUR have relatively stronger variabilities than other regions which might be related to the North Atlantic Oscillation. Variances of 30-year mean SATs are smaller by about half of the annual values.

(*a*) *Bayes factor*

The first approach to calculate the model weights w_k is to use normalized Bayes factors (BFs) as described in Min & Hense (2006*a*). Given observational data $\boldsymbol{d}$, the BF $B_{k\mathrm{r}}$ of the model M_k with respect to the reference model M_r is defined as the ratio of posterior odds to prior odds:

$$B_{k\mathrm{r}} = \frac{p(M_k|\boldsymbol{d})/p(M_\mathrm{r}|\boldsymbol{d})}{p(M_k)/p(M_\mathrm{r})} = \frac{l(\boldsymbol{d}|M_k)}{l(\boldsymbol{d}|M_\mathrm{r})}. \tag{3.4}$$

Selecting a different reference model has no effect on estimating w_k due to the normalization of the BFs. When two models are single distributions with no free parameters, the BF becomes identical to the likelihood ratio (Kass & Raftery 1995).

Assuming multivariate normal distribution for the observation and simulations, the likelihood can be expressed as

$$l(\boldsymbol{d}|M_k) = \frac{1}{\sqrt{(2\pi)^q}} \sqrt{\frac{\det A_k^{-1}}{\det \Sigma_k \det \Sigma_\mathrm{o}}} \exp\left(-\frac{1}{2}\Lambda_k\right), \tag{3.5}$$

where Σ_o and Σ_k are the covariance matrix of the observation and model simulation, respectively, $A_k = \Sigma_k^{-1} + \Sigma_\mathrm{o}^{-1}$, and $\Lambda_k = (\boldsymbol{d}-\boldsymbol{\mu}_k)^\mathrm{T}(\Sigma_k + \Sigma_\mathrm{o})^{-1}(\boldsymbol{d}-\boldsymbol{\mu}_k)$

which represents a generalized Mahalanobis distance between observation and model simulation (for more details refer to Min *et al.* 2004). We assume that Σ_o and Σ_k are identical to the covariance matrix Σ_{ctl} estimated from MME_PI.

(*b*) *Expectation–maximization algorithm*

Another convenient way taken from Raftery *et al.* (2005) is to maximize the log-likelihood function for the training dataset

$$l(w_1, \ldots w_K, \Sigma) = \sum_i \ln\left(\sum_{k=1}^{K} w_k g_k(\boldsymbol{d}|\boldsymbol{\mu}_{ki})\right), \tag{3.6}$$

by the expectation–maximization (EM) algorithm (Dempster *et al.* 1977). This algorithm is adapted for a problem that can be formulated with unobserved quantities z_{ki}. Here, we define $z_{ki}=1$ if model k is the best in realization i (see below) and $z_{ki}=0$ otherwise.

The EM algorithm is iterative and consists of two steps. In the first E (expectation) step, the current z_{ki} is estimated. The E step for iteration j is given by

$$\hat{z}_{ki}^{(j)} = \frac{w_k^{(j-1)} g(\boldsymbol{d}|\boldsymbol{\mu}_{ki}, \Sigma^{(j-1)})}{\sum_{m=1}^{K} w_m^{(j-1)} g(\boldsymbol{d}|\boldsymbol{\mu}_{mi}, \Sigma^{(j-1)})}. \tag{3.7}$$

In the M (maximization) step, weights and covariance matrix are estimated as follows:

$$w_k^{(j)} = \frac{1}{n} \sum_i \hat{z}_{ki}^{(j)},$$

$$\Sigma^{(j)} = \frac{1}{n} \sum_i \sum_{k=1}^{K} \hat{z}_{ki}^{(j)} (\boldsymbol{d} - \boldsymbol{\mu}_{ki})(\boldsymbol{d} - \boldsymbol{\mu}_{ki})^{\mathrm{T}},$$

where n is the number of realizations of each model. This method requires the same large number of realizations for every model. However, some models have only a single realization available. Thus, we need a way to expand the sample size. An adequate way can be to generate additional realizations by parametric resampling.

