Application of Bayesian Model Averaging in the Reconstruction of Past Climate Change Using PMIP3/CMIP5 Multimodel Ensemble Simulations

M. FANG
Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou, and University of Chinese Academy of Sciences, Beijing, China

X. LI
Cold and Arid Regions Environmental and Engineering Research Institute, Chinese Academy of Sciences, Lanzhou, and Center for Excellence in Tibetan Plateau Earth Sciences, Chinese Academy of Sciences, Beijing, China

(Manuscript received 1 November 2014, in final form 13 August 2015)

ABSTRACT

Climate change simulations based on climate models are inevitably uncertain. This uncertainty typically stems from parametric and structural uncertainties in climate models as well as climate forcings. However, combining model simulations with instrumental observations using appropriate statistical methods is an effective approach for describing this uncertainty. In this study, the authors applied Bayesian model averaging (BMA), a statistical postprocessing method, to an ensemble of climate model simulations from the Paleoclimate Modelling Intercomparison Project phase 3 (PMIP3) and phase 5 of the Coupled Model Intercomparison Project (CMIP5). Uncertainties, weights, and variances of individual model simulations were estimated from a training period using the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis dataset. The results presented here demonstrate that the BMA method is successful and attains a positive performance in this study. These results show that the selected proxy-based reconstructions and simulations are consistent with BMA estimates regarding climate variability in the past 1000 years, though differences can be found for some periods. The authors conclude that BMA is an effective tool for describing uncertainties associated with individual model simulations, as it accounts for the diverse capabilities of different models and generates a more credible range of past climate change over a relatively long-term period based on multimodel ensemble simulations and training data.

1. Introduction

Past climate changes can be inferred from both reconstructions (Jones et al. 1998; Mann et al. 2008, 2009) and simulations (Ammann et al. 2007; Jungclaus et al. 2010; Xiao et al. 2012). Climate model simulations can be used to understand the possible magnitudes and trends of climate variability and to recognize the relative contributions of various forcings to past climate change (Schmidt et al. 2011; Bothe et al. 2013). However, the numerous variables in the infinitely complex Earth system limit our potential understanding of its processes.

Indeed, current climate models only approximate complex physical, chemical, biological processes and include statistical approximations of other less well-understood processes. Uncertainty in such simulations is inevitable and generally stems from 1) parametric and structural uncertainties in the climate model and 2) the uncertainty in the forcing conditions.

Nonetheless, climate reconstruction methods can determine the relationship between changes in climate proxies, such as tree-ring widths, to climate variables of interest over the instrumental observation period (from approximately 1850 to the present). This relationship is extended backward to reconstruct climate variables for periods during which no instrumental observations of climate variables are available (Jones et al. 2009; Franke et al. 2011; Bhend et al. 2012). Climate reconstruction methods highly rely on the stationary relationship between the climate variable and proxy data (Franke et al. 2011; 2012).
Bayesian theory, is a statistical postprocessing method variability using appropriate statistical methods. Generating a more constrained range of past climate uncertainties associated with simulations and for CMIP5 as a suitable research dataset for characterizing climate change over the past 1000 years in PMIP3/CESM. However, we can consider the simulations of which climate model is best for simulating past climate 1850, we cannot use a quantitative method to evaluate measured data are not available for the period of 850–1700. Because observations from directly constructions that allow for detailed comparisons with climate reconstructions are not available for the period of 850–1850, and historical (1850–2005) climate change simulations, with various climate models that allow for detailed comparisons with climate reconstructions. Because observations from directly measured data are not available for the period of 850–1850, we cannot use a quantitative method to evaluate which climate model is best for simulating past climate change. However, we can consider the simulations of climate change over the past 1000 years in PMIP3/CMIP5 as a suitable research dataset for characterizing uncertainties associated with simulations and for generating a more constrained range of past climate variability using appropriate statistical methods.

Bayesian model averaging (BMA), which is based on Bayesian theory, is a statistical postprocessing method that uses multimodel ensemble simulations to yield probabilistic forecasts. It can supply not only an expectation (deterministic result) but also a probability distribution function (PDF) of any quantity of interest based on training data (Raftery et al. 2005). The BMA method has already been applied to a range of scientific problems—for example, climate change projections (Min et al. 2007; Smith et al. 2009; Bhat et al. 2011), soil moisture simulations (Tian et al. 2012), hydrologic predictions (Duan et al. 2010; Bhat et al. 2011), weather forecasting (Sloughter et al. 2007; Liu and Xie 2014), and economic forecasting (Faust et al. 2013)—and these studies have shown that BMA produces more accurate and reliable predictions than other multimodel ensemble averaging approaches. Detailed descriptions of the BMA method are shown in section 2.

