AGU PUBLICATIONS

JGR

Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1002/2015JD023781

Key Points:
- Probabilistic skill increase of MME over SME is more significant than the deterministic one
- Improved reliability and reduced overconfidence lead to the superiority of MME over SMEs
- A monotonic relationship between probabilistic resolution and correlation skill is found

Correspondence to:
X.-Q. Yang,
xqyang@nju.edu.cn

Citation:
Yang, D., X.-Q. Yang, Q. Xie, Y. Zhang, X. Ren, and Y. Tang (2016), Probabilistic versus deterministic skill in predicting the western North Pacific-East Asian summer monsoon variability with multimodel ensembles, *J. Geophys. Res. Atmos.*, *121*, 1079–1103, doi:10.1002/2015JD023781.

Received 10 JUN 2015
Accepted 31 DEC 2015
Accepted article online 5 JAN 2016
Published online 5 FEB 2016

Probabilistic versus deterministic skill in predicting the western North Pacific-East Asian summer monsoon variability with multimodel ensembles

Dejian Yang[1], Xiu-Qun Yang[1], Qian Xie[2], Yaocun Zhang[1], Xuejuan Ren[1], and Youmin Tang[3]

[1]CMA-NJU Joint Laboratory for Climate Prediction Studies, Institute for Climate and Global Change Research, School of Atmospheric Sciences, Nanjing University, Nanjing, China, [2]College of Meteorology and Oceanography, PLA University of Science and Technology, Nanjing, China, [3]Environmental Science and Engineering, University of Northern British Columbia, Prince George, British Columbia, Canada

Abstract Based on historical forecasts of three quasi-operational multimodel ensemble (MME) systems, this study assesses the superiority of coupled MME over contributing single-model ensembles (SMEs) and over uncoupled atmospheric MME in predicting the Western North Pacific-East Asian summer monsoon variability. The probabilistic and deterministic forecast skills are measured by Brier skill score (BSS) and anomaly correlation (AC), respectively. A forecast-format-dependent MME superiority over SMEs is found. The probabilistic forecast skill of the MME is always significantly better than that of each SME, while the deterministic forecast skill of the MME can be lower than that of some SMEs. The MME superiority arises from both the model diversity and the ensemble size increase in the tropics, and primarily from the ensemble size increase in the subtropics. The BSS is composed of reliability and resolution, two attributes characterizing probabilistic forecast skill. The probabilistic skill increase of the MME is dominated by the dramatic improvement in reliability, while resolution is not always improved, similar to AC. A monotonic resolution-AC relationship is further found and qualitatively explained, whereas little relationship can be identified between reliability and AC. It is argued that the MME's success in improving the reliability arises from an effective reduction of the overconfidence in forecast distributions. Moreover, it is examined that the seasonal predictions with coupled MME are more skillful than those with the uncoupled atmospheric MME forced by persisting sea surface temperature (SST) anomalies, since the coupled MME has better predicted the SST anomaly evolution in three key regions.

1. Introduction

The western North Pacific-East Asian summer monsoon (WNP-EASM) is a unique component of the grand Asian summer monsoon system [e.g., *Wang et al.*, 2001; *Wang and LinHo*, 2002]. Its interannual variability is often associated with floods, droughts, and other natural disasters that can critically influence human lives and economics over East and Southeast Asia. Thus, improving prediction of the WNP-EASM variability several months in advance is extremely important for decision making and risk management.

In the last decade, a variety of efforts have been made to develop complex general circulation models (GCMs) to predict the seasonal climate variability. Usually, there are two types of approaches for dynamical seasonal predictions, i.e., the so-called one-tier and two-tier approaches [e.g., *Barnston et al.*, 2003; *Kharin et al.*, 2009; *Palmer et al.*, 2004; *Weisheimer et al.*, 2009; *Saha et al.*, 2006, 2014; *Merryfield et al.*, 2013]. In the two-tier approach, stand-alone atmospheric general circulation models are integrated forward from observed atmospheric initial conditions while forced externally by a prescribed sea surface temperature (SST) that is from the persistence forecast, statistical models, or even dynamical oceanic models. In the one-tier approach, on the other hand, atmospheric and oceanic variables are incorporated into the coupled GCM and evolve together with time from the given initial states of both atmosphere and ocean. Since the observed climate signals and variability are usually governed in a framework of air-sea interaction [e.g., *Bjerknes*, 1969; *Rasmusson and Carpenter*, 1982; *Wang et al.*, 2000, 2003; *Rodwell and Folland*, 2002], it is believed to be a better way to simulate and predict the climate variability with coupled GCM [e.g., *Palmer et al.*, 2004; *Saha et al.*, 2006].

Seasonal climate prediction, in either one- or two-tiered approach, is inevitably subject to many error sources that can be generally grouped into two families: the uncertainties in initial conditions and the uncertainties in

model formulations. To quantify forecast uncertainties arising from these sources, different strategies have been proposed. To cope with the initial condition uncertainties, a straightforward and well-acknowledged strategy in seasonal predictions is to repeatedly integrate a model forward from multiple initial conditions that differ only slightly [e.g., *Palmer et al.*, 2004]. It is necessary to use this strategy to predict the WNP-EASM variability, since the atmospheric internal dynamics that is sensitive to atmospheric initial state plays an important role in generating seasonal climate anomalies over the western North Pacific-East Asian region [e.g., *Li et al.*, 2012; *Kosaka et al.*, 2012, 2013]. To address the model uncertainties, a multimodel ensemble (MME) strategy has been proposed since the end of last century [e.g., *Krishnamurti*, 1999; *Krishnamurti et al.*, 2000; *Doblas-Reyes et al.*, 2000; *Shukla et al.*, 2000; *Palmer et al.*, 2000; *Barnston et al.*, 2003]. The central argument of the MME idea lies in that through combining several quasi-independent single models, the errors may cancel out each other and the resulting ensemble could have a better coverage of the possible climate phase space [e.g., *Hagedorn et al.*, 2005]. As discussed in the literature [e.g., *Kang et al.*, 2002; *Chen et al.*, 2010; *Sperber et al.*, 2012], it is, in general, a challenging task to simulate and predict the Asian summer monsoon, since there are large uncertainties in the parameterizations of various physical processes associated with the monsoon. Therefore, it is expected that the MME strategy can effectively improve the forecast skill of the Asian summer monsoon.

Many previous studies have assessed the deterministic skills of different MME systems in historical predictions of the WNP-EASM variability [e.g., *Kang and Shukla*, 2006; *Wang et al.*, 2008a; *Chowdary et al.*, 2010; *Lee et al.*, 2011, 2013, 2014; *Li et al.*, 2012; *Tang et al.*, 2013; *Min et al.*, 2014; *Ma et al.*, 2015]. The results from these studies indeed indicate that some benefit could be received from compositing single-model ensembles (SMEs), but this benefit is sometimes cast into doubt by the fact that the MME does not always beat the best SME [*Kang and Shukla*, 2006; *Lee et al.*, 2011]. However, besides the deterministic manner, the forecasts can be made in a probabilistic way. A forecast in the form of probability can provide an indication of the likelihood for the event of interest to occur, a very useful message that is able to bring greater economic value for end-users than a single deterministic forecast with uncertain accuracy [e.g., *Richardson*, 2006]. It is worth noting that some studies have reported elsewhere that the MME seems to offer a more significant improvement when verified in a probabilistic way than in a deterministic way [e.g., *Doblas-Reyes et al.*, 2000]. For a full assessment and especially to unfold the MME's effectiveness, a probabilistic assessment is necessary.

Another important issue in dynamical seasonal prediction is the choice of one-tier coupled GCMs or two-tier uncoupled atmospheric GCMs. Although the general potential advantages of coupled GCMs are widely believed, their actual advantages could be compromised by their potentially larger systematic errors and higher computational expense. Therefore, the two-tier strategy is still used in several operational centers, especially for short-lead seasonal predictions. However, recent experimental studies have highlighted the necessity of using fully coupled GCMs in simulating and predicting the summer monsoon over the western North Pacific where local air-sea coupled processes are important in generating seasonal climate anomalies [e.g., *Wang et al.*, 2005; *Kug et al.*, 2008; *Chen et al.*, 2012; *Zhu and Shukla*, 2013]. Thus, it remains necessary to investigate the practical advantages of the current one-tier over two-tier operational MME forecasting systems, especially in terms of deterministic versus probabilistic skill measure.

In this study, we aim at exploring the superiorities of the one-tier coupled MME over each contributing SME as well as over the two-tier uncoupled atmospheric MME in predicting the WNP-EASM variability with probabilistic versus deterministic skill measure, using three quasi-operational ensemble forecast products. The paper is structured as follows. Section 2 describes the data used and the metrics of prediction skill. In section 3, the general features of coupled MME skills in predicting the WNP-EASM variability are assessed, the prediction skills of the MME and the SMEs are compared in detail, and the superiorities of MME over contributing SMEs are identified. An understanding of the MME' superiorities is presented in section 4. Section 5 compares the prediction skills between the coupled and uncoupled MMEs. The final section is devoted to conclusions and discussion.

2. Data and Method

2.1. Data

The data sets used in this study are taken from three quasi-operational ensemble forecast products, including the European one-tier Ensemble-Based Predictions of Climate Changes and their Impacts (ENSEMBLES) ensemble, the European one-tier Development of a European Multimodel Ensemble System for Seasonal to

Interannual Prediction (DEMETER) ensemble, and the Canadian two-tier Historical Seasonal Forecasting Project (Phase-2; HFP2) ensemble. The ENSEMBLES project involves five global coupled climate models from the UK Met Office (UKMO), Météo France (MF), the European Centre for Medium-Range Weather Forecasts (ECMWF), the Leibniz Institute of Marine Sciences at Kiel University (IFM-GEOMAR), and the Euro-Mediterranean Centre for Climate Change (CMCC-INGV) in Bologna. Compared to the previous version of the DEMETER project, the models involved in the ENSEMBLES project have been improved mainly in physical process parameterizations, initialization procedures, and resolution. The historical forecasts with the five models were performed for the 46-year period from 1960 to 2005. For each year, seasonal forecasts at 7 month lead were produced, starting on the first day of February, May, August, and November, respectively. Except for the CMCC-INGV model, the forecasts starting on 1 November with all the other four models were additionally extended with 14 month lead time. For each model, an ensemble of nine integrations starting from different initial conditions of atmosphere and ocean was generated for a forecast. A combination of these 9-member SMEs gives rise to a 45-member MME. A detailed description about the ENSMEBLES's models and historical forecasts can be found in *Weisheimer et al.* [2009].

The DEMETER project, described in detail in *Palmer et al.* [2004], involves seven global coupled models, of which five models are just the previous versions of the ENSEMBLES's models. As in *Alessandri et al.* [2011], the ensemble of these five models is chosen to be compared with the ENSEMBLES's MME. The historical forecasts with these five models were produced with 6 month lead time, starting yearly on 1 February, May, August, and November, respectively, for a common period of 1973–2001. Each SME has 9 members and the resulting MME has 45 members.

The HFP2 includes four Canadian global atmospheric models, which are the second and third generations of the general circulation models developed at the Canadian Centre for Climate Modeling and Analysis [*Boer et al.*, 1984; *McFarlane et al.*, 1992], the Global Environmental Multi-scale model [*Côté et al.*, 1998a, 1998b], and the reduced-resolution version of the global spectral model [*Ritchie*, 1991]. For each model, starting on the first day of each month and with different atmospheric initial conditions, 10 parallel integrations were run for 4 months, for the period from 1969 to 2001. As a result, the HFP2 MME has 40 ensemble members. The SST forcing for all the four atmospheric models is identical, specified as a persisting SST forcing, i.e., the sum of the monthly mean SST anomaly of the month just preceding the forecast period and the monthly varying climatological SST for the forecast period. Further details of the HFP2 setup can be found in *Kharin et al.* [2009].

The ENSEMBLES data set is only used in sections 3 and 4 to examine and understand the advantages of MME over the contributing SME, because it possesses relatively new models and the longest forecast lead time among the three MME suites. The DEMETER and HFP2 data sets are only used in section 5, where the comparison between the coupled and uncoupled MMEs is in particular concerned.

The verification data sets used in this study include the monthly mean 850 hPa zonal wind from the National Centers for Environmental Prediction/National Center for Atmospheric Research reanalysis [*Kalnay et al.*, 1996], monthly mean precipitation from the Climate Prediction Center merged analysis of precipitation [*Xie and Arkin*, 1996], and monthly mean SST from NOAA Extended Reconstructed SST (v3) [*Smith and Reynolds*, 2004].

To characterize the summer monsoon, seasonally averaged variables for June, July, and August (JJA) are targeted. The ensemble predictions starting on 1 May, 1 February, and 1 November in the preceding year are called one-, four-, and seven-month lead predictions, respectively. In addition, although a variety of monsoon indices were proposed to effectively extract the leading modes of the monsoon variability, they might not be the best choice to demonstrate the MME's advantages in predicting the variability, since those large-scale coherent modes are more probably able to "resist" the interference by noises and model errors. For instance, when the WNP-EASM index defined in *Wang et al.* [2008b] is used as the forecast target, the MME cannot defeat the best SME in deterministic or probabilistic prediction. Thus, the grid-point monsoon variables of 850 hPa zonal wind and precipitation are directly chosen as the prediction targets in this study.