Ideally different realizations of models should represent the whole range of internal variability. Assuming that internal variabilities in control and forced (ALL and ANTH) runs are identical, we can apply a parametric bootstrap technique as described in Efron & Tibshirani (1993). The idea is to assume that the data follow a parametric model which in our case is a multivariate normal distribution $N(\boldsymbol{\mu}, \Sigma)$. Then the parametric model has two parameters $\boldsymbol{\mu}$ (mean) and Σ (covariance). Here, we take $\boldsymbol{\mu}_k$ as mean and estimate the covariance Σ_{ctl} from the MME_PI control runs. Now we randomly draw a sample of size n from the multivariate normal distribution

$$N(\boldsymbol{\mu}_k, \Sigma_{ctl}) \rightarrow (\boldsymbol{\mu}_{k1}, \boldsymbol{\mu}_{k2}, \ldots, \boldsymbol{\mu}_{kn}).$$

With these resampled $\boldsymbol{\mu}_{ki}$, the log-likelihood function equation (3.6) is calculated and maximized as described above.

It has turned out that a relatively high number of realizations are necessary to produce stable results with respect to the magnitude of the weights. We choose $n=20\,000$ being a compromise between calculation time and stability of the results. The variation of the weights is about two orders of magnitude smaller than the weights themselves.

4. Results

(a) *Weighting factors*

Figure 2 shows distributions of weighting factors for the global and six continental area-averaged SATs for the 21 participating AOGCMs. The weighting factors are obtained based on BF and EM methods as described above. Results are also displayed for MME_ALL and MME_ANTH separately to see any significant difference between the two groups. BF and EM results show both similarities and differences. They share high-skilled models although there are exceptions for some regions and models. A major difference between the BF and EM results is that the weighting factors from BF are more evenly distributed over a larger number of models, while EM selects only a few best models and damps out the others. Higher weights to fewer models in EM reflects the fact that the model predictions are highly correlated and thereby models that contribute little additional information tend to have very low weights (Raftery *et al.* 2005). In contrast, BF weights are based on a generalized distance measure and so distributed across many models. One needs to note these different characteristics when applying BF or EM methods.

Comparing weights in figure 2, one can find only a few models which have consistently higher or lower weighting factors across the continental regions, for instance, GFDL-CM2.0, MRI-CGCM2.3.2, GISS-AOM, INM-CM3.0, CGCM3.1(t47), CGCM3.1(t63) and FGOALS-g1.0. The other models show very different combinations of weighting factors at different regions. This implies that there is no single best/worst model in simulating global and regional SAT changes, in accord with the concept of multi-model approach.

(b) *Global mean temperature predictions*

Using the standard deviations in table 1 and the weighting factors in figure 2, we obtain the weighted PDFs (BF and EM) of model-simulated future SATs and compare them with unweighted one (AEM). Figure 3 shows the results for the global mean annual SATs for the twenty-first century (2001–2099). From the middle of the twenty-first century onward, the PDFs exhibit a multi-modal structure which is strongest in the EM result. Three maximum densities in EM are delivered by the three models of largest weights: MIROC3.2(hi), MRI-CGCM2.3.2 and UKMO-HadGEM1 (figure 2). BF shows a broader response due to more distributed weighting factors. There are no physical reasons for the global mean SATs such as multiple equilibria to expect a multi-modal PDF response. The obvious explanation for the behaviour shown in figure 2 is undersampling especially with respect to the long-term response. The ensemble of opportunity yet seems to have not enough information to represent faithfully the future variability as long as

