APCC
APEC CLIMATE CENTER

TECHNICAL REPORT

APCC
APEC CLIMATE CENTER

PREFACE

It is our pleasure to present to you the APEC Climate Center (APCC)'s Technical Report 2012, which reports the core outcomes of our research activities from the past year.

Since 2005, APCC, as a hub of climate information in the Asia-Pacific region, has strived to share our analysis and prediction of abnormal climate and to apply this information to regional development. The Center has established the most extensive Multi-Model Ensemble (MME) system for seasonal prediction in the world through its international science network and has provided value-added products to various stakeholders. Recently, APCC has expanded its mandate to include enhancing the capacity of APEC member economies to respond effectively to climate change and variability through better application of climate information.

In 2012, APCC continued to make an effort to improve the quality and quantity of our short-term climate forecasts and our online climate information systems, as information dissemination tools. Additionally, APCC began its endeavor to produce more applicable climate information through interdisciplinary research among various sectors, such as agriculture and hydrology. The following technical report provides more information about our research outcomes from 2012.

In 2013, following APCC's goal to enhance socioeconomic well-being through better utilization of climate information, APCC will continue to improve the quality and accuracy of its climate information, recognizing that the utility of this information is only as good as its quality. We would like to make the best use of our research outcomes in various scientific and application areas. We welcome any feedback on this report or on our services.

My best and warmest regards to all of you.

Dr. Chin-Seung Chung
Director/APEC Climate Center

CONTENTS

APCC

TECHNICAL REPORT 2012-01

Application of Bayesian Model Averaging on Multi-Model Ensemble Seasonal Prediction

Dr. Hongwei Yang

ABSTRACT

To obtain the optimal weights for combining different model outputs in operational seasonal forecasts, Bayesian model averaging (BMA) was applied to the multimodel hindcast datasets at the Asia-Pacific Economic Cooperation (APEC) Climate Center (APCC). The weights were estimated according to the performance of individual members in simulating the given training data. Verification measurements such as the mean squared skill score (MSSS) and anomaly pattern correlation coefficient (ACC) were used to evaluate the forecast performance based on the observation. In terms of the MSSS, over the tropics, the BMA methods generally have broader areas with higher skills than the equal weight methods, whereas, over the global domain, the equal weight methods generally have larger areas with positive skills than the BMA methods. In terms of the Pearson ACC, the equal weight methods generally have higher skills than the BMA methods, whereas the BMA methods generally have higher skills than the equal weight methods in terms of the robust ACC.

1. INTRODUCTION

Seasonal forecasts are associated with the persistence of anomalies from slow fluctuations in surface temperature, soil moisture, and snow cover (Charney and Shukla 1981; Brankovic *et al.* 1994), as well as the stochastic nature of the climate system. Deterministic forecast is a practical way of predicting the atmospherical quantities that are of most concern to people.

The single-model ensemble approach, which can reduce the internal uncertainty of the model, is widely used in seasonal forecast. The multimodel ensemble (MME) method, which can reduce not only the internal uncertainty of an individual model but also the uncertainty between models (e.g., Kalnay and Ham 1989; Krishnamurti *et al.* 1999; Shukla *et al.* 2000; Wang *et al.* 2004), has surged to meet a great societal demand. An MME prediction system is used only by a few operational centers with different MME approaches. The International Research Institute for Climate and Society (IRI) MME prediction system is based on two approaches: a Bayesian method (Rajagopalan *et al.* 2002; Robertson *et al.* 2004) and a canonical variate technique method (Mason and Mimmack 2002). The Meteorological Service of Canada (MSC) uses a pooling method, which pools all the participating models into a single sample

with equal weights (e.g., Hagedorn *et al.* 2005). The Asia–Pacific Economic Cooperation (APEC) Climate Center (APCC) developed a probabilistic MME prediction system (Min *et al.* 2009) that is suitable when the weights are inconsistent between hindcast and forecast datasets, and with the individual model ensembles essentially differing in size. Besides the probabilistic MME forecast, APCC has employed other deterministic forecast methods, e.g., the simple composite methods (SCM), SPM (Kug *et al.* 2008), (multiple regression) MGR (Krishnamurti *et al.* 2000; Yun *et al.* 2003), and synthetic super ensemble (SSE) (Yun *et al.* 2005).

The SCM is a deterministic forecast scheme; it is an arithmetic mean of predictions based on participated individual models. In the SCM, it is assumed that a independent relationship exists among the participated models. This ensemble method has numerous applications such as climate change study, decadal climate prediction, and regional climate downscaling. Numerous studies (e.g., Hagedorn *et al.* 2005; Doblas-Reyes *et al.* 2005) showed that the SCM can generally produce a higher skill forecast compared with the involved individual models just by simple arithmetic average. This scheme maintains the model dynamics owing to the simple spatial filtering for each variable at all grid points. Additionally, the SCM has the common advantages and limitations of the individual forecast model; therefore, it could be used as a benchmark to evaluate other optimal MME schemes. The SCM scheme constructed with bias-corrected forecast is given by

$$S = \overline{O} + \frac{1}{N}\Sigma\,(M_i - \overline{M_i})$$

where M_i is the i^{th} model forecast at a certain time, $\overline{M_i}$ and $\overline{O}$ are the climatological mean of the i^{th} forecast and observation, respectively, and N is the number of the forecast models involved. In this scheme, the same weight of $1/N$ is assigned to each of the N member models over the entire grid regardless of their actual performance. Therefore, the result of the SCM is generated by combining bias-corrected anomalies from individual model forecasts. Skill improvement is achieved through the bias removal preprocesses as well as from the cancellation of the climate noise by ensemble averaging because of the independence of the individual models.

Second, the SSE method (Yun *et al.* 2005) is adopted as another official forecast method at APCC. Although the performance of both dynamic and statistic models were improved over the last few decades, the seasonal forecast skills are generally still low. MME forecasts rely on the statistical relationships established between historical forecasts and verification data (Chang *et al.*, 2000), implying that the MME forecast strongly depends on the historical performance of individual models. In the field of seasonal forecasts, numerous studies (Krishnamurti *et al.*, 1999, 2000a,b, 2001, 2003; Doblas-Reyes *et al.*, 2000; Pavan and Doblas-Reyes 2000; Stephenson and Doblas-Reyes 2000; Kharin and Zwiers 2002; Peng *et al.*, 2002; Stefanova and Krishnamurti, 2002; Yun *et al.*, 2003; Palmer *et al.*, 2004) have discussed various MME approaches for anomalous forecasting, such as the simple ensemble, the unbiased ensemble, and the superensemble forecasts:

$$E_b = \frac{1}{N}\sum(M_i - \overline{O})$$

$$E_c = \frac{1}{N}\sum(M_i - \overline{M_i})$$

$$S = \sum a_i(M_i - \overline{M_i})$$

Here, E_b is the simple ensemble mean, E_c, is the unbiased ensemble mean, S is the superensemble mean, M_i is the i^{th} model forecast from the N models, and $\overline{M_i}$ is the climatological mean of the i^{th} model forecast during the training period. $\overline{O}$ is the observed climatological mean during the training period, and a_i is the regression coefficient of the i^{th} model. The above three methods are different in that different means and weights are used in the ensemble. Both the unbiased ensemble and the superensemble contain no bias in the means because the seasonal climatologies are determined before the ensemble approach is used. The weights assigned to the individual models in the unbiased ensemble and the superensemble differ. A key property of the superensemble forecast is the training process carried out on the hindcast datasets. The skill of the superensemble forecast could be improved after the multimodel hindcasts are statistically corrected to reduce the systematic errors. The SSE method has the following steps. First, a transient dataset is produced

by determining the consistence of the spatial anomalous patterns between the verification dataset and the forecasts of individual models. This is a linear regression problem in the space of empirical orthogonal functions (EOFs). Then, the transient dataset is used as the ensemble member of the superensemble forecast. The algorithm is briefly introduced as follows:

Both the verification data (O) and the forecast of the individual models (M_i) can be decomposed in the EOF space, which describes the spatial and temporal variability:

$$O(x,t) = \sum \widetilde{O}_n(t)\phi_n(x)$$

$$M_i(x,T) = \sum \widetilde{M}_{i,n}(T)\varphi_{i,n}(x)$$

where $\widetilde{O}_n(t)$, $\widetilde{M}_{i,n}(t)$ and $\phi_n(x)$, $\varphi_{i,n}(x)$ are the principal component (PC) time series and the corresponding EOFs of the n^{th} mode for the verification data and the individual model forecast, respectively. The PCs represent the temporal evolution of spatial patterns during the training time (t) and the entire forecast time period (T). The consistent pattern can be then estimated between the verification and forecast data according to the PC time series during the training period. The linear regression relationship of the PC time series between the verification data and the individual model forecast can be written as

$$\widetilde{O}(t) = \sum \alpha_{i,n}\widetilde{M}_{i,n}(t) + \varepsilon_{i,n}(t)$$

Then, the time series of the verification data can be written as a linear combination of the time series of individual models (here, referred to as the predictors). To obtain the coefficients $\alpha_{i,n}$, the linear regression is calculated in the EOF space. The regression coefficients $\alpha_{i,n}$ are obtained such that the residual error is minimized. The seasonal cycle is removed before constructing of the covariance matrix from the PC time series of individual models. After the regression coefficients $\alpha_{i,n}$ are obtained, the new PC time series (approximating the PC time series of the verification data) can be written as

$$\widetilde{M}_i^{reg}(T) = \sum_n \alpha_{i,n}\widetilde{M}_{i,n}(T)$$

The final forecast can be reconstructed through the new PC time series and EOFs of the verification data

$$\widetilde{M}_i^{syn}(x,T) = \sum_n \widetilde{M}_{i,n}^{reg}(T)\phi_n(x)$$

The unique property of this approach is that the variance in the residual error between the verification data and each of the individual models in the EOF space is minimized.

The other approach used as an APCC official MME forecast is the MRG (Krishnamurti *et al.* 2000b; Yun *et al.* 2003) method. The conventional multimodel superensemble forecast (Krishnamurti *et al.*, 2000b) constructed from a bias-corrected forecast is given by

$$S = \overline{O} + \sum a_i (M_i - \overline{M_i})$$

where M_i is the i^{th} model forecast, $\overline{M_i}$ is the climatological mean of the i^{th} model forecast during the training period, $\overline{O}$ is the climatological mean of the verification data during the training period, and a_i is the regression coefficient obtained from the verification data and the hindcast of individual models during the training period. Because the anomalies $M_i - \overline{M_i}$ of each individual model are used in the ensemble approach, the multimodel superensemble forecast is not influenced by the systematic errors of the individual models. The regression technique is carried out at each spatial grid point, between the verification data and the hindcast of the individual models during the training period; thus the weight differs for different grids for the same model. To obtain the weights, the covariance matrix is built on the anomalies M' instead of on the total field. The seasonal cycle of the hindcast field is removed before establishing the covariance matrix:

$$C_{i,j} = \sum M'_{i,t} M'_{j,t}$$

where the summation is over the training period. i and j are respectively the i^{th} and i^{th} forecast models. Linear regression can establish a linear relationship between the vectors of two variables. First, suppose a set of linear equations exist,

$$C \bullet x = \widetilde{O'}$$

where $\widetilde{O'} = \sum O'_t M'_{j,t}$, is a $n \times 1$ vector containing the covariances of the verification data and the individual hindcasts for which a linear relationship has to be found, and O' is the seasonal mean-removed verification anomaly, C is the $n \times n$ covariance matrix, and x is a $n \times 1$ vector of regression coefficients (the unknowns). In the normal superensemble approach, the regression coefficients are solved by using Gauss–Jordan elimination. The covariance matrix C and O' are rearranged into a diagonal matrix C' and O'', respectively, and the solution is obtained as follows:

$$x^T = \left(\frac{\tilde{o}''_1}{C'_{11}}, \ldots, \frac{\tilde{o}''_n}{C'_{nn}} \right),$$

where the superscript T denotes the transpose. The Gauss–Jordan elimination for solving the regression coefficients for different hindcasts is not numerically robust. Problems arise if a zero pivot element exists on the diagonal; this will break the program by the exception of a zero denominator. On the other hand, if there are fewer equations than unknowns, the regression equation defines an underdetermined system such that there are more regression coefficients than the number of $\{o'_j\}$ values. Thus, there is no unique solution, and the covariance matrix is singular. Generally, the Gauss–Jordan elimination is not recommended because the singularity problem mentioned above is occasionally happened. In practice, when a singularity occurs, the superensemble forecast is replaced by a normal ensemble forecast.

The SVD method is another alternative for solving the regression coefficients for a set of multimodel hindcasts and for the verification data. In this method, the covariance matrix C is decomposed into a product of three matrixes. The covariance matrix C can be written as a sum of the outer products of the columns from matrix U and the rows from matrix V^T as follows:

$$C_{i,j} = (UWV^T)_{i,j} = \sum w_k U_{i,k} V_{j,k}$$

where U and V are $n \times n$ matrices that satisfy the orthogonal relations and W is a $n \times n$ diagonal matrix containing rank k real positive singular values (w_k) in

decreasing order. Because of the symmetry of the covariance matrix C, i.e., $C^T = VWU^T = UWV^T = C$ the left singular vector U is equal to the right singular vector V Therefore, the SVD method can also be referred to as a principal component analysis (PCA). The decomposition can be used to obtain the regression coefficients:

$$x = V \bullet \left[diag\left(\frac{1}{w_j} \right) \right] \bullet \left(U^T \bullet \tilde{Q}' \right)$$

The SVD method avoids the singular matrix problem that cannot be solved by the Gauss–Jordan elimination method.

For the SVD method, another technique is usually used when some of the w_j values are very low. The low w_j values will increase the residual error in the SVD method (Press *et al.* 1992). These low singular values can be discarded by setting them to zero. In other words, if most of the w_j values of the matrix C are low, then C will be better approximated by only a few high w_j singular values in the summation of the $C_{i,j}$.