The purposes of this study are as follows. 1) We aim to test the applicability and performance of BMA in the estimation of past climate change using PMIP3/CMIP5 multimodel ensemble simulations over a relatively long-term period; if we demonstrate its validity, many other interesting topics can be studied by BMA. 2) BMA results can be used to provide new insight into the comparison of proxy-based reconstruction and climate model simulations. 3) We attempt to explore a new understanding regarding climate change in the past 1000 years based on PMIP3/CMIP5 multimodel ensemble simulations using BMA. The novelty of our study is that BMA is applied to estimate climate change over a relatively long-term period rather than short-term forecasts.

This paper is organized as follows. The BMA method is described in section 2. Relevant information regarding the climate models, datasets, and implementation process used in our study are described in section 3. In section 4, the BMA results are presented and analyzed; in this section, we also compare the BMA results with specially selected proxy-based reconstructions and with other ensemble averaging techniques. In section 5, we discuss the strengths and weakness of BMA and provide some explanations and findings of the BMA results. Last, we present the conclusions of this study.

2. Bayesian model averaging

a. Basic concept

BMA is a postprocessing method based on Bayesian theory and is used for deriving the relative weights and variances of individual models in a multimodel ensemble. Raftery et al. (2005) extended BMA to forecasts from dynamical models. In BMA, the overall estimated PDF is a weighted average of the PDFs of individual simulations. The weights are the estimated posterior model probabilities; the weight for a given model
represents the simulation skill of that model during the training period relative to other models. The BMA deterministic result is a weighted average of the linear functions of the simulations, with the linear functions representing a bias-correction process for the original simulations. Detailed information on the BMA method is provided by Raftery et al. (2005).

Let \( f = f_1, \ldots, f_K \) denote an ensemble of simulations obtained from \( K \) climate models. Moreover, let \( y \) be the quantity of interest (i.e., surface air temperature in this study). The variable \( y^T \) represents the training data (or observed data). The BMA predictive model for dynamic ensemble simulations can then be expressed as follows:

\[
p[y | (f_1, \ldots, f_K, y^T)] = \sum_{k=1}^{K} w_k p_k[y | (f_k, y^T)],
\]

where each climate model simulation \( f_k \) (for \( k = 1, 2, \ldots, K \)) is associated with a conditional PDF \( p_k[y | (f_k, y^T)] \), which can be interpreted as the conditional PDF of \( y \) on \( f_k \) given that \( f_k \) is the best simulation in the multimodel ensemble. Moreover, \( w_k \) denotes the posterior probability that simulation \( k \) is the best; \( w_k \) is a nonnegative value that satisfies \( \sum_{k=1}^{K} w_k = 1 \) and is viewed as a weight that reflects an individual model’s relative contribution to the simulation skill over the training period. When the quantities of interest are temperature and sea level pressure, the conditional PDFs \( p_k[y | (f_k, y^T)] \) of the ensemble members are approximated by a normal distribution centered on a linear function of the original simulations \( a_k + b_k f_k \) with an ensemble-member-specific standard deviation of \( \sigma_k \):

\[
y | (f_k, y^T) \sim N(a_k + b_k f_k, \sigma_k^2). \tag{2}
\]

The values for \( a_k \) and \( b_k \) are bias-correction terms that are derived using a simple linear regression of \( y \) on \( f_k \) for each of the \( K \) ensemble members. Note that according to Raftery et al. (2005), bias correction can be achieved in various ways. However, selecting or evaluating the best bias-correction method was not a focus of this study; furthermore, in some cases, linear regression is a reasonable choice for individual cases (Raftery et al. 2005; Vrugt and Robinson 2007; Bhat et al. 2011; Tian et al. 2012; Liu and Xie 2014). Therefore, we used a linear method in accordance with the above cases. The BMA predictive mean can be computed as follows:

\[
E[y | (f_1, \ldots, f_K, y^T)] = \sum_{k=1}^{K} w_k (a_k + b_k f_k). \tag{3}
\]

Equation (3) provides a deterministic result whose performance can be compared with individual simulations in the ensemble or with the ensemble mean (Raftery et al. 2005). If we denote space and time with subscripts \( s \) and \( t \), respectively, such that \( f_{k,s,t} \) is the \( k \)th simulation in the ensemble at location \( s \) and time \( t \), then the associated variance in Eq. (3) can be computed as follows:

\[
\text{Var}[y_{s,t} | (f_{1,s,t}, \ldots, f_{K,s,t}, y^T)] = \sum_{k=1}^{K} w_k \left[ (a_k + b_k f_{k,s,t}) - \sum_{l=1}^{K} w_l (a_l + b_l f_{l,s,t}) \right]^2 + \sum_{k=1}^{K} w_k \sigma_k^2. \tag{4}
\]

The right side of Eq. (4) has two terms: the first term summarizes the between-forecast spread, and the second term measures the expected uncertainty condition that one of the simulations is best.