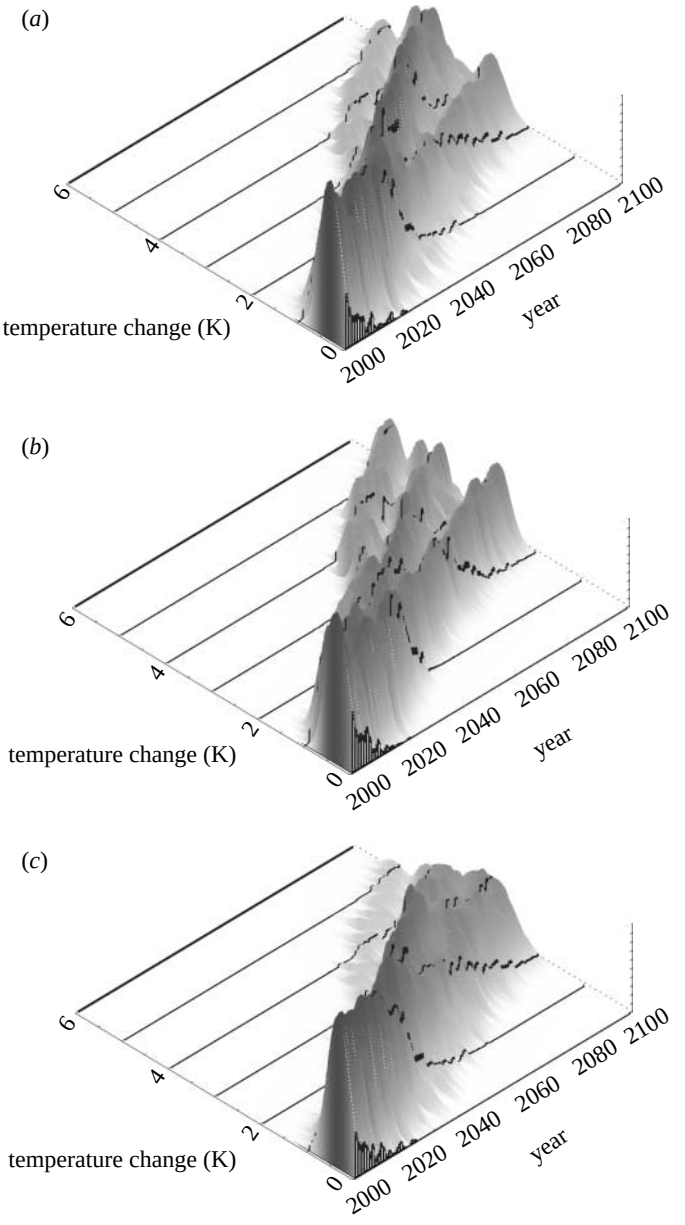

Figure 3. Time-series of PDF of global mean annual SATs for 2001–2099 with different weighing methods: (*a*) BF, (*b*) EM and (*c*) AEM. Temperatures are represented as anomalies with respect to 1970–1999 mean. See text for details.

the BMA-based calibrations are considered. At least this implies that skill-weighted predictions are highly sensitive to the method of evaluating ensemble members.

Figure 4 displays the annual time-series of multi-model-weighted ensemble means and their 5–95% percentiles of probabilistic predictions of global mean SATs. It can be seen that the mean values of BF and EM are very similar to each