The fourth method used at APCC is the stepwise pattern projection (SPPM) method (Kug *et al.* 2008), which is based on the statistical downscaling method. The SPPM technique is an improved version of the current CPPM method in the APCC MME methods. The differences between the two methods lie in the selection of the pre-predictor and the posterior prediction. The SPPM procedure has three steps: pre-predictor selection, pattern projection, and optimal choice of prediction. First, predictors are chosen based on the cross-validated correlation during the training period. The new predictor is reconstructed over the entire grid by using the top 100 predictors that correlated best with the predictand. Second, the covariance pattern is constructed using the verification pattern and the new predicted pattern. Then, the prediction is made by projecting the predicted pattern onto the covariance pattern. Third, to determine whether the selected predictand is acceptable, the predictand will be verified through double cross-validation with a given threshold on the correlation skill. If the prediction skill is less than the threshold, the predictor and the corresponding prediction will be ignored. All the acceptable predictions will be combined through the arithmetic average. It is shown that SPPM shows higher skills

over the regions where the individual model skill is generally poor. Wang *et al.* (2010) showed that the SPM generally has better skill than other MME schemes, although in some cases the SCM is better.

There are many studies of the ensemble forecast methods based on Bayesian inference, e.g., Coelho *et al.* 2006. But in recent times, the optimized ensemble method of Bayesian model averaging (BMA) (Hoeting *et al.* 1999) has been increasingly used in ensemble prediction. Raftery *et al.* (2005) successfully applied BMA to a 48-h regional weather forecast on surface temperature and sea level pressure. Min *et al.* (2007) conducted their climate change study based on BMA. Duan *et al.* (2007) applied BMA to hydrologic prediction and found that BMA prediction is generally superior to the best individual prediction method. Marrocu and Chessa (2008) found that BMA prediction is better than the SCMs in seasonal downscaling forced by reanalysis datasets.

Thus, we are motivated to apply BMA to the seasonal forecast products from different models at APCC. APCC has conducted both the deterministic and probabilistic seasonal forecast based on various ensemble methods developed in-house. The application of the BMA ensemble method to seasonal forecast will help us gain new knowledge.

BMA is a statistical approach for generating a weighted average of the ensemble members that outperform any single ensemble member. The weights are estimated according to the performance of individual members in simulating the given training data. The BMA forecast variance can analytically decompose into two components, corresponding to between-model variance and within-model variance.

By applying BMA to the seasonal forecast of temperature and precipitation of APCC datasets, we try to address the following questions: can the BMA ensemble method improve the seasonal forecast skill compared to the equal weight MME?

To answer the question, one-month lead winter low-level temperature and summer precipitation hindcast from multimodel outputs were used as the inputs for both the equal weight ensemble and the BMA weight ensemble in multiple years. The data, the evaluation method, and the experiments are described in Section 2

and Section 3. The application of BMA is discussed in Section 4. Section 5 presents the results of both MME methods. Section 6 is the conclusion, and the final section presents the discussion.

2. MODEL AND DATA

Winter (DJF (December-January-Feburary)) one-month lead temperature hindcast data from thirteen multimodel outputs were used in this MME study for both the equal weight ensemble and the BMA weight ensemble from 1983 to 2003, a total of 21 years. The model outputs include JMA, NCEP, PNU, SINT, UHT1, POAMA, APCC, BCC, CWB, GDAPS_F, MSC_GM2, MSC_GM3, MSC_SEF. Summer (JJA (June-July-August)) one-month lead precipitation hindcast data from fourteen models with an additional output of SUT1 besides that of the above models were selected by considering their common temporal expansion availability from 1983 to 2003. Precipitation in summer is more difficult to forecast than that in winter. The difficulty to forecast the 850 temperature is comparable in both seasons, only with the forecast skill of the summer temperature being slightly lower than that in winter.

BCC is the model output of the Beijing Climate Center of China, which is a T63L16 model using predicted SST with eight ensemble members. CWB is the model output of the Central Weather Bureau of Chinese Taipei, which is a T42L18 model using observed SST with 10 ensemble members. GDAPS_F is the model output of the Korea Meteorological Administration, which is a T106L21 model using predicted SST with 20 ensemble members. JMA is the model output of the Japan Meteorological Agency, which is a T95L40 model using predicted SST with five ensemble members. MSC_GM2 is the model output of the Meteorological Service of Canada, which is a T32L10 model using predicted SST with 10 ensemble members. MSC_GM3 is the model output of the Meteorological Service of Canada, which is a T63L32 model using predicted SST with 10 ensemble members. MSC_SEF is the model output of the Meteorological Service of Canada, which is a T95L27 model using predicted SST with 10 ensemble members. NCEP is the model output of the National Centers for Environmental Prediction of USA, which is a T62L64 model using predicted SST with 15 ensemble

members. PNU is the model output of the Pusan National University of Korea, which is a T42L18 model using predicted SST with five ensemble members. POAMA is the model output of the Centre for Australian Weather and Climate Research of the Bureau of Meteorology of Australia, which is a T63L21 model using predicted SST with 10 ensemble members. UHT1 is the model output of the University of Hawaii of USA, which is a T31L19 model using predicted SST with 10 ensemble members. SINT is the model output of FRCGC from Japan, which is a T106L19 model using predicted SST with nine ensemble members. APCC is the model output of APCC CCSM3, which is a T85L26 model using predicted SST with five ensemble members. SUT1 is the model output from the Seoul National University of Korea, which is a T42L21 model using predicted SST with six ensemble members.

For evaluating the forecast skill, we first select the mean squared skill score (MSSS) as one of the verification methods. Let x_{ij} and m_{ij} (i = 1,...,n) denote the time series of verification data and seasonal forecasts of individual models, respectively, for a grid point or station j over the period of verification. Then, their averages $\overline{x_j}$, $\overline{m_j}$ and their sample variances S_{xj}^2 and S_{fj}^2 are given by

$$\overline{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad \overline{m_j} = \frac{1}{n}\sum_{i=1}^{n} m_{ij}$$

$$s_{xj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2, \qquad s_{mj}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(m_{ij} - \overline{m_j})^2$$

The mean squared error of the forecasts is

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}(m_{ij} - x_{ij})^2$$

For the case of cross-validated climatology forecasts where forecast/observation pairs are reasonably temporally independent of each other, the mean squared error of climatology forecast (Murphy, 1988) is

$$MSE_{cj} = \frac{n-1}{n}s_{xj}^2 = \frac{1}{n}\sum_{i=1}^{n}(\overline{x_j} - x_{ij})^2$$

The MSSS for j is defined as one minus the ratio of the squared error of the forecasts to the squared error of the forecasts of climatology:

$$MSSS_j = 1 - \frac{MSE_j}{MSE_{cj}} = \frac{MSE_{cj} - MSE_j}{MSE_{cj} - 0} = \frac{A - A_{ref}}{A_{perfect} - A_{ref}}$$

For a certain domain, it is recommended that an overall MSSS be provided. This is computed as

$$MSSS = 1 - \frac{\sum_j w_j MSE_j}{\sum_j w_j MSE_{cj}}$$

where w_j is unity for verifications at stations and is equal to $\cos(\theta_j)$, where θ_j is the latitude at grid point j on latitude–longitude grids.

For either the MSSS or MSSSj, a corresponding root mean squared skill score (RMSSS) can be obtained easily from

$$RMSSS = 1 - (1 - MSSS)^{1/2}$$

$MSSS_j$ for forecasts fully cross-validated can be expanded as

$$MSSS_j = \left\{ 2\frac{s_{fj}}{s_{xj}} r_{mxj} - \left(\frac{s_{fj}}{s_{xj}}\right)^2 - \left(\frac{\left[\overline{m_j} - \overline{x_j}\right]}{s_{xj}}\right)^2 + \frac{2n-1}{(n-1)^2} \right\} \Bigg/ \left\{ 1 + \frac{2n-1}{(n-1)^2} \right\}$$

where r_{mxj} is the product–moment correlation of the forecasts and observations at point or station j.

$$r_{mxj} = \frac{\frac{1}{n}\sum_{i=1}^{n}(m_{ij} - \overline{m_j})(x_{ij} - \overline{x_j})}{s_{mj} s_{xj}}$$

The first three terms of the decomposition of the $MSSS_j$ re related to phase errors (through the correlation), amplitude errors (through the ratio of the forecast to observed variances), and overall bias error, respectively, of the forecasts. These terms

provide the opportunity for those wishing to use the forecasts for input into regional and local forecasts to adjust or weigh the forecasts as they deem appropriate. The last term takes into account the fact that the climatology forecasts are cross-validated as well.

The second verification method used in this study is an anomaly pattern correlation coefficient (ACC). It is used to quantify the spatial pattern correlation between the forecast and observed deviations from climatology (Miyakoda *et al.* 1972), containing aspects of forecast error and bias. The ACC measures pattern similarity, which leads to higher scores during periods of amplified versus zonal flow.

$$ACC = \frac{\sum_{n=1}^{N} m_n' o_n'}{\left[\sum_{n=1}^{N} (m_n')^2 \sum_{n=1}^{N} (o_n')^2\right]^{1/2}}$$

where $m'_n = m_m - c_n, o'_n = o_n - c_n$, m is the forecast, and o is the observed verification data. Here, the ACC is estimated between each MME mean and verification data as well as between each individual model (member) and verification data.

Beside the classic ACC (the Pearson ACC), the robust anomaly pattern correlation coefficient (the robust ACC), which is based on the robust statistic between the forecast anomaly and the anomaly from the verification data, is also used as the evaluation method in this study. The aim of robust methods is to ensure high stability in statistical inference under the deviations from the assumed distribution model. Less attention is devoted to the literature to robust estimators of association and correlation as compared to robust estimators of location and scale (Huber 1981; Hampel *et al.* 1986; Maronna and Yohai 1995, 2006; Rousseeuw and Driessen 1999; Woodruff and Rocke 1994; Maronna and Zamar 2002). However, it is necessary to study these problems owing to their widespread occurrence (e.g., estimation of the correlation and covariance matrices in regression and multivariate analysis, estimation of the correlation functions of stochastic processes) and because of the instability of classical methods in estimating the presence of outliers in the data.

3. EXPERIMENTAL DESIGN

In this study, we used the model anomalies to eliminate the climatological model biases (i.e., systematic errors in the climatology or mean bias of the forecast). For each individual model, the anomalies are estimated as departures from their climatology over the training period in a one-year-out-cross-validation manner (Wilks 1995). This implies that a new climatology is calculated at each cross-validation step, with the target year being withheld. The same procedure for estimating anomalies is applied to the observed dataset. All the anomalies are bias-corrected before validation and verification, by using a linear regression method.

The equal weight ensemble hindcast was carried out using two methods. One is the member pooling method named EQU_mem:

$$S = \overline{O} + \frac{1}{n}\sum(m_i - \overline{m_i})$$

where m_i, is the i^{th} member hindcast, $\overline{m_i}$ is the climatological mean of the i^{th} hindcast over the training period, and $\overline{O}$ is the climatological mean of the verification data over the training period.

The other method is the model pooling method named EQU_mod, which is identical to the SCM method mentioned in the introduction:

$$S = \overline{O} + \frac{1}{N}\sum(M_i - \overline{M_i})$$

where M_i is the i^{th} model hindcast and $\overline{M_i}$ is the climatological mean of the i^{th} hindcast over the training period. Equal weight was assigned to all ensemble members of the multimodel in the first ensemble method, whereas equal weight was assigned to all ensemble models in the second ensemble method.

The BMA ensemble hindcast was produced in three different ways. First, the ensemble member of each model was averaged as both training and forecast data before carrying out BMA. This method was named BMA_average and is as follows :

$$S = \overline{O} + \sum W_i (M_i - \overline{M_i})$$

where w_i is the BMA weight for the i^{th} model.

Second, the ensemble members of one model can be exchangeable and each member of that model has equal weights. It was named BMA_ex and is as follows:

$$S = \overline{O} + \sum \varpi_i (m_i - \overline{m_i})$$

where $\overline{w_i}$ is the BMA weight for the i^{th} member.

Third, all the ensemble members of the multimodel were distinguishably treated in the BMA method. It was named BMA_mem and is as follows:

$$S = \overline{O} + \sum w_i (m_i - \overline{m_i})$$

where w_i is the BMA weight for the i^{th} member.

The seasonal forecast of summer precipitation and the winter 850 mb temperature will be verified using the cross-validated approach based on the Climate Prediction Center Merged Analysis of Precipitation (CMAP, Xie and Arkin 1997) dataset and the 850 mb winter temperature obtained from the National Center for Environmental Prediction-Department of Energy reanalysis data (NCEP-R2, Kanamitsu *et al.* 2002).

4. APPLICATION OF BMA

BMA is a statistical approach for generating the ensemble mean that outperforms any single ensemble member (Hoeting *et al.* 1999). Raftery *et al.* (2005) successfully applied BMA to a 48-h regional weather forecast on surface temperature and sea level pressure. Min *et al.* (2007) carried out the climate change study based on BMA. Duan *et al.* (2007) applied BMA to hydrologic prediction and found that BMA prediction is generally superior to the best individual prediction. Sloughter *et al.* (2007, 2010) applied BMA to precipitation and wind speed forecast. BMA is different from other multimodel ensemble methods in that it provides a more realistic description of

the predictive uncertainty that accounts for both between-model variances and in-model variances. In this study, we applied BMA to the 21-year DJF low-level temperature data and JJA precipitation data from multimodel hindcast datasets at APCC. The approach is briefly described as follows.

Because each seasonal hindcast of the multimodel showed the seasonal climate behavior, the winter 850 mb temperature and summer precipitation of the hindcast datasets could be considered the estimates of the truth of seasonal climates. The posterior probability $p(M_k^w|O^w)$ is the probability of the temperature or precipitation simulated by the k-th model or member conditional on the verification temperature or precipitation, which reflected how well the hindcast mimicked the verification temperature or precipitation in the given places. The sum of these posterior probabilities is equal to one; thus, they could be regarded as a set of weights for measuring the skills of models or members. Then, the probability of the BMA ensemble mean of temperature or precipitation was the weighted sum of the probability of the temperature or precipitation from the ensemble models or members:

$$p(y) = \sum_{k=1}^{N} w_k p(y|M_k)$$

where y is the BMA ensemble mean of temperature or precipitation, $p(y|M_k)$ is the probability of simulated temperature or precipitation based on the k-th model or member, and the weight w_k is the posterior probability $p(M_k^w|O^w)$

The verification training data is the 850 mb winter temperature from the NCEP-R2 reanalysis and the CMAP global precipitation with both 2.5° horizontal resolution.