b. BMA weights and variances

Following Raftery et al. (2005) and Tian et al. (2012), we first estimated \( a_k \) and \( b_k \) via simple linear regression [shown in Eq. (5)] of \( y_{s,t} \) on \( f_{k,s,t} \) using the training data for 1948–2005. If the simulations have not yet been corrected for biases, then the estimation of \( a_k \) and \( b_k \) can be viewed as a very simple bias-correction process. Note that \( a_k \) and \( b_k \) are retained in Eq. (3), even when the simulations have been bias corrected (Raftery et al. 2005):

\[
y_{s,t} = a_k + b_k f_{k,s,t}. \tag{5}
\]

We estimate the values of \( w_k \) and \( \sigma_k^2 \), where \( k = 1, 2, \ldots, K \), from a calibration dataset using the maximum likelihood (ML) approach. Assuming that the simulation errors are independent in space and time, the log-likelihood function \( l \) for the BMA-estimation model defined in Eq. (3) is expressed as

\[
l(w_1, \ldots, w_K, \sigma_k^2) = \sum_{s,t} \log \left[ \sum_{k=1}^{K} w_k g_k(a_k + b_k f_{k,s,t}, \sigma_k^2) \right], \tag{6}
\]

where \( n \) denotes the total number of measurements in the training dataset. To achieve numerical stability and simplicity of the training process, in practice, this approach always maximizes the log-likelihood function rather than the likelihood function itself. Unfortunately, no analytical solutions that conveniently maximize Eq. (6) exist. Raftery et al. (2005) proposed the expectation-maximization (EM) algorithm to determine ML values for BMA weights and variances when the estimated PDFs of weather-related quantities are approximately
normal. It should be noted that in practice, the selection of an algorithm that estimates the weights and variances primarily depends on the variable types and computational complexity. Although the EM algorithm can only obtain a locally optimal solution, the method is simple, easy to implement, and computationally efficient. In particular, EM can produce more accurate estimates of the weights and variances of individual ensemble members when the quantity of interest follows a normal distribution. To implement the EM algorithm for the BMA method, an unobserved quantity $z_{k,s,t}$ was introduced. The quantity equals 1 if the ensemble member $k$ is the best simulation at space $s$ and time $t$; otherwise, $z_{k,s,t} = 0$. Hence, for each observation, only one quantity in the set $\{z_{1,s,t}, \ldots, z_{K,s,t}\}$ is equal to 1; all others are zero. After initializing the weights and variances for the individual ensemble members, the iterations of the EM algorithm alternate between an expectation step and a maximization step until convergence is achieved. The detailed steps are as follows.

1) Initialize the weights and variances of the individual ensemble members:

$$ j = 0, $$

$$ w_k^{(j)} = K^{-1}, \quad \text{and} \quad \sigma_k^{2(j)} = \frac{1}{n} \sum_{s,t} [y_{s,t} - (a_k + b_k f_{k,s,t})]^2. $$

2) Calculate the initial log likelihood:

$$ l[w_1^{(0)}, \ldots, w_K^{(0)}, \sigma_1^{2(0)}] = \sum_{s,t} \log \left\{ \sum_{k=1}^K w_k^{(0)} g_k[a_k + b_k f_{k,s,t}, \sigma_k^{2(0)}] \right\}. \quad (7) $$

3) Calculate the expectation $E$. In the expectation step, replace $j$ with $j + 1$. Then, the values of $z_{k,s,t}$ are reestimated given the current estimates of the weights and variances according to the following relationship:

$$ z_{k,s,t}^{(j)} = \frac{w_k^{(j)} g_k[a_k + b_k f_{k,s,t}, \sigma_k^{2(j-1)}]}{\sum_{l=1}^K w_l^{(j)} g_l[a_l + b_l f_{l,s,t}, \sigma_l^{2(j-1)}]}, \quad (8) $$

where the superscript $j$ refers to the $j$th iteration of the EM algorithm and $g_k[a_k + b_k f_k, \sigma_k^{2(j-1)}]$ is a normal density with a mean of $a_k + b_k f_{k,s,t}$ and a standard deviation of $\sigma_k^{2(j-1)}$ at $y_{s,t}$.

4) Perform the maximization $M$. In the maximization step, the values of the weights and variances are updated using the current estimates of $z_{k,s,t}^{(j)}$:

$$ w_k^{(j)} = \frac{1}{n} \sum_{s,t} z_{k,s,t}^{(j)} \quad \text{and} \quad \sigma_k^{2(j)} = \frac{\sum_{s,t} z_{k,s,t}^{(j)} [y_{s,t} - (a_k + b_k f_{k,s,t})]^2}{\sum_{s,t} z_{k,s,t}^{(j)}}. \quad (10) $$

Then, the updated log-likelihood function is as follows:

$$ l[w_1^{(j)}, \ldots, w_K^{(j)}, \sigma_k^{2(j)}] = \sum_{s,t} \log \left\{ \sum_{k=1}^K w_k^{(j)} g_k[a_k + b_k f_{k,s,t}, \sigma_k^{2(j)}] \right\}. \quad (11) $$

