RMetS
Royal Meteorological Society

An analysis of multi-model ensembles for seasonal climate predictions

Xiaoqin Yan[a] and Youmin Tang[a,b]*

[a]*Environmental Science and Engineering, University of Northern British Columbia, Prince George, BC, Canada*
[b]*State Key Laboratory of Satellite Ocean Environment Dynamics, Hangzhou, China*

*Correspondence to: Y. Tang, Environmental Science and Engineering, University of Northern British Columbia, 3333 University way, Prince George, V2N 4Z9, BC, Canada. E-mail: ytang@unbc.ca

In this study, the superiorities of the super-ensemble for seasonal climate predictions are investigated based on the 500 mb geopotential height (GPH500) hindcasts produced by four Canadian atmospheric seasonal climate prediction models. The investigations are carried out mainly in two aspects: (i) a comprehensive evaluation of predictions for each grid point over the global domain by deterministic, probabilistic and potential prediction skill measures; (ii) the empirical orthogonal function (EOF) and the Maximum Signal-to-Noise (MSN) EOF analyses in the Northern Hemisphere. It is found that improvements of the super-ensemble are mainly due to the increase of ensemble size in the mid-high latitudes and the offsets of model uncertainties in the tropical regions. Measures of temporal correlation coefficient (CORR), Brier skill score (BSS) and reliability (REL) are more affected by the ensemble size; whereas the relative root-mean-square error (RRMSE) and resolution (RES) are sensitive to the offsets of model uncertainties. The first EOF mode of ensemble mean is similar to the most predictable pattern derived by the MSN EOF method, but the latter has the temporal evolutions more associated with the oceanic boundary forcing. The super-ensemble shows advantages in both EOF and MSN EOF analyses. Copyright © 2012 Royal Meteorological Society

Key Words: deterministic prediction skill; probabilistic prediction skill; potential prediction skill; EOF; MSN EOF

Received 18 February 2011; Revised 6 July 2012; Accepted 9 July 2012; Published online in Wiley Online Library

Citation: Yan X, Tang Y. 2012. An analysis of multi-model ensembles for seasonal climate predictions. *Q. J. R. Meteorol. Soc.* DOI:10.1002/qj.2020

1. Introduction

Prompted by their great socio-economic effects, seasonal climate predictions have been produced in more than ten major forecast centres around the world, where a set of dynamical models are being used. Given the nonlinear evolution of the atmospheric system, the quality of the dynamical seasonal climate prediction is critically dependent on both the uncertainties of the initial conditions and the model formulations (Palmer, 2001). To reduce initial conditions uncertainties, different ensemble strategies and data assimilation algorithms have been proposed and applied practically (e.g. Epstein, 1969; Hoffman and Kalnay, 1983; Stockdale *et al.*, 1998; Tang *et al.*, 2005; Balmaseda and Anderson, 2008). To reduce model formulation uncertainties, the following methods could be used: (i) the multi-model approach (the super-ensemble method), which considers the structural inadequacy of individual models (e.g. Tracton and Kalnay, 1993; Vislocky and Fritsch, 1995; Atger, 1999; Krishnamurti *et al.*, 1999, 2000; DelSole, 2007; Peña and Van den Dool, 2008; DelSole *et al.*, 2012); (ii) the perturbed-parameter approach, in which the multiple uncertain parameters are considered (e.g. Stainforth *et al.*, 2005; Collins *et al.*, 2006); and (iii) the stochastic-physics

approach, which tries to resolve uncertainties in the subgrid processes (Shutts, 2005; Jin *et al.*, 2007; Shutts and Palmer, 2007). Doblas-Reyes *et al.* (2009) argued that the three methods are complementary to each other, but the multi-model approach is better than the other two when the lead times of climate predictions are shorter than 5 months, namely, the multi-model method is the best for seasonal forecasts.

The central argument of the multi-model ensemble superiority over the single model ensemble is that the former takes a holistic consideration of uncertainties from both the initial conditions and the model uncertainties (e.g. Palmer and Shukla, 2000; Palmer *et al.*, 2004), and our lack of understanding of atmospheric behaviour could possibly be offset by different assumptions of model framework. The multi-model approach has been applied in many aspects including short- and medium-range weather forecasting (e.g. Richardson, 2001; Mylne *et al.*, 2002; Eckel and Mass, 2005; Whitaker *et al.*, 2006; Candille, 2009), seasonal climate prediction (e.g. Krishnamurti *et al.*, 2000; Barnston *et al.*, 2003; Kumar *et al.*, 2003; Palmer *et al.*, 2004; Vitart, 2006), and climate change projection (e.g. AR4 of Intergovernmental Panel on Climate Change). By the analyses of some indices of regional average, the benefits of the multi-model approach have been displayed in the DEMETER (Development of a European Multi-Model Ensemble System for Seasonal to Interannual Climate Prediction) multi-model ensemble system (Palmer *et al.*, 2004; Hagedorn *et al.*, 2005), the CliPAS (Climate Prediction and its Application to Society) project (Wang *et al.*, 2009), and the El Niño/Southern Oscillation (ENSO) multi-model (CCSM, CFS) prediction (Kirtman and Min, 2009).

The objective of this study is to systematically explore the prediction skill of the super-ensemble in terms of the local predictions and the large-scale climate-mode predictions. The anomalies of the 500 mb geopotential height (GPH500) seasonal mean prediction are used as the analysis target. The GPH500 is a typical variable that represents the extratropical circulations. Its predictions also can be used to improve the seasonal climate predictions of precipitation and temperature by post-processing methods (Lin *et al.*, 2008). In detail, this study addresses two aspects: (i) the investigation of the spatial distribution of the prediction skill over the global domain, grid by grid; and (ii) the prediction skill of the large-scale climate modes in the Northern Hemisphere (NH: 20°N–90°N, 0°E–360°E). In both aspects, the performances of the super-ensemble are compared with those of each single model ensemble. The central questions that this study attempts to answer are: (i) whether and why the super-ensemble is necessarily superior to the single ensembles; (ii) whether and how the nature and merits of the super-ensemble are geophysically dependent; and (iii) whether the most predictable components are more significant in the super-ensemble than in single model ensembles. There were some similar studies focusing on question (i) in the literature, but questions (ii) and (iii) have been little addressed in the climate prediction community.

This article is composed of six sections. In section 2, the data and methods used in this work are introduced. Section 3 compares the super-ensemble against each single ensemble in the global domain, in terms of the deterministic, probabilistic and potential forecast skills. The reasons responsible for the superiorities of the multi-model ensemble in local predictions are also discussed. In section 4, the potential predictability is investigated based on the 'perfect model' scenario. In chapter 5, the large-scale climate modes derived by the super-ensemble and each single model ensembles are presented and compared by using the Empirical Orthogonal Function (EOF) method and the Maximum Signal-to-Noise ratio (MSN) EOF method. The merits of the super-ensemble in predictions of climate modes are also discussed. Finally, section 6 presents the summary and conclusions.

2. Data and methods

2.1. Data

The ensemble forecasts used in this study are from the second phase of the Historical Seasonal Forecasting Project (HFP2), which aims to study the seasonal climate predictability. The seasonal hindcasts were produced by four Canadian atmospheric models: the second and third generations of the general circulation models (GCM2 and GCM3) developed at the Canadian Centre for Climate Modeling and Analysis (CCCma) (Boer *et al.*, 1984; McFarlane *et al.*, 1992), the Global Environmental Multi-scale model (GEM: Côté *et al.*, 1998a, 1998b) developed at Recherche en Prévision Numerique (RPN) in Montreal, Canada, and the reduced-resolution version of the global spectral model (SEF: Ritchie, 1991). For each model, an ensemble of ten parallel integrations of 4-month duration, with initial conditions at 12-hour intervals of the National Centers for Environmental Prediction-National Center for Atmospheric Research (NCEP-NCAR) reanalysis (Kalnay *et al.*, 1996) preceding the first day of the forecast seasons, was produced from the beginning of each month for the period from January 1969 to December 2001. The sea surface temperature (SST) in HFP2 was taken as the sum of the SST anomaly of the month preceding the forecasts and the monthly varying climatological SST, i.e. the same SST anomaly persists through the whole 4-month forecast periods. Detailed instructions for the initial conditions and boundary forcing of these models can be found in Lin *et al.* (2008) or Kharin *et al.* (2009). For validation, the NCEP reanalysis dataset (Kalnay *et al.*, 1996) over the same time period was used as the observation. In order to get rid of model bias, the climatology of both hindcasts and observation data were removed individually, and all data used in this work were the anomalies of GPH500.

2.2. Skill scores and analysis methods

2.2.1. Deterministic skill scores

To evaluate the deterministic prediction skill of the ensemble products, two measures are used in this study: the temporal correlation coefficient (CORR) and the relative

root-mean-square error (RRMSE):

$$CORR_{pq}(t) = \frac{\sum_{i=1}^{N}\left(z_i^p(t)-\overline{z_i^p}\right)\left(z_i^q-\overline{z_i^q}\right)}{\left\{\sum_{i=1}^{N}\left(z_i^p(t)-\overline{z_i^p}\right)^2\right\}^{\frac{1}{2}}\left\{\sum_{i=1}^{N}\left(z_i^q-\overline{z_i^q}\right)^2\right\}^{\frac{1}{2}}}, \quad (1)$$

$$RRMSE = 1-\frac{RMSE_{pq}}{RMSE_{\mathrm{ref}}} = 1-\frac{\left\{\frac{1}{N-1}\sum_{i=1}^{N}\left(z_i^p-z_i^q\right)^2\right\}^{\frac{1}{2}}}{\left\{\frac{1}{N-1}\sum_{i=1}^{N}\left(z_i^q-\overline{z_i^q}\right)^2\right\}^{\frac{1}{2}}}, \quad (2)$$

where z is the anomalies of GPH500, the superscripts p and q represent prediction and observation, respectively, t is the lead time, the overbar indicates the average over the entire time period, and N is the total number of samples. Larger CORR and RRMSE indicate good forecasts. The statistical test for the difference of two CORRs, obtained by the super-ensemble prediction and by a single ensemble separately, both against observation, is evaluated by Steiger's Z-test (Meng *et al.*, 1992). RRMSE was used instead of the root-mean-square error (RMSE) to remove the effect of climatological variability at different locations.

2.2.2. *Probability skill scores*

For evaluations of the probabilistic forecast skill, the standardized Brier Skill Score (BSS) and its two components of reliability (REL) and resolution (RES) are calculated at each grid in the global domain. Higher BSS values suggest better forecasts. With a perfect score of 1, BSS lies in the range of $[1-1/\{P_c(1-P_c)\},\ 1]$, where P_c is the climatological frequency of the forecast event. Positive BSS values indicate that the model forecasts are better than the climatological forecasts. REL reflects the statistical consistency between the forecast probability and the mean observed frequencies over a long-term period (Toth *et al.*, 2003; Wilks, 2006). Smaller REL values suggest more reliable probabilistic forecasts. RES describes the ability of the forecast system to resolve the set of sample events into subsets with characteristically different outcomes by quantifying the difference between the conditional observed frequencies and climatological frequency (Murphy, 1973). Larger RES values suggest better resolution ability of forecast systems.

The reliability diagram (Murphy, 1985) can visually reveal the dependence of the relative frequency of an observed event on the forecast probability. The joint distribution of the forecast probability and the corresponding observation frequencies is close to the diagonal line in the reliability diagram if the forecast system is reliable. But it is impractical to present the reliability diagram at each grid point given the large number of grid points over the global domain, for which the REL is calculated instead. Similarly, the sharpness histogram can be used to characterize the resolution in a reliable forecast system. It can visually show a tendency of one system to forecast the extreme probabilities near 0 or 1 rather than values gathering around the mean. Since the sharpness histogram just shows an attribute of the forecasts alone without correspondence to observations, it is only meaningful if the forecast system is reliable.

2.2.3. *Potential prediction skill*

The potential prediction skill (PCORR) is an upper limit of the model prediction skill usually obtained with the assumption of the 'perfect model scenario', under which each ensemble member could be viewed as the real observation. The PCORR is obtained by averaging all correlation coefficients between an individual ensemble member (viewed as the observation) against the ensemble mean of remaining members (Peng *et al.*, 2005):

$$PCORR(t) = \frac{1}{K}\sum_{k=1}^{K} \frac{\sum_{i=1}^{N}\left(z_i^{p,k}-\overline{z_i^{p,k}}\right)\left(z_i^{p,ens}(t)-\overline{z_i^{p,ens}}\right)}{\left\{\sum_{i=1}^{N}\left(z_i^{p,k}-\overline{z_i^{p,k}}\right)^2\right\}^{\frac{1}{2}}\left\{\sum_{i=1}^{N}\left(z_i^{p,ens}(t)-\overline{z_i^{p,ens}}\right)^2\right\}^{\frac{1}{2}}}, \quad (3)$$

$$z_i^{p,ens}(t) = \frac{1}{K-1}\sum_{j=1,j\neq k}^{K} z_i^{p,j}(t), \quad (4)$$

where z^p is the prediction of the anomalies of GPH500 at a fixed lead time t, the superscripts k and *ens* represents the prediction of ensemble member k and the ensemble mean of the remaining members, the overbar indicates the average over the entire time period, and N is the total number of samples.

Due to the uncertainties of model frameworks, the atmospheric predictability is usually model dependent. Comparisons of PCORR among different models will be reasonable only if the practical prediction skills (CORR) of those models are comparable (e.g. a perfect PCORR of 1 would be meaningless if the CORR is 0). For a particular model, the closer the CORR approaches the PCORR, the closer the actual prediction skill approaches the upper limit (or better prediction). Thus, the PCORR can be regarded as an indicator of the accuracy of predictions, offering a means to estimate prediction skill when the observations are not available. The difference between the potential prediction skill and the practical prediction skill is quantified by their relative errors:

$$REE = \frac{PCORR-CORR}{CORR}\times 100\%. \quad (5)$$

A smaller REE value implies a better prediction.