Under the assumption that the conditional distribution $p(y|M_k)$ of temperature is a normal distribution, we calculated the weights for temperature through the expectation–maximization algorithm (e.g., Hoeting *et al.* 1999 and Duan *et al.* 2007). Usually we use the log-likelihood function to maximize the likelihood function instead of the likelihood function itself. We denote

$$\theta = \left[\{ w_{k,} \sigma_k, k = 1, 2, \ldots, K \} \right]$$

The log-likelihood function can be approximated as

$$l(\theta) = \log\left(\sum_{k=1}^{k} w_k \bullet p_k(y \mid M_k)\right)$$

Obviously, it is impossible to solve the analytical solution of θ, and therefore, the numerical computational method must be used. Following the recommendation of Raftery *et al.* (2005), we used the expectation–maximization algorithm to calculate the maximum of $l(\theta)$ The expectation–maximization algorithm treats the maximum likelihood issue as a "missing data" problem. The "missing data" is a virtual variable without any real physical meaning, but a variable introduced mathematically to solve the mathematical problem. In this study, a virtual variable $Z_{k,t}$ is introduced. If the k-th model or member is the best estimation at time t, $Z_{k,t} = 1$; otherwise, $Z_{k,t} = 0$. At any given time t, there is only one value of $Z_{k,t}$ equal to 1, while all the others are equal to zero. The expectation–maximization algorithm alternates between the expectation step and the maximization step. It starts with an initial guess $\theta^{(0)}$ for parameter θ. In the expectation step, $Z_{k,t}$ is estimated given the current guess of θ. In the maximization step, θ is estimated given the current values of $Z_{k,t}$. The expectation–maximization steps are repeated until certain convergence criteria are satisfied. The expectation–maximization algorithm is illustrated as the following five steps (Duan *et al.* 2007).

Step 1- Initialization:

$$\text{Let } I = 0,\ w_{k,I} = \frac{1}{K},\ \sigma^2_{k,I} = \frac{1}{K}\sum_{t=1}^{T}\frac{(\sum_{k=1}^{K}(O_t^w - M_{k,t}^w)^2)}{T}$$

where T is the total number of datapoints in the training datasets.

Step 2 - Initializing the log-likelihood function:

$$l(\theta_I) = \log\left(\sum_{k=1}^{K} w_k \cdot p_k(y \mid M_k)\right)$$
$$= \log\left(\sum_{k=1}^{K} w_k \cdot g(O_t^w \mid M_{k,t}^w, \sigma_{k,I})\right)$$

where $g(\cdot)$ is the Gaussian distribution.

Step 3 - Computing the expectation:

For I = I+1, k = 1,2,3,4, and t = 1,2,···,T,

$$Z_{k,t,I} = \frac{g\left(O_t^w \mid M_{k,t}^w, \sigma_{k,I-1}\right)}{\sum_{k=1}^{K} g\left(O_t^w \mid M_{k,t}^w, \sigma_{k,I-1}\right)}$$

Step 4 - Computing the maximization:

$$\text{Calculate the weight } w_{k,I} = \frac{1}{T}\sum_{t=1}^{T} Z_{k,t,I}$$

$$\text{Update the variance } \sigma_{k,I}^2 = \frac{\sum_{t=1}^{T} Z_{k,t,I} \cdot (O_t^w - M_{k,t}^w)^2}{\sum_{t=1}^{T} Z_{k,t,I}}$$

Step 5 - Calculating the criteria:

By updating the logarithmic likelihood in step 2, if $l(\theta_I) - l(\theta_{I-1}) < \eta$, where η is a given tolerance number, we get what we want and stop; otherwise, we go back to step 3.

For a detailed description of the expectation–maximization algorithm, readers are referred to McLachlan and Krishnan (1997).

For precipitation, the assumption of normality is not suitable any more. First, the precipitation is positive and usually it is zero. Second, the probability distribution of the positive part is highly skewed. In this study, we employed the model developed by Sloughter *et al.* (2007):

$$P(y \mid M_k) = P(y = 0 \mid M_k) I[y = 0] + P(y > 0 \mid M_k) g_k(y \mid M_k) I[y > 0]$$

where y is the cube root of the precipitation; the indicator function $I[]$ is equal to 1 if the condition in brackets is satisfied and is equal to 0 otherwise. The probability $P(y = 0 \mid M_k)$ is the probability of no precipitation occurring given the forecast M_k The probability $P(y > 0 \mid M_k)$ is the probability of positive precipitation occurring given the forecast M_k. Those two probabilities satisfy the following logistic regression model:

$$\text{log it } P(y = 0 \mid M_k) \equiv \log \frac{P(y = 0 \mid M_k)}{P(y > 0 \mid M_k)} = a_{0k} + a_{1k} M_k^{1/3} + a_{2k} \delta_k$$

where a_{0k}, a_{1k} and a_{2k} are the regression cofficients. δ_k is equal to 1 if $M_k = 0$ and is equal to 0 otherwise. The conditional PDF $g_k(y \mid M_k)$ is a gamma distribution given that the cube root precipitation y is positive:

$$g_k(y \mid M_k) = \frac{1}{\beta_k^{\alpha_k} \Gamma(\alpha_k)} y^{\alpha_k - 1} \exp(-y / \beta_k)$$

The parameters of the gamma distribution can be determined by M_k from

$$\mu_k = b_{0k} + b_{1k} M_k^{1/3}$$

and

$$\sigma_k^2 = c_{0k} + c_{1k} M_k$$

where $\mu_k = \alpha_k \beta_k$ is the mean of the gamma distribution and $\sigma_k^2 = \alpha_k \beta_k^2$ is its variance. We calculate the log-likelihood function to maximize the likelihood function. The expectation–maximization algorithm used for the temperature was used to estimate the parameters required. For a detailed description of the expectation–maximization algorithm used for precipitation, readers are referred to Sloughter *et al.* (2007).

5 RESULT

5.1 Temperature

To evaluate the hindcast skill of the BMA methods and the equal weight methods, we first checked their performances in terms of the MSSS.

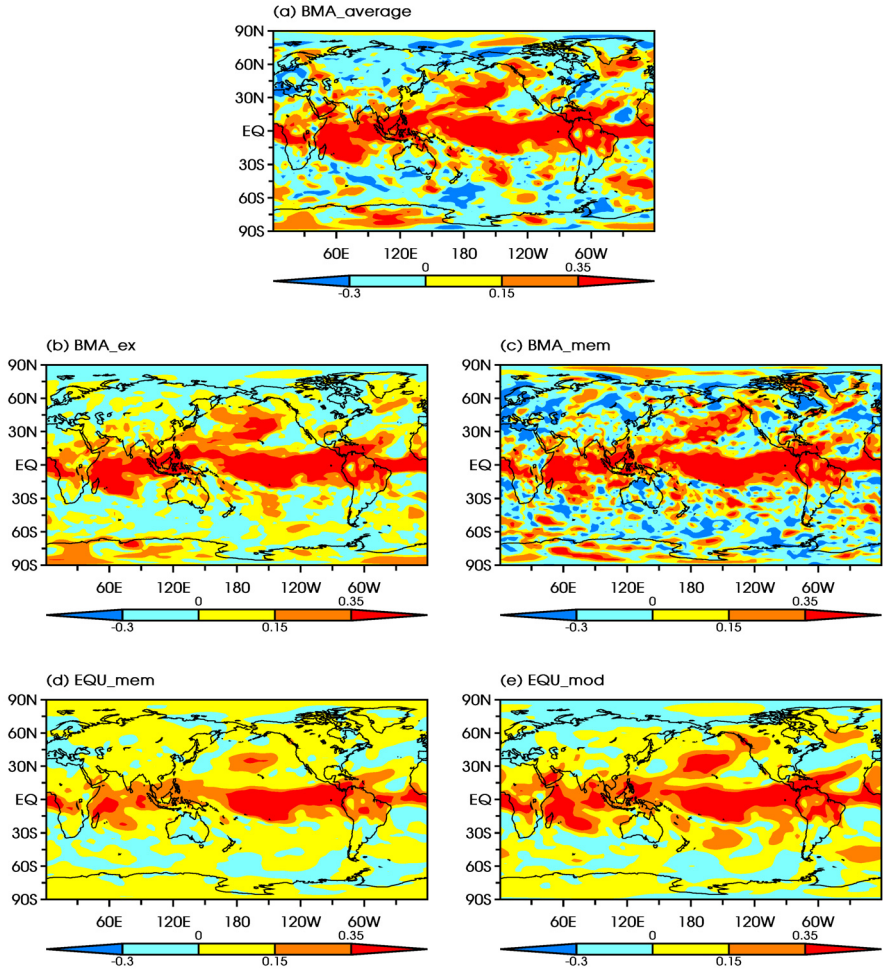


Figure 1 Mean squared skill score (MSSS) of one-month lead DJF temperature hindcast at 850 mb over the global domain from 1983 to 2003. (a), (b), (c), (d), and (e) are obtained from the MME hindcast of the BMA_average, the BMA_ex, the BMA_mem, the EQU_mod, and the EQU_mem methods, respectively. The verification data is the NCEP-R2 reanalysis.

Figure 1 shows the MSSS for the global winter temperature hindcast. Over the Maritime Continent, the MSSS is the highest in the MME of the BMA_average method, whereas it is the lowest in the MME of the EQU_mem method. Over the tropical Atlantic, the BMA methods have a higher MSSS than the equal weight methods. Over the tropical Pacific, the BMA methods show broader areas with a high MSSS than the equal weight methods. Over the North Pacific, the BMA_average method shows the largest area with an MSSS higher than 0.35 compared with the other MME methods. Over the Indian Ocean, the BMA methods generally outperform the equal weight methods, whereas the EQU_mod method may have broader areas with an MSSS higher than 0.15. Generally, the BMA methods have higher skills than the equal weight methods over the tropical areas but have lower skills over the extratropical areas than the equal weight methods. Over the global domain, the BMA_average method and the BMA_mem methods show many locations with an MSSS lower than −0.3, whereas the other three methods do not show an MSSS lower than −0.3. The BMA_mem method frequently shows discontinuities in areas with positive or negative skill. The EQU_mem method shows the largest areas with positive skills compared to the other methods. The EQU_mod method shows broader areas with an MSSS higher than 0.35 over the tropical areas than the EQU_mem method. Over the tropical areas (-15S to 15N), the overall MSSSs (see definition in section 2) are 0.348, 0.350, 0.293, 0.265, and 0.324 for the BMA_average, the BMA_ex, the BMA_mem, the EQU_mem, and the EQU_mod methods, respectively. Over the Nino3 region (-5S-5N, 90W-150W), the overall MSSSs are 0.704, 0.720, 0.667, 0.593, and 0.637 for the BMA_average, the BMA_ex, the BMA_mem, the EQU_mem, and the EQU_mod methods, respectively.

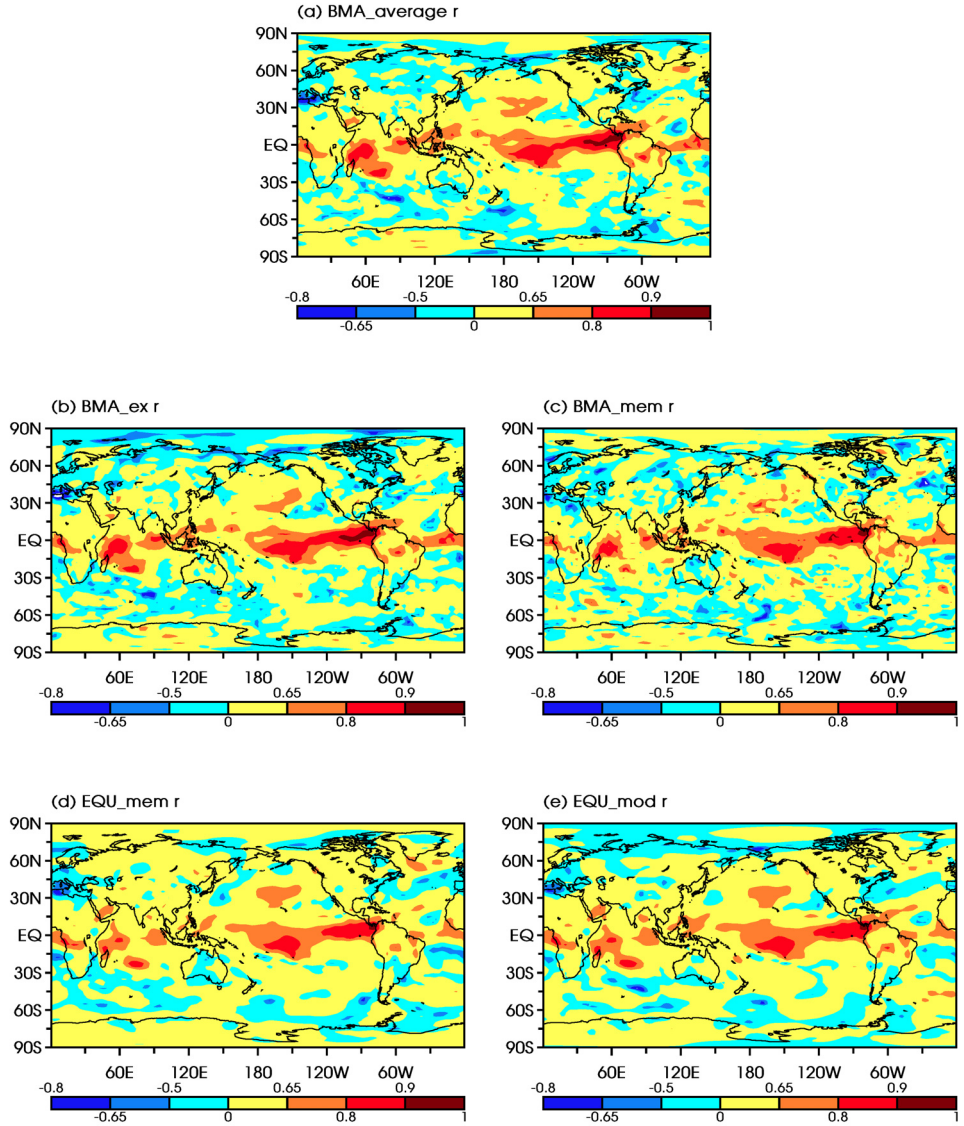


Figure 2 Correlation component of the decomposed MSSS of one-month lead DJF temperature hindcast at 850 mb over the global domain from 1983 to 2003. (a), (b), (c), (d), and (e) are obtained from the MME hindcast of the BMA_average, the BMA_ex, the BMA_mem, the EQU_mod, and the EQU_mem methods, respectively. The verification data is the NCEP-R2 reanalysis.