Furthermore, convergence testing is performed as follows: if $l[w_1^{(j)}, \ldots, w_K^{(j)}, \sigma_k^{2(j)}] - l[w_1^{(j-1)}, \ldots, w_K^{(j-1)}, \sigma_k^{2(j-1)}] \leq \varepsilon$ (where $\varepsilon$ is a small, predefined tolerance), then the iteration is halted. Otherwise, the iteration continues until the $E$ and $M$ steps converge. The BMA weights and variances are achieved when the EM algorithm reaches a convergent state. The EM method is easy to implement because the optimal value can be achieved using only steps 3 and 4. Furthermore, the maximization step is designed such that the weights are always positive and add up to one.

3. Datasets and implementation process

a. Datasets

The PMIP3/CMIP5 climate model simulations, which provide the necessary temperature information, are available from “historical” and “past 1000 yr” experiments (Schmidt et al. 2011; Taylor et al. 2012). The climate models and information associated with the models are listed in Table 1. (The data are available online from http://pcmdi9.llnl.gov/.)

The analysis was performed using the following models: BCC_CSM1.1 (Wu et al. 2010), CCSM4 (Landrum et al. 2013), GISS-E2-R (Schmidt et al. 2014), IPSL-CM5A-LR (Hourdin et al. 2013), HadCM3 (Gordon et al. 2000), and the current generation of MPI-ESM-P (Giorgetta et al. 2013; Jungclaus et al. 2013). The six climate models are widely employed within the context of climate change and are representative of the progress in the field of climate model development.

We note that the foundation of the BMA method is model training. In this study, each grid point in the globe participates in the training process. However, BMA is not applicable when training data do not exist for some regions. Because the climate model
simulations of PMIP3/CMIP5 have global coverage, the training data must also have global coverage (land and ocean), in accordance with our design. We used the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalysis data as the training data. NCEP–NCAR reanalysis is produced from a state-of-the-art data assimilation system using data from 1948 to the present with global coverage (land and ocean), and the output is commonly used as a baseline reference for a variety of computations (Kalnay et al. 1996). (The NCEP–NCAR reanalysis data are available online from https://climatedataguide.ucar.edu/climate-data/ncep-ncar-r1-overview.)

To evaluate the reliability and applicability of the BMA method for global and Northern Hemisphere (NH) temperature series, we used the HadCRUT4 dataset, a common reference dataset and a collaborative product of the Met Office Hadley Centre and the Climatic Research Unit at the University of East Anglia (Morice et al. 2012). The HadCRUT4 dataset can be used to understand climate change from 1850 to the present at global, hemispheric, and regional scales; it is also used as a standard dataset to assess the simulation skills of climate models and to evaluate the performances of proxy-based reconstructions. However, a fact of HadCRUT4 is missing data for some locations, particularly at high latitudes and over oceans. We note that in this study, we used only the hemispheric and global anomaly series from HadCRUT4 relative to the reference period of 1961–90. (This dataset is available online from http://www.metoffice.gov.uk/hadobs/hadcrut4/.)

The various data grids used for this study require that the data are interpolated to a common horizontal grid of spectral triangular truncation at wavenumber 21 (T21; ~5°); it should be noted that this resolution is in line with the resolution of Bothe et al. (2013). We used free software [i.e., Climate Data Operators (CDO) software, developed by the Max Planck Institute for Meteorology] to perform the interpolation.

b. Implementation process

First, we selected the simulations of the six models and the NCEP–NCAR reanalysis data for the period of 1948–2005 to estimate the weights and variances of the individual models. Second, we selected the simulations of the six models over 1850–2005 to separately generate global and NH temperature changes based on BMA. Then, we computed root-mean-square errors (RMSEs) and correlation coefficients between the deterministic results of BMA and the HadCRUT4 dataset to evaluate the performance and applicability of BMA. The 90% confidence intervals of the BMA estimates are shown and compared with the multimodel ensemble simulations. Finally, we used BMA to generate global and NH temperatures for the past 1000 years (850–1849). Moreover, comparison of the BMA results and the proxy-based reconstructions were conducted to evaluate the consistency of the two results.

Currently, two relatively simple approaches are frequently used to combine simulations from multimodel ensembles: the equally weighted ensemble average (EWEA) treats all of the models equally, and the correlation-coefficient-weighted ensemble average (CCWEA) assumes that some models in the ensemble are superior to others. In general, assigning weights to individual models in the multimodel ensemble depends on their simulation skills. In this study, the simulation skills are derived by calculating the temporal correlation coefficients between the model simulations and NCEP–NCAR data at every grid point globally. The results of EWEA and CCWEA can be
used for comparison with BMA results to evaluate the performance of BMA.