2.2.4. *EOF and MSN EOF methods*

EOF analysis has been widely applied in atmospheric studies. It finds the spatial patterns of variability and their temporal variation, and gives a measure of the 'importance' of each pattern by the explained variance. Differences between the EOF method and PCA (Principal Component Analysis) method are quite confusing in the literature. Some authors (Richman, 1986) define the two methods differently; some authors (Preisendorfer and Mobley, 1988; Peixoto and Oort, 1992) refer to PCA and EOF methods as the same. In this study, the phrases EOF and PCA are used interchangeably. The patterns were referred to as the EOF modes; the associated time series were referred to as PCs (principal components).

Different from the EOF analysis, the MSN EOF analysis is essentially a discrimination issue aiming to maximize the ratio of signal over noise (SNR) (Fukunaga, 1990; Schneider and Griffies, 1999). The MSN EOF analysis introduced by Allen and Smith (1997) is an optimal method that can derive the signal correctly by removing the influences of noise. It is used in finite ensemble forecasts to derive signals driven by the external forcing (Venzke *et al.*, 1999; Sutton *et al.*, 2000; Straus *et al.*, 2003; Huang, 2004; Hu and Huang, 2007; Liang *et al.*, 2009). In ensemble forecasts, the response of the signal to the external forcing can be represented by the ensemble mean anomalies, which are the potentially predictable components; whereas the noise (overall internal variability) is estimated by the average ensemble spread, which is essentially unpredictable due to the atmosphere's chaotic internal dynamics (Straus and Shukla, 2002; Straus *et al.*, 2003; Kang *et al.*, 2004; Tippett and Giannini, 2006; Liang *et al.*, 2009). The signal and noise are theoretically independent when the ensemble size is infinite (e.g. Venzke *et al.*, 1999; Sutton *et al.*, 2000). When the ensemble size is finite, the estimation of the signal is often contaminated by the noise. The goal of MSN EOF is to exclude the impact of noise as much as possible while the signal is extracted.

The MSN EOF is essentially the same as the most Predictable Component Analysis (PrCA), which is based on information theory to maximize predictive information, the difference between the entropy of the forecast distribution and the entropy of the climatology distribution (Schneider and Griffies, 1999). Both the MSN EOF and PrCA methods are equivalent to the discriminant analysis given that the two methods, though from different perspectives, can be understood to search for a best linear combination of variables that separates the signal (entropy of the forecast) and the noise (entropy of the climatology) as much as possible (DelSole and Tippett, 2007).

In this study, the phrases MSN EOF and PrCA are used interchangeably. The MSN EOF method finds the spatial pattern, or weight matrix providing an optimized filter to discriminate the signal and noise, the time series reflecting the temporal evolution of the dominant mode of the signal, and the spatial pattern characterizing the spatial distribution of the dominant mode of signal. In this study, the pattern of discrimination is referred to as the filter pattern, the associated time series are referred to as the PrCs, and the spatial distribution of the dominant mode of optimized signal is referred to as the most predictable pattern.

According to the Raleigh Quotient theorem, the maximization of SNR leads to a generalized eigenvalue–eigenvector problem (DelSole and Tippett, 2007). Practically, the number of grid points is always much larger than the number of total samples in climate studies, thus the covariance matrix of noise, denoted by Σ_N, is usually not full-rank, leading to a solution of ill-conditioned inversions. To solve this issue, the SNR is optimized in a truncated EOF space, in which the pre-whitening and regularization techniques were both used to make Σ_N an identity matrix, and whiten the covariance matrix of signal (Σ_S) simultaneously. A further EOF analysis is then applied to the whitened signal covariance matrix. In this study, the MSN EOF analysis mainly follows the algorithm by Venzke *et al.* (1999). The algorithm is briefly summarized as follows:

- Make the covariance matrix of noise (Σ_N) identity, namely,

$$\mathbf{D}^{-1/2}\mathbf{E}^T\Sigma_N\mathbf{E}\mathbf{D}^{-1/2} = \mathbf{I}. \tag{6}$$

  D and **E** are the eigenvalue and eigenvector matrices of Σ_N. $\mathbf{E}\mathbf{D}^{-1/2}$ is a transformation matrix that makes the covariance matrix of noise (Σ_N) identity.
- Whiten the signal covariance matrix by the transformation matrix $\mathbf{E}\mathbf{D}^{-1/2}$:

$$\Sigma_{WS} = \mathbf{D}^{-1/2}\mathbf{E}^T\Sigma_S\mathbf{E}\mathbf{D}^{-1/2}. \tag{7}$$

- Apply the EOF analysis to the whitened signal covariance matrix Σ_{WS} to obtain the optimized SNRs in descending order:

$$\begin{aligned} SNR &= \mathbf{T}^T\Sigma_{WS}\mathbf{T} \\ &= \mathbf{T}^T\mathbf{D}^{-1/2}\mathbf{E}^T\Sigma_S\mathbf{E}\mathbf{D}^{-1/2}\mathbf{T}, \end{aligned} \tag{8}$$

  where **T** is the matrix of eigenvectors of the whitened signal covariance matrix; the matrix $\mathbf{U} = \mathbf{E}\mathbf{D}^{-1/2}\mathbf{T}$ contains the filter patterns.
- Project the ensemble mean on the filter patterns to obtain the predictable components PrCs:

$$PrCs = \mathbf{X}_s\mathbf{U}. \tag{9}$$

  The most predictable component is the one corresponding to the largest SNR. All PrCs are temporally orthogonal (uncorrelated) with each other.
- Obtain the corresponding predictable patterns **V** by projecting the ensemble mean on the PrCs:

$$\mathbf{V} = \mathbf{X}_x^T\mathbf{X}_s\mathbf{U}. \tag{10}$$

3. Actual prediction skill

In this section, the actual prediction skills of all model ensembles are evaluated using the deterministic and probabilistic skill metrics discussed in section 2. To enlarge the sample size, all predictions initialized at each calendar month for the period from 1969 to 2002 are used to form a total sample size of 33*12. For brevity, the seasonal mean prediction (the average over forecasts from 2 to 4 months) is used to evaluate model skills. The emphasis of this study is to reveal the geophysical dependence of the superiorities of the super-ensemble over the global domain.

3.1. *Deterministic prediction skill*

For each single ensemble, the ensemble mean over all members is used to evaluate the deterministic prediction skill. For the super-ensemble, initially we attempted to construct a weighted average of the four single ensemble means for its deterministic prediction. Both the linear weighted and the nonlinear weighted combinations were carried out using linear regression and neural network. Sets of cross-validation experiments show that the equally weighted average of the four single ensemble means is the best construction for the deterministic prediction of the super-ensemble for the seasonal prediction of GPH500 anomalies. This might result from the insufficient samples

for the training in the cross-validation scheme. However, it was argued in DelSole *et al.* (2012) that the approach of unequal weighting of forecasts may contribute to advantages of multi-model ensemble only in relatively limited regions of the globe, which was also supported by the multi-model forecasts through ridge regression (DelSole, 2007; Peña and Van den Dool, 2008). Thus, the equally weighted average of the four ensemble means is used to construct the deterministic prediction for the super-ensemble in this study.

Figure 1 shows the global distributions of CORR (a)–(d) and RRMSE (e)–(h) of GCM2, GCM3, GEM and SEF. The CORR and RRMSE of all models reach maximum in the tropical regions and decrease poleward rapidly to the mid-high latitudes in both hemispheres. At the approximately same latitudes, regions of the Pacific-North America (PNA) (Wallace and Gutzler, 1981) and Antarctic Dipole (ADP: 55°S-75°S, 180°-120°W) (Yuan and Martinson, 2000, 2001) have higher CORR and RRMSE skills. Among all models, SEF has the poorest deterministic prediction skill.

Figure 2 shows the differences of CORR (a)–(d) and RRMSE (e)–(h) between the super-ensemble and each single ensemble. The significant differences of CORR at 95% confidence level by Steiger's Z-test are shaded in Figure 2(a)–(d). Improvements of the super-ensemble are more visible at the mid-high latitudes than in the tropical regions. This is because the atmospheric uncertainties are much higher in the mid-high latitudes, where the deficiency of individual models can be more significantly offset by the super-ensemble. But the super-ensemble outperforms SEF over almost the whole global domain. The performance of single ensembles may be placed in descending order: GCM3, GCM2, GEM and SEF.

Figure 2(e)–(h) shows the RRMSE difference between the super-ensemble and individual ensembles. Unlike the CORR difference (Figure 2(a)–(d)), the RRMSE difference is mostly positive, indicating that the super-ensemble always has better RRMSE skills than single ensembles. Similar to CORR, the super-ensemble has better RRMSE skill than SEF in the global domain. For other single model ensembles, the super-ensemble has higher RRMSE in the following regions: (i) the middle latitudes of East Asia, the mid-high latitudes of western Europe, southwest North America, northeast Australia and the eastern tropical Pacific for GCM2; (ii) some sporadic regions of the middle and low latitudes such as the central Pacific, southwest North America and central South America for GCM3; and (iii) over the Indian Ocean, southwest Australia, a northeast–southwest oriented band over the North Pacific and high latitudes of western North America for GEM.

In terms of the deterministic forecasts, the superiority of the super-ensemble is both metric measure dependent and geophysical location dependent. Generally, the super-ensemble is a little better than the best single ensemble but much better than the worst single ensemble. For CORR, the superiority of the super-ensemble is mainly at the mid-high latitudes, whereas for RRMSE, the super-ensemble also shows merit in the tropical regions.

3.2. *Probabilistic forecast skill*

To perform probabilistic skill analysis, three mutually exclusive events (BN: below normal; NN: near normal; AN: above normal) were defined at each grid point based on the corresponding observation from 1969 to 2002. The climatology used here is the sample climatology. A number of methods have been developed to construct super-ensemble probabilistic forecasts, including the count-based unweighted method, count-based weighted method, Gaussian unweighted method and Gaussian adjusted method (e.g. Kharin and Zwiers, 2003; Kharin *et al.*, 2009). Kharin *et al.* (2009) discussed these methods and performed the probabilistic prediction of super-ensemble using normalized ensemble members and terciles to define categories. In this study, we performed the probabilistic prediction analysis of super-ensemble using the count-based unweighted method (i.e. raw ensemble predictions of individual models) and the terciles as the category definition.

To examine the possible impacts of the construction methods of super-ensemble on probabilistic skills, we performed two statistical experiments: (i) the ensemble members of individual models were normalized as in Kharin *et al.* (2009) but quartiles were used to define the category of events; and (ii) the definition of events used the terciles but ensemble members were not normalized. The results show that the BSS and RES derived from normalized ensemble data are larger for the low latitudes and tropical regions but smaller for the mid-high latitudes, compared with those from un-normalized ensemble data (not shown). However, these differences are small, and under 0.05. We also examined the difference of probabilistic skill scores when the quartiles are used instead of terciles in the above experiments, and found that the differences are also very small for AN and BN events. But terciles division leads to much worse BSS in the lower latitudes and tropical regions. Both experiments show comparable level of REL. Since the terciles are more commonly used in the community, we used the count-based unweighted construction method and terciles as the definition of event category in this study.

The global distributions of BSS, REL and RES scores, derived from the super-ensemble prediction, are shown in Figure 3 for BN, NN and AN, respectively. As can be seen in this figure, positive values of BSS (Figure 3(a)–(c)) dominate the tropical domain for all categories, suggesting that super-ensemble prediction is always better than climatological prediction. However, the improvement of super-ensemble prediction relative to climatological prediction is very subtle for mid-high latitudes, with the values of BSS less than 0.1, especially for NN. The significant improvement occurs in the tropical regions, as shaded in Figure 3. Apparently, the ensemble prediction skill decreases with latitude, with the best skill located around the central equatorial Pacific. A sharp variation in BSS occurs at the boundary, around 30°N/S. Compared with the regions of the same latitudes, the PNA region and the ADP region of both BN and AN events are occupied by larger BSS, suggesting the large-scale climate modes having better prediction skill than general atmospheric flows. A comparison between BSS in Figure 3 with CORR in Figure 1 reveals their good consistency. For events defined in this study, BN has the best BSS skills and NN the worst. This is due mainly to the different contribution of RES for different events as indicated in Figure 3(g)–(i). As can be seen from the global distribution of REL and RES in Figure 3, the item REL (Figure 3(d)–(f)) has little variation with different events. However, the RES (Figure 3(g)–(i)) has a significant variation with events in tropical regions, with the RES of NN obviously smaller

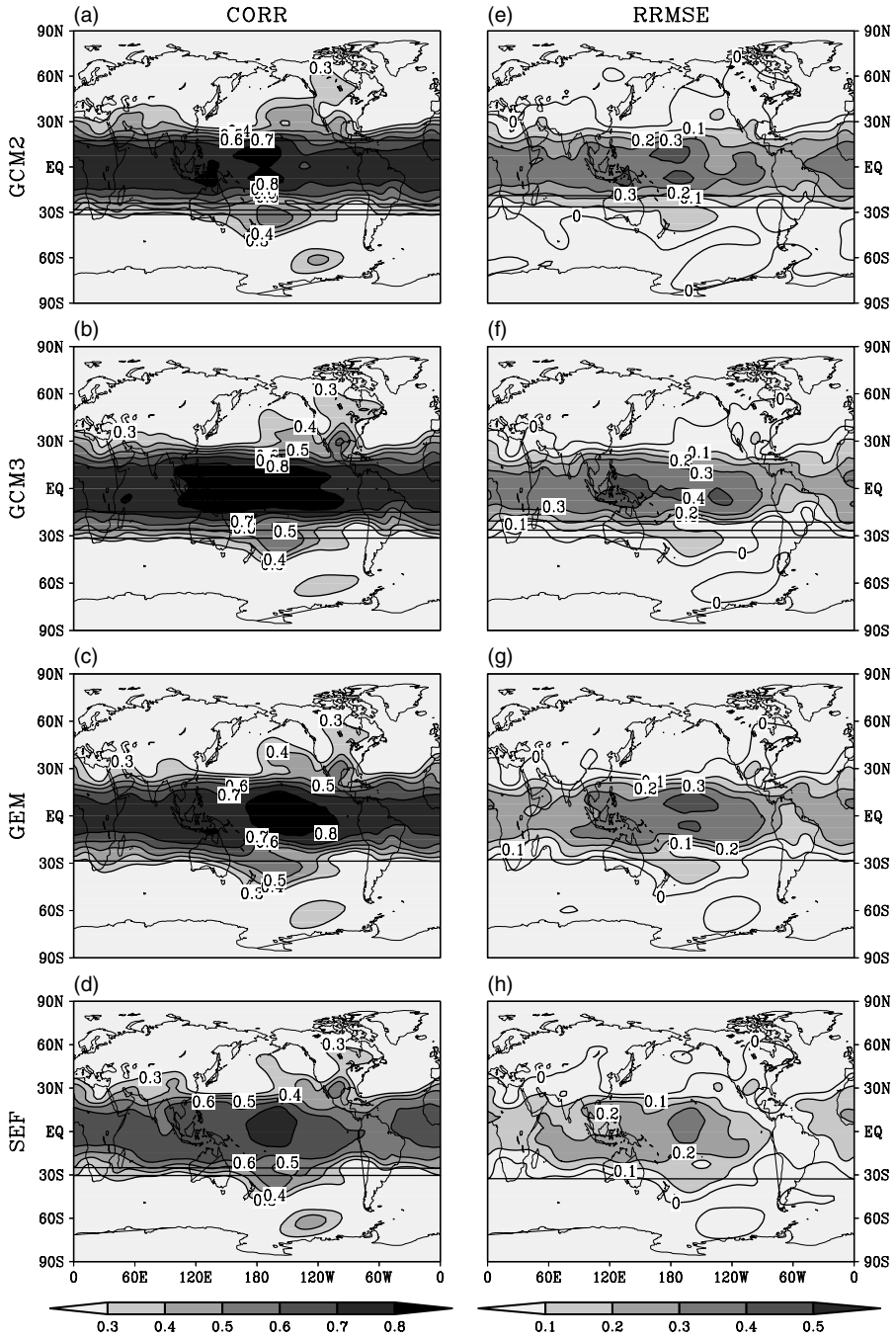

Figure 1. CORR (a)-(d) and RRMSE (e)-(h) of GCM2, GCM3, GEM and SEF.