2.2. Metrics of Prediction Skill

Conventional metrics used to measure forecast skill includes the anomaly correlation (AC) for deterministic forecasts and Brier skill score (BSS) for probabilistic forecasts. At each grid point both the single-model hindcasts and observations are normalized with respect to their own local climatologies. Then the resulting standardized anomalies are used to calculate forecast skills. To avoid overfitting, the skills are calculated in a cross-validated mode using the so-called leave-one-out method; i.e., the anomaly for a certain year is obtained with

respect to the climatological mean that is evaluated only over the remaining years. For the MME, forecasts are built on the grand ensemble of the cross-validated single-model standardized anomalies. All the single-model trajectories are assembled indistinguishably. For probabilistic forecast, three target events: below-normal, near-normal, and above-normal, are defined based on the terciles of the observed climatology. The forecast probability of an event is estimated as the fraction of ensemble members that forecast the event.

AC is a measure of the linear association between prediction and observation, formulated as

$$\mathrm{AC} = \frac{\frac{1}{N}\sum_{j=1}^{N} x_j^p x_j^o}{\sqrt{\frac{1}{N}\sum_{j=1}^{N} x_j^p x_j^p}\sqrt{\frac{1}{N}\sum_{j=1}^{N} x_j^o x_j^o}}, \tag{1}$$

where x_j^p denotes the predicted anomaly based on the ensemble mean, x_j^o denotes the corresponding observed anomaly, and N is the size of the samples. As a skill measure, AC is positively oriented and negative AC is interpreted as worse than climatology. The summation in equation (1) can be over time-only to calculate a local skill or over both time and space to calculate an "area-aggregated" skill [e.g., *Déqué*, 1997; *Saha et al.*, 2006]. In this study, the two types of skills are both calculated.

The definition of BSS is based on the Brier score (BS), which measures the accuracy of probability forecasts for a predefined event, as expressed by [e.g., *Wilks*, 2011]

$$\mathrm{BS} = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - o_j\right)^2, \tag{2}$$

where y_j denotes the forecast probability; o_j represents the corresponding observed outcome for the event, taking the value 1 if the event actually occurs and 0 otherwise; and N is the sample size. As a mean square error measure, the BS is negatively oriented; the larger, the worse. BS can be further decomposed into three terms, known as reliability, resolution, and uncertainty, which can be expressed as [e.g., *Wilks*, 2011]

$$\mathrm{BS} = \underbrace{\frac{1}{N}\sum_{k=1}^{K} N_k\left(y_k - \overline{o}_k\right)^2}_{\text{reliability}} - \underbrace{\frac{1}{N}\sum_{k=1}^{K} N_k\left(\overline{o}_k - \overline{o}\right)^2}_{\text{resolution}} + \underbrace{\overline{o}\left(1-\overline{o}\right)}_{\text{uncertainty}}, \tag{3}$$

$$\equiv \mathrm{BS_{REL}} - \mathrm{BS_{RES}} + \mathrm{BS_{UNC}}$$

where the probability space is partitioned into K bins, N_k is the number of probability forecasts collected into the kth probability bin, y_k represents the forecast probability for the bin k, $\overline{o}_k$ is the conditional observed frequency associated with the forecast probability y_k, and $\bar{o}$ is the unconditional observed frequency, or climatological probability of the event.

Reliability quantifies how well the forecast probabilities are consistent with the corresponding observed frequencies. For example, a probability forecast of 0.7 is called reliable if and only if in all cases where the issued probability forecasts for an event are 0.7, this event indeed occurs in 70% of these cases. A forecast system is reliable if and only if all the forecast probabilities are reliable. However, reliability alone is not sufficient for a probabilistic forecast system to be skillful. Apparently, always using the historical climatological probability as forecasts would in principle result in a perfect reliability. However, these forecasts provide no extra information beyond the historical climatological information. Resolution is just such a measure of the extra information, through quantifying the variability of observed frequencies around the climatological probability. The term of uncertainty just indicates the "a priori" information provided by the climatology. Actually, the BS of climatological forecasts turns out to be equal to the uncertainty, i.e., $\mathrm{BS_{clim}} = \mathrm{BS_{UNC}}$.

Based on the BS and taking the climatology as the reference forecast, BSS, as a relative measure of probabilistic skill, can be uniquely defined as [e.g., *Wilks*, 2011]

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS_{clim}}}. \tag{4}$$

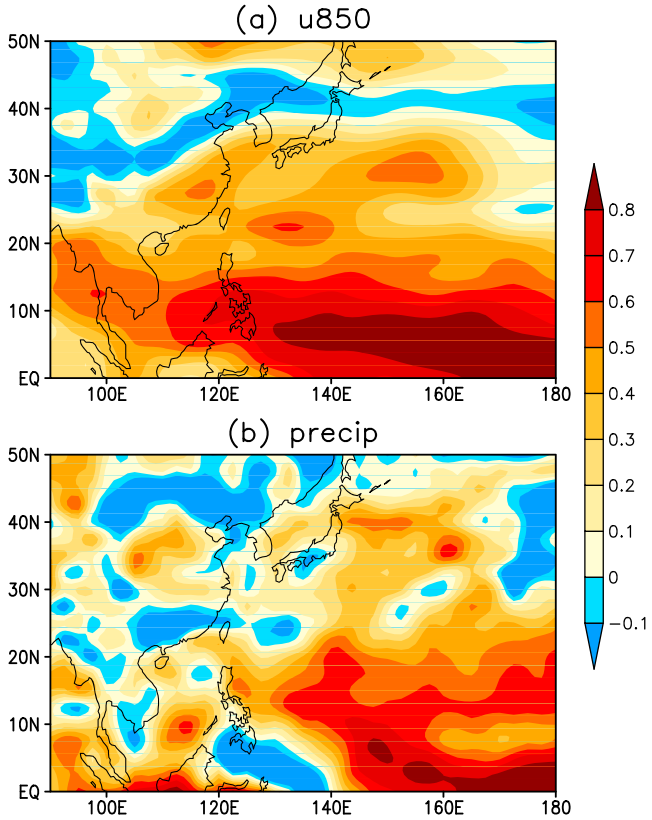

Figure 1. Spatial distributions of the anomaly correlation (AC) skill for the ENSEMBLES's MME predictions at one-month lead of JJA (a) u850 anomaly (1960–2005) and (b) precipitation anomaly (1979–2005).

Unlike the BS, BSS is positively oriented and contains a built-in comparison with the climatological forecast. Zero BSS indicates a skill equivalent to a climatological forecast, and positive (negative) BSS means better (worse) than climatology. After invoking equation (3) and $BS_{clim} = BS_{UNC}$, BSS can be further expressed as [*Kharin and Zwiers*, 2003]

$$BSS = \frac{BS_{RES}}{BS_{UNC}} - \frac{BS_{REL}}{BS_{UNC}} \equiv BSS_{RES} - BSS_{REL}. \quad (5)$$

As in *Kharin and Zwiers* [2003], we refer to the "standardized" reliability and resolution terms of the BS as the reliability and resolution components of the BSS, which are diagnosed in this study. For the tercile-based categorical events considered here, BSS_{REL} and BS_{REL} (also, BSS_{RES} and BS_{RES}) only differ by a factor of 4.5. Only the forecasts with BSS_{RES} greater than BSS_{REL} are better than the climatological forecasts. The concepts of reliability and resolution are not only restricted within the framework of BSS. Actually, they are two basic and independent attributes of any probabilistic forecast system [e.g., *Toth et al.*, 2006]. BSS_{REL} (BS_{REL}) and BSS_{RES} (BS_{RES}) are only specific measures of reliability and resolution, and the former is negatively oriented. BSS is often preferred to quantify the probabilistic forecast skill because it is an integrated measure of both reliability and resolution.

Similar to AC, area-aggregated BSS and its components are also calculated in this study. To avoid the possible fictitious skill arising from the spatially varying climatology, the very categorical event for which the forecast probabilities and the corresponding observed binary outcomes are spatially aggregated is still defined for each grid cell based on respective long-term climatologies [*Hamill and Juras*, 2006].

3. Superiorities of MME Over Contributing SMEs

3.1. General Features of MME Forecast Skills

Figure 1 displays the spatial distributions of the AC skills of the ENSEMBLES's MME for the JJA 850 hPa zonal wind (u850) and precipitation predictions at one-month lead over the WNP-EASM region. The spatial pattern of the AC skills for the u850 prediction, as shown in Figure 1a, is characterized roughly by two zonally elongated belts located in the tropics and subtropics, respectively, bounded around 20°N. The tropical skills are overwhelmingly stronger than the subtropical skills. The strong tropical skill is primarily determined by more impact of tropical SST variability and less impact of internal atmospheric variability [e.g., *Charney and Shukla*, 1981; *Shukla*, 1998]. The tropical SST forcing exists not only in the remote El Niño–Southern Oscillation (ENSO) region but also in neighboring regions of the western North Pacific [e.g., *Wang et al.*, 2000] and North Indian Ocean [e.g., *Xie et al.*, 2009]. The forcing in the latter regions tends to play a more immediate role in regulating the WNP-EASM variability [e.g., *Wang et al.*, 2000; *Xie et al.*, 2009]. Such a forcing could be partly linked to the ENSO and partly resulted from the local air-sea interaction in the Indo-western Pacific that is independent of the remote ENSO [e.g., *Wang et al.*, 2013; *Kosaka et al.*, 2013].

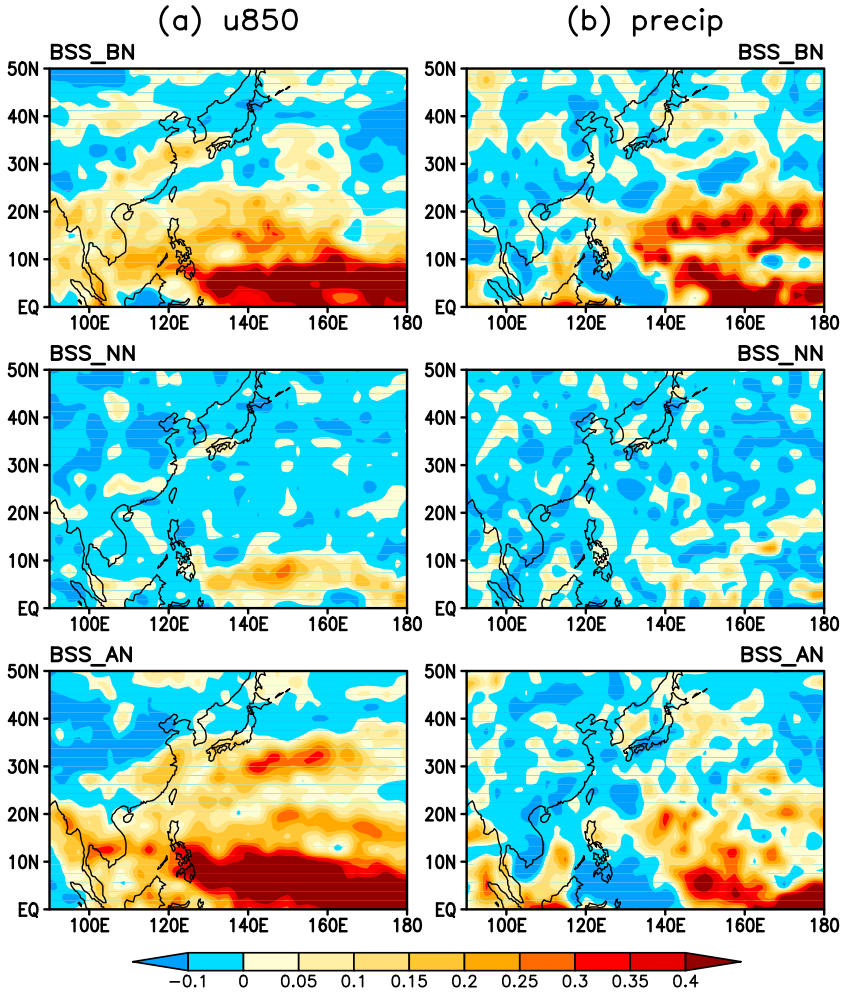

Figure 2. Spatial distributions of the Brier skill score (BSS) for the ENSEMBLES's MME predictions at one-month lead of JJA (a) u850 anomaly (1960–2005) and (b) precipitation anomaly (1979–2005). The subplots in the top, middle, and bottom plots are for the below-normal (BN), near-normal, and above-normal (AN) categorical events, respectively.

The AC skills for the precipitation prediction (Figure 1b) exhibit a different spatial pattern from that for the u850 prediction, with large skills along the equator and to the east of the Philippines and relatively weak skills over the subtropics and midlatitudes. Like the u850 prediction, the precipitation prediction skill is also much better in the tropics than in the subtropics. However, the precipitation prediction skill is not as good as that of the u850, which would be probably due to some very localized or small-scale processes that generate those precipitation variabilities.