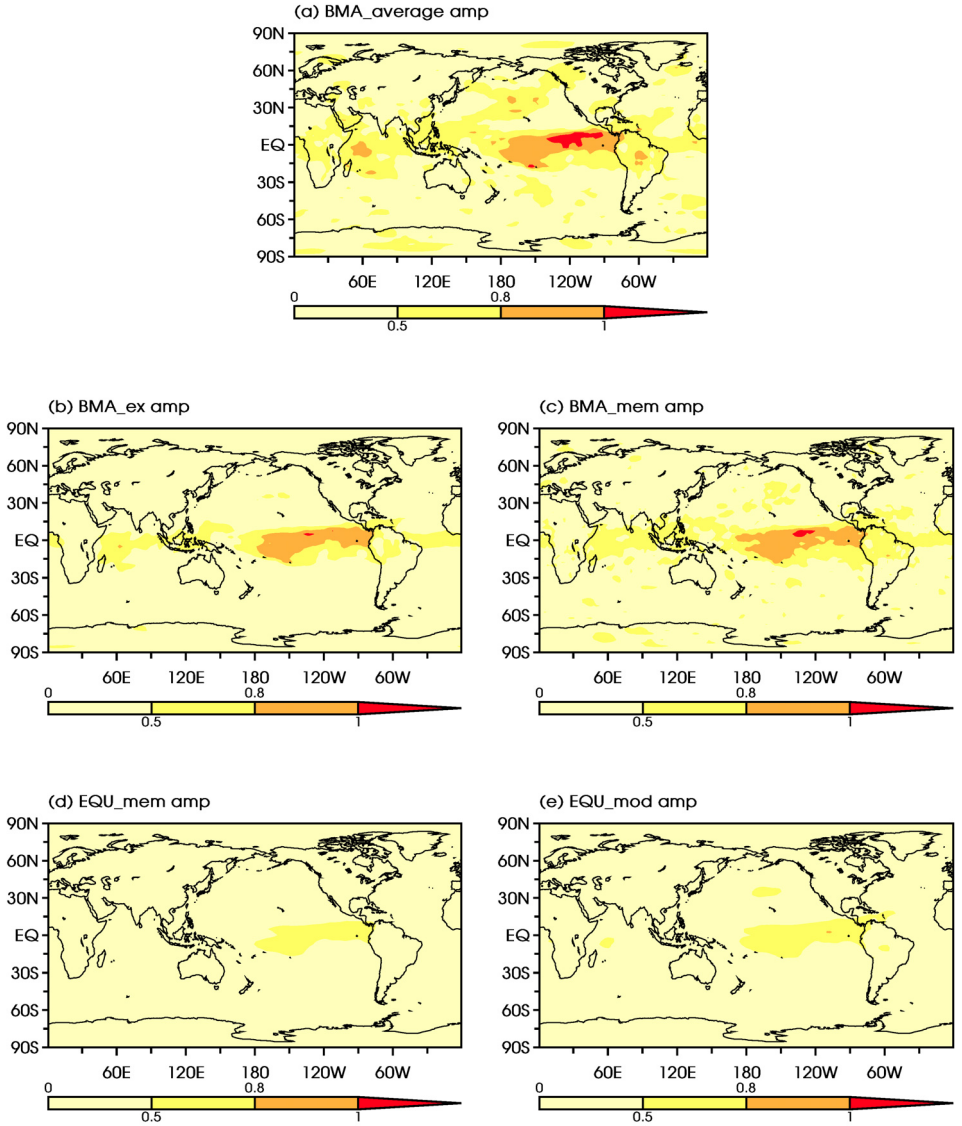


Figure 3 Amplitude component of the decomposition of MSSS of one-month lead DJF temperature hindcast at 850 mb over the global domain from 1983 to 2003. (a), (b), (c), (d), and (e) are obtained from the MME hindcast of the BMA_average, the BMA_ex, the BMA_mem, the EQU_mod, and the EQU_mem methods, respectively. The verification data is the NCEP-R2 reanalysis.

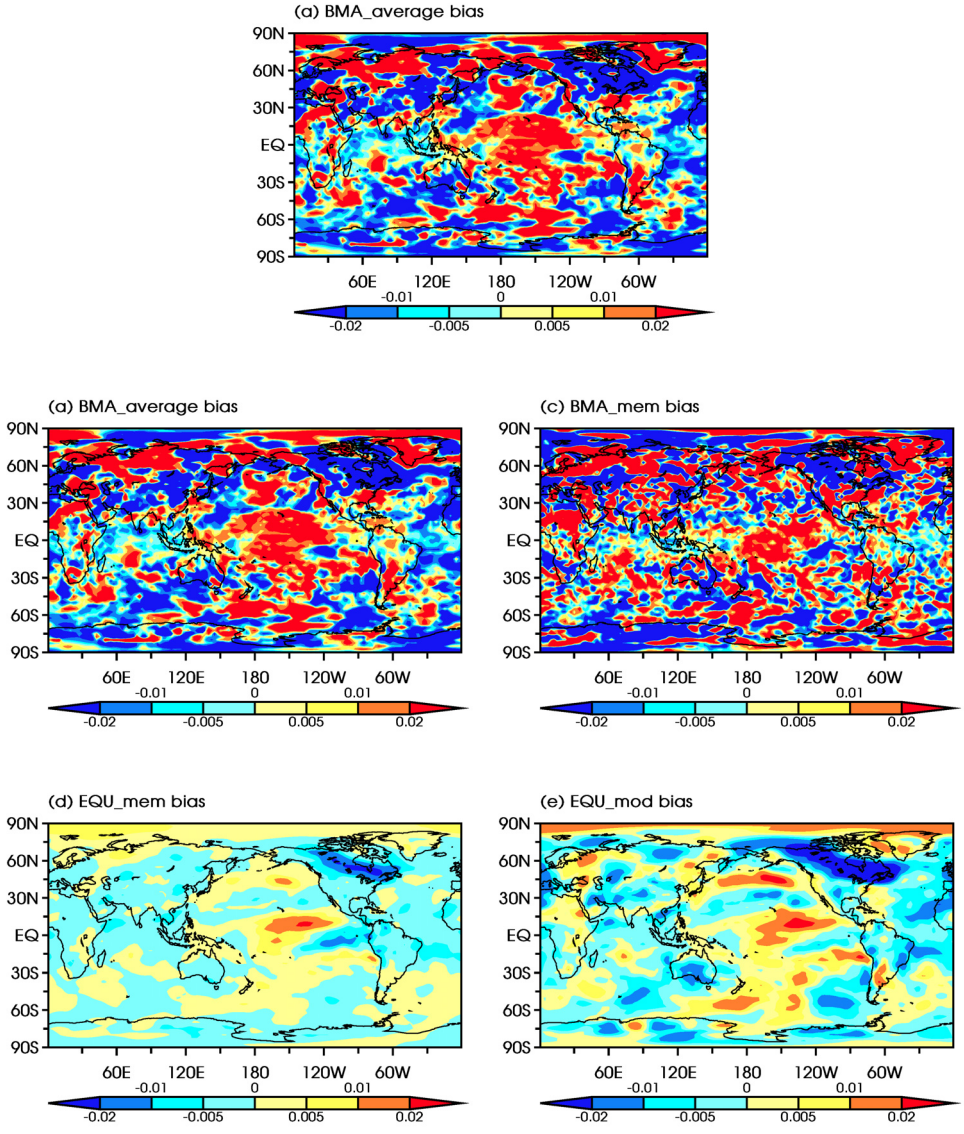
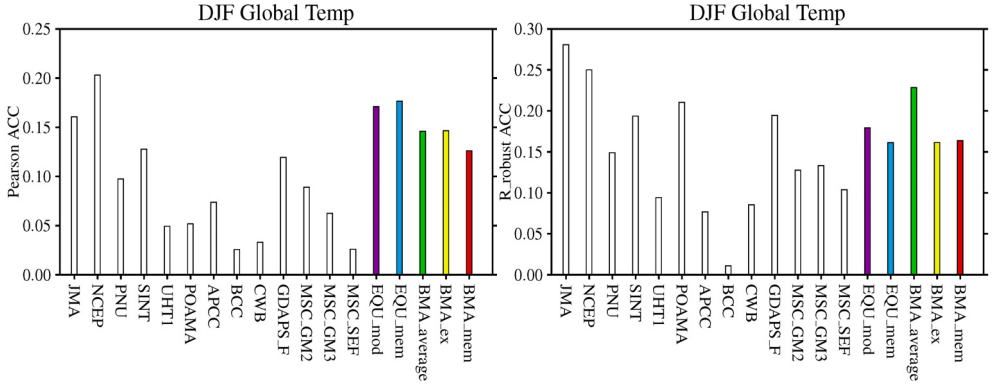
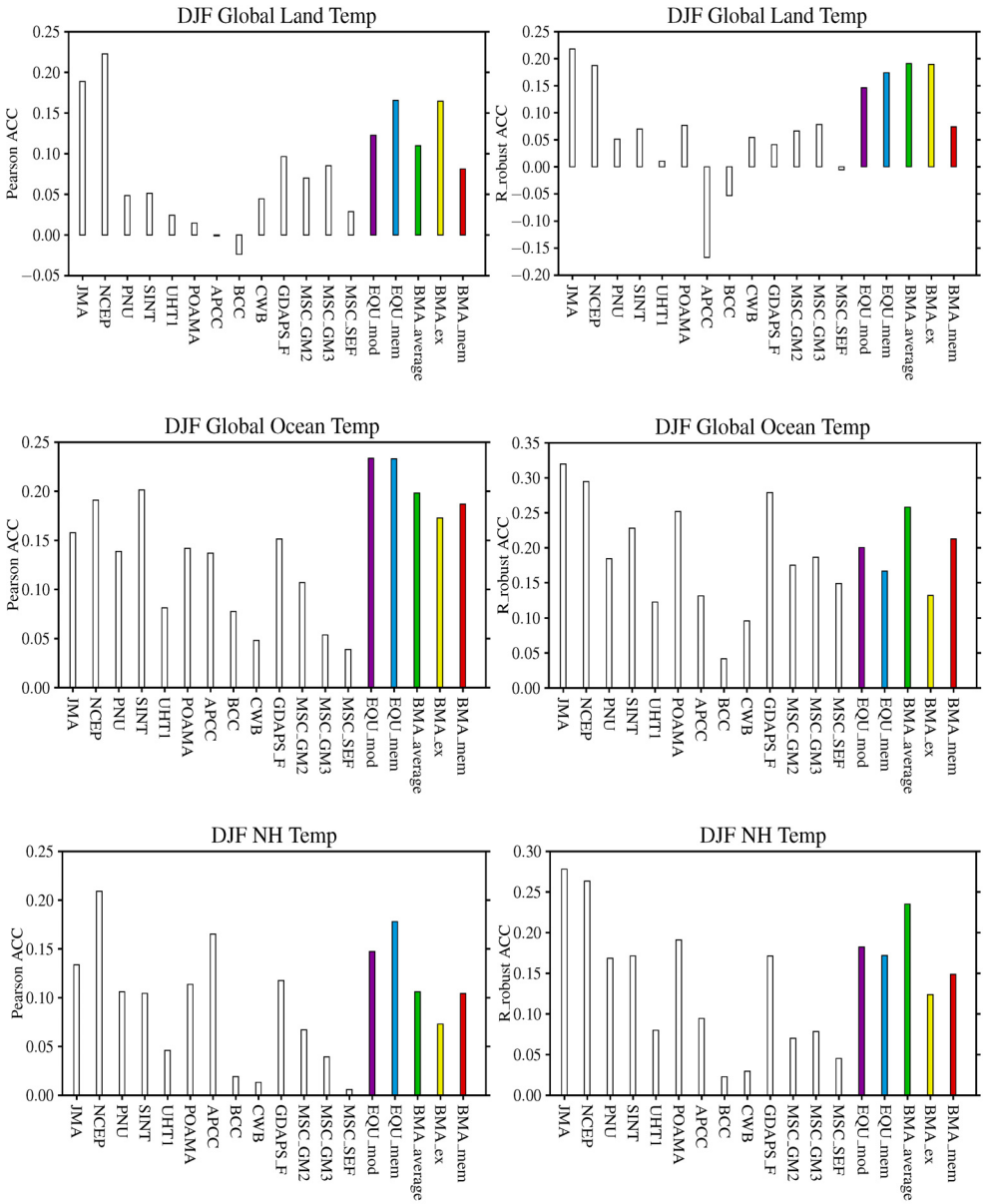


Figure 4 Overall bias component of the decomposed of MSSS of one-month lead DJF temperature hindcast at 850 mb over the global domain from 1983 to 2003. (a), (b), (c), (d), and (e) are obtained from the MME hindcast of the BMA_average, the BMA_ex, the BMA_mem, the EQU_mod, and the EQU_mem methods, respectively. The verification data is the NCEP-R2 reanalysis.