than the RES of BN and AN. Because a smaller REL value suggests larger reliability and a smaller RES indicates worse resolution, the RELs in Figure 3(d)–(f), with values smaller than 0.1 for most grids, indicate that the forecasts are reliable in most regions. Thus, the BSS skill is dominated by the RES. The small values of RES in mid-high latitudes lead to small BSS there. The good probabilistic skill with the super-ensemble is probably due to the increase of ensemble size and the error offsets among individual model ensembles.

Similar analysis was also applied to individual model ensembles. It was found that individual model ensembles have relatively poor BSS skill (not shown). The positive values mainly occurred in tropical regions between 30°S and 30°N. In the mid-high latitudes, BSS values are usually negative, i.e. single-model ensemble predictions are usually worse than climatology forecasts. It is especially true for the category of NN where the single-model ensemble predictions generate positive BSS values only in a very limited area of the tropical region.

Shown in Figures 4–6 are the differences of probabilistic prediction skills between super-ensemble and individual model ensembles. As can be seen in Figure 4, the super-ensemble has higher BSS values than any other individual model. In fact, the individual model ensembles had negative BSS values for most regions (not shown). The degree of improvement in BSS by the super-ensemble varies with regions, models and categories of event, in a range of 0.01–0.2. Like CORR and RRMSE, the super-ensemble has the maximum improvement of BSS relative to SEF, and then to GCM2, GEM and GCM3. It also has the largest improvement to BSS for the category of NN. Since the BSS is determined by REL and RES (Wilks, 2006), either a decrease

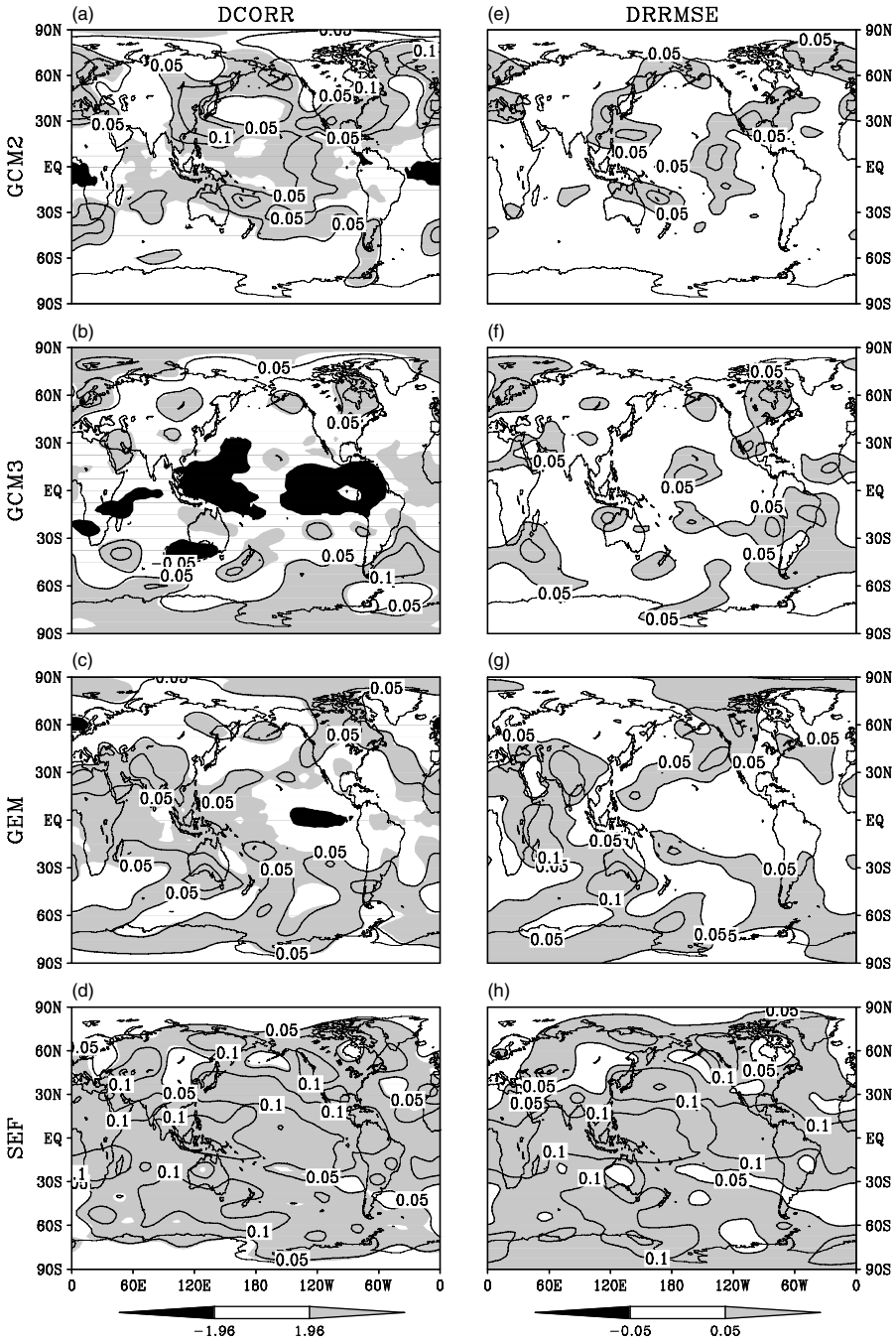

Figure 2. Difference of CORR (a)-(d) and RRMSE (e)-(h) between the super-ensemble and GCM2, GCM3, GEM and SEF.

in REL or an increase in RES can lead to an improvement in BSS. Below, we will examine in detail the REL and RES, the reliability and resolution – two most important attributes of probabilistic predictions.

As mentioned above, the smaller the REL is, the more reliable is the prediction. Practically, the REL is much smaller than the RES for weather forecasts (e.g. Jolliffe and Stephenson, 2003; Atger, 2004). For the seasonal mean forecasts of GPH500 anomalies, this may also be true. For example, it was found here that the REL is one order smaller than the RES as shown in Figure 3. Figure 5 shows the differences of REL of super-ensemble against individual model ensemble, indicating clearly the improvement of super-ensemble to REL as indicated by the negative values over the model domain for all categories of event (BN, NN and AN), especially in the mid-high latitudes and for NN.

Different from the REL, larger RES values indicates higher resolution. The RES measures the ability of a prediction system to resolve predictions into different categories of events. Figure 6 shows the differences of RES between super-ensemble and individual model ensemble. In contrast to the REL, the RES has relatively little improvement by super-ensemble, especially for GEM, GCM2 and GCM3. Even for SEF, the super-ensemble improves the RES only for the tropical regions. These facts suggest that the reliability score may be much more sensitive to ensemble size or uncertainties of initial conditions or model biases than the resolution score, resulting in a great advantage of super-ensemble for the REL. Cheng *et al.* (2010) also found similar conclusions for ENSO ensemble predictions and concluded that the RES is little sensitive to the construction of ensemble predictions. In other words, the merits of a good

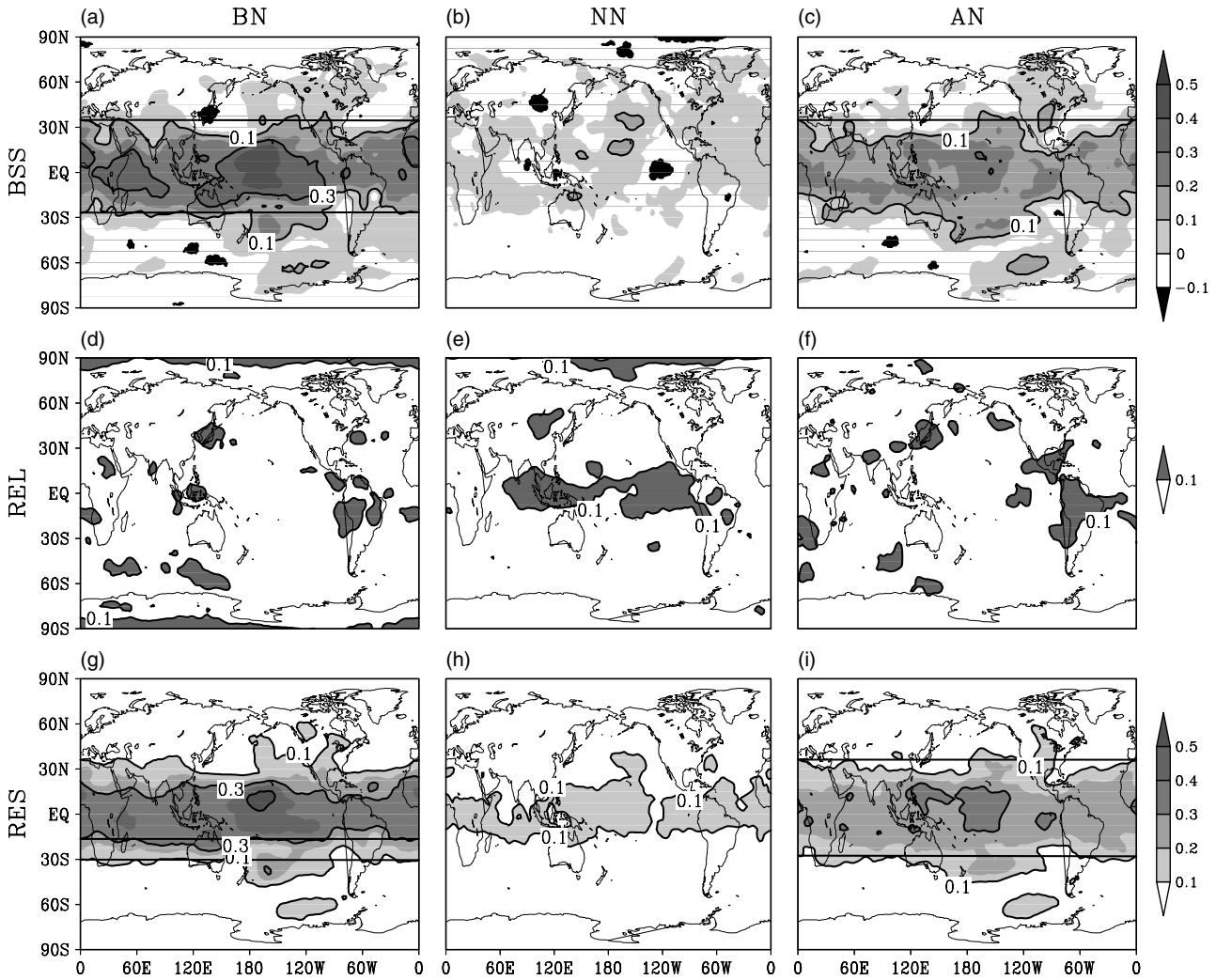

Figure 3. BSS (a)-(c), REL (d)-(f) and RES (g)-(i) of the super-ensemble for BN (a), (d), (g), NN (b), (e), (h) and AN (c), (f), (i).

super-ensemble prediction system may be mainly reflected in the reliability score.

To further examine the REL and the RES, we will analyse the reliability histogram and sharpness histogram. Because of the large number of model grid points, it is impractical to plot these histograms for each point. Here, we used two indices of GPH500 anomalies, obtained by the averages over two regions: the North America (NA: 240°E–300°E, 30°N–50°N) and the tropical Pacific (TR: 120°E–240°E, 15°S–15°N). Figure 7 and Figure 8 are their reliability and sharpness histograms for the categories of BN (a)–(b), NN (c)–(d) and AN (e)–(f). Figure 7 shows that super-ensemble has significantly better REL in both regions for BN and AN, as suggested by the fact that the joint distribution of the observed frequencies and the forecast probabilities of super-ensemble is more centred on the diagonal line than any other individual model. In North America, individual model ensembles show similar reliability skill to each other; however, in the tropical Pacific, the REL skill varies with individual models with a range of forecast probabilities (P_{fi}) from 0.3 to 0.7, with SEF best, followed by GEM and GCM2. For the super-ensemble, the REL skill is comparable for different locations (e.g. $\mathrm{REL}_{NA} = 0.02$; $\mathrm{REL}_{TR} = 0.03$). The sharpness histograms (Figure 8) indicates that the ensemble predictions in the tropical Pacific are much sharper than those for North America, leading to different BSS scores in the two regions (i.e. $\mathrm{BSS}_{NA} = 0.04$; $\mathrm{BSS}_{TR} = 0.35$). Again, the difference of sharpness among individual model ensembles is small, suggesting that the ability of a seasonal climate prediction system to predict the extreme probabilities (e.g. 0 or 1) is difficult to improve by super-ensemble approach. In other words, super-ensemble has a very limited ability to improve forecast resolution. However, this conclusion might be prediction system dependent. For example, Hagedorn *et al.* (2005) reported a good improvement of resolution skill by super-ensemble when the seasonal prediction target is the average air temperature at 2 metres over the tropical band (30°S–30°N), where the super-ensemble consists of seven coupled models. It is interesting to further examine the resolution of the super-ensemble using other prediction targets, which is currently in progress.