Figure 2 shows the BSS skills of the ENSEMBLES MME for the JJA u850 and precipitation predictions at one-month lead over the WNP-EASM region. As can be seen, the BSS skill is very low for near-normal events. For the above- and below-normal events, the BSS skills are geographically sensitive for both the u850 and the precipitation, with spatial patterns similar to those of the corresponding AC skills shown in Figure 1. The locations of large BSS basically match those of large AC. This AC-BSS resemblance in spatial pattern was also noticed in *Wang et al.* [2009] for the seasonal predictions of global temperature and precipitation using the Climate Prediction and its Application to Society MME system. We will find in the next section that this spatial resemblance is mainly due to the high coherence between AC and BSS's resolution component (rather than its reliability component). The reliability component of BSS is much more uniformly distributed within the region of interest. A more detailed discussion on the similarity between AC and resolution is provided in section 4.2.

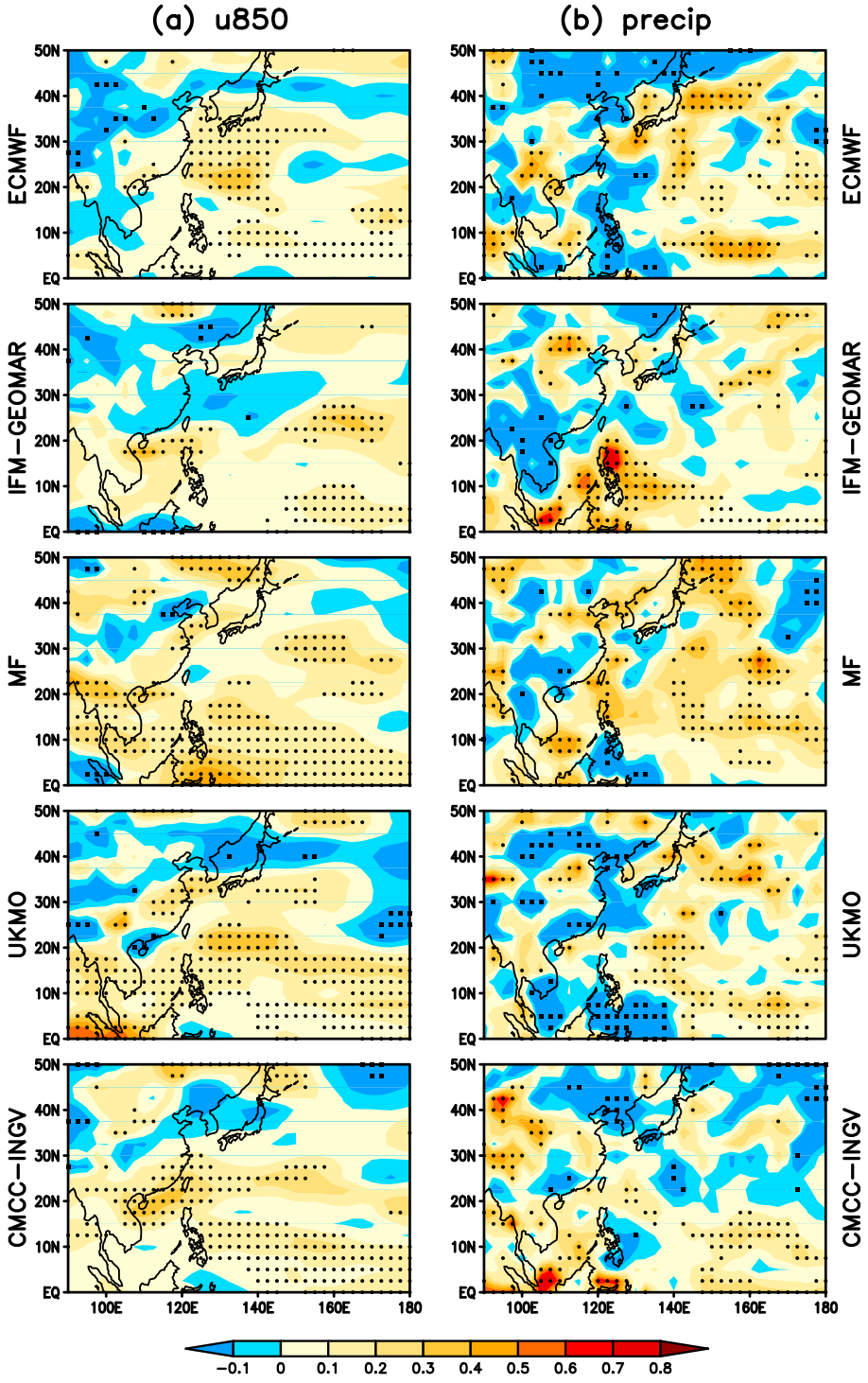

Figure 3. Differences in the AC skill between the MME and each of the SMEs for the ENSEMBLES's predictions at one-month lead of JJA (a) u850 anomaly (1960–2005) and (b) precipitation anomaly (1979–2005). Dots (squares) indicate that the AC skill differences are significantly positive (negative) at the 10% significance level with a two-sided bootstrap significance test.

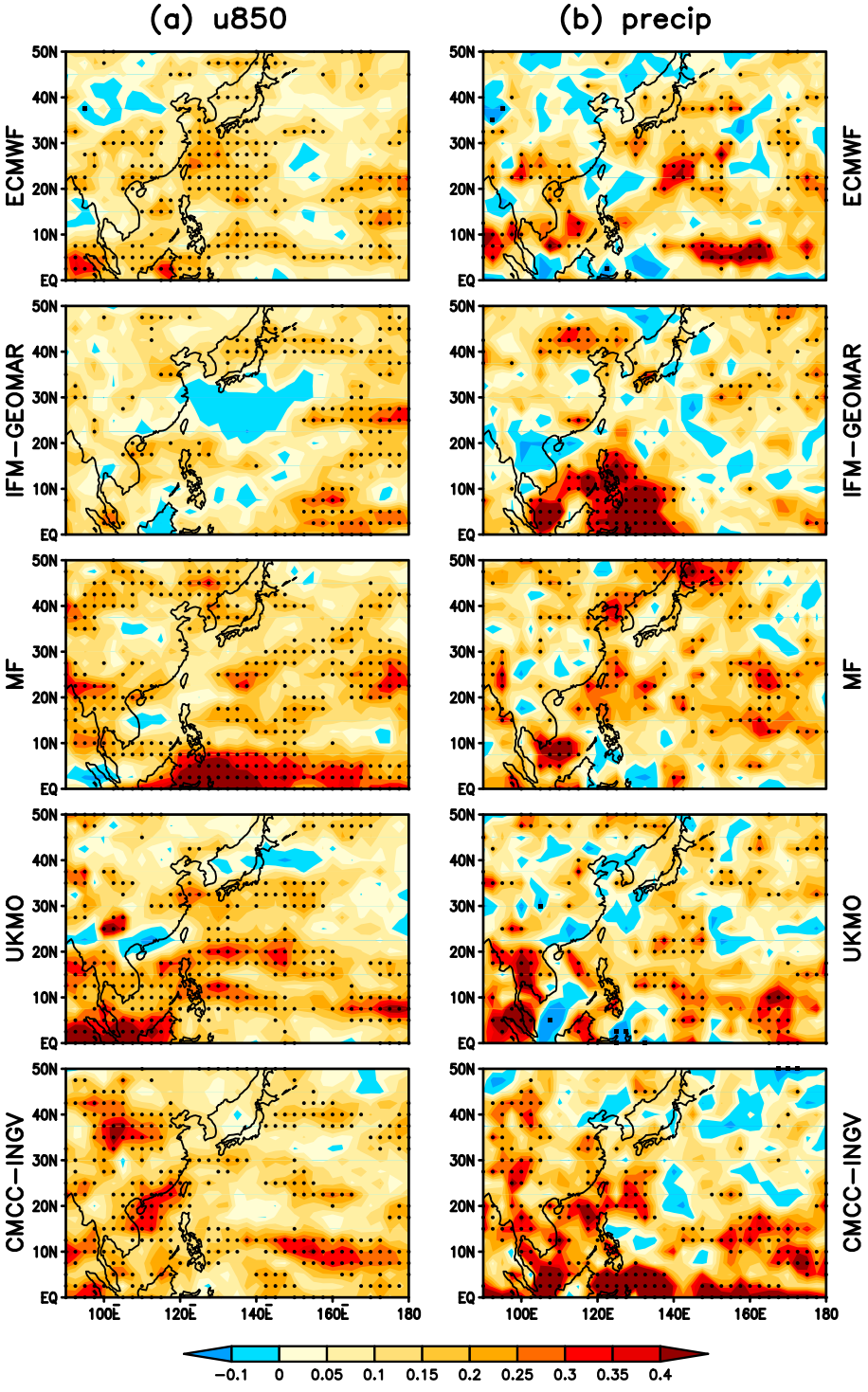

Figure 4. As in Figure 3 but for BSS. The results shown here are for an average over the above- and below-normal events.

Also interestingly, there is an asymmetry between the BSS skills for the above-normal and below-normal events. It might be due to the asymmetry in the SST anomalies or the asymmetry in the atmospheric response to the SST anomalies [*Kharin and Zwiers*, 2003].

3.2. Skill Difference Between MME and Each of the SMEs

In this section, the prediction skills are in detail compared between MME and each contributing SME. For simplicity, the results shown here are only for an average over the above- and below-normal events.

Figures 3 and 4 show the skill differences of the u850 and precipitation predictions at one-month lead between MME and each of the contributing SMEs for AC and BSS, respectively. A two-sided bootstrap test as introduced in *Hogan and Mason* [2011] is used for statistical significance test. Since no data from overlapping or adjacent seasons are pooled and therefore strong serial correlation seems unlikely, the usage of the simple bootstrapping method without addressing the effects of serial correlation is appropriate. The significance at a 10% level is marked in Figures 3 and 4. As seen in Figure 3, statistically significantly positive AC difference occupies large areas for each single model for both u850 and precipitation. However, there are also certain areas where the difference appears significantly negative, signifying that the MME prediction with AC skill measure is worse than the predictions of some SMEs. This feature is especially prominent for the precipitation prediction, although it also occurs in subtropical regions for the u850. In sharp contrast to AC skill differences, the negative BSS differences, as displayed in Figure 4, are basically not statistically significant at the 10% level, indicating the superiority of MME over each of the contributing SMEs in probabilistic prediction.

The prediction skills of the MME and the SMEs are further compared in the area-aggregated way. For this purpose, we aggregated the skills for two domains: the tropical domain (90°–160°E, 0°–20°N) and the subtropical domain (90°–160°E, 20°–50°N). Due to the more samples used, the area-aggregated skills would be less affected by sampling error and statistically much more robust than the grid-point skills. As seen in Figure 5, in the area-aggregated way, the MME outperforms all the SMEs not only in BSS but also in AC for the one-month lead predictions of both precipitation and u850 over both regions. However, the BSS witnesses more dramatic improvement than the AC. Most strikingly, the negative BSS skills for the precipitation as well as the subtropical u850 measured in each of the contributing SMEs are avoided and even upgraded to a level that is positive in the MME. However, this profound turnabout of forecast quality does not occur for the AC. In addition, even for the prediction of the tropical u850 for which most of the SMEs already have a positive BSS, the BSS improvement is still more significant. In this case, a calculation of relative improvement by $[(SS_{SME} - SS_{MME})/SS_{MME} \times 100]\,\%$ (where SS represents positively oriented skill score) reveals that the BSS obtains a much larger relative improvement (up to 80%) than the AC (about 18%) on average.

The merit of the MME in BSS skill is also seen in the predictions of u850 and precipitation at four- and seven-month leads, as shown in Figure 6. In Figure 6, the BSSs of the SMEs are mostly below zero and many of them are severely negative, lower than −0.05 and even −0.1. The averaged BSS skills of SMEs are negative even for the predictions of the tropical u850. However, the BSS skills of the MME are much better, either far exceeding or not so worse than the "zero-skill" of the climatological forecasts. The MME probabilistic forecasts are not only more skillful than the SMEs on average but also than the best SME in all cases. In contrast, the MME is not much beneficial to AC skill. For example, the MME cannot beat the best SME for the four (seven)-month lead prediction of the tropical (subtropical) precipitation.