4. Results and analysis

a. Global temperature over 1948–2005

The trained weights of individual models are shown in Fig. 1. To understand the uncertainties after using BMA, we specially display the PDFs of BMA estimations and individual model simulations for 1950, 1960, 1970, 1980, 1990, and 2000 (Fig. 2). According to Fig. 1, the weights of individual models are commonly less than 0.3, and apparent differences among the weights can also be found. Regarding global temperatures, individual model simulations have relatively low skill in the training period and contribute considerably but unevenly (none of the weights is equal to 0, but all weights are not equal to each other) to the BMA results. In addition, Fig. 2 indicates that different model simulations express a variety of uncertainties regarding past climate change because the PDFs of individual climate models vary across the multimodel ensemble without obvious clustering. Moreover, not all of the simulated ranges of climate change from individual climate models can completely cover the observations, and in some years, observations even fall entirely outside some simulated ranges. Note that the estimated PDFs of BMA provide a more reliable description of the total estimated uncertainty than the raw ensemble, leading to sharper and better PDFs for the probabilistic estimate of climate change in 1950, 1960, 1970, 1980, 1990, and 2000. In addition, the deterministic results of BMA are closer to the reference values (e.g., NCEP–NCAR reanalysis), and the 90% confidence intervals of BMA completely contain the NCEP–NCAR reanalysis data. The above results and analysis reflect the following: 1) the model simulations disagree somewhat with each other, and simulating skills vary greatly from individual to individual; 2) using BMA can effectively constrain the simulated PDFs of the multimodel ensemble to the observations and yield a more constrained confidence interval of climate change; and 3) the results offer a point estimation that is closer to the reference.

Although the deterministic result of BMA is only a by-product, it can also be treated as a supplementary but effective metric for evaluating BMA performance by comparing it with individual model simulations in the ensemble and the ensemble mean (Raftery et al. 2005). Therefore, we compared the deterministic result of BMA with the individual model simulations as well as with the results of EWEA and CCWEA. First, we implemented grid-to-grid correlation analysis between the individual model simulations (annually resolved temperature) and the NCEP–NCAR reanalysis as well as between the results of EWEA, CCWEA, and BMA and the NCEP–NCAR reanalysis from 1948 to 2005 (Fig. 3). Second, absolute errors between the climatologies of the individual model simulations and the NCEP–NCAR reanalysis, as well as those between the climatologies of the results of EWEA, CCWEA, and BMA and the NCEP–NCAR reanalysis from 1948 to 2005, were
Further statistics for correlation coefficients and absolute errors are shown in Table 2 and 3, respectively.

From Figs. 3a–f, the individual model simulations commonly show positive correlation coefficients over Antarctica, the equatorial western Pacific, the central and southern Pacific Ocean, much of the American continents, and the mid-Atlantic Ocean. However, negative correlation coefficients are found in many other regions, particularly in the North Atlantic. Furthermore, based on Table 2, the median correlation coefficients for individual model simulations are 0.26 or less, and many of the correlation coefficients in Figs. 3a–f are negative. Based on Figs. 3g,h, obvious and expected improvements in the correlation coefficients occur for the results of EWEA and CCWEA; specifically, the median correlation coefficients are 0.32 or greater, which are higher than those of any single model (Table 3). However, we also found that the results of the two methods have relatively large errors over northern cold areas, some southern cold areas, and small areas of the mid-Pacific Ocean. Based on Fig. 3i, the deterministic result of BMA has very small absolute errors globally, except for small regions in northern cold areas and Southeast Asia. Another substantial improvement is that the median of the absolute errors is 0.44, which is dramatically lower than that of any single model, EWEA, or CCWEA (Table 3).

Based on the results and analysis presented above, we find that the deterministic results of BMA, albeit a by-product, outperformed all the ensemble members and the ensemble average (e.g., EWEA and CCWEA) in terms of correlations and absolute errors.

b. Global and NH temperature during 1850–2005

Figure 5 shows the annual global and NH temperature anomalies over 1850–2005. The RMSEs and correlation coefficients between the anomalies and HadCRUT4 are shown in Table 4. It should be noted that the RMSEs and correlation coefficients are computed from nonfiltered anomalies to evaluate high-frequency variability.