Figure 2 shows the correlation component of the decomposed MSSS, which indicates the phase errors. Figure 3 shows the amplitude error indicated through the ratio of the forecast to observed variances. Figure 4 shows the overall bias error. All definitions can be found in Section 2. Combining these three figures, we found that the correlations are larger in the BMA_average and the BMA_ex methods than those of the other experiments over the Indian Ocean, the Maritime Continent, and the large part of the Tropical Pacific. Over the Atlantic, all the BMA methods show higher correlation than the equal weight methods. A higher correlation contributes to a larger positive skill in terms of the MSSS. Over the tropical areas, all the BMA methods show a larger amplitude than the equal weight methods. In particular, over North Pacific, the BMA_average and the BMA_mem methods show an amplitude larger than 0.5, whereas the other methods show a very low amplitude. Generally, all MMEs show an amplitude of less than 1 over most of the global area, except that the BMA methods show equivalent variances of verification data over some part of East Pacific. The hindcast fidelity is held when the amplitude tends to unity. Generally, the overall biases in the equal weight methods are lower than those in the BMA methods. The BMA methods show a larger spatial heterogeneity in the overall bias, which may be caused by the short validation period.

The ACC is another verification method usually used in seasonal forecast.




DJF Global Land Temp
Pearson ACC
R_robust ACC
DJF Global Ocean Temp
DJF NH Temp
JMA
NCEP
PNU
SINT
UHT1
POAMA
APCC
BCC
CWB
GDAPS_F
MSC_GM2
MSC_GM3
MSC_SEF
EQU_mod
EQU_mem
BMA_average
BMA_ex
BMA_mem

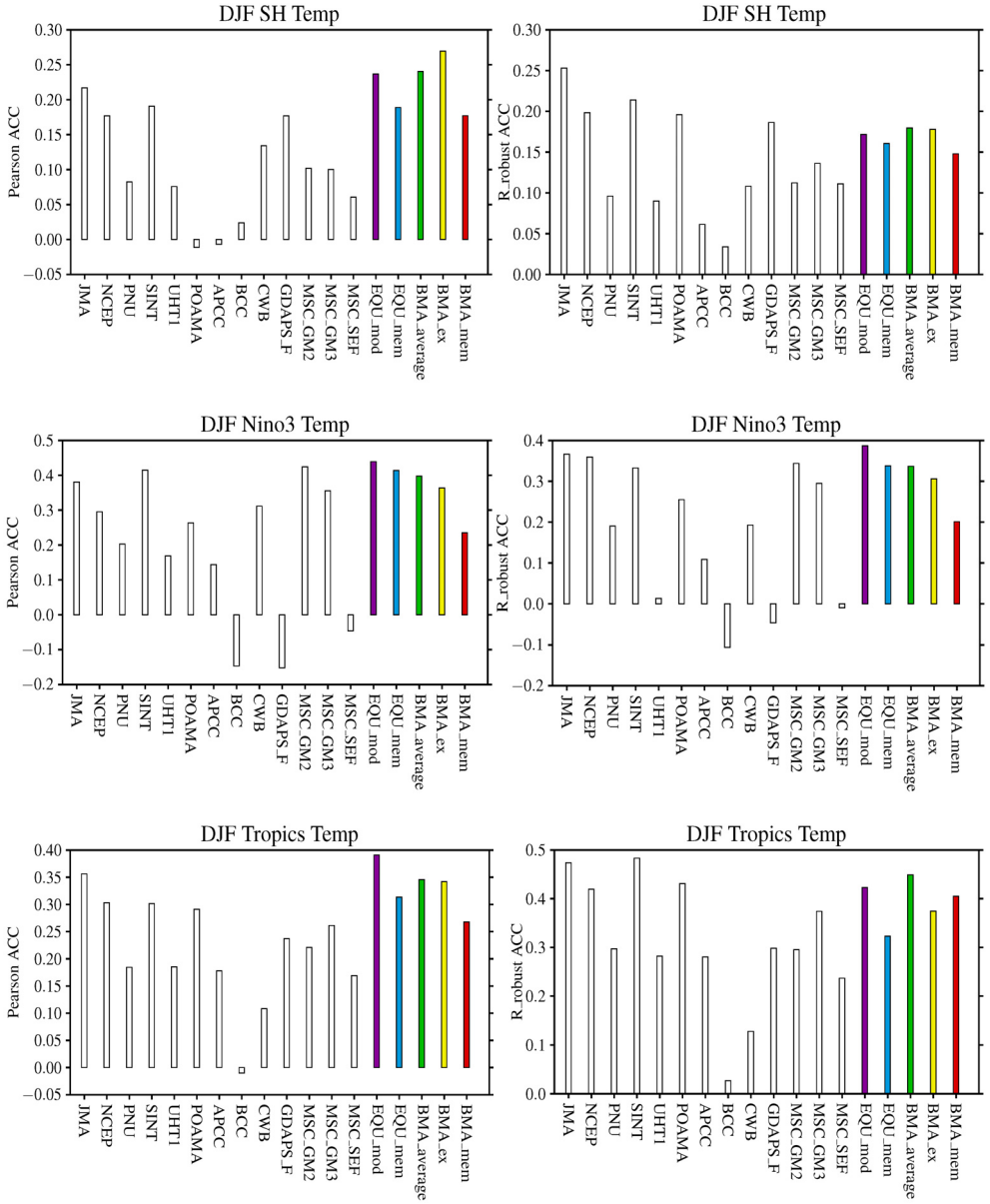


Figure 5 Pearson (left) and robust (right) anomaly pattern correlation coefficients between hindcast and verification temperature at 850 mb for DJF over the global domain, the global land, the global ocean, the North Hemisphere, the South Hemisphere, the Nino3 region, and the tropics from thirteen individual models and five MME methods. The one-month lead DJF temperature hindcast was carried out from 1983 to 2003. The verification data is the temperature from the NCEP-R2.

The ACC is shown in figure 5, which measures the pattern similarity between the hindcast and the verification data.

Globally, the equal weight methods outperform most of the other experiments except the hindcast of the NCEP model, which is from the same climate center of the verification data. Even though the BMA_average and the BMA_ex methods are worse than the equal weight methods, they are still better than most of the individual models. The BMA_mem method is the worst one among the MME methods, but it is still better than most of the individual models. In terms of the robust ACC, the BMA_average method is the best among the MME methods. All the MME methods are better than most of the individual models.

Over the global land, the EQU_mem and the BMA_ex methods have almost the same skills. The equal weight methods, the BMA_average, and the BMA_ex methods are better than most of the individual models. As mentioned in the previous paragraph, the highest skill of the NCEP hindcast may be obtained from the similar background model of the NCEP-R2 reanalysis dataset. In terms of the robust ACC, the BMA_average and the BMA_ex methods are better than the equal weight methods. Except for the JMA and the NCEP models, the other individual models have less than half the skills of the above four MME methods. The BMA_mem method has comparable skill with most of the individual models.

Over the global ocean, in terms of the Pearson ACC, the equal weight methods are the best among all the experiments. All the BMA methods are better than most of the individual models. In terms of the robust ACC, the BMA_average method is the best among the five MME methods. None of the MME methods is the best among all the experiments, whereas the BMA_average method is better than most of the individual models. The BMA_mem method is also slightly better than the EQU_mod method.

Combining the cases of the global region, the global land, and the global ocean, for both the Pearson ACC and the robust ACC, generally the global hindcast difference for most of the experiments dominated by the different skills over the ocean region but the land region.

Over the North Hemisphere, the equal weight methods produce the best hindcast in terms of the Pearson ACC among the five MME methods. The BMA methods do not have better skills than most of the individual models. On an average, the BMA methods have comparable skills to the individual models. In the robust ACC, the BMA_average method has the highest skill among the MME methods. The EQU_mod method is better than most of the individual models.

Over the South Hemisphere, in terms of the Pearson ACC, the BMA_ex method shows the highest skill than the other MME methods and all the individual models. The BMA_average method and the EQU_mod method have comparable skills and both of them are better than the individual models. The EQU_mem method and the BMA_mem method are better than most of the individual models. In terms of the robust ACC, five individual models have higher skills than all the MME methods. The BMA_average and the BMA_ex methods are slightly better than the equal weight methods. All the MME methods have higher skills than the most of the individual models.

Over the Nino3 region, all the Pearson ACCs of the equal weight methods are slightly better than the BMA_average and the BMA_ex methods. The EQU_mod method has the best skill among all the experiments. The BMA_average and the BMA_ex methods are better than most of the individual models. In terms of the robust ACC, the EQU_mod method is the best among all the experiments. The EQU_mem and the BMA_average methods are slightly worse than the EQU_mod method but better than most of the individual models. The BMA_ex method is also better than most of the individual models.

Over the tropics, in terms of the Pearson ACC, the EQU_mod method shows the highest skill among all the experiments. The BMA_average and the BMA_ex methods are better than the EQU_mem method and most of the individual models except the JMA model. In terms of the robust ACC, the BMA_average method is the best among all the MME methods. However, none of the MME methods is the best among all the experiments. All the BMA methods are better than most of the individual models.

Most of the experiments show high skills over the tropical region, which has a higher predictability than other regions.

5.2 Precipitation

The hindcast skill of precipitation in terms of the MSSS of different MME methods are also shown in figure 6. Over the Maritime Continent, the MSSS is the highest in the BMA_average method, and in general, the BMA methods have higher skills than the equal weight methods. Over the tropical Atlantic, the BMA_average method has the largest area with an MSSS higher than 0.35, whereas the EQU_mem method has the smallest area with such high skill. The BMA_average and the BMA_mem methods have relatively narrow areas with positive skills; however, the BMA_ex and the equal weight methods have larger areas around this high skill region. Over the tropical Pacific, the BMA methods show broader areas with an MSSS higher than 0.35 than the equal weight methods, especially over the eastern tropical Pacific where the equal weight methods only show a skill score little more than 0.35. However, similar to the tropical Atlantic, the equal weight methods have broader areas with positive skills than the BMA methods. Over the eastern Philippine Sea, the BMA_average method also has the highest MSSS than other methods. However, over the extratropical area, the equal weight methods have larger areas with a positive MSSS than the BMA methods. Except for the BMA_ex method, the spatial heterogeneity of the MSSS in the BMA_average and the BMA_mem methods is very high, which is probably caused by the short validation period. Over the tropical area from 90E to 0W between −10S and 10N, the overall MSSSs are 0.261, 0.269, 0.242, 0.200, and 0.254 for the BMA_average, the BMA_ex, the BMA_mem, the EQU_mem, and the EQU_mod methods, respectively. Over the Nino3 region (-5S–5N, 90W–150W), the overall MSSSs are 0.416, 0.433, 0.384, 0.270, and 0.318 for the BMA_average, the BMA_ex, the BMA_mem, the EQU_mem, and the EQU_mod methods, respectively.

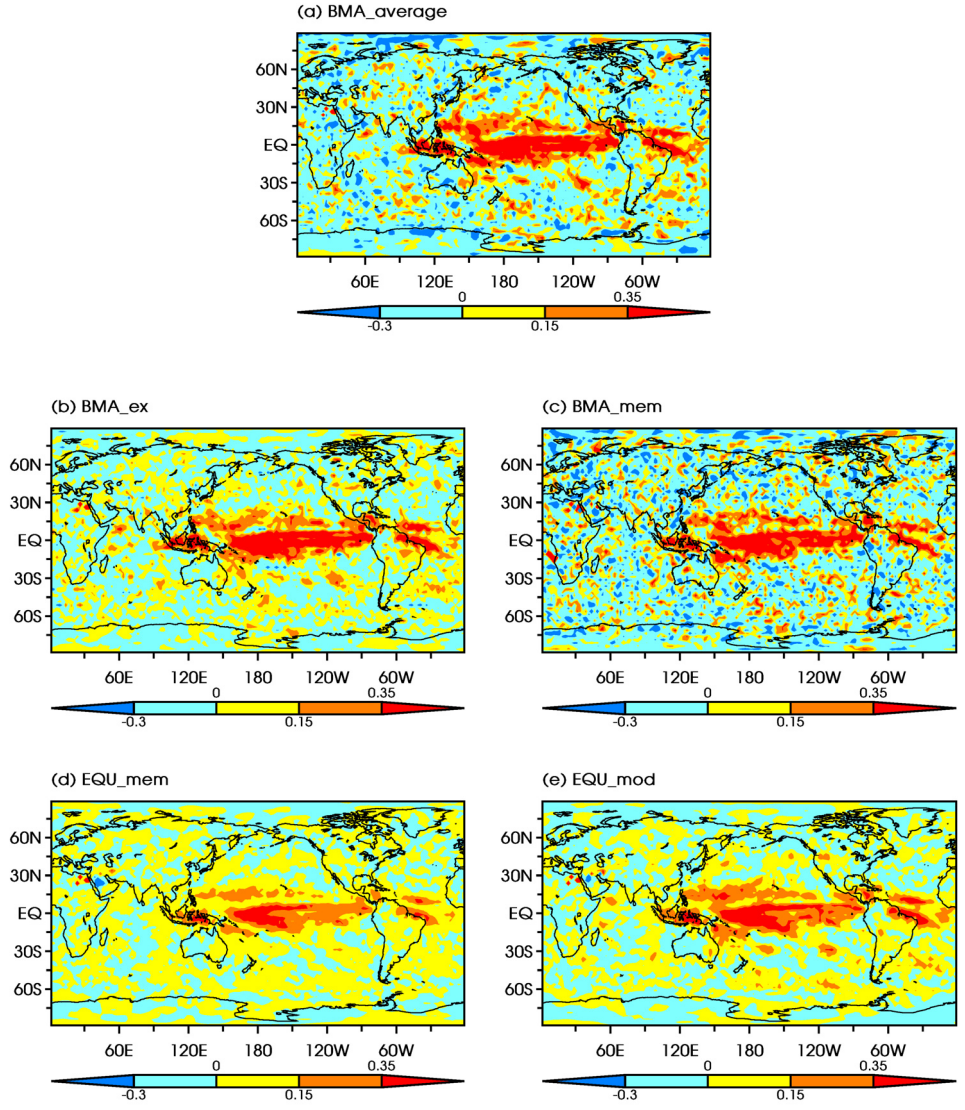


Figure 6 Mean squared skill score (MSSS) of one-month lead JJA precipitation hindcast over the global domain from 1983 to 2003. (a), (b), (c), (d), and (e) are obtained from the MME hindcast of the BMA_average, the BMA_ex, the BMA_mem, the EQU_mod, and the EQU_mem methods, respectively. The verification data is CMAP data.