3.3. Role of Model Diversity and Ensemble Size

It has been reported that the MME superiorities in seasonal prediction could be from the use of different models which sample uncertainties in model formulations as well as from the increase of ensemble size which samples uncertainties in initial conditions, as compared to the SMEs [e.g., *Doblas-Reyes et al.*, 2000; *Kharin et al.*, 2009; *Yan and Tang*, 2013; *DelSole et al.*, 2014; *Becker et al.*, 2015]. It is interesting to examine whether the observed MME advantages here arise at least in part from model diversity or are purely a result of the increased ensemble size. For this purpose, we calculated the forecast skills of a size-reduced MME consisting of only nine members, which are drawn from the SMEs in the manner described as follows. First, we randomly select one member from each of the five contributing SMEs. Second, we randomly select four SMEs without replacement from the five SMEs and then randomly select one member from each of these four SMEs. These size-reduced MME skills are also plotted in Figures 5 and 6. It can be seen that for the tropical predictions, the size-reduced MME's prediction skills still outperform the SMEs' ones on average, while the

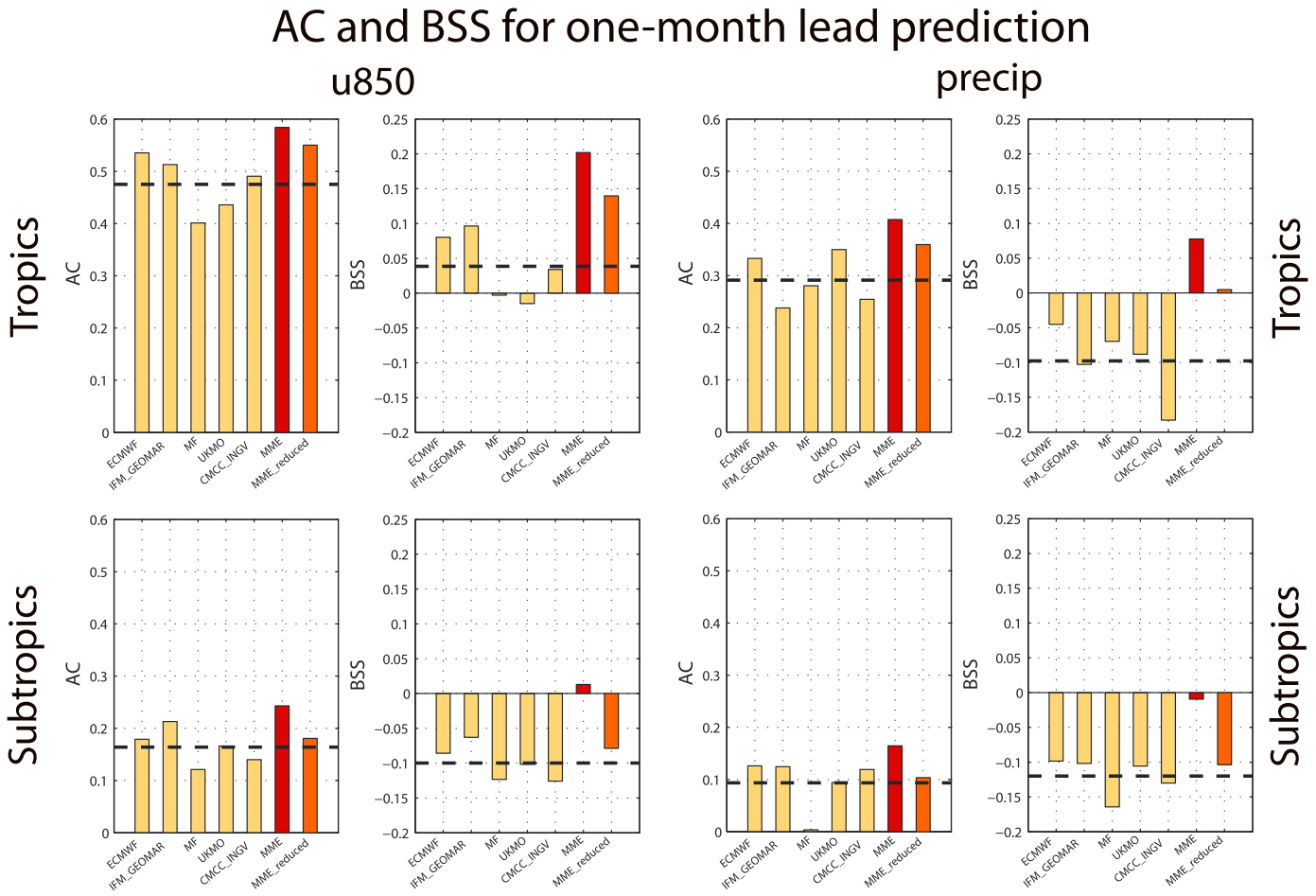

Figure 5. Area-aggregated AC and BSS skills for the ENSEMBLES's predictions at one-month lead of JJA u850 anomaly (1960–2005, left two panels) and precipitation anomaly (1979–2005, right two panels) over the tropical domain (90°–160°E, 0°–20°N; top row) and the subtropical domain (90°–160°E, 20°–50°N; bottom row). The light red bars represent the skills of the size-reduced MME (see main text). Thick dashed lines show the average of all SMEs' skills. The probabilistic skills shown here are for an average over the above- and below-normal events.

size reduction indeed degrades the skills of the original MME, in terms of both AC and BSS measures. In addition, in terms of the BSS measure, the size-reduced MME's prediction skills even outperform the best SME. This result evidences that over the tropical domain of the WNP-EASM, the MME's superiority does benefit from the model diversity rather than only from the increased ensemble size. For the subtropical predictions, however, the MME skills are degraded to a level that is basically not better than the skills of the SMEs, no matter which skill measure is considered, indicating that the MME improvements are mainly by the augmented ensemble size. Based on an investigation of the HFP2 MME skill in predicting the global 500 hPa geopotential height, *Yan and Tang* [2013] also found that the MME improvements are mainly due to the increase of ensemble size in the middle-high latitudes and the offsets of model uncertainties in the tropical regions.

In summary, the MME's versus SMEs' results indicate that the MME skill improvement is strongly forecast-format-dependent. As diagnosed by the BSS measure, the probabilistic forecast skill is improved very impressively and systematically. However, the deterministic forecast skill, measured by AC, is on the whole only marginally and not as systematically improved. An important reason for the MME's probabilistic forecast skill improvement comes from the model diversity, which is in particular significant in the tropical domain of the WNP-EASM.

4. Understanding the Superiorities of MME Over SMEs

4.1. Reliability Versus Resolution Skill

In this section, we try to understand why the MME skill has a forecast-format-dependent improvement, that is, why the probabilistic rather than deterministic forecast skill can be improved significantly. As described in

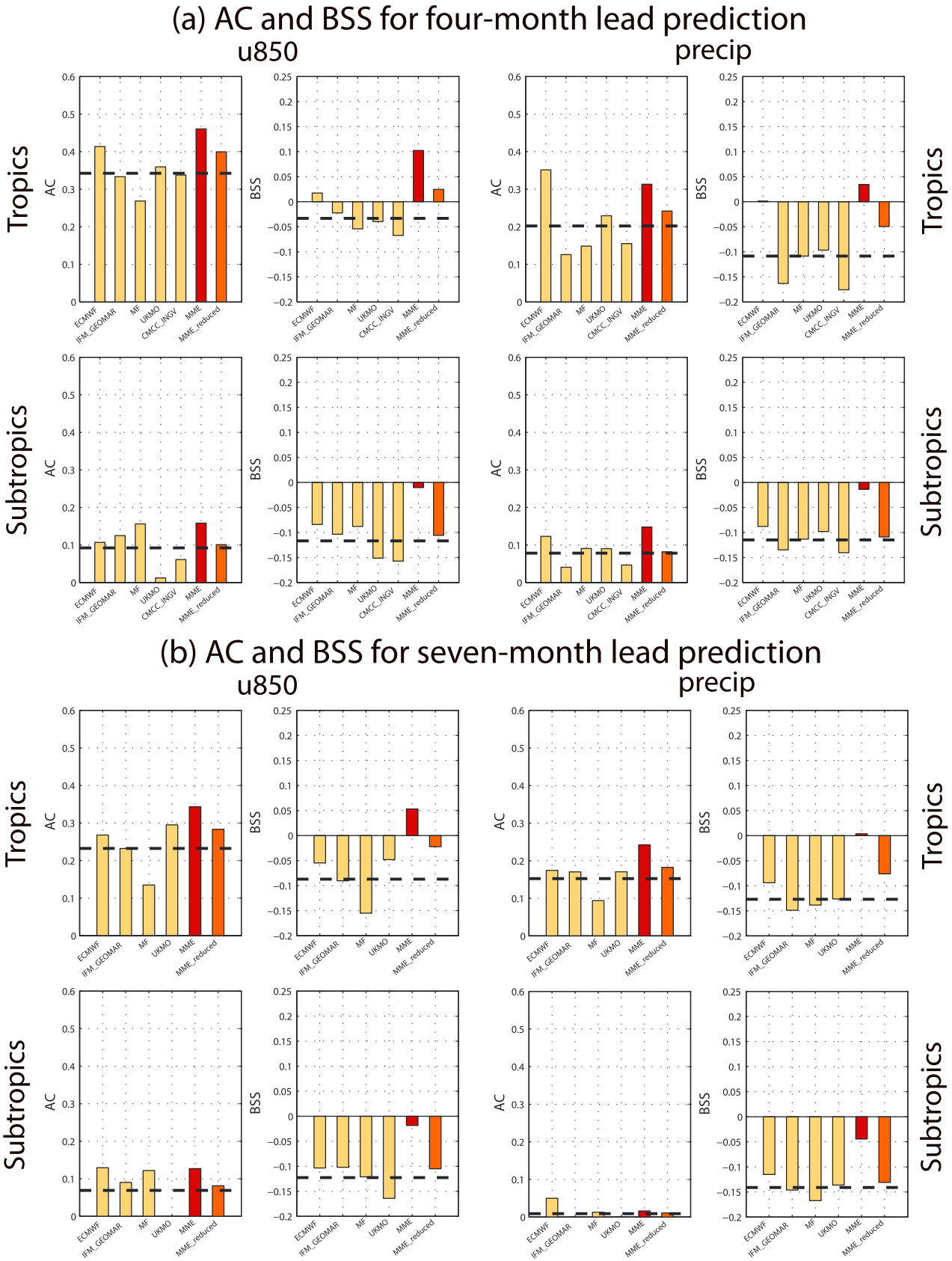

Figure 6. As in Figure 5 but for the predictions at (a) four-month lead and (b) seven-month lead.

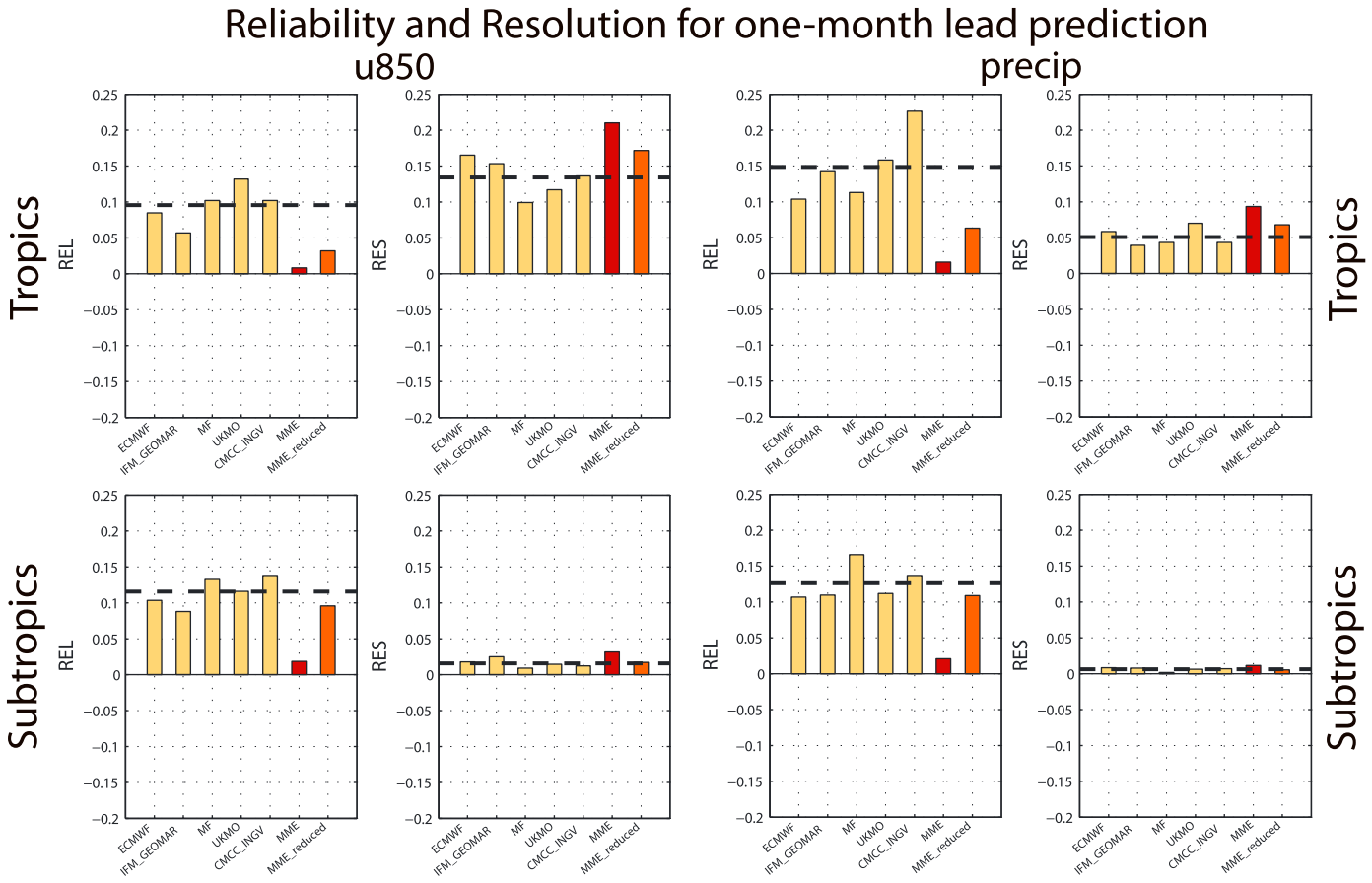

Figure 7. Reliability (REL) and resolution (RES) components of the area-aggregated BSS for the ENSEMBLES predictions at one-month lead of JJA u850 anomaly (1960–2005) and precipitation anomaly (1979–2005) over the tropical domain (90°–160°E, 0°–20°N; top row) and the subtropical domain (90°–160°E, 20°–50°N; bottom row). The light red bars represent the skills of the size-reduced MME (see main text). Thick dashed lines show the average of all SMEs' skills. The probabilistic skills shown here are for an average over the above- and below-normal events.

section 2.2, the BSS is composed of reliability and resolution. Here we examine how the two components themselves vary from the SMEs to the MME, ultimately giving rise to the significant improvement in the BSS.