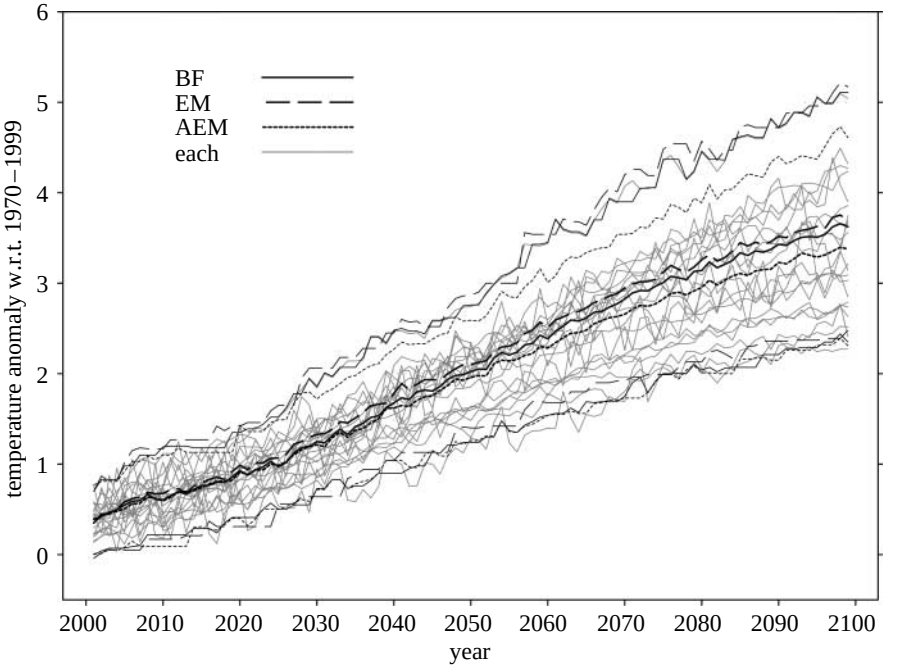

Figure 4. Multi-model average (thick) and its 5–95% percentile (thin lines) of global mean annual SAT predictions for 2001–2099 with BF, EM and AEM under SRES A1B scenarios. Light grey solid lines represent predictions of 21 participating models.

other and larger than AEM where the difference increases in time with a maximum about 0.3 K by the end of the twenty-first century. The 5% percentiles of the three predictions are very close to each other while the 95% percentiles of BF and EM are larger than that of AEM, indicating the broadened PDFs in the upper tail due to the Bayesian weighting. These characteristics are also found in figure 3 showing enhanced densities in the upper branch of PDF time-series of BF and EM.

(*c*) *Continental-scale temperature predictions*

We apply the same technique to continental regions using area-averaged 30-year mean (2070–2099) SATs. Standard deviations of 30-year mean SATs in table 1 and the same weighting factors as in figure 2 are used for this prediction. We take the long-term means to avoid the noisy patterns which arise on interannual time-scales and to simplify comparisons across the regions and weighting methods. Figure 5 shows the PDFs of SAT predictions over six continental regions using BF, EM and AEM methods. The 5% percentile, weighted multi-model mean, and 95% percentile are depicted on top of each PDF plot. Weighted PDF patterns in 2070–2099 are not Gaussian in all three methods. Comparisons to the 2010–2039 predictions (dashed lines) which are more similar to a Gaussian distribution reveal that 21 multi-models are insufficient to sample reasonably the large range of inter-model uncertainties for the late twenty-first century. Hence, their PDFs produce fine structures which can vary highly according to the composition of the sample.