3.3. *Reasons for superiorities of super-ensemble*

It has been reported that the good skill achieved by super-ensemble could be due to either the error compensation among different individual ensemble systems or to the large ensemble size from multi-models (e.g. Hagedorn *et al.*, 2005; Kharin *et al.*, 2009). The real merit of super-ensemble should be related to the former rather than the latter, since a large ensemble size can also be achieved

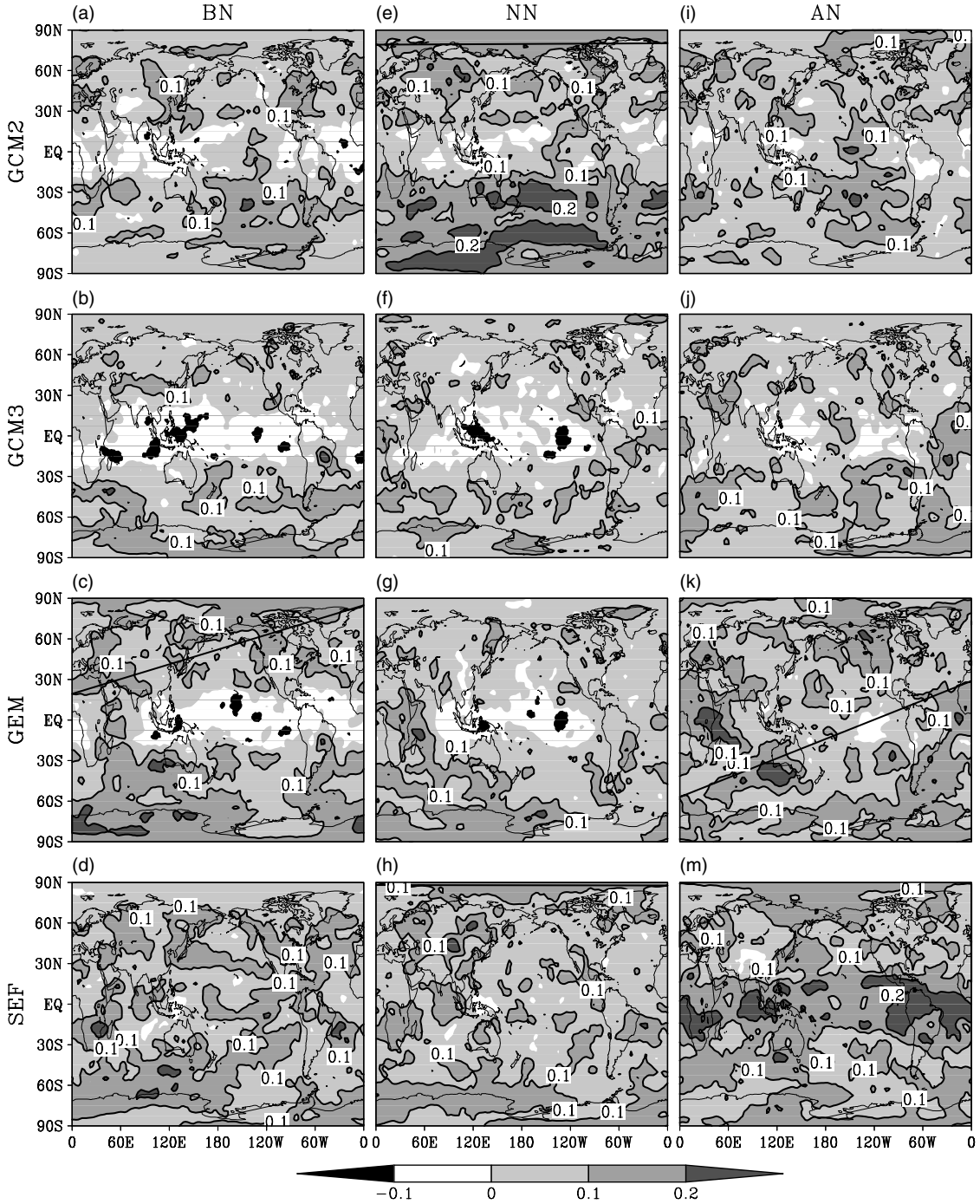

Figure 4. BSS difference between the super-ensemble and GCM2 (a), (e), (i), GCM3 (b), (f), (j), GEM (c), (g), (k) and SEF (d), (h), (m) for BN (a)-(d), NN (e)-(h) and AN (i)-(m).

by a single model. Therefore, it is interesting to examine whether the better prediction skills shown in the above super-ensemble analysis are due simply to the increase in ensemble size. For this purpose, we performed a bootstrap experiment (referred to as BEXP), namely: (i) constructing the super-ensemble of 10 members, the same size as that of a single model, by random selection from all 40 ensemble members; (ii) applying all analyses above (evaluations of the deterministic and probabilistic prediction skill) to the BEXP ensemble; (iii) repeating (i) and (ii) 1000 times, and then averaging them for stable statistics.

Figure 9 shows the difference of CORR and RRMSE between the BEXP and single model ensemble. As can be seen in Figure 9(a)–(d), the BEXP has a correlation skill better than SEF, but similar to GCM2, GCM3 and GEM. However, in terms of RRMSE (Figure 9(e)–(h)), the BEXP seems much worse than individual model ensembles in the tropical regions. A comparison between Figure 9 and Figure 2 leads to the following conclusions: (i) the super-ensemble can significantly improve correlation skill for poor individual ensembles (e.g. SEF here), but is of little use for good individual ensembles (e.g. GEM); (ii) the super-ensemble is little effective in improving RRMSE-based skills if the potential impact of ensemble size is removed. In other words, the significant improvement of super-ensemble to RRMSE shown in Figure 2 is actually simply due to the increase of ensemble size, rather than a real compensation of model errors.

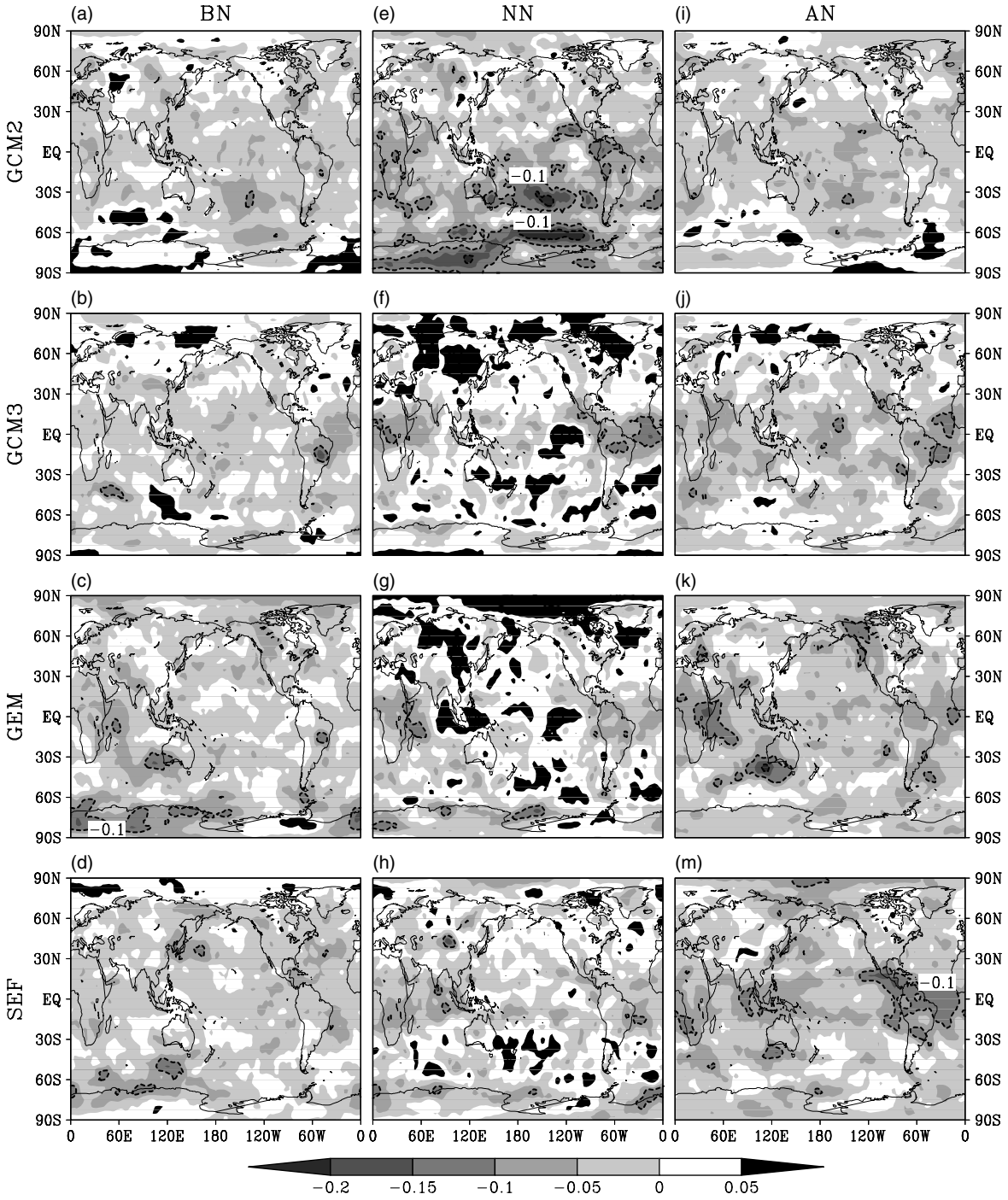

Figure 5. REL difference between the super-ensemble and GCM2 (a), (e), (i), GCM3 (b), (f), (j), GEM (c), (g), (k) and SEF (d), (h), (m) for BN (a)-(d), NN (e)-(h) and AN (i)-(m).

Shown in Figure 10 is the BSS (a)–(c), REL (d)–(f) and RES (g)–(h) skills of the BEXP for BN, NN and AN. A comparison between Figure 10 and Figure 3 reveals that the decrease in ensemble size leads to a decrease in probabilistic prediction skills to a different extent, which depends on regions and skill measures. The significant decrease occurred at the high latitudes for BSS and REL. For example, for BN and AN, the BEXP has worse seasonal predictions than climatology forecasts in high latitudes ((a) and (c)) even in PNA and ADP regions. However, the RES in high latitudes is relatively less impacted by the ensemble size, especially for NN and AN. These results suggest that the real merits of super-ensemble might be only reflected in the RES, and the BSS and REL in tropical regions. The significant improvement of BSS and REL of super-ensemble at mid-high latitudes as shown in Figures 4 and 5 are probably due to a simple increase in ensemble size.

A further analysis is to compare the BEXP against individual models for BSS, REL and RES. It was found that the superiorities of the super-ensemble over individual ensemble significantly weaken when the ensemble size decreases for these probabilistic skills, especially for REL. The difference between the BEXP and individual ensembles (not displayed) shows that the BEXP with 10-member super-ensemble still has higher BSS and RES skill than SEF, but no such significant difference in REL as shown in Figure 5(d).

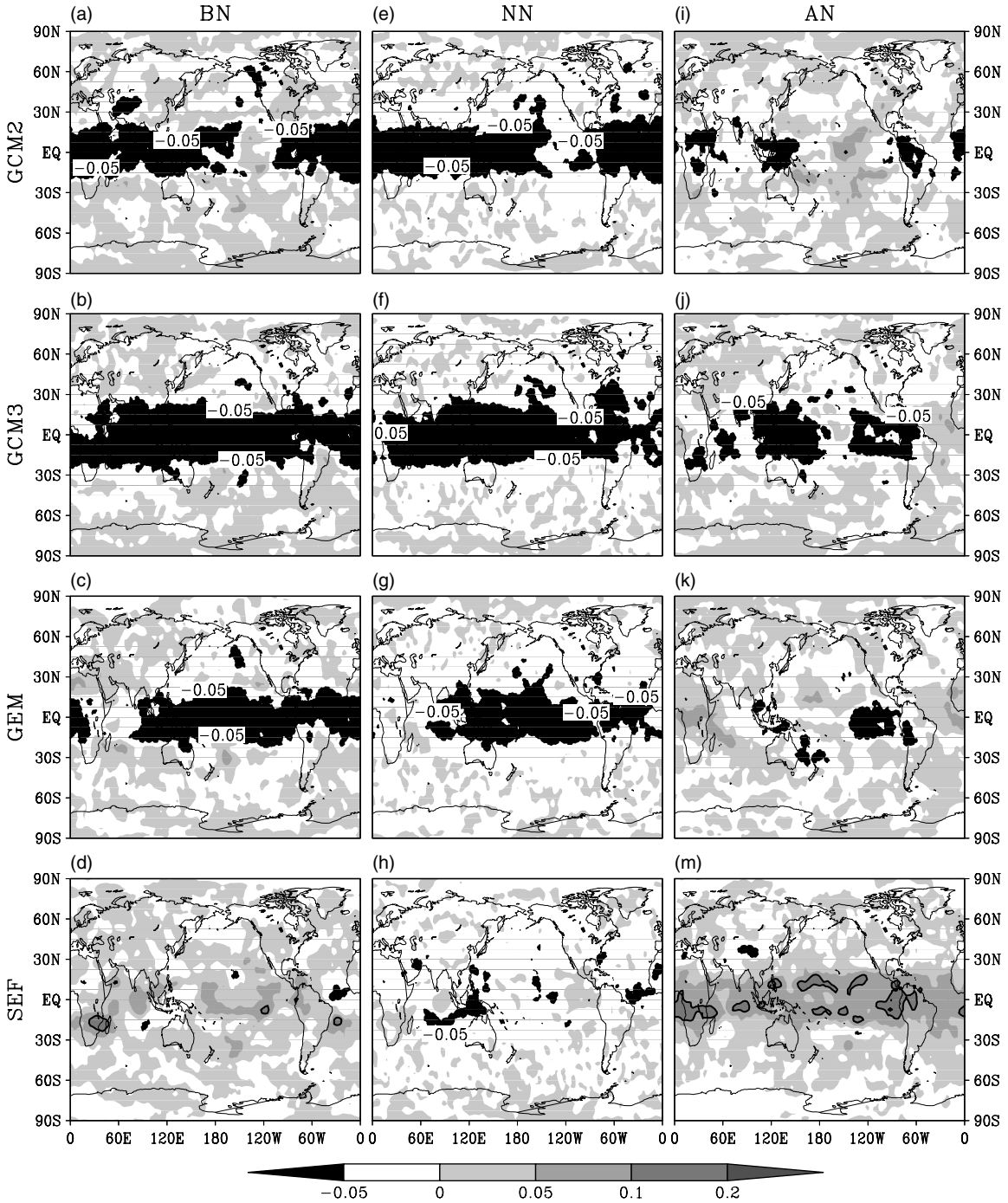

Figure 6. RES difference between the super-ensemble and GCM2 (a), (e), (i), GCM3 (b), (f), (j), GEM (c), (g), (k) and SEF (d), (h), (m) for BN (a)-(d), NN (e)-(h) and AN (i)-(m).