Figure 7 shows the reliability and resolution components of the BSS for the predictions of u850 and precipitation at one-month lead. Both reliability and resolution are improved in the MME relative to any SME. However, the reliability has a much more significant improvement than the resolution, especially for the predictions of the precipitation and the subtropical u850. As a result, the considerable BSS improvement seen in Figure 5 is dominated by the improvement in reliability.

Figure 8 is the same as Figure 7 but for the predictions at four- and seven-month leads. As in the one-month lead prediction, the reliability improvement here is also very significant and makes a major contribution to the improvement in the BSS. Unlike the one-month lead case, however, the MME resolution does not always exceed that of the best single model, such as in the four-month lead prediction of the tropical precipitation.

The reliability and resolution components of the BSS for the size-reduced MME predictions are also displayed in Figures 7 and 8. For the subtropical predictions, the size-reduced MME shows little improvement over the SMEs in both reliability and resolution. For the tropical predictions, the reliability of the size-reduced MME is significantly better than that of SMEs, whereas the resolution is little improved.

Figures 9a and 9c are the reliability diagrams, the observed frequency against the forecast probability, for the one-month lead predictions of the above-normal u850 and below-normal precipitation in the tropics, respectively, which can further address the improvement of reliability in the MME. The diagonal dotted line is a perfect reliability line on which the forecast probability equals the observed frequency. As can be seen,

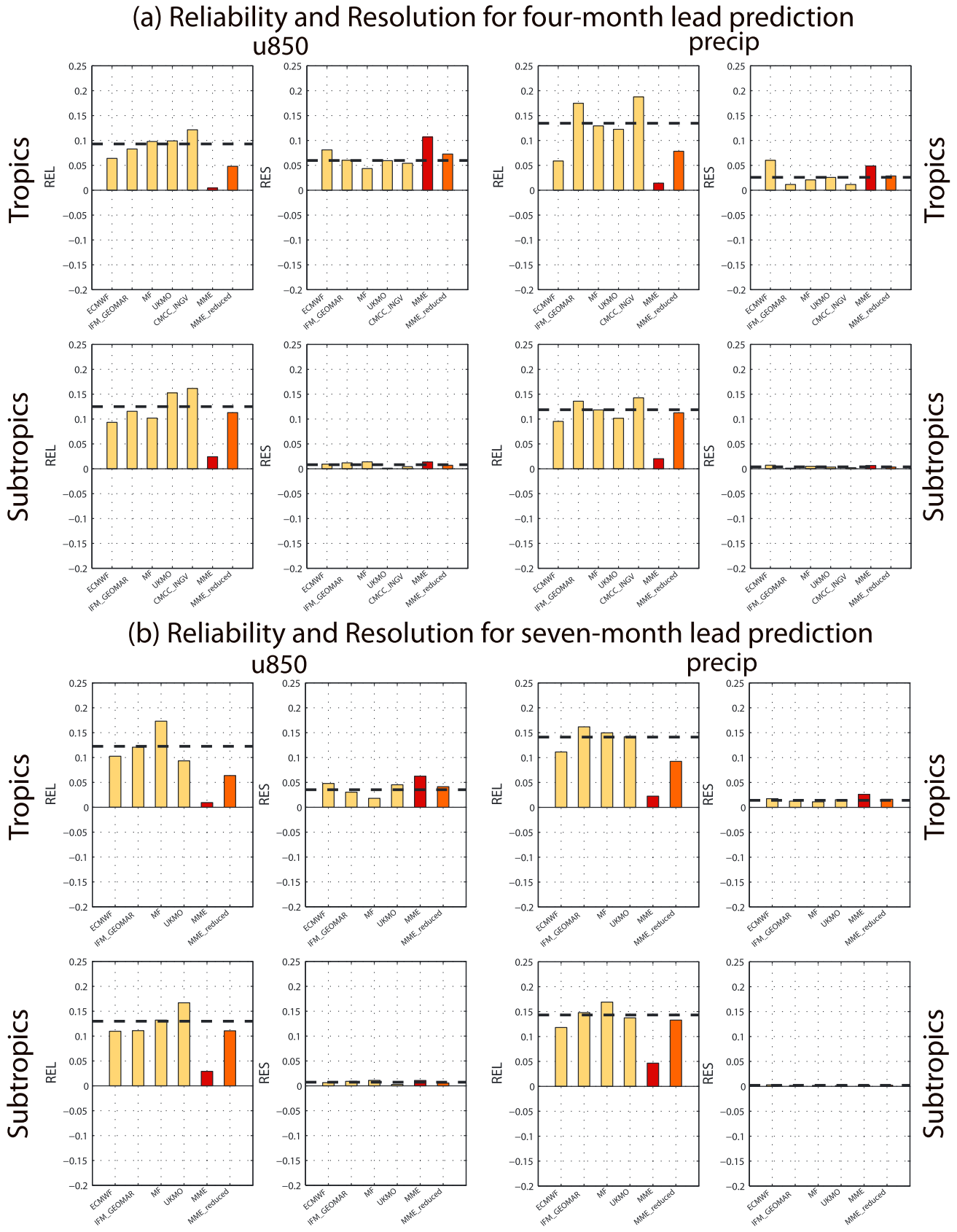

Figure 8. As in Figure 7 but for the predictions at (a) four-month lead and (b) seven-month lead.

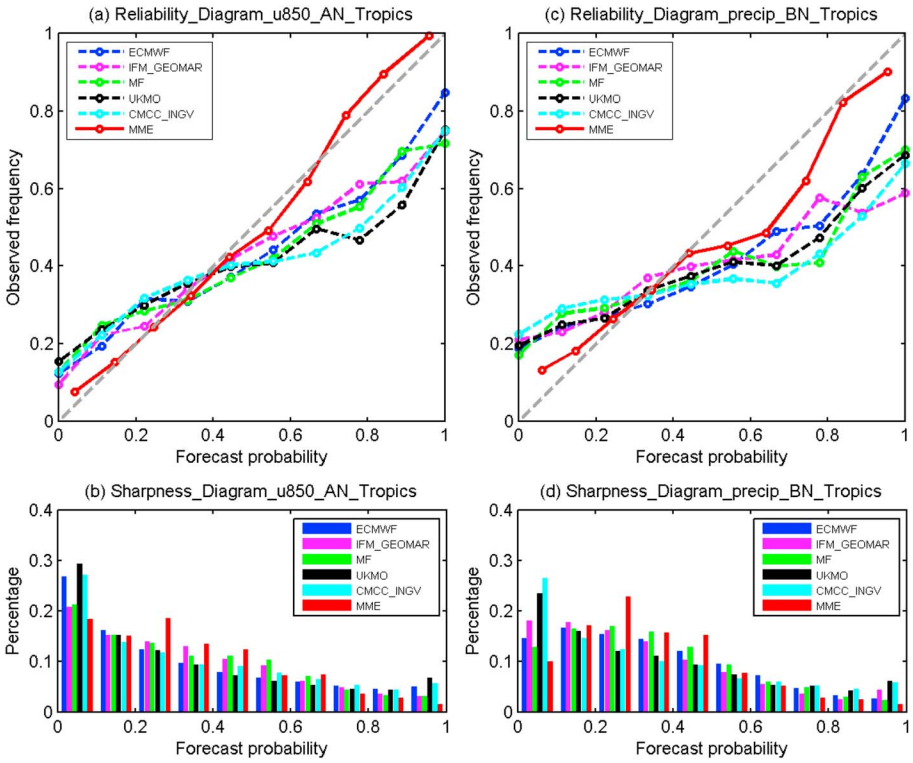

Figure 9. (a and c) Reliability and (b and d) sharpness diagrams for the ENSEMBLES predictions at one-month lead of JJA u850 anomaly (1960–2005) and precipitation anomaly (1979–2005) over the tropical domain (90°–160°E, 0°–20°N). Left column shows the u850's reliability and sharpness diagrams for the above-normal (AN) categorical event, while right column shows the precipitation's reliability and sharpness diagrams for the below-normal (BN) event.

the curve for each of the SMEs has a slope smaller than one and intersects the diagonal near the climatological probability (0.33). Higher (lower)-than-climatology forecast probabilities tend to be followed by insufficient (excessive) observed frequencies. On the other hand, the curve for the MME is much closer to the perfect reliability line, indicating a much better match between the forecast probabilities and observed frequencies. Figures 9b and 9d further show the sharpness diagrams, in which the percentages of probability forecasts falling into different probability bins are displayed. As can be seen, the MME percentages are lower than those of SMEs in extreme bins, whereas the MME percentages are larger than those of SMEs in the bins centered close to the climatological probability, indicating a smaller sharpness for the MME predictions. This is not odd and does not necessarily point to a weakness for the MME. Through averaging conflicting signals and widening the ensemble spread, the sharpness is necessarily reduced in the MME. However, these processes are just in favor of an improvement in reliability [e.g., *Barnston et al.*, 2003; *Kirtman et al.*, 2014]. A further discussion of the improvement in reliability is given in final section.

4.2. Relationship Between the Resolution and AC Skills

In the preceding sections, we found that the MME has much less significant improvement in the AC and resolution skills than in the reliability skill. A typical example is the case of the tropical precipitation prediction at four-month lead, in which the AC skill of MME is worse than that of the best single model, the ECMWF model (Figure 6a). Surprisingly, the MME's resolution skill is also worse than that of the best single model (Figure 8a). One question naturally arises as to why there appears to be such an AC-resolution similarity. Below, we try to answer this question by demonstrating that there is a strong and consistent relationship between AC and resolution.

First of all, a careful inspection of AC subplots in Figures 5 and 6 and resolution subplots in Figures 7 and 8 reveals that they are likely to covary consistently. For instance, a consistent rank can be obtained when the SMEs and MME skills are measured by AC or by resolution, both tending to decrease synchronously with lead time and from the tropics to subtropics. In fact, the covarying relationship between AC and resolution can be

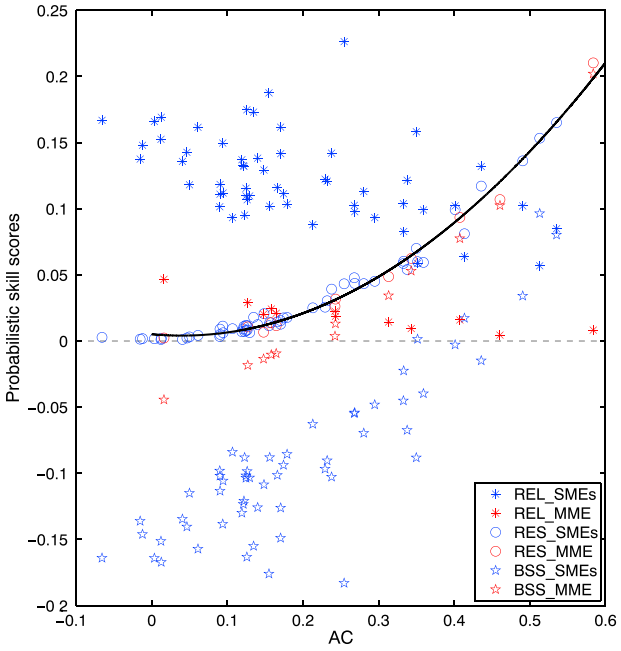

Figure 10. Scatterplots of resolution, reliability, and BSS (*y* axis) versus AC (*x* axis) with using all the data for the ENSEMBLES-predicted u850 anomaly (1960–2005) and precipitation anomaly (1979–2005), area-aggregated over the tropical and subtropical domains for both SMEs and MME at one-, four-, and seven-month leads. Circles, asterisks, and stars separately denote the resolution, reliability, and BSS. The red points are for the MME. The thick black curve is a quadratic polynomial fit to data of AC (only nonnegative values are used) and resolution. The probabilistic skills shown here are for an average over the above- and below-normal events.

easily identified in Figure 10, in which the scatterplots of resolution versus AC are displayed. Surprisingly, the scatters of resolution versus AC for both SMEs and MME are distributed in a very regular way. There seems to be a monotonic, quadratic-like relationship between resolution and AC. As can be seen, the resolution values tend to concentrate in the vicinity of the curve. The coefficient of determination of the polynomial fit is 0.99, demonstrating a remarkable "goodness of fit." The nonlinearity in the AC-resolution relationship is probably due to two reasons. First, while AC scales as the first power of the forecast variables, the resolution scales as the square, and second, the mapping from physical variables to probabilistic variables is nonlinear. The scatterplots of reliability and BSS against AC are also shown in Figure 10 for comparison. As shown in the figure, the scatter patterns for reliability and BSS versus AC are complicated. It is difficult to derive any relationship between reliability (or BSS) and AC.