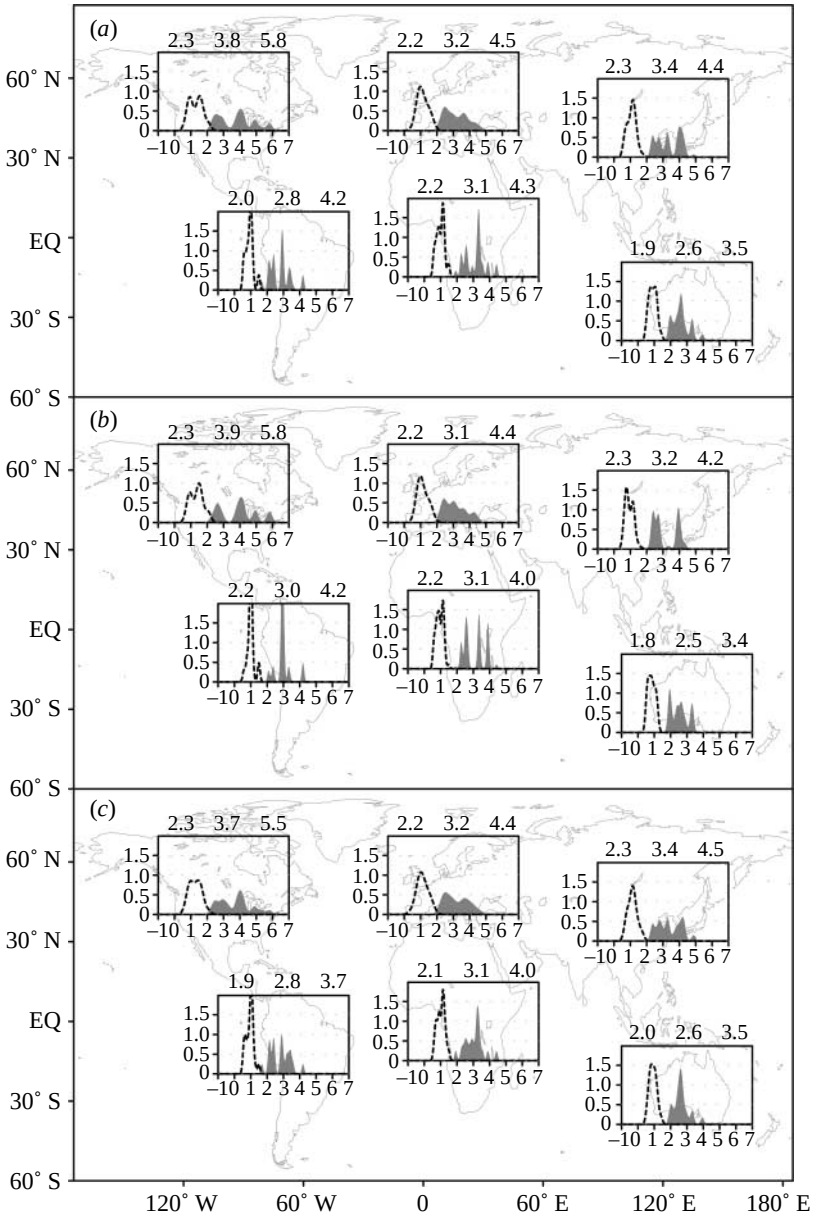

Figure 5. Probabilistic predictions (PDFs) of area-averaged 30-year (2070–2099, filled bars) mean SATs over six continental regions obtained from (*a*) BF, (*b*) EM and (*c*) AEM methods. In each plot, horizontal axis is temperature change relative to 1970–1999 and vertical axis is probability density. Numbers on top of each plot represent 5% percentile, weighted multi-model mean and 95% percentile (from left to right). Dashed lines display the PDFs for 2010–2039.

Although there is the possibility of multiple flow regimes on a regional scale, the multi-modality illustrated in figure 5 cannot be regarded as a realistic modelling of a probabilistic climate change. Multi-modality appears more clearly in BF and EM; for instance, EM has two maximum peaks near 2.2 and 4.0 K over ASI and

marked three peaks in AFR. Nevertheless, there is little change over EUR and AUS. Across the regions, BF and EM patterns are very close, which is already reflected by the similar weighting factors in figure 2.

In terms of multi-model averages, the differences between BF and EM are very small (less than 0.2 K). The effect of Bayesian weighting is not found clearly even in 5–95% percentiles in these 30-year mean SATs. SAM is the region where largest changes appear (0.5 K increase in the 95% percentile). Annual SAT prediction for the continental regions supports this result (not shown). In short, unlike the global mean, we get only a little effect of Bayesian weighting for continental SATs. Besides the problems arising from under-sampling as discussed in the global scale, the regional results also point to problems in estimating the weights. Here, we treated each of six regions independently from each other. A more promising approach would be to combine the SATs of each region and their temporal development into a common space–time vector and evaluate the weights under the spatio-temporal correlation implied by the observations and simulation.

5. Concluding remarks

The BMA technique is applied to the twenty-first century SAT changes simulated by the multi-model AOGCM ensembles of IPCC AR4 to produce probabilistic predictions of global and regional SATs. This approach provides a way of observationally constrained prediction of PDFs by using weighting factors which are obtained through evaluating models for the last 50 years of the twentieth century. This training is based on long-term temporal components (Legendre degrees from LP1 to LP4) to eliminate the noise on shorter time-scales.