Relative to other individual model ensembles, the BEXP has less advantage, and may even be worse. For example, the GCM2, GEM and GCM3 ensemble can actually outperform the BEXP for some regions.

In summary, the simple increase in ensemble size is an important contribution to the advantages of super-ensemble. This is especially obvious for mid-high latitudes and for BSS and REL. Thus, one should be cautious in interpreting the superiorities of super-ensemble in improving seasonal climate probabilistic prediction in mid-high latitudes. However, the super-ensemble has inherent advantages in tropical seasonal climate prediction. We also carried out a detailed analysis for two indices: the average over tropical regions (TR) and the average over the North America region (NA). The results confirmed the above summary, namely, there is a large difference of BSS and REL between the super-ensemble and the BEXP for the NA (not shown). However, for the TR and for RES, the BEXP shows similar skills to the super-ensemble, suggesting that model uncertainties responsible for RES and the tropical prediction, rather than ensemble size, can indeed be improved by super-ensemble.

4. Potential prediction skill

In preceding sections, we discussed the deterministic skills and probabilistic skills of super-ensemble predictions. In this section, we will analyse the potential prediction skill

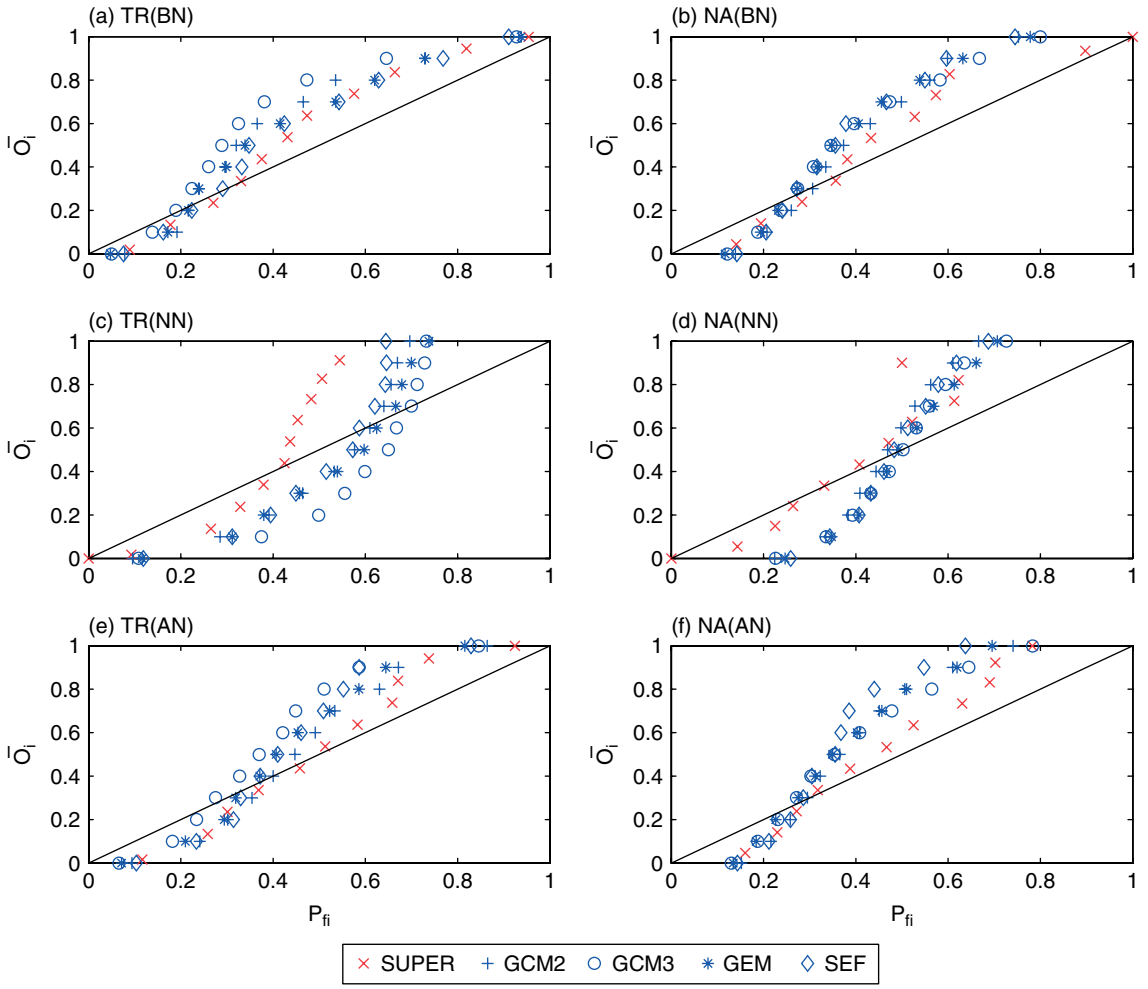

Figure 7. Reliability diagrams of TR and NA regions for BN (a), (b), NN (c), (d) and AN (e), (f).

under 'perfect model scenario' of all ensembles and examine the possible advantages of the super-ensemble by measures of PCORR and REE introduced in section 2.2.3.

Figure 11(a)–(e) shows PCORR under 'perfect model scenario' of all ensembles. A comparison between Figure 11(b)–(e) and Figure 1(a)–(d) reveals that the potential prediction skills are similarly distributed as the actual prediction skills, with sharp transition zones at approximate 30°N/S in both hemispheres and relatively higher prediction skills in both PNA and ADP regions. The spatial correlation between the distributions of PCORR and CORR are 0.66, 0.86, 0.85, 0.84 and 0.87 respectively for the super-ensemble, GCM2, GCM3, GEM and SEF.

Figure 11(f)–(j) shows REE between the PCORR and their corresponding CORR for all ensembles. REE of all ensembles are mostly positive, indicating that the potential prediction skills are generally higher than the actual prediction skills. In the global domain, the REE are smallest in tropical regions, and increase poleward in both hemispheres. The REE of each single ensemble is quite large in the global domain, indicating a large difference between the practical prediction skill and potential forecast skill. Compared with individual ensembles, REE skill of the super-ensemble is better in particular in mid-high latitudes. In other words, the potential prediction skill of the super-ensemble can be a better reflection of the actual prediction skill. Thus, when a predictor of the prediction accuracy is asked for before observations are available, the best choice would be the potential prediction skill of the super-ensemble.

A question raised by Figure 11 is why the PCORR of a multi-model ensemble (MME) is smaller than for an individual ensemble. Similar results were also found in other works (e.g. Yang *et al.*, 2012). This is most likely due to two reasons: (i) small ensemble size in a single ensemble may underestimate the ensemble spread (noise), which can be greatly alleviated by MME; and (ii) the model bias of a single ensemble can contribute spurious components to the signal, which can be offset by MME. Thus, the potential predictability of a single ensemble is often overestimated.

5. EOF and MSN EOF analyses

In this section, we will explore the climate variability and climate predictability by applying the EOF and MSN EOF methods to ensemble forecasts. Emphasis will be placed on the comparison between the EOF and MSN EOF methods, and between the super-ensemble and each single ensemble. The former will identify the relationship between the variability and predictability whereas the latter will examine the superiorities of the super-ensemble over individual ensembles.

In the analysis below, the sample climatology is used as the real (population) climatology. The prediction is the GPH500 anomalies with the seasonal cycle removed. Analyses focus on the seasonal mean (the average of 30–120 days) forecasts

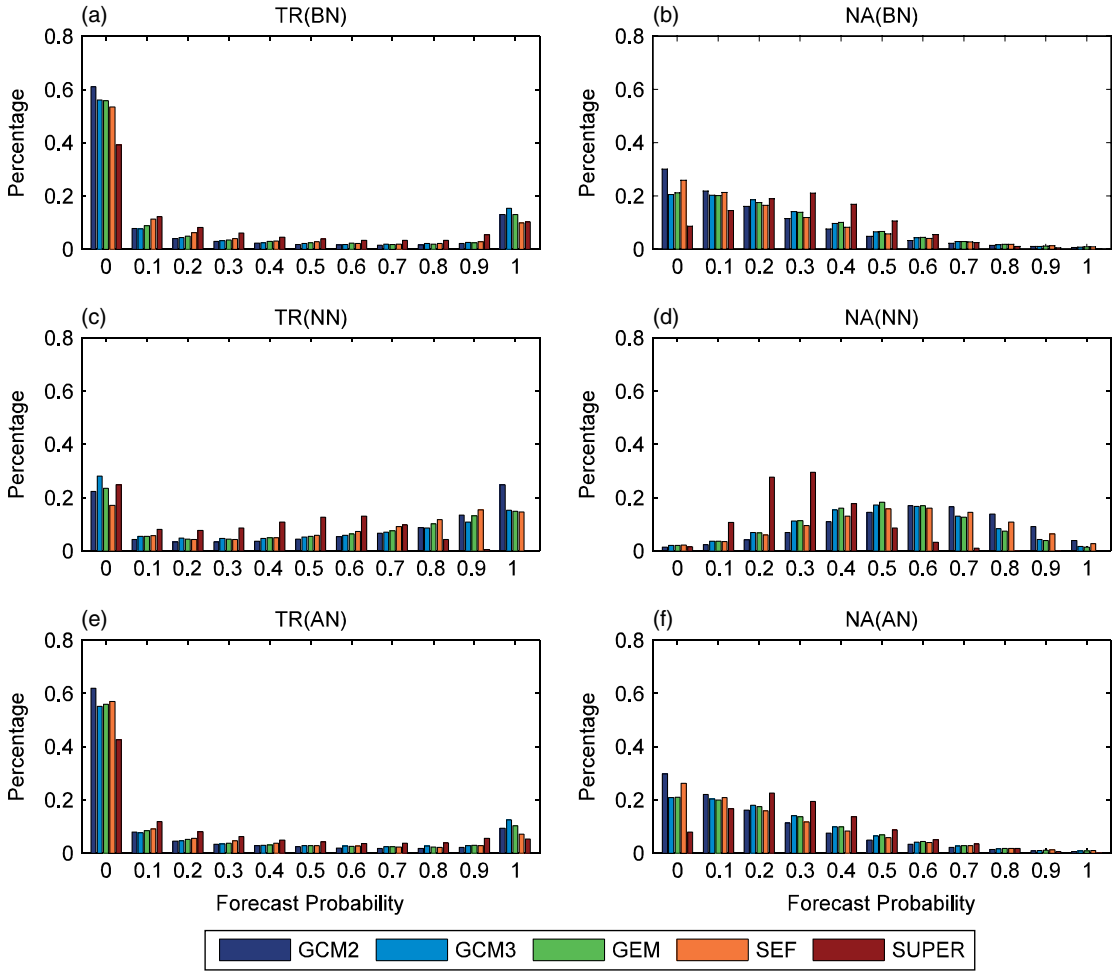

Figure 8. Sharpness histograms in TR and NA regions for BN (a), (b), NN (c), (d) and AN (e), (f).

in winter (December–February (DJF) mean) for the period from 1969 to 2002. The super-ensemble is formed by pooling together all ensemble members available.

In the MSN EOF analysis, we should be careful to choose the appropriate number of modes to make the projection basis adequate and the pre-whitened signal covariance matrix well conditioned. It has been found in our analysis that the SNR gradually increases with the number of modes, especially for the first 70 modes. Typically, the SNR becomes stable after 80 modes. In this study, we kept the first leading 80 modes in the pre-whitening process in the MSN EOF analysis.

5.1. *EOF analysis*

Often the ensemble mean, called signal, is regarded as the external variability induced by different boundary forcings (e.g. Straus and Shukla, 2002; Straus *et al.*, 2003; Honda *et al.*, 2005; Liang *et al.*, 2009; Matsumura *et al.*, 2010), and considered to be potentially predictable. The deviations from the ensemble mean are regarded as the internal variability, referred to as noise that is considered to be potentially unpredictable (e.g. Venzke *et al.*, 1999). The noise is typically due to the chaotic or stochastic components in the weather and climate systems. Theoretically, the resultant signal and noise are independent if the ensemble size is infinite (Venzke *et al.*, 1999; Sutton *et al.*, 2000). Given the limited ensemble size available in reality, the signal and noise may be dependent in either temporal or spatial scales, or in both. To check the dependence of signal and noise in detail given finite ensemble size, the EOF decomposition is applied to covariance matrices of both signal and noise, respectively.

Figure 12(a)–(e) shows the spatial patterns of the first EOF mode of ensemble mean predictions (signals) in winter for the super-ensemble and individual ensembles, respectively. The variance explained is shown in the upper-right corner of each map. In winter, all the first EOF modes are PNA-like patterns distributed with the largest active centres in the North Pacific. Both GCM2 and GCM3 show intense active centres above East Asia and western Europe. GEM and SEF are more consistent with the super-ensemble. The spatial correlations between the first EOF mode of the super-ensemble and GCM2, GCM3, GEM and SEF are 0.83, 0.87, 0.95 and 0.96. These spatial patterns are roughly consistent with the pattern of the SST-forced signals as reported in Matsumura *et al.* (2010).