Clearly, from Table 4, the BMA-based global temperature series has the smallest RMSE and the highest
correlation coefficient, which is the same situation for NH. Based on Fig. 5, all of the global and NH temperature series derived from EWEA, CCWEA, BMA, and the individual model simulations generally exhibit a global warming trend from 1850 to 2005 and cooling during 1900–30 and 1960–80. The global and NH climate change trends in Fig. 5 are consistent with those recorded in HadCRUT4. According to the RMSEs and correlation coefficients in Table 4, we find that the results of EWEA, CCWEA, and BMA are better constrained than those of the individual model simulations; however, the EWEA and CCWEA methods are less well constrained than the BMA method. Based on Fig. 5, we can clearly observe that the 90% confidence intervals of the BMA estimations can completely contain HadCRUT4 and are narrower than the ranges of the multimodel ensemble simulations themselves; some individual model simulations fall partly outside the 90% confidence interval of the BMA estimation. The above findings indicate that the BMA-based method can effectively provide an ensemble constraint on the uncertainties of the raw ensemble and is also effective in constraining the model ensemble toward the observed state outside of the training period.

c. Global temperatures from 850 to 1849

Figure 6 shows the global temperature anomalies for the past 1000 years produced by the BMA results, proxy-based global temperature reconstructions [based on Jones et al. (1998, hereinafter Jones1998) and Mann et al. (2008, hereinafter Mann2008), respectively], and individual model simulations. The deterministic result of BMA roughly exhibits a warm anomaly during 850–1250 and a cool anomaly from 1251 to 1849. Detailed analyses of Jones1998 and Mann2008 are not shown here; please see the related references. The correlation coefficients between the deterministic result of BMA and Jones1998 and Mann2008 are 0.53 and 0.60, respectively, during 1000–1849 and are both significant at the 99% confidence level; please note that the correlation coefficients were computed using nonsmoothed anomalies. From Fig. 6 (top) we find that the two proxy-based reconstructions are essentially contained in the 90% confidence interval of the BMA estimation. Therefore, we think that the two reconstructions are overall consistent with the range of the BMA estimation, though some differences can be found when we compare them with the deterministic result of BMA. Figure 6 (bottom) shows that the 90% confidence interval of the BMA estimation is narrower than the range of the multimodel ensemble simulations because some individual model simulations fall partly outside the 90% confidence interval of the BMA estimation. According to the above analysis, the BMA-based result approximately agrees with the reconstructions of Jones1998 and Mann2008, as based on the correlation coefficients and the 90% confidence interval of the BMA estimation. In addition, in Fig. 6 (top), the two
proxy-based reconstructions and the BMA deterministic result commonly show warmer temperatures during the period of 850–1250, though the two proxy-based reconstructions are more or less warmer than afterward until 1450. Furthermore, the three series agree on a colder period in the second half of the last 1000 years. The three series do not show universal climate variability or the trend during the period of 1250–1600.

d. NH temperature from 850 to 1849

Figure 7 shows the NH temperature anomalies for the past 1000 years produced by the BMA result, proxy-based NH temperature reconstructions [based on Esper et al. (2002), Mann and Jones (2003), Moberg et al. (2005), D’Arrigo et al. (2006), Hegerl et al. (2006), and Mann2008], and individual model simulations. In Fig. 7, the deterministic result of BMA exhibits a warm period during 850–1250 and a cold period from 1251 to 1849, respectively. Detailed analyses of the references cited above are not shown here; please see the related references. In Fig. 7 (top), the correlation coefficients between the deterministic result of BMA and Mann2008, Mann and Jones (2003), Moberg et al. (2005), D’Arrigo et al. (2006), Esper et al. (2002), and Hegerl et al. (2006) for 850–1849 are 0.70, 0.78, 0.69, 0.46, 0.53, and 0.63, respectively; all of the correlation coefficients were computed using nonsmoothed anomalies and are significant at the 99% confidence level. Clearly, differing past climate variability can be observed from proxy-based reconstructions.

In addition, although some proxy-based reconstructions fall partly outside the bounds of the 90% confidence interval of BMA estimate in certain periods, we still think that those proxy-based reconstructions are consistent with our BMA estimates for the past 1000 years. A substantial finding from Fig. 7 (bottom) is that the 90% confidence interval of the BMA estimation is narrower than the range of the multi-model ensemble simulations.

Based on the results presented above, although differences existed among the temperature series in terms of the warming and cooling rates per century, all of the temperature series well reproduced the cold phase and the warm phase of the past 1000 years. Thus, we believe that the BMA method is effective for estimating climate change over the past 1000 years based on PMIP3/CMIP5 multimodel ensemble simulations. We also may conclude that BMA is capable of generating an estimate of a probable range for climate change in the NH over the past 1000 years.

In addition, based on Fig. 7 (top), the seven reconstructed temperature series not only indicate that the period of 850–1250 is in a warmer stage but also demonstrate that a colder period universally occurs in the second half of the last 1000 years. However, we should note that universal climate variability and the trend during the 1250–1600 period cannot be found based on the seven series.

From Figs. 6 and 7, the correlation coefficient between the deterministic result of BMA on a global scale and
the deterministic result of BMA in the NH is 0.93 for the period of 850–1849, which is statistically significant at the 99% confidence level. The average trend and lowest temperature during the warm stage and cool stage do not appear to differ between the two spatial scales. Comprehensive, in the past millennium, the two spatial scales essentially show similar climate changes. However, we also want to note that even for global reconstructions, the data are mainly with regard to the NH.