In order to consider the influence of inter-model and internal variability systematically when estimating the weighting factors, we apply two estimation methods for the weights: BF and EM algorithm. The BMA based upon the BF approach takes the likelihood ratio which is an exponential function of a generalized Mahalanobis distance between observation and model simulation. Hence, it filters out low-skilled models more effectively than a mean-square error-based approach (Min & Hense 2006*a*). The BMA based upon the EM algorithm, which is the version suggested by Raftery *et al.* (2005) for mesoscale weather forecasting, tends to select only a few high-skilled models and leave out the other models more strongly than the BF-based method.

When applied to global and regional SAT anomalies, both BMAs exhibit comparable results in mean and higher level quantiles. Especially, the occurrence of multi-modal PDF for the global mean SATs suggests a severe undersampling of the inter-model variability at least for the long-term projections for the second half of the twenty-first century. Another possible or additional explanation could be that the weighting factors which are obtained from simulations and observations for the second half of the twentieth century could not be used beyond a horizon of similar time length. Indications for this conjecture are the prediction of the quasi-unimodal albeit non-Gaussian PDFs for the first half of the twenty-first century.

The results presented here demonstrate that observationally constrained probabilistic climate change predictions using BMA are feasible and can provide more information than the raw ensemble. However, a straightforward application of the BMA relevant for ensemble weather forecasting is not possible and might be highly dependent on the method of measuring weighting factors in the training period even if we have a large enough multi-model ensemble to construct the probabilistic predictions. Comprehensive measure of model skills based either on space–time vectors of SAT or on multiple variables (e.g. temperature and sea level pressure) might be useful to produce more robust weighting factors and hence more reliable probabilistic predictions of global and regional climate changes.

This study was supported by the German Research Foundation (DFG) with grant He1916/8. We are grateful to two anonymous reviewers for their clarifying comments. We also acknowledge the modelling groups for making their model output available as part of the World Climate Research Program's (WCRP's) CMIP3 multi-model dataset, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) for collecting and archiving this data and the WCRP's Working Group on Coupled Modelling (WGCM) for organizing the model data analysis activity. The WCRP CMIP3 multi-model dataset is supported by the Office of Science, US Department of Energy.

References

Allen, M. 1999 Do-it-yourself climate prediction. *Nature* **401**, 642. (doi:10.1038/44266)

Allen, M. R. & Stainforth, D. A. 2002 Toward objective probabilistic climate forecasting. *Nature* **419**, 228. (doi:10.1038/nature01092a)

Allen, M. R., Stott, P. A., Mitchell, J. F. B., Schnur, R. & Thomas, L. D. 2000 Quantifying the uncertainty in forecasts of anthropogenic climate change. *Nature* **407**, 617–620. (doi:10.1038/35036559)

Cubasch, U., Meehl, G. A., Boer, G. J., Stouffer, R. J., Diz, M., Noda, A., Senior, C. A., Raper, S. & Yap, K. S. 2001 Projections of future climate change. In *Climate change 2001: the scientific basis* (eds J. T. Houghton, Y. Ding, D. J. Griggs, M. Noguer, P. J. van der Linden & D. Xiaoxu), p. 944. Cambridge, UK: Cambridge University Press.

Dempster, A. P., Laird, N. M. & Rubin, D. B. 1977 Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc. B* **39**, 1–38.

Efron, B. & Tibshirani, R. J. 1993 *An introduction to the bootstrap*. Monographs on statistics and applied probability 57. London, UK: Chapman and Hall.