Figure 13 (left panel) shows the corresponding PCs of the first EOF modes. The observed PC is obtained by projecting the observation (i.e. NCEP reanalysis) onto the corresponding EOF modes. A good agreement between the predicted and the observed PCs is expected for a skilful prediction. Therefore the comparison between the two PCs in Figure 13 can reveal the prediction accuracy. As can be seen, there are large discrepancies between the forecast and observation. Table 1 presents the temporal correlation (AC) and root-mean-square error (RMSE) between the two PCs.

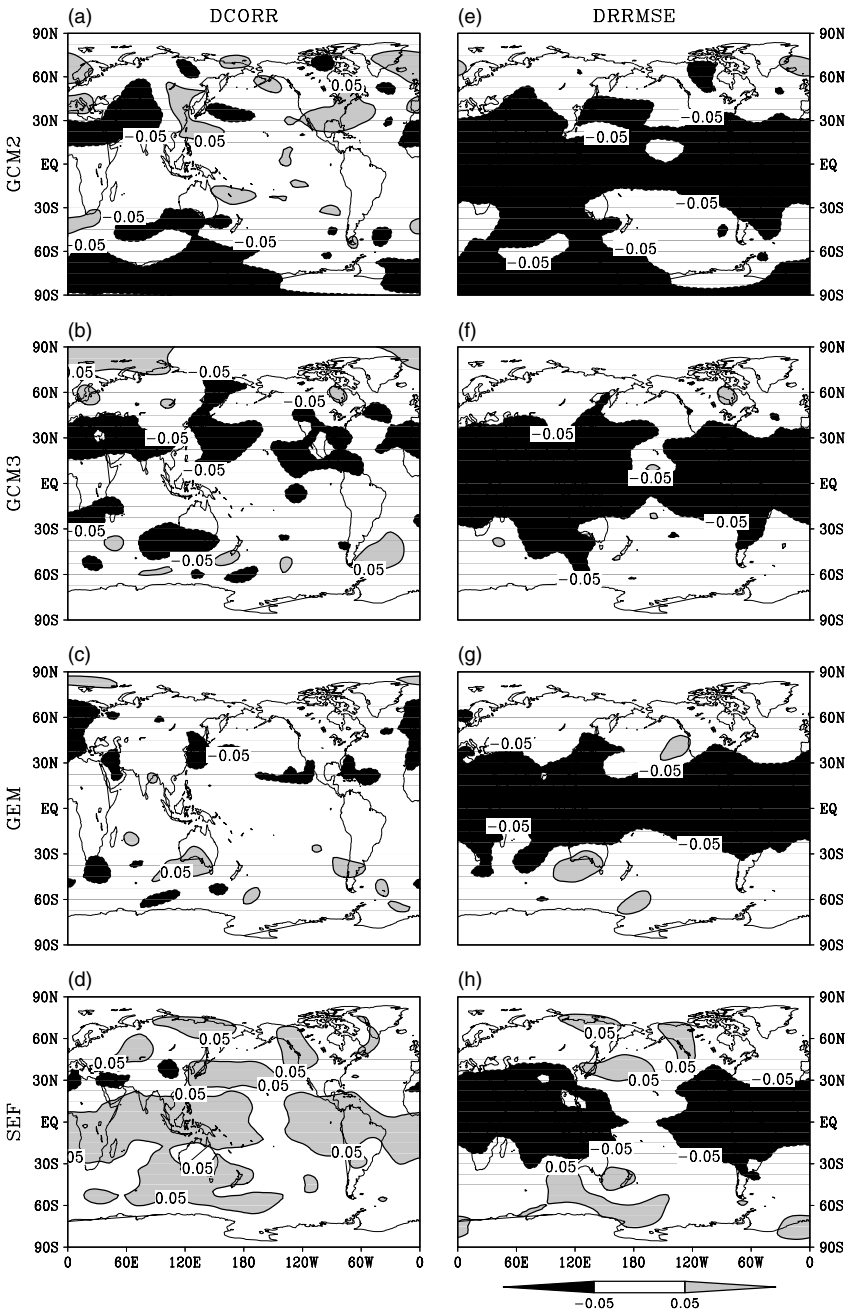

Figure 9. Difference of CORR (a)-(d) and RRMSE (e)-(h) between the BEXP and each single ensemble.

As can be seen, ACs of all model ensembles including the super-ensemble are negative, indicating that these models all have a poor prediction at the seasonal time-scale for the large-scale variability characterized by the PC.

It is possible that the external variability forced by boundary conditions may not be well extracted here by the EOF analysis due to the limited ensemble size. For example, the external variability of the extratropical atmosphere is expected to be larger in warmer ENSO years (El Niño years) than in neutral ENSO years, because the larger SST anomaly in the central and eastern tropical Pacific can dramatically affect the extratropical atmospheric circulations in the Northern Hemisphere through teleconnection mechanisms. However, Figure 13 does not show that the amplitudes of PCs in El Niño years (e.g. 1972/1973; 1982/1983; 1991/1992) are systematically larger than those in neutral years, suggesting that the first EOF mode of the ensemble mean still includes some noise that contaminates the signal and finally lowers the amplitudes of PCs. A further discussion of this issue will be given in the MSN EOF analysis.

5.2. *MSN EOF analysis*

As mentioned above, the MSN EOF analysis seeks the maximum ratio of signal over noise, and produces the most predictable pattern. To explore the improvement of the super-ensemble in characterizing the most predictable patterns, the MSN EOF method is applied to the super-ensemble and individual ensembles respectively. The observed data are also projected on the first filter pattern to obtain the observed PrCs. Similar to the EOF analysis, a good consistency between the predicted PrC and observed PrC is expected for a skilful prediction.

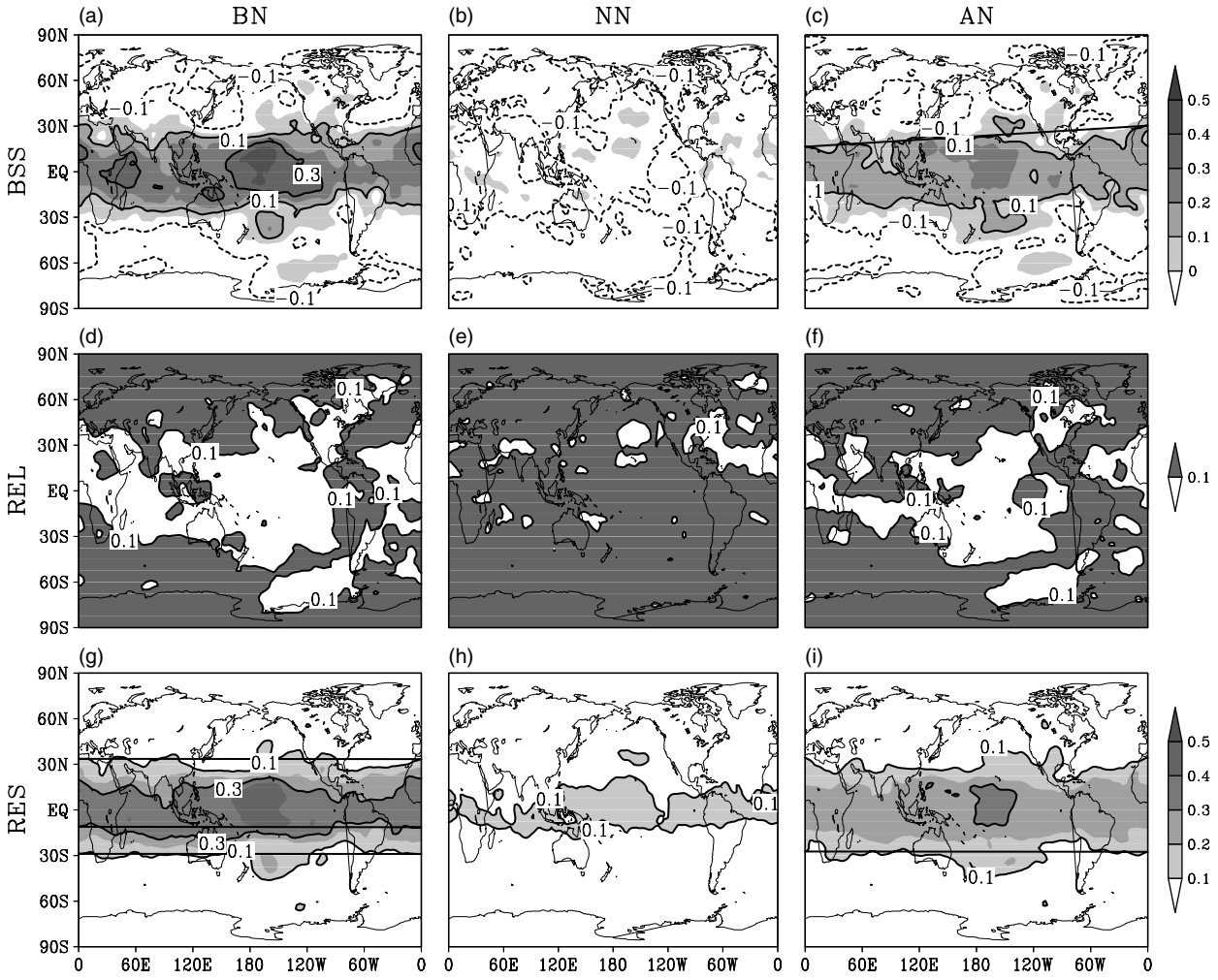

Figure 10. BSS (a)-(c), REL (d)-(f) and RES (g)-(i) of BEXP for BN, NN and AN.

Table 1. AC and RMSE between the observed and forecasted PCs and PrCs in winter.

PCs	SUPER	GCM2	GCM3	GEM	SEF	BEXP2
AC	−0.3544	−0.0497	−0.2762	−0.3610	−0.2408	−0.1020
RMSE	13.37	9.324	9.760	10.05	9.529	9.617
PrCs	SUPER	GCM2	GCM3	GEM	SEF	BEXP2
AC	0.8806	0.8010	0.8611	0.8487	0.7912	0.8731
RMSE	9.740	23.55	25.40	19.99	26.78	6.1514

Displayed in Figure 12(f)−(j) are the PrCA spatial patterns, i.e. the most predictable patterns in winter. The value of SNR is shown in the upper-right corner. The corresponding PrCs are displayed in Figure 13 (right panel). In winter, the most predictable structure is the PNA-like pattern in all ensembles, although there are some subtle model-dependent features. Comparison between Figure 12(a)−(e) and Figure 12(f)−(j) reveals that the most predictable patterns are similar to the first EOF modes of the ensemble mean anomaly predictions. This indicates that the ensemble mean anomaly can indeed spatially represent the signal to some extent. The significant relation of the ensemble mean to the potential predictability also was found in some previous studies (e.g. Straus and Shukla, 2002; Tang *et al.*, 2008).

Figure 13 (right panel) shows that the forecasted PrCs are in good agreement with the observation counterparts. The large amplitudes of the predicted PrC correspond well with El Niño events such as those in 1972/1973; 1982/1983; 1986/1987; 1997/1998 etc. This suggests that the seasonal climate predictability is probably due to ENSO forcing. The phase and amplitude consistency between the prediction and observation can be clearly seen from Table 1, which summarizes the AC and RMSE of the PrCs. All ACs are statistically significant at the 95% confidence level. The level of CORR of the super-ensemble is comparable to that of the single ensembles, whereas the super-ensemble has much reduced RMSE skills. The RMSE between the predicted and observed PrCs, mostly contributed by their magnitude differences, may be due to the erroneous estimation of the transient diabatic forcing (e.g. the release of latent heat of the convective precipitation) in the atmospheric general circulation models (AGCMs). In the NH, the transient diabatic forcing has minor effects on the spatial structures of atmospheric circulation, but plays an important role in the regulation of their temporal behaviour (Sardeshmukh and

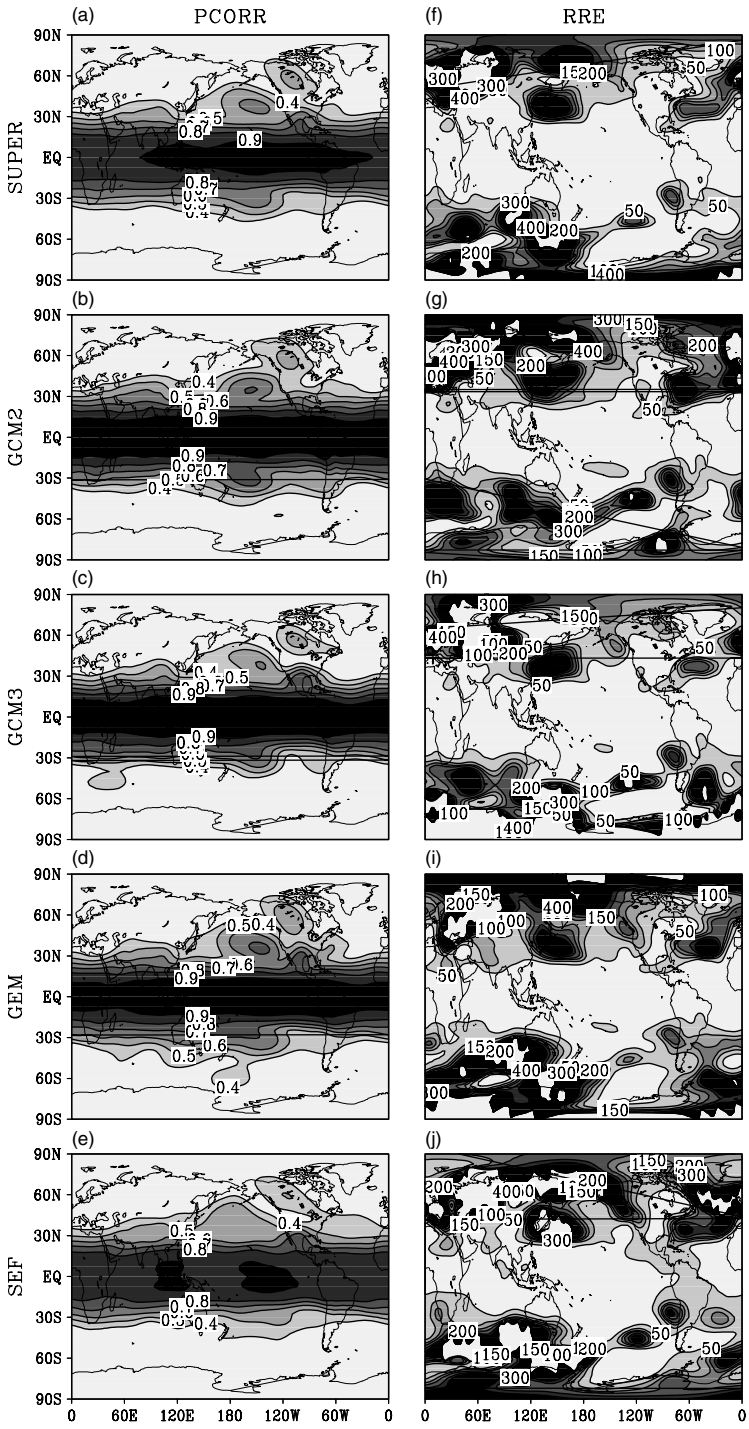

Figure 11. PCORR (a)-(e) and RRE (f)-(j) of SUPER, GCM2, GCM3, GEM and SEF.