It should be noted that we also found that such a monotonic AC-resolution relationship exists in the winter predictions (figure not shown). Recently, *Wang et al.* [2009] and *Alessandri et al.* [2011] have also tried to explore the possible relationship between deterministic and probabilistic skills in seasonal climate prediction. In evaluating the MME hindcasts, *Wang et al.* [2009] found a nonlinear relationship between AC and BSS for the prediction of the tropical precipitation. With the ENSMEBLES MME hindcasts, *Alessandri et al.* [2011] further showed that AC has a better relationship with resolution than with reliability for the prediction of the global surface air temperature. However, their relationships of AC-BSS and AC-resolution are far from what could be characterized as a monotonic relationship and therefore are much weaker than the AC-resolution relationship identified here.

Here we further provide a qualitative explanation for the observed strong AC-resolution relationship as follows. Let p be a random variable for the forecast probability and O represent the corresponding observed outcome for the event, one for occurrence and zero for nonoccurrence. In the limit of infinite sample size, the resolution can be formally written as [e.g., *Palmer et al.*, 2000]

$$\mathrm{BSS_{RES}} = \frac{9}{2}\int_0^1 f(p)\left[P(O=1|p) - P(O=1)\right]^2 dp, \tag{6}$$

where $P(|)$ and $P()$ separately denote conditional and unconditional probabilities and $f(p)$ is a probability density function (PDF) of p such that $f(p)dp$ is the relative frequency of the forecast probabilities lying in the bin $[p, p+dp]$. In equation (6), the integration measures the averaged quadratic difference (also the variance of the conditional probabilities) between the conditional and unconditional probabilities. On the other hand, according to probability theory, the event occurrence is statistically independent of the probabilistic forecast, if and only if the conditional and unconditional probabilities are equal for all the forecast probabilities. Strong statistical dependence corresponds to large deviations of the conditional probability from the unconditional probability, and vice versa. It therefore seems that the resolution could be understood in terms of the statistical dependence between the probabilistic forecast and the event occurrence. A similar point of view

was also proposed by *Bröcker* [2015]. Further, if the underlying forecast distributions are Gaussian and the ensemble variance changes little from case to case (these approximations are suitable for seasonal mean variables [e.g., *Kumar et al.*, 2000; *Tang et al.*, 2008; *Yang et al.*, 2012]), the condition p for $P(O=1|p)$ could be expressed equivalently as a condition on the ensemble mean μ_f only. Then, given the PDF of μ_f (denoted $f(\mu_f)$), equation (6) can be rewritten as

$$\mathrm{BSS_{RES}} = \frac{9}{2}\int_{-\infty}^{\infty} f(\mu_f)\left[P\left(O=1\middle|\mu_f\right) - P(O=1)\right]^2 d\mu_f, \tag{7}$$

which indicates that the resolution could be further understood in terms of the statistical dependence between the ensemble mean and the event occurrence. Since $P(O=1|\mu_f)$ and $P(O=1)$ are, after all, determined respectively by the underlying conditional PDF $f(x|\mu_f)$ and unconditional PDF $f(x)$, where x is the original continuous predictand, the resolution would be ultimately dominated by the statistical dependence between μ_f and x, whose strength depends on how large the difference between $f(x|\mu_f)$ and $f(x)$ is [e.g., *DelSole*, 2005]. The above understanding of the resolution in terms of the statistical dependence between forecast (μ_f) and observation (x) facilitates us to further understand the observed resolution-AC correspondence, since the differences and similarities between the general concepts of statistical dependence and linear correlation are widely discussed in statistical and predictability literature [e.g., *DelSole*, 2004, 2005]. Specifically, statistical dependence is generally not tantamount to linear correlation, but for joint normally distributed variables they are intrinsically equivalent. On the seasonal time scale considered here, a joint-Gaussian distribution would be approximately satisfied for the forecast and observation [e.g., *Tippett et al.*, 2014]. As such, a good AC-resolution agreement seems to be theoretically expected, as reported by *Kharin and Zwiers* [2003]. However, their theoretical derivation is based on a perfect model assumption, which requires the observation to be statistically indistinguishable from the model ensemble members. Since the skills calculated here are the real skills that make use of the actual observations rather than the perfect model skills that do not, a more generalized theoretical consideration without requiring the perfect model assumption should be made for further understanding the observed AC-resolution relationship. We leave it for future study.

5. Skills of Coupled Versus Uncoupled MMEs

In this section, the skills of the coupled ENSEMBLES and DEMETER MMEs are compared with that of the uncoupled HFP2 MME, in predicting the WNP-EASM variability and the underlying SST conditions. While there are many studies targeting on comparing the prediction performances of coupled and uncoupled models for outside the WNP-EASM region [e.g., *Graham et al.*, 2005; *Guérémy et al.*, 2005; *Kumar et al.*, 2008; *Stefanova et al.*, 2012; *Landman et al.*, 2012; *Misra et al.*, 2014; *Li and Misra*, 2014; *Tang et al.*, 2014], most of the previous comparison studies for the WNP-EASM region have been mainly focused on the monsoon simulation, rather than on the monsoon prediction [e.g., *Fu et al.*, 2002; *Aldrian et al.*, 2005; *Hu et al.*, 2012; *Huang et al.*, 2012; *Fang et al.*, 2013; *Ham et al.*, 2014; *Zou and Zhou*, 2013; *Song and Zhou*, 2014; *Cha et al.*, 2015]. Here we present an assessment about the practical advantages of the current one-tier over two-tier operational MME systems in predicting the WNP-EASM.

Figure 11 shows the spatial distributions of the AC skills in the ENSEMBLES, DEMETER, and HFP2 MMEs for the one-month lead predictions of JJA u850 and precipitation over the period of 1979–2001, respectively. For ENSEMBLES, the u850 skill shown here is greater than that for the period of 1960–2005 (Figure 1). For precipitation, the shown skill is little different from that in Figure 1 due to their similar time coverage. For DEMETER, the prediction skills for u850 and precipitation show no large differences from those in ENSEMBLES. For the two-tier system of HFP2, however, the skill is substantially lower than those in ENSEMBLES and DEMETER, especially in the tropical region south of 20°N for both u850 and precipitation. The overwhelming dominance of ENSEMBLES and DEMETER over HFP2 is almost unaffected, even after the best model of ECMWF is removed from the former two systems and the total ensemble size is consequently reduced to 36, smaller than the latter's (figure not shown).

Figure 12 shows the MME skills in the area-aggregated way. Here AC, BSS, reliability, and resolution are all shown. The advantages of the coupled MMEs over the uncoupled MME are reflected basically in each variable, region, and skill measure. The disadvantage of the uncoupled MME in probabilistic forecasts is embodied in both reliability and resolution. Figure 13 further shows the reliability and sharpness diagrams for the three MME predictions of the above-normal u850 and below-normal precipitation in the tropics.

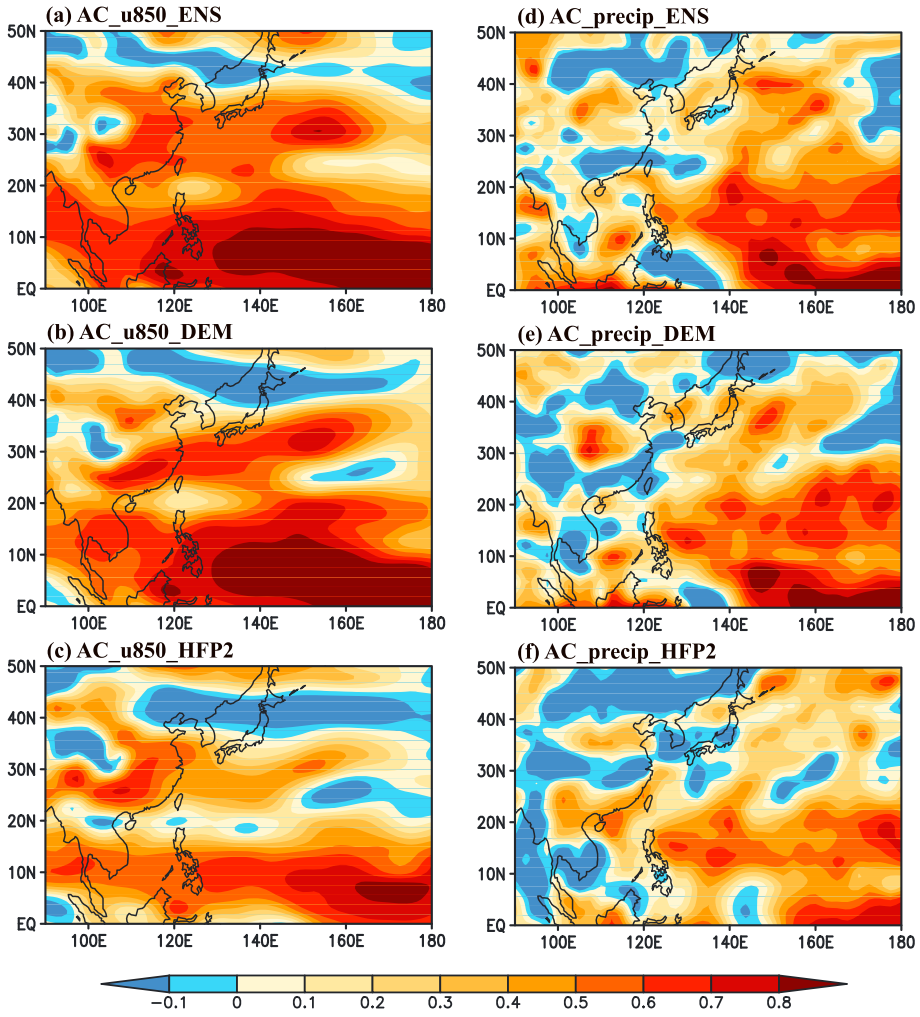

Figure 11. Spatial distributions of the AC skills of the (top row) ENSEMBLES (ENS), (middle row) DEMETER (DEM), and (bottom row) HFP2 MMEs for the predictions of (left column) JJA u850 and (right column) precipitation at one-month lead for the period 1979–2001.

Despite the fact that the curve for the uncoupled MME is actually less flat than those for the corresponding single models (figure not shown), it is still far from the perfect reliability line, which is in contrast to the situation for the coupled MMEs. Accordingly, the sharpness for the uncoupled MME predictions is larger than that for the coupled MMEs.

The remarkable advantage of the one-tier MME systems over the two-tier system seems not to be completely explained by the differences in their atmospheric model physics. For example, ENSEMBLES only marginally outperforms DEMETER despite the fact that the former has improved atmospheric model physics compared to the latter as mentioned in section 2.1. On the other hand, the atmospheric model physical schemes of DEMETER might not be as good as those of HFP2, for the former is older than the latter. Many of the advantages seen in the coupled MMEs would be ascribed to their great potentials in representing important air-sea coupled processes in the seasonal prediction of the WNP-EASM. Broadly speaking, the benefits of including air-sea coupling may include an improvement in the underlying SST forecast skills in key regions, which are further translated into atmospheric skills [e.g., *Goddard and Mason*, 2002; *Graham et al.*, 2005; *Guérémy et al.*, 2005; *Li et al.*, 2008], and a better simulation of atmospheric variability, which is unrelated to the SST skill and occurs even without an improved SST skill [e.g., *Wang et al.*, 2005; *Zhu and Shukla*, 2013].