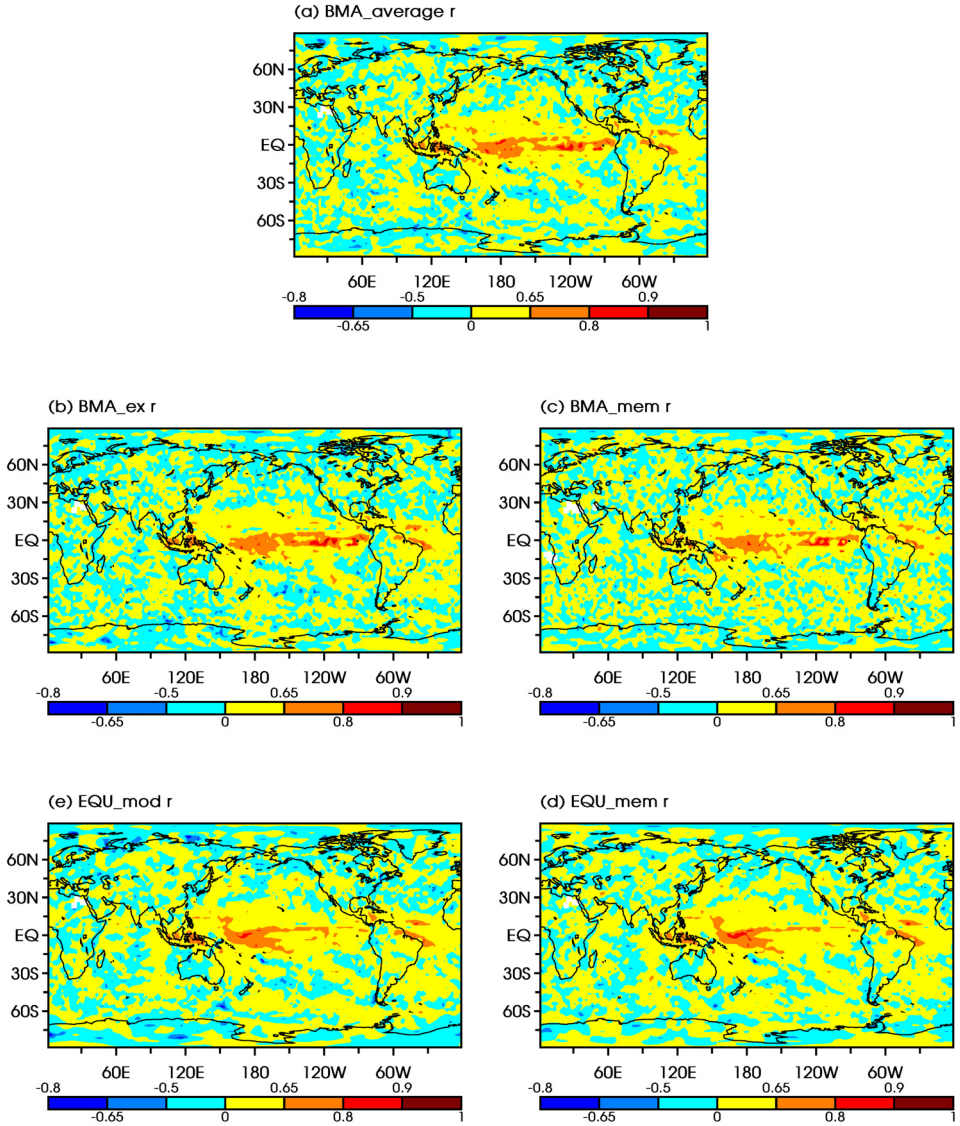


Figure 7 Correlation component of the decomposed MSSS of one-month lead JJA precipitation hindcast over the global domain from 1983 to 2003. (a), (b), (c), (d), and (e) are obtained from the MME hindcast of the BMA_average, the BMA_ex, the BMA_mem, the EQU_mod, and the EQU_mem methods, respectively. The verification data is CMAP data.

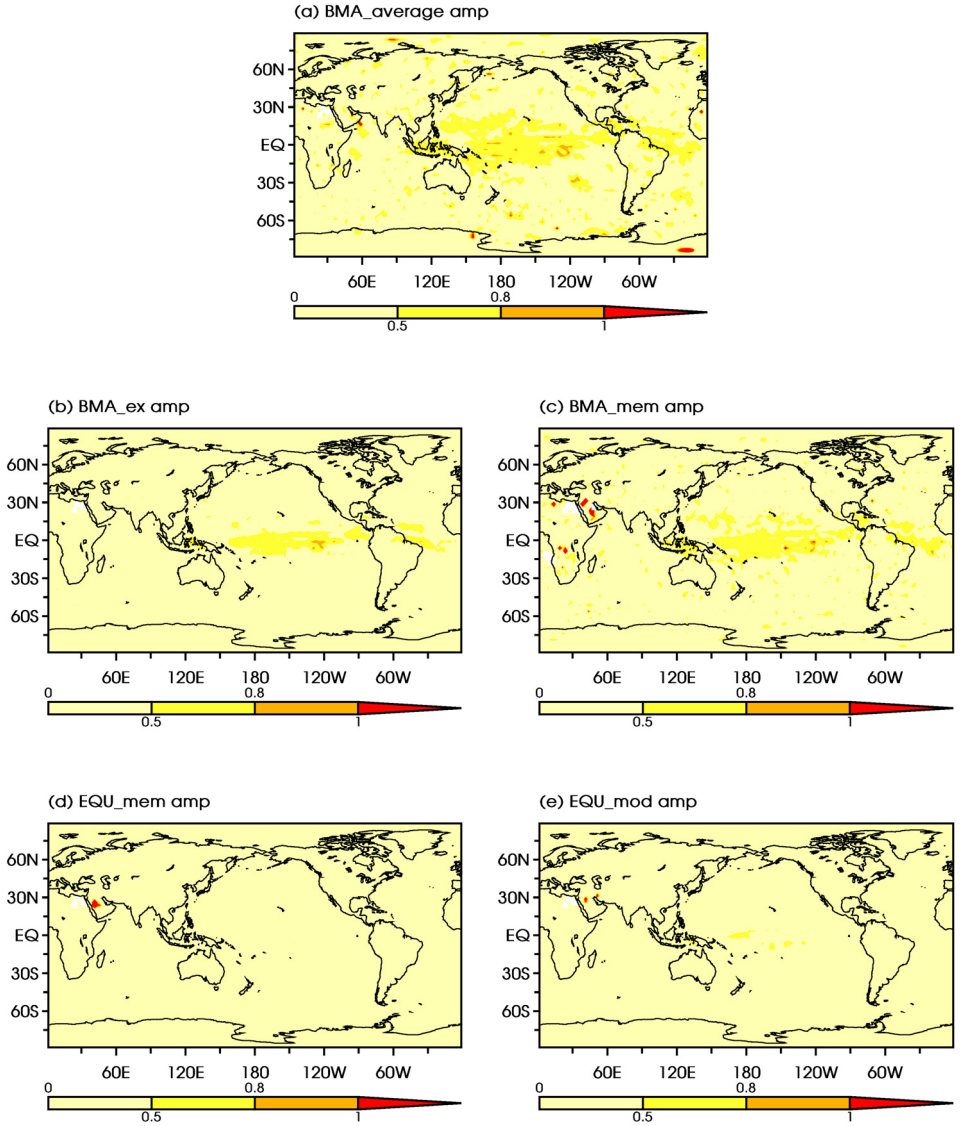


Figure 8 Amplitude component of the decomposed MSSS of one-month lead JJA precipitation hindcast over the global domain from 1983 to 2003. (a), (b), (c), (d), and (e) are obtained from the MME hindcast of the BMA_average, the BMA_ex, the BMA_mem, the EQU_mod, and the EQU_mem methods, respectively. The verification is data CMAP data.

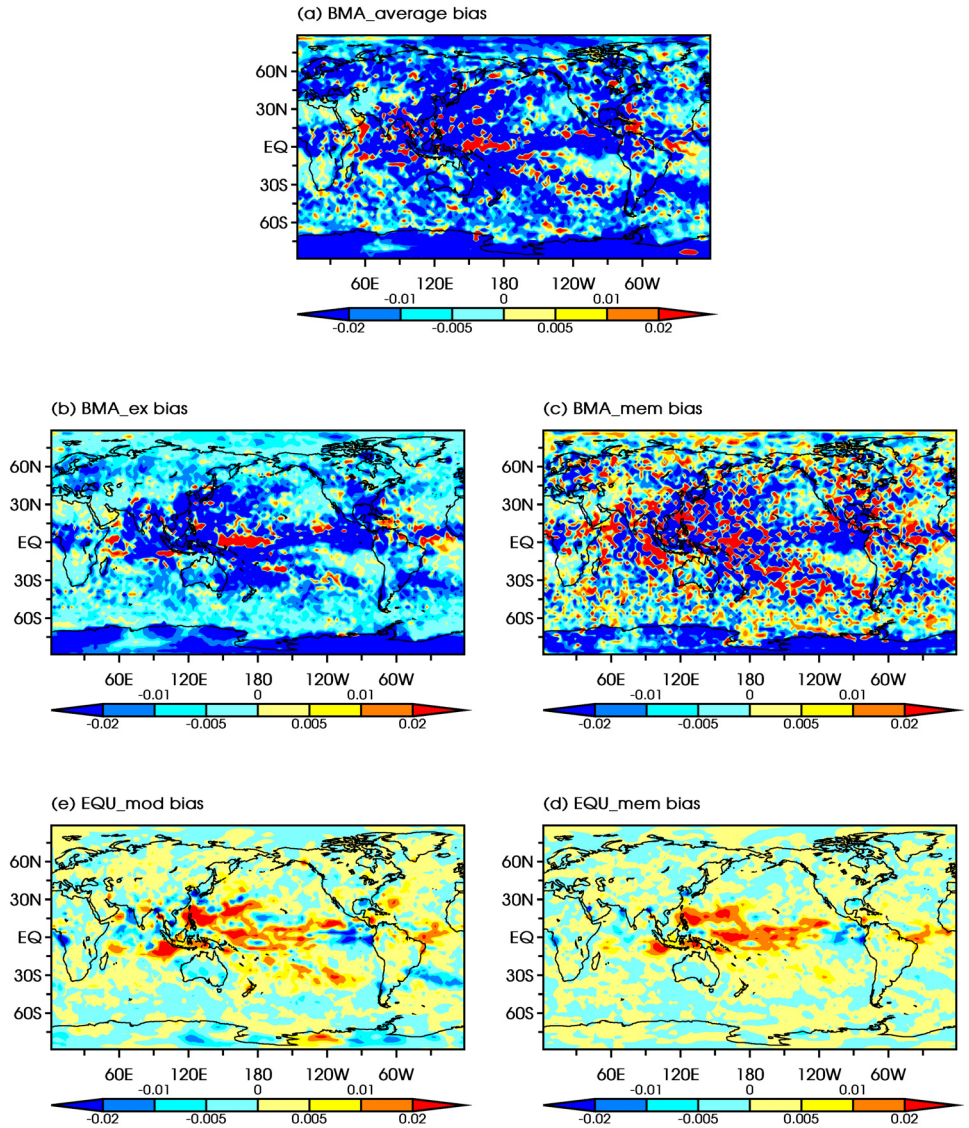
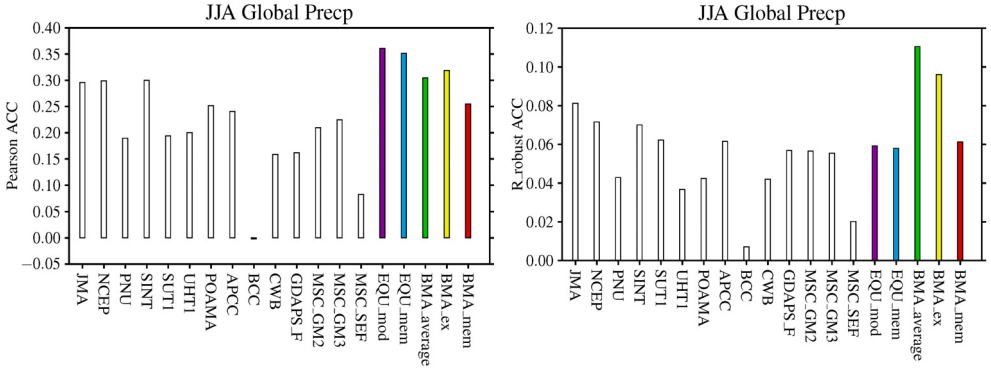