Sura, 2007; Yasui and Watanabe, 2010). To some extent, the estimation of the transient diabatic forcing in the AGCMs can be improved by the super-ensemble.

5.3. *Reasons for benefits of the super-ensemble*

As in section 3.3, we will use a bootstrap experiment to explore the influences of the ensemble size on both EOF and MSN EOF analyses. This experiment, called BEXP2, is designed in the same way as that in section 3.3: (i) the ensemble with 10 members is constructed by random selections from all single model ensembles; (ii) the EOF and MSN EOF methods are applied to the BEXP2 ensemble respectively; (iii) steps (i) and (ii) are repeated 1000 times, and the averages are presented. The PCs and PrCs of BEXP2 are displayed in Figure 14. Shown in Table 1 (last column) are the AC and RMSE between the forecasted and observed PCs (PrCs) of BEXP2.

As in Figure 13 (left panel), the forecasted and observed PCs of BEXP2 are quite different (Figure 14). In contrast with the PCs, the discrepancies between the forecasted and observed PrCs are much smaller in BEXP2. The ENSO forcing seems well reflected in PrCs, e.g. the higher amplitudes of PrCs occur in El Niño years (1972–1973, 1982–1983, 1986–1987 and 1997–1998). As shown in Table 1, the ACs of BEXP2 are comparable to those of

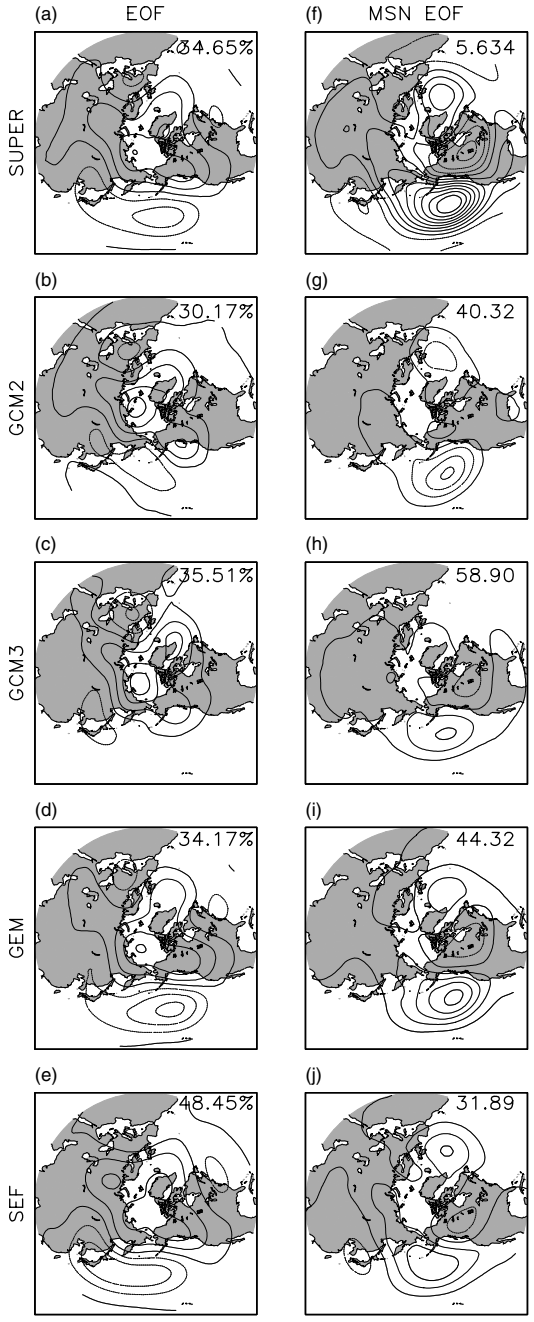

Figure 12. Spatial distributions of the first EOF modes (a)-(e) and the most predictable patterns (f)-(j) of the winter seasonal mean GPH500 anomalies.

the super-ensemble, suggesting that the ensemble size does not much impact the correlation skill. However, the RMSE of BEXP2 are smaller than those of the single ensembles (equal ensemble size), suggesting that the offsets of model uncertainties occur in the super-ensemble.

The SNR of the BEXP2 is 33.13 in winter, which is larger than SNRs of both the super-ensemble and individual ensembles. To examine the variation of SNR, SNRs are calculated using ensembles with changing ensemble size by a set of bootstrap experiments. The SNR is found to be highly sensitive to the ensemble size, i.e. it decreases quickly with the increase of ensemble size (not shown). Given equal ensemble size of the BEXP2 and the single ensembles, the larger SNR of the BEXP2 may reflect the error offsets among different model frameworks. The SNR measures the maximum ratio of signal over noise. Mathematically the optimized SNR equals the largest eigenvalue of the whitened signal covariance matrix. A matrix that has a larger rank often corresponds to a smaller variance accounted for by the first eigenvalue. This may explain why the SNR of the super-ensemble is smaller than those of individual ensembles as shown in the upper-right corner of Figure 12, since the former has much larger ensemble size.

6. Discussion and conclusions

In this study, the superiorities of the super-ensemble are comprehensively evaluated using 500 mb geopotential height anomalies of multiple model ensembles. First, the predictions at each grid point over the global domain were evaluated in terms of the deterministic, probabilistic and potential forecast skill. Second, the most predictable components associated with large-scale climate modes were analysed using both EOF and MSN EOF methods.

It was found that for predictions of the seasonal mean anomalies, the improvements of the super-ensemble are measure, location and model dependent. In terms of the deterministic forecast, the super-ensemble is a little better than the best single model, especially at mid-high latitudes where the atmospheric uncertainty is large. For the correlation skill, the superiority of the super-ensemble is mainly at mid-high latitudes, whereas for the RRMSE skill, the super-ensemble is better in the tropical regions. For the probabilistic forecast, the super-ensemble improves BSS and REL greatly at mid-high latitudes, but improves the RES only in the tropical regions. It has been found that the REL is much more sensitive to the ensemble size, whereas the RES to some extent is sensitive to the uncertainties of model frameworks. In addition, it has been found that the potential prediction skills PCORR represents well the actual prediction skill CORR. Compared with the single ensembles, the super-ensemble enables the potential prediction skill best consistent with the actual prediction skill.

Both the EOF and MSN EOF analyses can theoretically characterize the most predictable patterns. However, given limited ensemble size, the principal components can hardly represent the temporal variations of the external forcing. In contrast, the most predictable components derived by the MSN EOF method can well capture the temporal variations of the external forcing, in particular the ENSO forcing. Given limited ensemble size, the super-ensemble has a better prediction of the PrCs than single model ensembles.

The bootstrap experiments show that, geophysically, the improvements of the super-ensemble at mid-high latitudes are mainly due to the increase of ensemble size. In contrast, the improvements of the super-ensemble in tropical regions are mainly due to the offsets of model uncertainties. Measures of CORR, BSS and REL are very sensitive to the increase of ensemble size, whereas RRMSE and RES are more sensitive to the offsets of model uncertainties. Given limited ensemble size, the offsets of model uncertainties can lead to a better prediction for the principal components and the most predictable components.

As found in this study, the super-ensemble indeed has advantages over the single model ensembles. However, the improvements from the super-ensemble are not very impressive. This is probably due to the following two reasons: (i) the sample size is not sufficient; and (ii) the single model has already achieved very skilful predictions for the seasonal mean GPH500 anomalies. It is expected

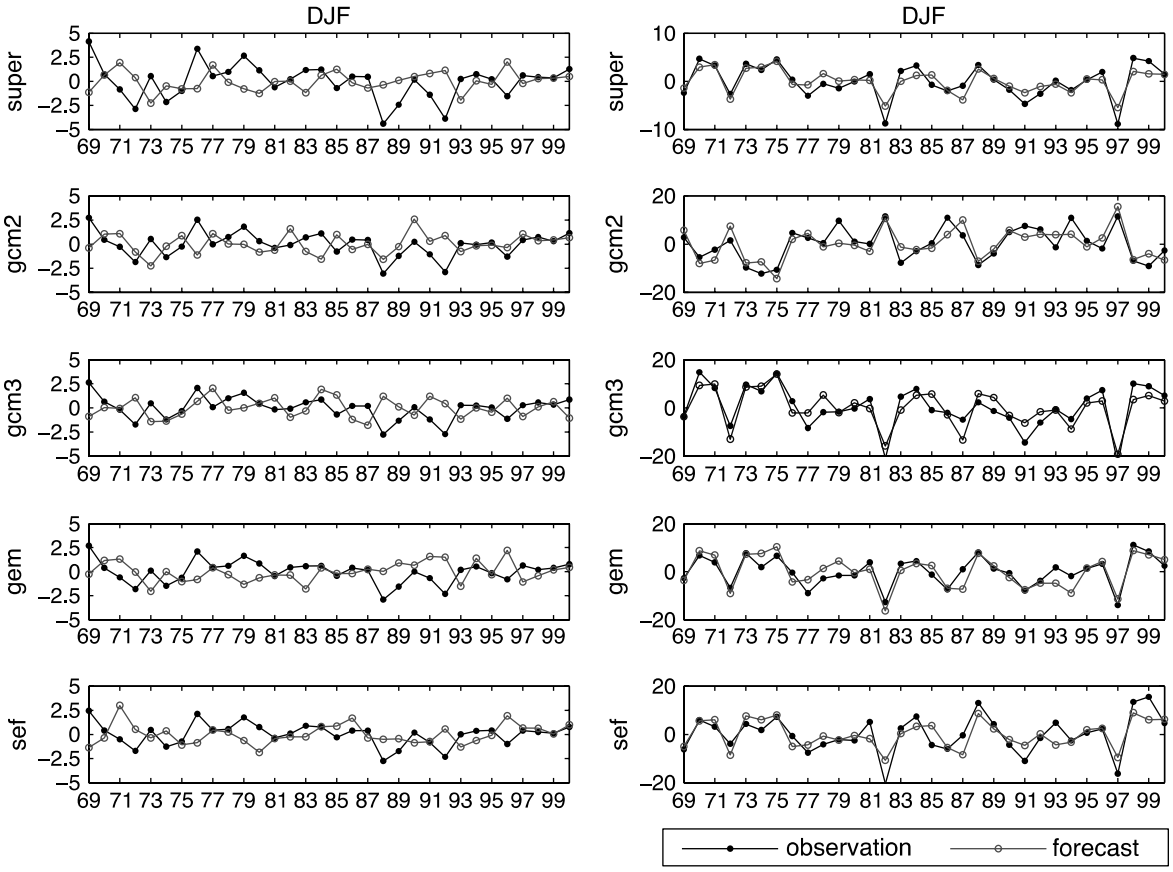

Figure 13. The principal components (PC) (left panels) and the most predictable components (PrC) (right panels) of the seasonal mean GPH500 anomalies in winter.

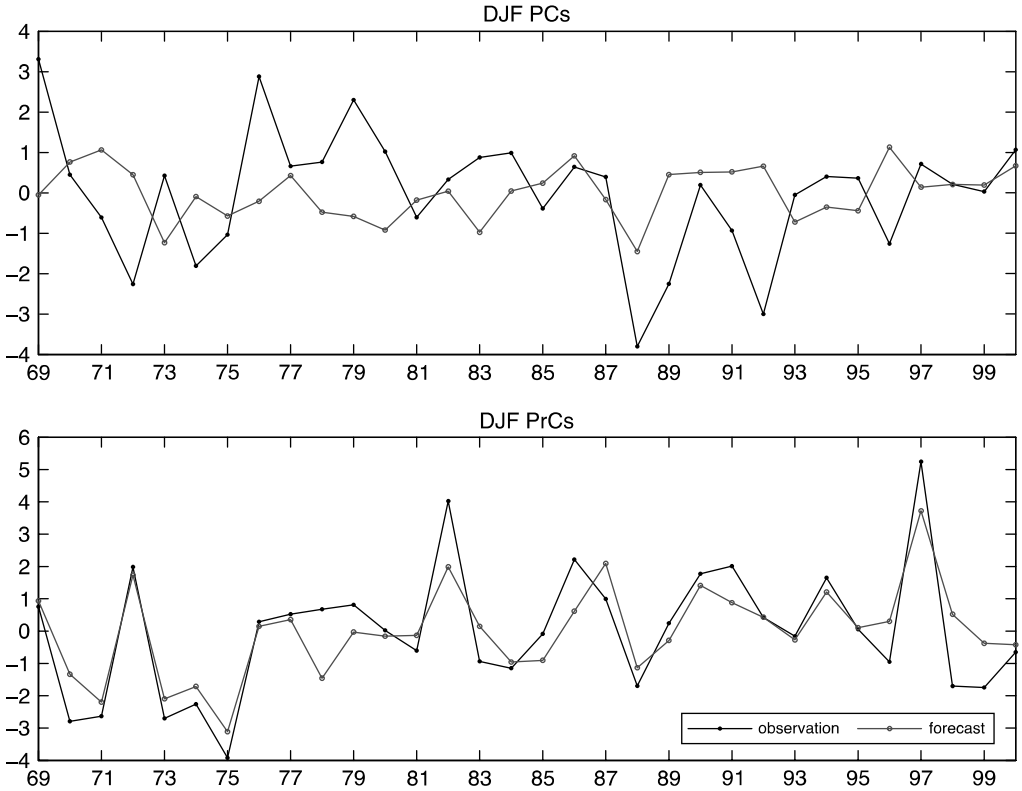

Figure 14. The principal components (PC) and the most predictable components (PrC) of BEXP2.

that the superiorities of the super-ensemble may be much more obvious if other variables, such as precipitation and surface temperature, are considered, using more ensemble members. A further study using European Centre for Medium-Range Weather Forecasts products of DEMETER and ENSEMBLE is under way. Finally, it should be recognized that the further improvements of forecast ability of the super-ensemble still rely on the refinements of the single ensemble and refinements of individual models, especially for the parametrizations and understanding of the physical processes at mid-high latitudes. The demand of improving models should not be obviated by the multiple model approach.