Considering the dominant role of SST forcing in sustaining monsoon seasonal predictability [e.g., *Charney and Shukla*, 1981; *Yang et al.*, 1998; *Webster et al.*, 1998; *Hassan et al.*, 2004; *Kang and Shukla*, 2006; *Yang*

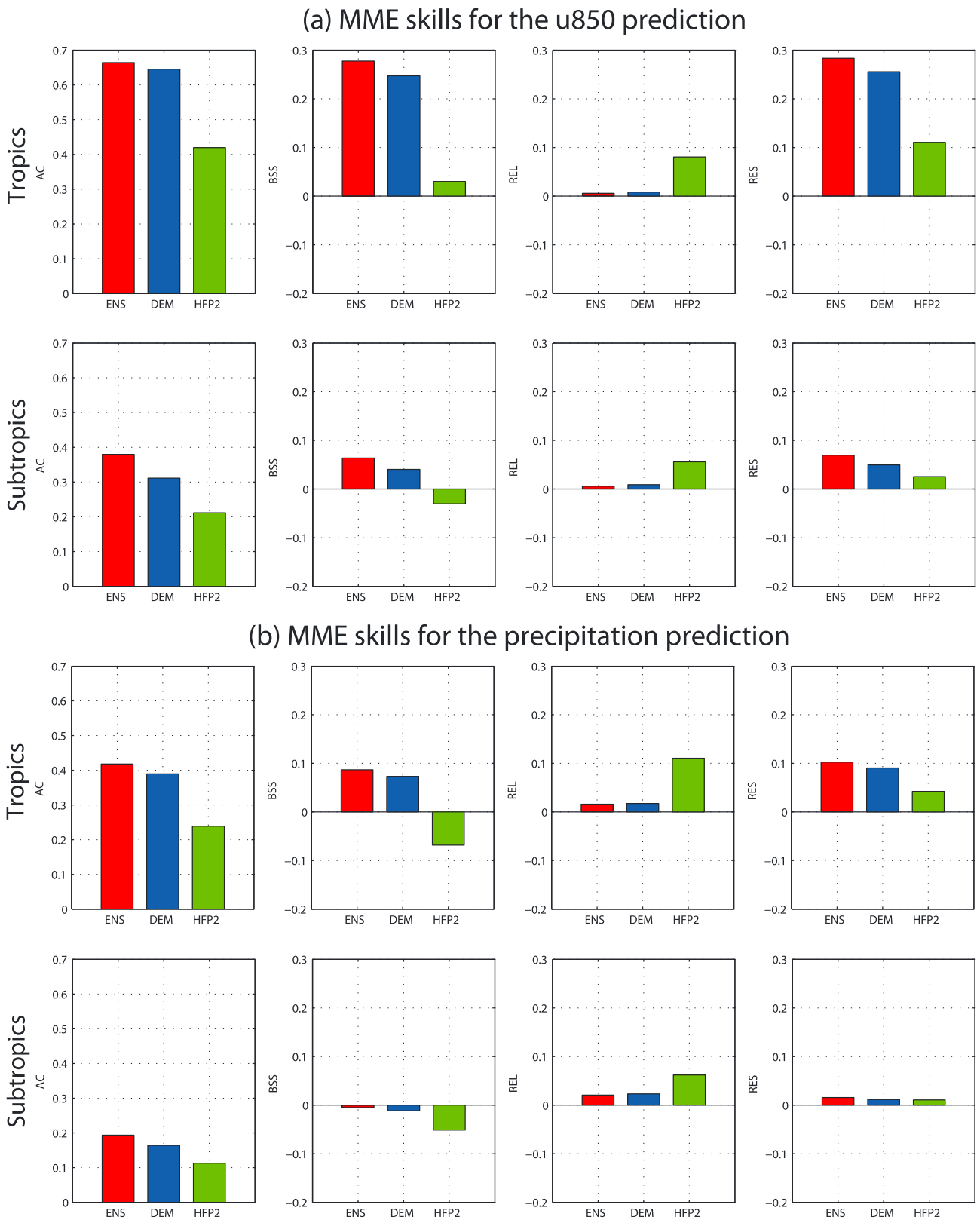

Figure 12. Area-aggregated skills (AC, BSS, reliability, and resolution) of the ENSEMBLES (ENS), DEMETER (DEM), and HFP2 MMEs for the predictions of (top two panels) JJA u850 and (bottom two panels) precipitation over the tropical and subtropical domains at one-month lead for the period 1979–2001. The probabilistic skills shown here are for an average over the above- and below-normal events.

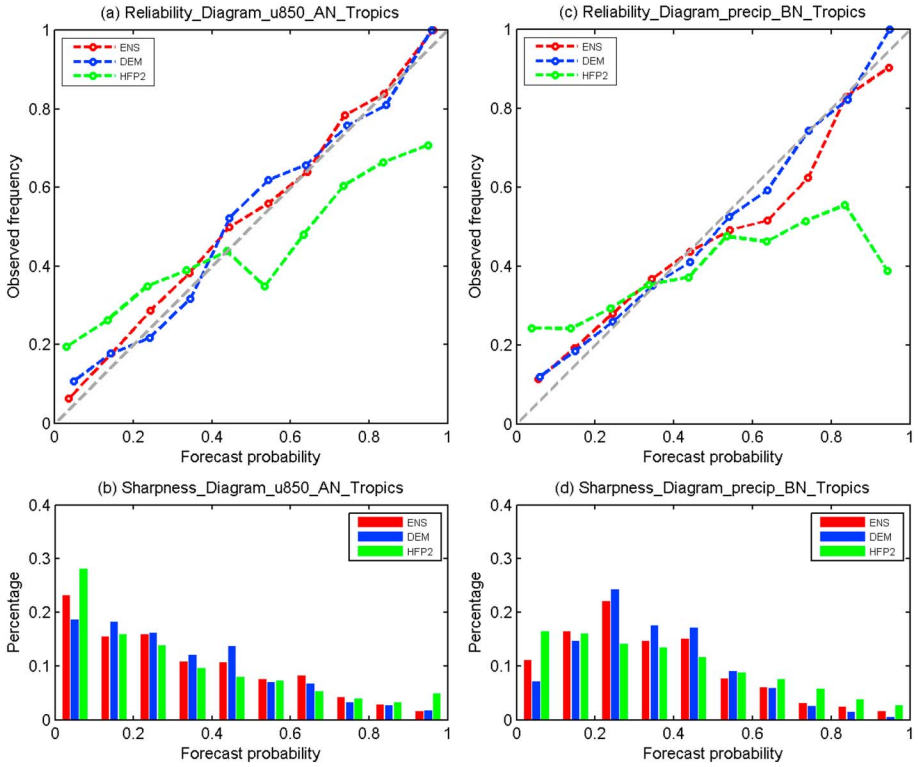

Figure 13. (a and c) Reliability and (b and d) sharpness diagrams for the predictions of JJA u850 and precipitation over the tropical domain at one-month lead for the period 1979–2001 with the ENSEMBLES (ENS), DEMETER (DEM), and HFP2 MMEs. Left column shows the u850′ reliability and sharpness diagrams for the above-normal (AN) categorical event, while right column shows the precipitation′ reliability and sharpness diagrams for the below-normal (BN) event.

et al., 2012], we now focus on examining the SST prediction skills in the three MMEs. Figure 14 shows the spatial distributions of one-month lead AC skills of JJA SST anomaly for the three MMEs. The reason for not showing probabilistic skills is that HFP2 persisted the observed SST anomaly as the forecast and therefore lacks an estimate of the forecast uncertainty, failing to produce any forecast probability different from zero or one. As shown in Figure 14a, for ENSMEBLES, high skill (greater than 0.7) mainly appears in the tropical central-eastern Pacific, North Indian Ocean, western North Pacific, and the South Pacific Convergence Zone in the Southern Hemisphere. Interestingly, the first three regions are just the main tropical regions, where the SST anomalies play crucial roles in providing predictability for the WNP-EASM, as mentioned in section 3.1. For DEMETER (Figure 14b), the pattern of the SST anomaly skill is very similar to that for ENSEMBLES, while the amplitude of the skill is somewhat smaller. Compared to DEMETER, the stronger SST anomaly skill seen in ENSEMBLES may be due to the wider use of subsurface assimilation for ocean initialization, as argued in *Alessandri et al.* [2011]. However, the two coupled MMEs′ SST anomaly skills can still be counted as comparable, especially in contrast to the HFP2 SST anomaly skill shown in Figure 14c. The latter (i.e., the persistence skill) appears strikingly weaker in the key regions mentioned above.

Although the method of forecasting SST anomaly for HFP2 is too simple, just the persistence forecast, the disadvantage of the SST anomaly persistence skill over the Indo-western Pacific region may have a universal significance for many more two-tier systems. As argued in previous studies [e.g., *Kumar et al.*, 2005; *Wang et al.*, 2005; *Wu and Kirtman*, 2005; *Wu et al.*, 2006], it is in the Indo-western Pacific sector that air-sea coupled processes can play crucial roles in controlling the evolution of underlying SST anomaly. Therefore, it is likely that the SST anomaly prediction skill over this region is inherently flawed in any statistical forecast model, in which during the phases of training and predicting only the oceanic information is used. Similar point was also argued in *Graham et al.* [2005] for the seasonal prediction over the North Atlantic and European regions.

Another interesting feature in Figure 14 is that the one-month lead persistence prediction of SST anomaly is much worse than the coupled model counterpart in the equatorial central-eastern Pacific, which seems not

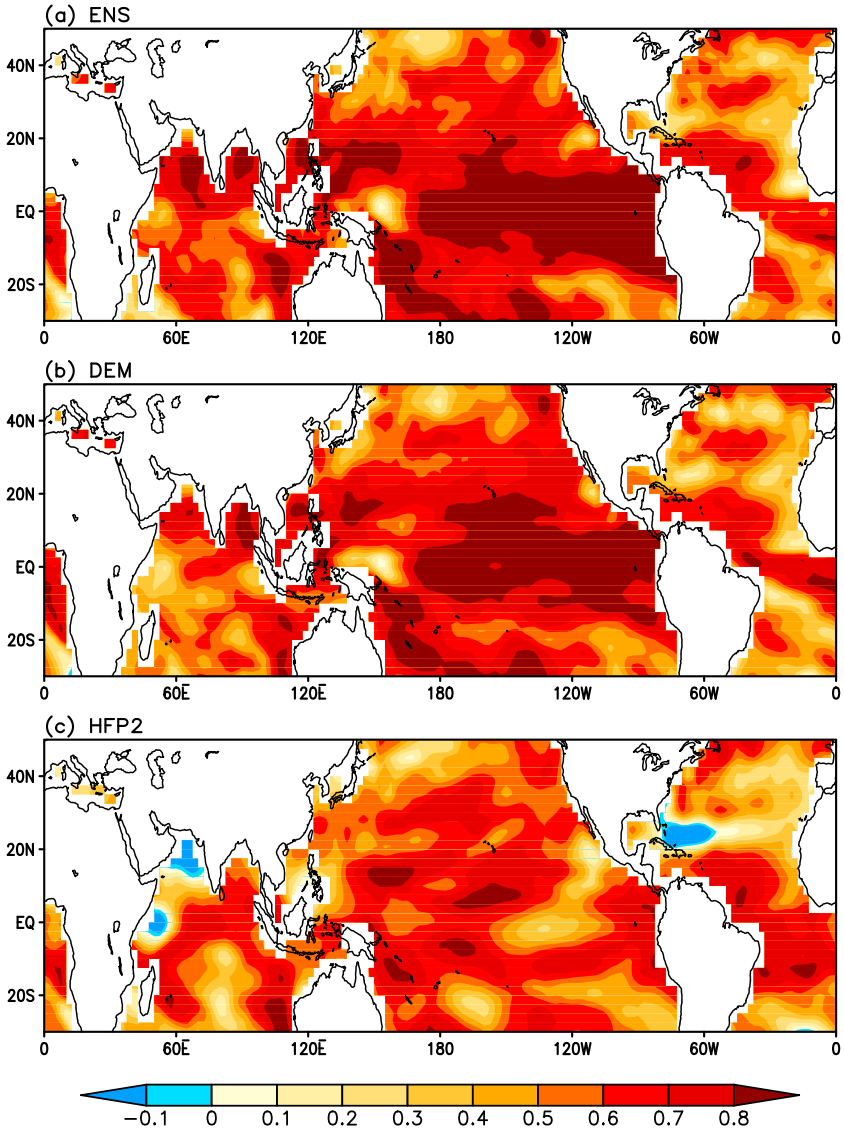

Figure 14. Spatial distributions of the AC skills of the (top) ENSEMBLES (ENS), (middle) DEMETER (DEM), and (bottom) HFP2 MMEs for the prediction of JJA SST anomaly at one-month lead for the period 1979–2001.

to be consistent with some previous works [e.g., *Younas and Tang*, 2013]. This is because Figure 14 only focuses on the SST anomaly prediction of the summer season when the SST anomaly has typically the weakest signals in the region. It has been reported that there is a significant seasonal dependence of the SST anomaly persistence skill in the ENSO region [e.g., *Jin et al.*, 2008; *Wang et al.*, 2009; *Alessandri et al.*, 2010]. Indeed, we found that the persistence skill of the SST anomaly during winter at one-month lead is as skillful as the coupled predictions over this region (figure not shown).

6. Conclusions and Discussion

6.1. Conclusions

In this study, based on the historical forecasts of the ENSEMBLES, DEMETER, and HFP2 ensembles, we have investigated the advantages of the coupled MME over the contributing SMEs and over the uncoupled atmospheric MME, in predicting the seasonal variability of the WNP-EASM. The metrics of prediction skill includes the deterministic measure of AC and the probabilistic measure of BSS together with its two components,

reliability and resolution. Two representative variables, the 850 hPa zonal wind and precipitation, are chosen to characterize the monsoon variability.

With the ENSEMBLES coupled historical forecasts, the prediction skills of the WNP-EASM variability and the superiorities of the MME over the SMEs have been assessed. We find that the superiorities of the MME over the SMEs in predicting the WNP-EASM variability are strongly forecast-format-dependent. In terms of probabilistic forecast, the BSS measure shows that the MME is much more skillful than any of the SMEs. However, in terms of deterministic forecast, the AC measure shows that the MME advantages are usually only marginal and not systematic. For some cases, the deterministic forecast skills of MME are even worse than those of some SMEs. In addition, a further analysis of the size-reduced MME skill reveals that the MME improvements in the tropics arise both from the model diversity inherent in the MME approach and from the increase in ensemble size, while those in the subtropics seems to be mainly due to the increase in ensemble size.

To understand the forecast-format-dependent MME superiorities in BSS rather than AC skills, we have further diagnosed the two components of BSS, reliability and resolution, and find that the MME improvements in BSS are dominantly due to the improvement in reliability. In contrast, the resolutions in most cases are improved only to a significantly lesser extent or even not improved at all. This feature is similar to AC. It is the significant reliability improvement that makes the improvements in AC and BSS so different.