Giorgi, F. & Bi, X. 2005 Updated regional precipitation and temperature changes for the 21st century from ensembles of recent AOGCM simulations. *Geophys. Res. Lett.* **32**, L21 715. (doi:10.1029/2005GL024288)

Greene, A. M., Goddard, L. & Lall, U. 2006 Probabilistic multimodel regional temperature change projections. *J. Clim.* **19**, 4326–4343. (doi:10.1175/JCLI3864.1)

Hoeting, J. A., Madigan, D. M., Raftery, A. E. & Volinsky, C. T. 1999 Bayesian model averaging: a tutorial (with discussion). *Stat. Sci.* **14**, 382–401. (doi:10.1214/ss/1009212519)

Jones, P. D. & Moberg, A. 2003 Hemispheric and large-scale surface air temperature variations: an extensive revision and an update to 2001. *J. Clim.* **16**, 206–223. (doi:10.1175/1520-0442(2003)016<0206:HALSSA>2.0.CO;2)

Karoly, D. J., Braganza, K., Stott, P. A., Arblaster, J. M., Meehl, G. A., Broccoli, A. J. & Dixon, K. W. 2003 Detection of a human influence on North American climate. *Science* **302**, 1200–1203. (doi:10.1126/science.1089159)

Kass, R. E. & Raftery, A. E. 1995 Bayes factors. *J. Am. Stat. Assoc.* **90**, 773–795. (doi:10.2307/2291091)

Min, S. K. & Hense, A. 2006*a* A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models. *Geophys. Res. Lett.* **33**, L08 708. (doi:10.1029/2006GL025779)

Min, S. K. & Hense, A. 2006*b* A Bayesian assessment of climate change using multimodel ensembles. Part I: global mean surface temperature. *J. Clim.* **19**, 3237–3256. (doi:10.1175/JCLI3784.1)

Min, S. K. & Hense, A. 2007 A Bayesian assessment of climate change using multimodel ensembles. Part II: regional and seasonal mean surface temperatures. *J. Clim.* **20**, 2769–2790.

Min, S.-K., Hense, A., Paeth, H. & Kwon, W.-T. 2004 A Bayesian decision method for climate change signal analysis. *Meteorol. Z.* **13**, 421–436. (doi:10.1127/0941-2948/2004/0013-0421)

Murphy, J. M., Sexton, D. M. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M. & Stainforth, D. A. 2004 Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature* **430**, 768–772. (doi:10.1038/nature02771)

Nakicenovic, N. & Swart, R. (eds) 2000 *Special report on emissions scenarios*, p. 612. Cambridge, UK: Cambridge University Press.

Piani, C., Frame, D. J., Stainforth, D. A. & Allen, M. R. 2005 Constraints on climate change from a multi-thousand member ensemble of simulations. *Geophys. Res. Lett.* **32**, L23 825. (doi:10.1029/2005GL024452)

Raftery, A. E., Gneiting, T., Balabdaoui, F. & Polakowski, M. 2005 Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**, 1155–1174. (doi:10.1175/MWR2906.1)

Stainforth, D. A. *et al.* 2005 Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* **433**, 403–406. (doi:10.1038/nature03301)

Stott, P. A. 2003 Attribution of regional-scale temperature changes to anthropogenic and natural causes. *Geophys. Res. Lett.* **30**, 1728. (doi:10.1029/2003GL017324)

Stott, P. A. & Kettleborough, J. A. 2002 Origins and estimates of uncertainty in predictions of twenty-first century temperature rise. *Nature* **416**, 723–726. (doi:10.1038/416723a)

Stott, P. A., Kettleborough, J. A. & Allen, M. R. 2006 Uncertainty in continental-scale temperature predictions. *Geophys. Res. Lett.* **33**, L02 708. (doi:10.1029/2005GL024423)

Tebaldi, C., Smith, R. L., Nychka, D. & Mearns, L. O. 2005 Quantifying uncertainty in projections of regional climate change: a Bayesian approach to the analysis of multimodel ensembles. *J. Clim.* **18**, 1524–1540. (doi:10.1175/JCLI3363.1)

Zhang, X., Zwiers, F. W. & Stott, P. A. 2006 Multimodel multisignal climate change detection at regional scale. *J. Clim.* **19**, 4294–4307. (doi:10.1175/JCLI3851.1)

Zwiers, F. W. & Zhang, X. 2003 Toward regional scale climate change detection. *J. Clim.* **16**, 793–797. (doi:10.1175/1520-0442(2003)016<0793:TRSCCD>2.0.CO;2)