Acknowledgements

This work was supported by Canada NSERC Discovery Grant and the Open Grant of State Key Laboratory of Satellite Ocean Environment Dynamics.

References

Allen MR, Smith LA. 1997. Optimal filtering in singular spectrum analysis. *Phys. Lett. A* **234**: 419–428.

Atger F. 1999. The skill of ensemble prediction systems. *Mon. Weather Rev.* **127**: 1941–1953.

Atger F. 2004. Estimation of the reliability of ensemble-based probabilistic forecasts. *Q. J. R. Meteorol. Soc.* **130**: 627–646.

Balmaseda MA, Anderson D. 2009. Impact of initialization strategies and observations on seasonal forecast skill. *Geophys. Res. Lett.* **36**: L01701, DOI: 10.1029/2008GL035561.

Barnston AG, Mason SJ, Goddard L, Dewitt DG, Zebiak SE. 2003. Multimodel ensembling in seasonal climate forecasting at IRI. *Bull. Am. Meteorol. Soc.* **84**: 1783–1796.

Boer GJ, McFarlane NA, Laprise R, Henderson JD, Blanchet J-P. 1984. The Canadian Climate Centre spectral atmospheric general circulation model. *Atmos.–Ocean* **22**: 397–429.

Candille G. 2009. The multiensemble approach: The NAEFS example. *Mon. Weather Rev.* **137**: 1655–1665.

Cheng YJ, Tang YM, Jackson P, Chen DK, Deng ZW. 2010. Ensemble construction and verification of the probabilistic ENSO prediction in the LDEO5 model. *J. Climate* **23**: 5476–5497.

Collins M, Booth BBB, Harris GR, Murphy JM, Sexton DMH, Webb MJ. 2006. Towards quantifying uncertainty in transient climate change. *Clim. Dyn.* **27**: 127–147.

Côté J, Desmarais J-G, Gravel S, Méthot A, Patoine A, Roch M, Staniforth A. 1998a. The operational CMC–MRB Global Environmental Multiscale (GEM) model. Part II: Results. *Mon. Weather Rev.* **126**: 1397–1418.

Côté J, Gravel S, Méthot A, Patoine A, Roch M, Staniforth A. 1998b. The operational CMC–MRB Global Environmental Multiscale (GEM) model. Part I: Design considerations and formulation. *Mon. Weather Rev.* **126**: 1373–1395.

DelSole T. 2007. A Bayesian framework for multimodel regression. *J. Climate* **20**: 2810–2826.

DelSole T, Tippett MK. 2007. Predictability: Recent insights from information theory. *Rev. Geophys.* **45**: RG4002, DOI: 10.1029/2006RG000202.

DelSole T, Yang XS, Tippett MK. 2012. Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Q. J. R. Meteorol. Soc.*, DOI: 10.1002/qj.1961.

Doblas-Reyes FJ, Weisheimer A, Déqué M, Keenlyside N, McVean M, Murphy JM, Rogel P, Smith D, Palmer TN. 2009. Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Q. J. R. Meteorol. Soc.* **135**: 1538–1559.

Eckel FA, Mass CF. 2005. Aspects of effective mesoscale, short-range ensemble forecasting. *Weather and Forecasting* **20**: 328–350.

Epstein ES. 1969. Stochastic dynamic prediction. *Tellus* **21**: 739–759.

Fukunaga K. 1990. *Introduction to Statistical Pattern Recognition*, 2nd edition. Academic Press.

Hagedorn R, Doblas-Reyes FJ, Palmer TN. 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting. I: Basic concept. *Tellus* **57A**: 219–233.

Hoffman RN, Kalnay E. 1983. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus* **35A**: 100–118.

Honda M, Kushnir Y, Nakamura H, Yamane S, Zebiak SE. 2005. Formation, mechanisms, and predictability of the Aleutian–Icelandic Low seesaw in ensemble AGCM simulations. *J. Climate* **18**: 1423–1434.

Hu Z-Z, Huang BH. 2007. The predictive skill and the most predictable pattern in the tropical Atlantic: The effect of ENSO. *Mon. Weather Rev.* **135**: 1786–1806.

Huang BH. 2004. Remotely forced variability in the tropical Atlantic Ocean. *Clim. Dyn.* **23**: 133–152.

Jin F-F, Lin L, Timmermann A, Zhao J. 2007. Ensemble-mean dynamics of the ENSO recharge oscillator under state-dependent stochastic forcing. *Geophys. Res. Lett.* **34**: L03807, DOI: 10.1029/2006GL027372.

Jolliffe IT, Stephenson DB. 2003. *Forecast Verification: A practitioner's guide in atmospheric science*. John Wiley & Sons.

Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetmaa A, Reynolds E, Jenne R, Joseph D. 1996. The NCEP/NCAR 40-year reanalysis project. *Bull. Am. Meteorol. Soc.* **77**: 437–472.

Kang I-S, Lee J-Y, Park C-K. 2004. Potential predictability of summer mean precipitation in a dynamical seasonal prediction system with systematic error correction. *J. Climate* **17**: 834–844.

Kharin VV, Zwiers FW. 2003. Improved seasonal probability forecasts. *J. Climate* **16**: 1684–1701.

Kharin VV, Teng QB, Zwiers FW, Boer GJ, Derome J, Fontecilla JS. 2009. Skill assessment of seasonal hindcasts from the Canadian historical forecast project. *Atmos.–Ocean* **47**: 204–223.

Kirtman BP, Min DH. 2009. Multimodel ensemble ENSO prediction with CCSM and CFS. *Mon. Weather Rev.* **137**: 2908–2930.

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

Krishnamurti TN, Kishtawal CM, Zhang Z, LaRow T, Bachiochi D, Williford E, Gadgil S, Surendran S. 2000. Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate* **13**: 4196–4216.

Kumar TSVV, Krishnamurti TN, Fiorino M, Nagata M. 2003. Multimodel superensemble forecasting of tropical cyclones in the Pacific. *Mon. Weather Rev.* **131**: 574–583.

Liang JY, Yang S, Hu Z-Z, Huang BH, Kumar A, Zhang ZQ. 2009. Predictable patterns of the Asian and Indo-Pacific summer precipitation in the NCEP CFS. *Clim. Dyn.* **32**: 989–1001.

Lin H, Brunet G, Derome J. 2008. Seasonal forecasts of Canadian winter precipitation by postprocessing GCM integrations. *Mon. Weather Rev.* **136**: 769–783.

McFarlane NA, Boer GJ, Blanchet J-P, Lazare M. 1992. The Canadian Climate Centre second-generation general circulation model and its equilibrium climate. *J. Climate* **5**: 1013–1044.

Matsumura S, Huang G, Xie S-P, Yamazaki K. 2010. SST-forced and internal variability of the atmosphere in an ensemble GCM simulation. *J. Meteorol. Soc. Jpn* **88**: 43–62.

Meng X-L, Rosenthal R, Rubin DB. 1992. Comparing correlated correlation coefficients. *Psychol. Bull.* **111**: 172–175.

Murphy AH. 1973. A new vector partition of the probability score. *J. Appl. Meteorol.* **12**: 595–600.

Murphy AH. 1985. Probabilistic weather forecasting. Pp 337–377 in *Probability, Statistics, and Decision Making in the Atmospheric Sciences*. Westview Press.

Mylne KR, Evans RE, Clark RT. 2002. Multi-model multi-analysis ensembles in quasi-operational medium-range forecasting. *Q. J. R. Meteorol. Soc.* **128**: 361–384.

Palmer TN. 2001. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. R. Meteorol. Soc.* **127**: 279–304.

Palmer TN, Shukla J. 2000. Editoral to DSP/PROVOST special issue. *Q. J. R. Meteorol. Soc.* **126**: 1989–1990.

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

Peixoto JP, Oort AH. 1992. *Physics of Climate*. American Institute of Physics: New York.

Peña M, van den Dool H. 2008. Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. *J. Climate* **21**: 6521–6538.

Peng P, Zhang Q, Kumar A, van den Dool H, Wang W, Saha S, Pan H. 2005. 'Variability, predictability and prediction of DJF climate in NCEP Coupled Forecast System (CFS).' *AGU, Spring Meeting*, abstract A43B-03.

Preisendorfer RW, Mobley CD. 1988. *Principal Component Analysis in Meteorology and Oceanography*. *Developments in Atmospheric Science* **17**. Elsevier: Amsterdam.
Richardson DS. 2001. Ensembles using multiple models and analyses. *Q. J. R. Meteorol. Soc.* **127**: 1847–1864.
Richman MB. 1986. Rotation of principal components. *J. Climatol.* **6**: 293–335.
Ritchie H. 1991. Application of the semi-Lagrangian method to a multilevel spectral primitive-equations model. *Q. J. R. Meteorol. Soc.* **117**: 91–106.
Sardeshmukh PD, Sura P. 2007. Multiscale impacts of variable heating in climate. *J. Climate* **20**: 5677–5695.
Schneider T, Griffies SM. 1999. A conceptual framework for predictability studies. *J. Climate* **12**: 3133–3155.
Shutts GJ. 2005. A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.* **131**: 3079–3102.
Shutts GJ, Palmer TN. 2007. Convective forcing fluctuations in a cloud-resolving model: Relevance to the stochastic parameterization. *J. Climate* **20**: 187–202.
Stainforth DA, Aina T, Christensen C, Collins M, Faull N, Frame DJ, Kettleborough JA, Knight S, Martin A, Murphy JM, Piani C, Sexton D, Smith LA, Spicer RA, Thorpe AJ, Allen MR. 2005. Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature* **433**: 403–406.
Stockdale TN, Anderson DLT, Alves JOS, Balmaseda MA. 1998. Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. *Nature* **392**: 370–373.
Straus DM, Shukla J. 2002. Does ENSO force the PNA? *J. Climate* **15**: 2340–2358.
Straus DM, Shukla J, Paolino D, Schubert S, Suarez M, Pegion P, Kumar A. 2003. Predictability of the seasonal mean atmospheric circulation during autumn, winter, and spring. *J. Climate* **16**: 3629–3649.
Sutton RT, Jewson SP, Rowell DP. 2000. The elements of climate variability in the tropical Atlantic region. *J. Climate* **13**: 3261–3284.
Tang YM, Kleeman R, Moore AM. 2005. Reliability of ENSO dynamical predictions. *J. Atmos. Sci.* **62**: 1770–1791.
Tang YM, Lin H, Moore AM. 2008. Measuring the potential predictability of ensemble climate predictions. *J. Geophys. Res.* **113**: D04108, DOI: 10.1029/2007JD008804.
Tippett MK, Giannini A. 2006. Potentially predictable components of African summer rainfall in an SST-forced GCM simulation. *J. Climate* **19**: 3133–3144.
Toth Z, Talagrand O, Candille G, Zhu YJ. 2003. Probability and ensemble forecasts. Pp 137–163 in *Forecast Verification: A practitioner's guide in atmospheric science*, Jolliffe IT, Stephenson DB (eds). John Wiley & Sons, Ltd.
Tracton MS, Kalnay E. 1993. Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Weather and Forecasting* **8**: 379–398.
Venzke S, Allen MR, Sutton RT, Rowell DP. 1999. The atmospheric response over the North Atlantic to decadal changes in sea surface temperature. *J. Climate* **12**: 2562–2584.
Vislocky RL, Fritsch JM. 1995. Improved model output statistics forecasts through model consensus. *Bull. Am. Meteorol. Soc.* **76**: 1157–1164.
Vitart F. 2006. Seasonal forecasting of tropical storm frequency using a multi-model ensemble. *Q. J. R. Meteorol. Soc.* **132**: 647–666.
Wallace JM, Gutzler DS. 1981. Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Weather Rev.* **109**: 784–812.
Wang B, Lee J-Y, Kang I-S, Shukla J, Park C-K, Kumar A, Schemm J, Cocke S, Kug J-S, Luo J-J, Zhou T, Wang B, Fu X, Yun W-T, Alves O, Jin EK, Kinter J, Kirtman B, Krishnamurti T, Lau NC, Lau W, Liu P, Pegion P, Rosati T, Schubert S, Stern W, Suarez M, Yamagata T. 2009. Advance and prospectus of seasonal prediction: Assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). *Clim. Dyn.* **33**: 93–117.
Whitaker JS, Wei X, Vitart F. 2006. Improving week-2 forecasts with multimodel reforecast ensembles. *Mon. Weather Rev.* **134**: 2279–2284.
Wilks D. 2006. *Statistical Methods in the Atmospheric Sciences*, 2nd edition. Academic Press.
Yang D, Tang Y, Zhang Y, Yang X. 2012. Information-Based Potential Predictability of the Asian Summer Monsoon in a Coupled Model, *J. Geophys. Res.*, DOI:10.1029/2011JD016775.
Yasui S, Watanabe M. 2010. Forcing processes of the summertime circumglobal teleconnection pattern in a dry AGCM. *J. Climate* **23**: 2093–2114.
Yuan XJ, Martinson DG. 2000. Antarctic sea ice variability and its global connectivity. *J. Climate* **13**: 1697–1717.
Yuan XJ, Martinson DG. 2001. The Antarctic dipole and its predictability. *Geophys. Res. Lett.* **28**: 3609–3612.