5. Discussion

The EWEA-based results of our study show relatively large discrepancies with the observations because EWEA cannot consider the differences among the various models regarding the ability to simulate actual climate elements. The CCWEA method can take into account the differences among various models with respect to the ability to simulate actual climate elements. This method assigns distinct weights to individual models according to the correlation coefficients between simulated and observed data over an entire time period; however, the CCWEA method still shows a limited ability to reproduce the observed or reanalyzed climate change. Although CCWEA considers the differences among climate models, the deviation between the simulated and observed data are not corrected; hence, this deviation will naturally propagate to the final result.

BMA is a method that can assess the performance of individual models and assign weights to models. Moreover, the BMA-estimated PDF is a weighted average of the PDFs centered on the linear function of the simulations of \( a_k + b_k f_k \), for \( k = 1, \ldots, K \), rather than the \( f_k \) simulations directly. Estimating \( a_k \) and \( b_k \) is a simple bias-correction process, which can also be considered to be a very simple form of model output statistics. The result is a credible estimate of the past climate variability. In our study, the BMA-estimated PDFs were better adjusted than the multimodel ensemble, and the 90% confidence intervals of the BMA estimation can completely contain the observations or reanalysis data. In addition, the BMA deterministic results had lower RMSEs and higher correlation coefficients than any of the individual model simulations or the ensemble average (e.g., EWEA and CCWEA), though the ensemble average was better than any of the individual models. A comparison of the BMA-based results and proxy-based reconstructions during the period of 850–1849 demonstrates that although some reconstructions fall outside the 90% confidence interval of BMA estimation in certain periods, the BMA results are consistent with the selected proxy-based reconstructions.

Because of the lack of instrument-measured temperatures, we cannot assess the performance of the BMA method in estimating temperatures for the period of 850–1849. However, we do think that the BMA method is feasible and applicable based on the satisfactory performance in estimating historical climate change. Furthermore, an interesting issue we want to note is that the training of the simulation data in BMA is to a great extent similar to traditional reconstruction approaches because they commonly use training data (observations) and depend on the observational period. In fact, fundamentally, a dependency is produced by various statistical methods based on samples that always contain instrumental observations, simulations, or proxies. Specifically, traditional reconstruction approaches always use training data (observations) and proxy data.
(e.g., tree-ring width) to statistically construct a relationship between observations and proxy data related to climate variables in an instrumental period with strong anthropogenic forcing and then reconstruct past climate change by extending this relationship back over the preinstrumental period. BMA is a statistical postprocessing method that utilizes training data (observations) and multimodel ensemble simulations to estimate weights and variances of individual ensemble members in the instrumental period with strong anthropogenic forcing and then uses those weights and variances to estimate climate change based on multimodel ensemble simulations over the preinstrumental period. Hence, we think that the application of both methods backwardly extends some statistical relationship or dependency. As a follow-up to this study, one may extend the BMA method to describe the uncertainties associated with proxies and proxy-based reconstructions. A general issue is that we cannot quantitatively evaluate the reliability of reconstructions and simulations in the preinstrumental period because we have no instrumental observations prior to 1850. Nonetheless, we can obtain more valuable information regarding past climate change through intercomparison of reconstructions and simulations, and the 90% confidence interval of BMA estimation can supply a meaningful reference for this issue.

In principle, the foundation of the BMA method is model training; thus, the estimate performance of the BMA method depends on raw multimodel ensemble simulations, training data, and a training algorithm. As the bias-correction method and the parameter estimation method of the training algorithm always affect the estimate performance of the BMA approach (Schmeits and Kok 2010; Tian et al. 2012; Liu and Xie 2014), more numerical experiments need to be performed by comparing the estimate performance of the BMA approach using a variety of bias-correction methods and parameter estimation methods. The simple linear bias-correction method in the BMA–EM algorithm,

TABLE 4. RMSEs and correlation coefficients (CC) between the annual global and NH temperature of the individual model simulations, the results of EWEA, CCWEA, and BMA and the annual global average temperature of HadCRUT4 from 1850 to 2005. All of the correlation coefficients are statistically significant at the $\alpha = 0.01$ level. All anomalies are computed with respect to the climatology of 1961–90.