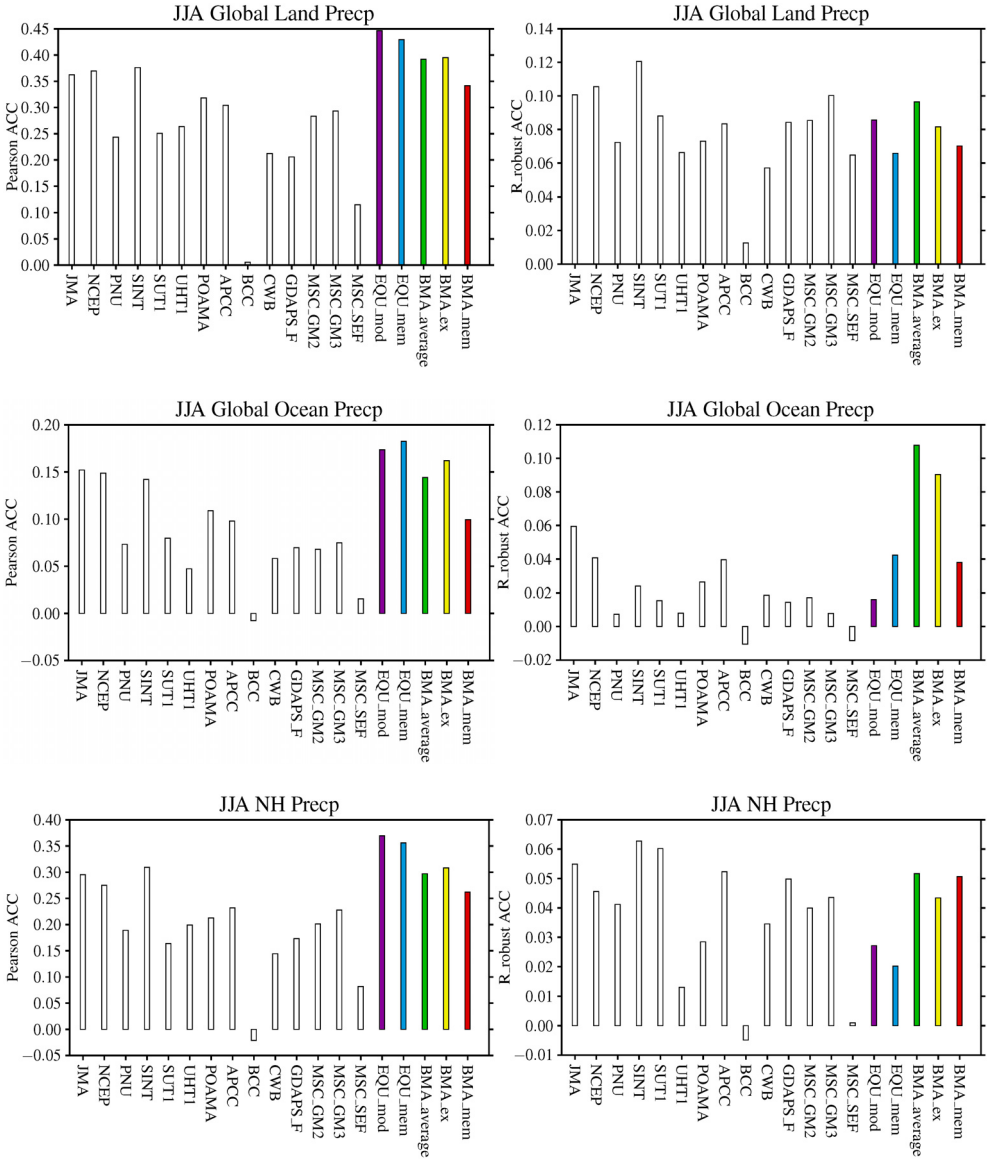
Figure 9 Overall bias component of the decomposed MSSS of one-month lead JJA precipitation hindcast over the global domain from 1983 to 2003. (a), (b), (c), (d), and (e) are obtained from the MME hindcast of the BMA_average, the BMA_ex, the BMA_mem, the EQU_mod, and the EQU_mem methods, respectively. The verification data is CMAP data.

Figure 7 shows the correlation component of the decomposition of the MSSS, which indicates the phase errors. Figure 8 shows the amplitude error indicated through the ratio of the forecast to verification variances. Figure 9 shows the overall bias error. Combining these three figures, we found that the correlations are larger in the BMA_average and the BMA_ex methods than those in the other experiments over the tropical Pacific, especially over the eastern tropical Pacific, compared to the equal weight methods, which almost only have low correlation in this region. Over the extratropicals, the equal weight methods only have slightly larger areas with positive skill than the BMA methods. A high correlation contributes to the positive skill in terms of the MSSS.

Over the tropical Pacific and the tropical Atlantic, the BMA methods show a larger amplitude than the equal weight methods. The equal weight methods only have a small amplitude over almost the global domain. The hindcast fidelity is held when the amplitude tends to unity.

Generally, the overall biases in the equal weight methods are lower than those in the BMA methods. The BMA methods, especially the BMA_average and the BMA_mem methods, show a large spatial heterogeneity in the overall bias, which may be caused by the short validation period. The equal weight methods only show large overall biases over the Maritime Continent, the tropical Pacific, and the northern west tropical Pacific.




JJA Global Land Precp
Pearson ACC
R_robust ACC
JJA Global Ocean Precp
JJA NH Precp
JMA
NCEP
PNU
SINT
SUT1
UHT1
POAMA
APCC
BCC
CWB
GDAPS_F
MSC_GM2
MSC_GM3
MSC_SEF
EQU_mod
EQU_mem
BMA_average
BMA_ex
BMA_mem

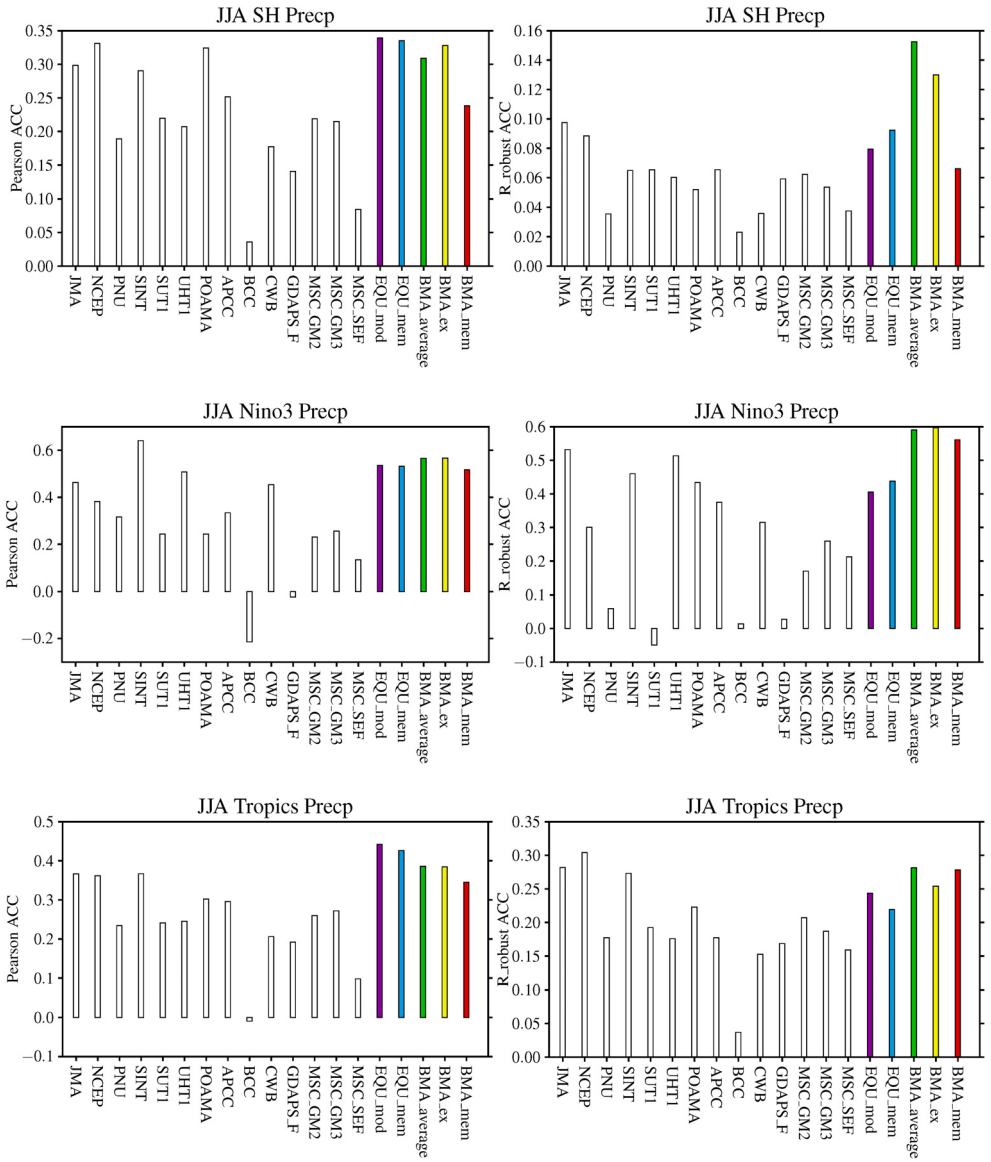


Figure 10 Pearson (left) and robust (right) anomaly pattern correlation coefficient between hindcast and verification precipitation for JJA over the global domain, the global land, the global ocean, the North Hemisphere, the South Hemisphere, the Nino3 region, and Tropics from fourteen individual models and five MME methods. The one-month lead JJA precipitation hindcast was carried out from 1983 to 2003. The verification data is the CMAP seasonal precipitation.

The ACC is shown in figure 10, which measures the pattern similarity between hindcast and verification data.

Globally, the EQU_mod method has the highest skill in terms of the Pearson ACC and the EQU_mem method is slightly worse than the EQU_mod method. The BMA_ex and the BMA_average methods follow the equal weight methods in terms of the Pearson ACC. All the four MME methods outperform the individual models. However, in terms of the robust ACC, the BMA_average and the BMA_ex methods have the highest skills, almost double that of the equal weight methods. All the BMA methods are superior to the equal weight methods. In terms of individual models, the JMA, NCEP, and SINT are the best three models in this case.

Over the global land, both the equal weight methods show the highest Pearson ACC than the BMA methods and the individual models. The BMA_average and the BMA_ex methods have comparable skills to each other. Both the BMA_average and the BMA_ex methods are also superior to the individual models. However, in terms of the robust ACC, even though the BMA_average and the EQU_mod methods are better than most of the individual models, none of the MME methods is the best. The BMA_average method is still better than the equal weight methods in terms of the robust ACC. However compared with the global case, both the Pearson and the robust ACC increased, which is probably because reliable gauge data was used over land in the verification precipitation.

Over the global ocean, in terms of the Pearson ACC, the equal weight methods are still the best compared with the BMA methods and the individual models. The BMA_ex method is also better than all the individual models. The BMA_average method is better than most of the individual models. In terms of the robust ACC, both the BMA_average and the BMA_ex methods are superior to the other MME methods and the individual models by a large extent.

Combining the cases of the global region, the global land, and the global ocean, in terms of the robust ACC, most of the difference in the global skills between the BMA methods and the equal weight methods arise from the ocean region.

Over the North Hemisphere, both the equal weight methods show the highest

Pearson ACC than the BMA methods and the individual models. The BMA_average and the BMA_ex methods have comparable skills to each other. All the BMA methods are superior to most of the individual models. However, in terms of the robust ACC, even though the BMA_average and the BMA_mem methods are better than most of the individual models, none of the MME methods is the best. All the BMA methods are much better than the equal weight methods.

Over the South Hemisphere, in terms of the Pearson ACC, the equal weight methods are still the best compared with the BMA methods and the individual models. The skills of the BMA_average, the BMA_ex, and even the BMA_mem methods are still better than most of the individual models. In terms of the robust ACC, both the BMA_average and the BMA_ex methods are superior to the other MME methods and the individual models by a large extent. Although, the equal weight methods are worse than the BMA_average and the BMA_ex methods, they still outperform most of the individual models.

Over the Nino3 region (-5S–5N, 90W–150W), the BMA_average and the BMA_ex methods are better than the equal weight methods. All the MME methods outperform most of the individual models. In terms of the robust ACC, all the BMA methods have high skills close to 0.6, which are better than both the equal weight methods and the individual models. Over the tropics, the equal weight methods are better than the BMA methods in terms of the Pearson ACC.

The BMA_average and the BMA_ex methods are better than all the individual models. The BMA_mem method is better than most of the individual models. In terms of the robust ACC, all the BMA methods are better than the equal weight methods. However, none of the MME methods is the best.

6. CONCLUSION

In order to obtain the optimal weights for combining different model outputs in seasonal forecasts, Bayesian model averaging (BMA) was applied to the multimodel hindcast datasets at the Asia–Pacific Economic Cooperation (APEC) Climate Center

(APCC). The weights were estimated according to the performance of individual members in simulating the given training data. In this study, we used the model anomalies to eliminate the climatological model biases (i.e., systematic errors in the climatology or mean bias of the forecast). For each individual model, the anomalies were estimated as departures from their climatology over the training period in a one-year out cross-validation manner (Wilks 1995).

Over the Maritime Continent and the tropical Atlantic, one or more BMA methods showed the highest MSSS. Over the tropical Pacific and North Pacific, the BMA methods showed broader areas with a higher MSSS than the equal weight methods. Over the Indian Ocean, the BMA methods generally outperform the equal weight methods, whereas the EQU_mod method may have broader areas with positive skills. Generally, the BMA methods had higher skills over the tropical areas, especially in the Nino3 region, but had lower skills over the extratropical areas than the equal weight methods. Generally, the MSSS of the BMA methods showed high spatial heterogeneity.

The decomposition of the MSSS provided more information in detail. Over the Indian Ocean, the Maritime Continent, the large part of Tropical Pacific, and the Atlantic, the BMA methods generally showed a higher correlation with the verification data, which contributed to the larger positive skill for the MSSS. The BMA methods generally showed a larger amplitude than the equal weight methods over the Tropicals and North Pacific, and even showed equivalent variances of the verification data over some part of East Pacific. The hindcast fidelity was held when the amplitude tended to unity. Generally, the overall biases in the equal weight methods were lower than those in the BMA methods.

For the anomaly pattern correlation coefficient (ACC), over different regions such as the global domain, the global land, the global ocean, the North Hemisphere, and the tropics, the equal weight methods were better than the BMA methods in terms of the Pearson ACC, and one or more BMA methods were better than most of the individual models. In terms of the robust ACC, one or more BMA methods were the best among the MME methods. Over the South Hemisphere, one of the BMA methods was the best in terms of the Pearson ACC. Further, none of the MME methods is the best in terms of the robust ACC. Over the Nino3 region, for both the Pearson

and the robust ACCs, the equal weight methods are slightly better than the BMA methods. Generally, the global hindcast difference for most of the experiments was dominated by the skills over the ocean region but the land region. The model JMA was generally better than the individual models and sometimes even better than the MME methods. In some cases, the NCEP model showed the highest skill, which may be attributed to the similarity of the model background with the verification data.