<table>
<thead>
<tr>
<th></th>
<th>GISS</th>
<th>CCSM4</th>
<th>BCC</th>
<th>HadCM3</th>
<th>IPSL</th>
<th>MPI-ESM</th>
<th>EWEA</th>
<th>CCWEA</th>
<th>BMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global</td>
<td>RMSE</td>
<td>0.37</td>
<td>0.31</td>
<td>0.24</td>
<td>0.28</td>
<td>0.24</td>
<td>0.27</td>
<td>0.22</td>
<td>0.21</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.86</td>
<td>0.84</td>
<td>0.7</td>
<td>0.85</td>
<td>0.82</td>
<td>0.85</td>
<td>0.89</td>
<td>0.88</td>
</tr>
<tr>
<td>NH</td>
<td>RMSE</td>
<td>0.59</td>
<td>0.41</td>
<td>0.29</td>
<td>0.4</td>
<td>0.35</td>
<td>0.48</td>
<td>0.32</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>CC</td>
<td>0.83</td>
<td>0.81</td>
<td>0.68</td>
<td>0.83</td>
<td>0.8</td>
<td>0.81</td>
<td>0.85</td>
<td>0.84</td>
</tr>
</tbody>
</table>
although applicable in our study, is not optimal because the observed data and the simulated data could not be arbitrarily (or previously) fitted with a linear relationship in the instrumental age with strong anthropogenic forcing. This is why Raftery et al. (2005) suggested employing other advanced bias-correction methods in further applications of BMA. For the BMA–EM algorithm used in this study, the uncertainties involved in inaccurately predefined tolerance, iteration length, and length of the training period may give rise to suboptimal, or even nonoptimal, weights and variances associated with individual model simulations. Suboptimal or nonoptimal weights indicate that we inaccurately estimated the relative contributions of individual model simulations to the BMA results, whereas suboptimal or nonoptimal variances indicate that we underestimated or overestimated the uncertainties of individual model simulations. Suboptimal or nonoptimal weights and variances mean that we eventually cannot achieve better BMA results through Eqs. (3) and (4). Moreover, it should be noted that the weighting operators in EWEA, CCWEA, and BMA somewhat reduce the effect of the simulation-specific internal variability.

6. Conclusions

Current climate models are imperfect representations of the climate system, and different climate models have strengths for capturing different aspects of the climate system. It is desirable to have an approach that can take advantage of the diverse skill of different climate models. The BMA approach has been recently proposed as a statistical postprocessing method for producing probabilistic forecasts from a multimodel ensemble and for producing more constrained and reliable results than the currently available multimodel ensemble techniques (Raftery et al. 2005).

We investigated the potential of a BMA-based method for estimating global and NH temperatures during 850–1849 and 1850–2005 using PMIP3/CMIP5 multimodel ensemble simulations and the NCEP–NCAR reanalysis dataset. The results presented here demonstrate that the method is successful and attains a positive performance, as based on the RMSEs and correlation coefficients between the deterministic results of BMA and the HadCRUT4 dataset and proxy-based reconstructions, including comparisons between the 90% confidence intervals of BMA and the ranges of multimodel ensemble simulations. In addition, the selected proxy-based reconstructions fall almost entirely within the 90% confidence intervals of BMA; the selected proxy-based reconstructions and the BMA deterministic results commonly show warmer temperatures during the period of 850–1250 and agree on a colder period in the second half of the last 1000 years on a global scale and in the NH, although some proxy-based reconstructions are more or less warmer than afterward until 1450. Hence, we can conclude that those
selected reconstructions are consistent with the 90% confidence intervals of BMA estimations with regard to climate variability over the past 1000 years. These findings can be considered evidence of the reliability and applicability of the BMA method for estimating relatively long-term climate changes in the past. Furthermore, the method may be of use for constraining the uncertainty in estimates of past climate variability.

Neukom et al. (2014) recently found that no globally coherent warm phase occurred during the preindustrial (1000–1850) era. However, based on our results, we find that the global temperature variations over past 1000 years were almost similar to the NH variations in terms of trends, warming magnitudes, and cooling magnitudes. Regardless, our results do not allow us to contradict the findings of Neukom et al. (2014) because even for global reconstructions, the data are derived mainly from the NH. For climate change over the past 1000 years, we think that the various data sources and methods employed by different researchers may result in different interpretations that can be treated as a comprehensive dataset for understanding the possible range of past climate change.

Finally, it should be emphasized that although we selected only six climate model simulations regarding air temperature from PMIP3/CMIP5 globally and in the NH, we suggest that the BMA method can easily be applied to other specific areas (e.g., Europe, East Asia, and the Tibetan Plateau) and can consider more climate model simulations and various climate variables for a relatively long-term period.

Acknowledgments. We acknowledge the World Climate Research Programme’s Working Group on Coupled Modelling, which is responsible for CMIP. We thank the climate modeling groups for providing model output. We thank Oliver Bothe for offering valuable suggestions regarding the climate data processing. We thank Jianguo Liu for providing the BMA algorithm. The authors also would like to thank all contributors for their assistance, as well as the three anonymous reviewers and editors for their constructive comments and suggestions. This work is jointly supported by the National Science Foundation of China (NSFC) project (Grants 91425303 and 91225302), the Chinese Academy of Sciences Action Plan for West Development Program project [Remote Sensing Data Products in the Heihe River Basin: Algorithm Development, Data Products Generation and Application Experiments (KZCX2-XB3-15)], and the NSFC Heihe Watershed Allied Telemetry Experimental Research project (Grant 91125001).

REFERENCES