For precipitation, over the Maritime Continent, the BMA_average method had the best skill in the MSSS, and the BMA methods are generally better than most of the individual models. Over the tropical Pacific and the tropical Atlantic, one or more BMA methods showed broader areas with a higher MSSS than the equal weight methods, whereas globally, the equal weight methods generally had broader areas with positive skills than the BMA methods.

For the decomposition, over the tropical Pacific, one or more BMA methods showed larger correlations with the verification data than the equal weight methods, whereas the equal weight methods had slightly broader areas with positive skill than the BMA methods over the extratropicals. A higher correlation contributed to a positive larger skill to the MSSS. Over the tropical Pacific and the tropical Atlantic, the BMA methods showed a larger amplitude than the equal weight methods. The hindcast fidelity was held when the amplitude tended to unity. Generally, the overall biases in the equal weight methods were lower than those in the BMA methods. The equal weight methods only show large overall biases over the Maritime Continent, the tropical Pacific, and the northern west tropical Pacific.

Over the global domain, the global ocean, the South Hemisphere, and the tropics, the equal weight methods are the best in terms of the Pearson ACC, whereas one or more BMA methods are the best in terms of the robust ACC. Over the global land and the North Hemisphere, the equal weight methods are the best in terms of the Pearson ACC, whereas one BMA method is better than the equal weight methods in terms of the Pearson ACC. Over the Nino3 region, one or more BMA methods are better than the equal weight methods in terms of both the ACCs.

7. DISCUSSION

As is well known, the ensemble climate prediction of global models from different centers yields better results than any individual prediction (e.g., Kalnay and Ham 1989; Krishnamurti *et al.* 1999; Shukla *et al.* 2000; Wang *et al.* 2004) because the MME represents the uncertainties not only in the initial conditions but also in the models. Recently, the BMA MME method is popularly used in weather and hydrologic forecasting. The aim of this study is to investigate the superiority and shortcoming of the BMA MME method in seasonal forecast with the APCC datasets. The hindcast of the equal weight methods and the BMA methods are verified based on the MSSS and ACC.

In terms of the ACC, generally, for most of the experiments, the scale of the robust ACC and the Pearson ACC of temperature are comparable, whereas there is a big difference between the two ACCs of precipitation, implying that outliers are favorably produced in precipitation forecast. Thus, the numerous outliers in the original hindcast generally reduce skill in the robust ACC instead of in the Pearson ACC. The robust ACC shows the correlated relationships for the major amount of points between hindcast and verification data. However, for precipitation in the Nino3 region, the relatively small difference between these two ACCs in the BMA methods indicates that the BMA methods produce much less outliers than the equal weight methods. The major difference in the robust ACC between the BMA methods and the equal weight methods arises from the ocean region instead of land.

The availability of hindcast datasets from individual models is limited, and hence, only 21-year data is used in the BMA methods. In the BMA method, the weights are estimated according to the performance of individual members in simulating the given training data. However, if the training period is short, the weights may only make the cost function become the local minimum. Thus, the weights may not be the optimal results in practice, which is indicated from the large spatial heterogeneity of the skill in the BMA methods.

The cross-validated pointwise BMA algorithm requires huge computational resources, which limits our experiment to just two variables. The rigorous test of

the new MME method should be carried out for different lead times, different seasons, and as for as many variables. One individual model may be better than the BMA methods in some cases, whereas the BMA methods show superiority throughout all cases because the best individual model in one case may be not the best in another case.

Our study shows that the seasonal forecast skills is strongly dependent on the verification matrix. A good seasonal forecast in one measurement may not necessarily be good in another standard. The robust ACC may be another option for reviewing the seasonal forecast.

REFERENCES

Brankovic C., T. N. Palmer, and L. Ferranti, 1994. Predictability of seasonal atmospheric variations. J. Climate, 7:217–237.

Chang, Y., Schubert, S. D. and Suarez, M. J. 2000. Boreal winter predictions with the GEOS- 2 GCM: the role of boundary forcing and initial conditions. Q. J. R. Meteorol. Soc. 126, 2293–2321.

Charney J. G. and J. Shukla, 1981. Predictability of monsoons. Monsoon Dynamics, pages 99–109. Cambridge University Press. Editors: Lighthill, J. and Pearce, R.

Coelho, C. A. S., D. B. Stephenson, F. J. Doblas-Reyes, M. Balmaseda, A. Guetter, and G. J. Van Oldenborgh, 2006. A Bayesian approach for multi-model downscaling: seasonal forecasting of regional rainfall and river flows in South America. Meteor. Appl., 13, 73-82

Doblas-Reyes, F. J., Déqué, M. and Piedelievre, J.-P. 2000. Multi-model spread and probabilistic forecasts in PROVOST. Q. J. R. Meteorol. Soc. 126, 2069–2087.

Doblas-reyes F. J., Renate Hagedorn, and T.N. Palmer (2005), The rationale behind the success of multi-model ensembles in seasonal foreasting - II. Calibration and combination. Tellus 57A:234-252

Duan Q, Ajami NK, Gao X, Sorooshian S (2007) Multi-Model ensemble hydrologic prediction using bayesian model averaging. Advances in Water Resources 30:1371-386

Hagedorn R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting--I. Basic concept. Tellus, 57A, 219–233.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986) Robust statistics. The approach based on influence functions. New York: John Wiley

Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: A tutorial. Statistical Science 14:382-417

Huber, P. J. (1981) Robust Statistics, New York: John Wiley

Kanamitsu M, Ebisuzaki W, Woollen J, Yang S-K, Hnilo JJ, Fiorino M, Potter GL (2002) NCEP-DOE AMIP-II Reanalysis (R-2). Bull Am Meteor Soc 83:1631-1643

Kharin, V. V. and Zwiers, F. W. 2002. Notes and correspondence: Climate predictions with multi-model ensembles. J. Climate 15, 793–799.

Kug JS, Lee JY, Kang IS, Wang B, Park CK (2008) Optimal multi-model ensemble method in seasonal climate prediction. Asia-Pacific J Atmos Sci 44:259-267

Krzysztofowicz R., 1983: Why should a forecaster and a deci- sion maker use Bayes theorem. Water Resour. Res., 19, 327–336

Kalnay E, Ham M (1989) Forecasting forecast skill in the Southern Hemisphere. Extended Abstracts, Third Int. Conf. on Southern Hemisphere Meteorology and Oceanography, Buenos Aires, Argentina, Amer Meteor Soc 24–27

Krishnamurti TN, Kishtawal CM, LaRow TE, Bachiochi DR, Zhang Z, Williford CE, Gadgil S, Surendran S (1999) Improved weather and seasonal climate forecasts from multi-model superensemble. Science 285:1548–1550

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T. E., Bachiochi, D. R. and co-authors. 2000a. Improving tropical precipitation forecasts from a multi-analysis superensemble. J. Climate 13, 4217–4227

Krishnamurti TN, Kishtawal CM, Shin DW, Williford CE (2000b) Multi-model superensemble forecasts for weather and seasonal climate. J Clim 13:4196-4216

Krishnamurti, T. N. and Sanjay, J. 2003. A new approach to the cumulus parametrization issue. Tellus 55A, 275–300

Maronna, R., and V. J. Yohai (1995) The Behavior of the Stahel-Donoho Robust Multivariate Estimator. Journal of the American Statistical Association 90 (429), 330–341

Maronna R. A. and R. H. Zamar (2002) Robust estimates of location and dispersion of high dimensional datasets. Technometrics 44 (4), 307–317

Maronna, R., and Yohai, D. M. V. (2006) Robust Statistics. Theory and Methods. New York: John Wiley & Sons

Marino Marrocu1, and Piero A. Chessa (2008) A multi-model/multi-analysis limited area ensemble: calibration issues. Meteorological Applications. 15(1), 171–179

Miyakoda, K., Hembree, G. D., Strickler, R. F., Shulman, I., 1972: Cumulative results of extended forecast experiments. 1. Model performance for winter cases. Mon. Wea. Rev., 100, 836-855

Murphy A. H., 1977: The value of climatological, categorical, and probabilistic forecasts in the cost–loss ratio situation. Mon. Wea. Rev., 105, 803–816

Murphy A. H., 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. Mon. Wea. Rev., 16, 2417-2424

Mason S. J. and G. M. Mimmack, 2002. Comparison of some statistical methods of probabilistic forecasting of ENSO. J. Climate, 15:8–29

Min Y.-M., V. N. Kryjov, and C.-K. Park, "A Probabilistic Multimodel Ensemble Approach to Seasonal Prediction," Weather and Forecasting, 24

Min S-K, Daniel S, Andreas H (2007) Probabilistic climate change predictions applying Bayesian model averaging. Phil Trans R Soc, A, 365:2103–2116

Marrocu, M., and P. A. Chessa, 2008: A multimodel / multianalysis limited-area ensemble: Calibration issues. Meteor. Appl., 15, 171–179

Onogi K, Tsutsui J, Koide H *et al* (2007) The JRA-25 reanalysis. J Meteorol Soc Jpn 85:369–432

Palmer, T. N., A. Alessandri, U. Andersen, P. Cantelaube, M. Davey, P. Délécluse, M. Déqué, E. Díez, F. J. Doblas-Reyes, H. Feddersen, R. Graham, S. Gualdi, J.-F. Guérémy, R. Hagedorn, M. Hoshen, N. Keenlyside, M. Latif, A. Lazar, E. Maisonnave, V. Marletto, A. P. Morse, B. Orfila, P. Rogel, J.-M. Terres, and M. C. Thomson, 2004: Development of a european multimodel ensemble system for seasonal-to-interannual prediction (DEMETER). Bull. Amer. Meteor. Soc, 85, 853–872

Pavan, V. and Doblas-Reyes, J. 2000. Multi-model seasonal hindcasts over the Euro-Atlantic: skill scores and dynamic features. Climatic Dyn. 16, 611–625

Peng, P., Kumar, A., Van den Dool, A. H. and Barnston, A. G. 2002. An analysis of multi- model ensemble predictions for seasonal climate anomalies. J. Geophys. Res. 107, 4710

Press, W. H., S. A. Teukolsky, W. T. Vettering, and B. P. Flannery, 1992: Numerical Recipes in Fortran. 2d ed. Cambridge University Press, 963 pp.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using bayesian model averaging to calibrate forecast ensembles. Mon Wea Rev 133:1155-174

Rajagopalan B., U. Lall, and S. E. Zebiak, 2002. Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. Mon. Wea. Rev., 130:1792–1811

Robertson A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. Mon. Wea. Rev., 132, 2732–2744

Rousseeuw P. J. and K. van Driessen (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223

Shukla J, Anderson J, Baumhefner D *et al* (2000) Dynamical seasonal prediction. Bull Am Meteorol Soc 81:2593–2606

Sloughter J. M., A. E. Raftery, T. Gneiting, and C. Fraley (2007) Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. Monthly Weather Review, 135: 3209-3220

Stefanova, L. and Krishnamurti, T. N. 2002. Interpretation of seasonal climate forecast using Brier skill score, FSU superensemble, and the AMIP-1 data set. J. Climate 15, 537–544

Stephenson, D. B. and Doblas-Reyes, F. J. 2000. Statistical methods for interpreting Monte Carlo ensemble forecasts. Tellus 52A, 300–322

Thompson J. C., 1962: Economic gains from scientific advances and operational improvement sin meteorological prediction. J. Appl. Meteor., 1, 13–17

Wang B, Kang I-S, Lee J-Y (2004) Ensemble simulations of Asian– Australian monsoon variability by 11 AGCMs. J Clim 17:803–818

Wang B, Kang IS, Shula J, Lee JY, *et al.* (2010) Improvement of APCC seasonal prediction and assessment of characteristics and forecast of 2009/2010 climate anomalies. Final report of APCC international research project 2010, APEC Climate Center, Korea

Wilks D. S. (1995) Statistical methods in the atmospheric sciences: An introduction. Academic Press. 467 PP

Woodruff D. L. and D. M. Rocke (1994) Computable robust estimation of multivariate location and shape on high dimension using compound estimators. Journal of the American Statistical Association, 89, 888–896

Xie P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observation, satellite estimates, and numerical model outputs. Bull. Amer. Meteor. Soc., 78, 2539–2558

Yun, W.-T., Stefanova, L. and Krishnamurti, T. N. 2003. Improvement of the superensemble technique for seasonal forecasts. J. Climate 16, 3834–3840

R. A. Maronna and V. J. Yohai (1995) The Behavior of the Stahel-Donoho Robust Multivariate Estimator. Journal of the American Statistical Association 90 (429), 330–341

P. J. Rousseeuw and K. van Driessen (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41, 212–223

D. L. Woodruff and D. M. Rocke (1994) Computable robust estimation of multivariate location and shape on high dimension using compound estimators. Journal of the American Statistical

Association, 89, 888–896.

R. A. Maronna and R. H. Zamar (2002) Robust estimates of location and dispersion of high-dimensional datasets. Technometrics 44 (4), 307–317.

Yun WT, Stefanova L, Mitra AK, Kumar W, Dewar W, Krishnamurti TN (2005) A multi-model superensemble algorithm for seasonal climate prediction using DEMETER forecasts. Tellus 57A: 280-289

Yun WT, Stefanova L, Krishnamurti TN (2003) Improvement of the superensemble technique for seasonal forecasts. J Clim 16: 3834-3840

APCC TECHNICAL REPORT 2012-01

- Application of Bayesian Model Averaging on Multi-Model Ensemble Seasonal Prediction
- Decadal Change of Variability and Predictability of Two Types of ENSO
- Assessment of Relationship between EL Nino and Indian Summer Monsoon Rainfall
- Long-lead MME Extreme Drought Prediction
- Assessment of APCC Multi-Model Ensemble Predictions

APEC Climate Center

12, Centum 7-ro, Haeundae-gu, Busan 612-020,
Republic of Korea
Tel: +82-51-745-3900 Fax: +82-51-745-3949
www.apcc21.org

비매품
9 788997 333363 94500
ISBN 978-89-97333-36-3
ISBN 978-89-97333-35-6 (세트)