It is striking that both AC and resolution are not always improved in the MME. The possible cause for such a similarity is that AC and resolution are strongly related to each other. A monotonic relationship between AC and resolution is identified. In contrast, the relationship between AC and reliability is very poor. A qualitative explanation of the excellent AC-resolution relationship is further proposed. It is argued that the resolution could be understood in terms of the statistical dependence between forecast and observation, which would have a good coherence with the linear correlation (AC) if variables are joint normally distributed.

The prediction skills of the coupled ENSEMBLES and DEMETER MMEs and the uncoupled atmospheric HFP2 MME forced by persisted SST anomalies have also been compared. It is found that both deterministic and probabilistic skills of the coupled MMEs in predicting the WNP-EASM variability are much higher than those of the uncoupled MME. Analyses reveal that the much better skills of the coupled MME predictions are associated with their better skills in predicting the underlying SST anomalies over the ENSO region, western North Pacific, and North Indian Ocean.

6.2. Discussion

In this study, we have shown that the MME improvement in AC and resolution is generally less effective. The ultimate reasons deserve to be discussed. In nature, AC and resolution are insensitive to systematic biases and are measures of the inherent predictive ability of models [e.g., *Toth et al.*, 2006]. As such, their values cannot exceed the level indicated by the predictability limit of the real monsoon system. AC and resolution for the SMEs are further limited by their deficient ability in capturing the predictable portion of the real monsoon system. Model error is a crucial factor giving rise to this deficiency. Additionally, insufficient averaging within the small-size ensemble would leave a certain amount of random noise in the ensemble means, degrading the SMEs' ability in reproducing the real signal. The multimodel composite could improve the AC and resolution by means of error cancelation. However, while the multimodel averaging is indeed capable of cancelling out the random noises that are largely independent, it may not always quite effective in reducing the model errors, because in reality the model errors are not completely independent [e.g., *Yoo and Kang*, 2005], limiting the magnitude of AC and resolution improvement. At last, we stress that if the utmost skill level implied by the predictability limit is low, there is not much potential to improve AC and resolution in the first place. In this situation, the MME approach naturally cannot be of much help. This argument may provide another clue to the inability of the MME in improving the skills of subtropical predictions at long leads.

Unlike resolution, reliability is not slaved to the statistical dependence between forecast and observation. Contrarily, it (or more accurately *unreliability*) measures the conditional probability biases. In principle, there should be no barrier at the physical level in reducing these biases, which may provide an overall explanation of the strong reliability improvement. Statistically speaking, these probability biases are ultimately determined by the discrepancy between the underlying forecast probability distribution and the conditional distribution for observation corresponding to the forecast. What causes the disagreement between the two distributions and how the MME can reduce this disagreement need to be discussed. If the normality

of the two distributions is assumed, then the probabilistic reliability can only be impacted by the disagreements in their means and variances. If we further assume that the ensemble mean μ_f and observation x satisfy a joint-Gaussian distribution and the ensemble variance σ_f^2 does not vary much from case to case, the diagnoses of the biases in mean and variance become straightforward. In this case, the parameters of the Gaussian conditional distribution for the observation can be determined through a linear regression procedure with μ_f and x as the independent and dependent variables, respectively [*Tippett et al.*, 2014]. Specifically, the mean μ_r and variance σ_r^2 of this conditional distribution can be expressed as $\mu_r = \beta\mu_f$ and $\sigma_r^2 = (1-\rho^2)\sigma_x^2$, where β and ρ are the regression and correlation coefficients and σ_x^2 is the variance of x. In other words, the mean is given by the linear regression and the variance equals to the error variance of the linear regression. After relevant calculations for the aggregated forecast cases in sections 3 and 4, we find that the SMEs are universally biased in both mean and variance. Specifically, they are overconfident, characterized by $\beta < 1$ and $\sigma_f^2 < \sigma_r^2$, which indicates that the forecast signals are too large and the ensemble spread is too small. This overconfidence for the SMEs is due to a lack of accounting for model uncertainty and small ensemble size [e.g., *Barnston et al.*, 2003]. However, the overconfidence is much alleviated in the resulting MME. This improvement should benefit from both the model diversity and the increase in ensemble size. The powerful ability of the MME in reducing the overconfidence seems less subject to the possible mutual dependence between the SMEs, since even for the cases where the MME cannot beat the best single model in AC, the overconfidence is still effectively lessened. In short, the reduction of overconfidence is the possible reason for the MME reliability improvement. *Weigel et al.* [2008] also highlighted the importance of reducing overconfidence for the MME to improve probabilistic forecast skill. Further investigation is needed in order to understand such a reason.

Acknowledgments

This work is jointly supported by the National Natural Science Foundation of China under grants 41305085 and 41330420 and by Jiangsu Collaborative Innovation Center for Climate Change. The ENSEMBLES and DEMETER data sets are available from http://apps.ecmwf.int/datasets/. The HFP2 data set is downloaded from http://cis.apcc21.org/. Authors would be grateful to three anonymous reviewers for their constructive comments and suggestions to improve the manuscript.

References

Aldrian, E., D. Sein, D. Jacob, L. D. Gates, and R. Podzun (2005), Modelling Indonesian rainfall with a coupled regional model, *Clim. Dyn.*, *25*, 1–17, doi:10.1007/s00382-004-0483-0.

Alessandri, A., A. Borrelli, S. Masina, P. Di Pietro, A. F. Carril, A. Cherchi, S. Gualdi, and A. Navarra (2010), The INGV-CMCC seasonal prediction system: Improved ocean initial conditions, *Mon. Weather Rev.*, *138*, 2930–2952.

Alessandri, A., A. Borrelli, A. Navarra, A. Arribas, M. Déqué, P. Rogel, and A. Weisheimer (2011), Evaluation of probabilistic quality and value of the ENSEMBELS multimodel seasonal forecasts: Comparison with DEMETER, *Mon. Weather Rev.*, *139*, 581–607, doi:10.1175/2010MWR3417.1.

Barnston, A. G., S. J. Mason, L. Goddard, D. G. DeWitt, and S. E. Zebiak (2003), Multimodel ensembling in seasonal climate forecasting at IRI, *Bull. Am. Meteorol. Soc.*, *84*, 1783–1796, doi:10.1175/BAMS-84-12-1783.

Becker, E. J., H. Van den Dool, and Q. Zhang (2015), Probabilistic forecasting with NMME, *Sci. Technol. Infusion Clim. Bull.*, *43–44*. [Available at http://www.nws.noaa.gov/ost/climate/STIP/39CDPW/39cdpw-EBecker.pdf.]

Bjerknes, J. (1969), Atmospheric teleconnections from the equatorial Pacific, *Mon. Weather Rev.*, *97*(3), 163–172.

Boer, G. J., N. A. McFarlane, R. Laprise, J. D. Henderson, and J. P. Blanchet (1984), The Canadian Climate Centre spectral atmospheric general circulation model, *Atmos.-Ocean*, *22*(4), 397–429, doi:10.1080/07055900.1984.9649208.

Bröcker, J. (2015), Resolution and discrimination–two sides of the same coin, *Q. J. R. Meteorol. Soc.*, *141*, 1277–1282, doi:10.1002/QJ.2434.

Cha, D. H., C. S. Jin, J. H. Moon, and D. K. Lee (2015), Improvement of regional climate simulation of East Asian summer monsoon by coupled air–sea interaction and large-scale nudging, *Int. J. Climatol.*, doi:10.1002/joc.4349.

Charney, J. G., and J. Shukla (1981), Predictability of monsoons, in *Monsoon Dynamics*, edited by J. Lighthill and R. P. Pearec, pp. 99–110, Cambridge Univ. Press, Cambridge, U. K., doi:10.1017/CBO9780511897580.009.

Chen, H., T. Zhou, R. B. Neale, X. Wu, and G. J. Zhang (2010), Performance of the new NCAR CAM3.5 in East Asian summer monsoon simulations: Sensitivity to modifications of the convection scheme, *J. Clim.*, *23*(13), 3657–3675, doi:10.1175/2010jcli3022.1.

Chen, M., W. Wang, A. Kumar, H. Wang, and B. Jha (2012), Ocean surface impacts on the seasonal-mean precipitation over the tropical Indian Ocean, *J. Clim.*, *25*(10), 3566–3582, doi:10.1175/jcli-d-11-00318.1.

Chowdary, J. S., S.-P. Xie, J.-Y. Lee, Y. Kosaka, and B. Wang (2010), Predictability of summer northwest Pacific climate in 11 coupled model hindcasts: Local and remote forcing, *J. Geophys. Res.*, *115*, D22121, doi:10.1029/2010JD014595.

Côté, J., S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth (1998a), The operational CMC-MRB global environmental multiscale (GEM) model. Part I: Design considerations and formulation, *Mon. Weather Rev.*, *126*(6), 1373–1395.

Côté, J., J.-G. Desmarais, S. Gravel, A. Méthot, A. Patoine, M. Roch, and A. Staniforth (1998b), The operational CMC-MRB global environmental multiscale (GEM) model. Part II: Results, *Mon. Weather Rev.*, *126*(6), 1397–1418.

DelSole, T. (2004), Predictability and information theory. Part I: Measures of predictability, *J. Atmos. Sci.*, *61*, 2425–2440, doi:10.1175/1520-0469(2004)061<2425:PAITPI>2.0.CO;2.

DelSole, T. (2005), Predictability and information theory. Part II: Imperfect forecasts, *J. Atmos. Sci.*, *62*, 3368–3381, doi:10.1175/JAS3522.1.

DelSole, T., J. Nattala, and M. K. Tippett (2014), Skill improvement from increased ensemble size and model diversity, *Geophys. Res. Lett.*, *41*, 7331–7342, doi:10.1002/2014GL060133.

Déqué, M. (1997), Ensemble size for numerical seasonal forecasts, *Tellus, Ser. A*, *49*(1), 74–86.

Doblas-Reyes, F. J., M. Deque, and J.-P. Piedelievre (2000), Multi-model spread and probabilistic seasonal forecasts in PROVOST, *Q. J. R. Meteorol. Soc.*, *126*, 2069–2088, doi:10.1256/smsqj.56704.

Fang, Y. J., Y. C. Zhang, A. N. Huang, and B. Li (2013), Seasonal and intraseasonal variations of East Asian summer monsoon precipitation simulated by a regional air-sea coupled model, *Adv. Atmos. Sci.*, *30*(2), 315–329.

Fu, X. H., B. Wang, and T. Li (2002), Impacts of air-sea coupling on the simulation of mean Asian summer monsoon in the ECHAM4 model, *Mon. Weather Rev.*, *130*, 2889–2904.
Goddard, L., and S. J. Mason (2002), Sensitivity of seasonal climate forecasts to persisted SST anomalies, *Clim. Dyn.*, *19*(7), 619–632, doi:10.1007/s00382-002-0251-y.
Graham, R., et al. (2005), A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model, *Tellus, Ser. A*, *57*(3), 320–339, doi:10.1111/j.1600-0870.2005.00116.x.
Guérémy, J.-F., M. Déqué, A. Braun, and J.-P. Piedelièvre (2005), Actual and potential skill of seasonal predictions using the CNRM contribution to DEMETER: Coupled versus uncoupled model, *Tellus, Ser. A*, *57*, 308–319.
Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting. Part I: Basic concept, *Tellus, Ser. A*, *57*, 219–233.
Ham, S., S. Y. Hong, and S. Park (2014), A study on air–sea interaction on the simulated seasonal climate in an ocean–atmosphere coupled model, *Clim. Dyn.*, *42*(5–6), 1175–1187.
Hamill, T. M., and J. Juras (2006), Measuring forecast skill: Is it real skill or is it the varying climatology?, *Q. J. R. Meteorol. Soc.*, *132*, 2905–2923.
Hassan, A. S., X.-Q. Yang, and S.-S. Zao (2004), Reproducibility of seasonal ensemble integrations with ECMWF GCM and its association with ENSO, *Meteorol. Atmos. Phys.*, *86*(3–4), 159–172.
Hogan, R. J. and I. B. Mason (2011), Deterministic forecasts of binary events, in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd ed., edited by I. T. Jolliffe and D. B. Stephenson, John Wiley, Chichester, U. K., doi:10.1002/9781119960003.ch3.
Hu, Y., Z. Zhong, X. Liu, and Y. Zhu (2012), Influence of air– sea interaction on the simulation of the East Asian summer monsoon: A case study, *Dyn. Atmos. Oceans*, *53–54*, 1–16.
Huang, Q., S. X. Yao, and Y. C. Zhang (2012), Analysis of local air-sea interaction in East Asia using a regional air-sea coupled model, *J. Clim.*, *25*, 767–776, doi:10.1175/2011JCLI3783.1.
Jin, E., et al. (2008), Current status of ENSO prediction skill in coupled ocean–atmosphere models, *Clim. Dyn.*, *31*(6), 647–664, doi:10.1007/s00382-008-0397-3.
Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, and J. Woollen (1996), The NCEP/NCAR 40-year reanalysis project, *Bull. Am. Meteorol. Soc.*, *77*(3), 437–471.
Kang, I.-S., and J. Shukla (2006), Dynamic seasonal prediction and predictability, in *The Asian Monsoon*, edited by B. Wang, chap. 15, pp. 585–612, Springer, New York, doi:10.1007/3-540-37722-0_15.
Kang, I.-S., et al. (2002), Intercomparison of the climatological variations of Asian summer monsoon precipitation simulated by 10 GCMs, *Clim. Dyn.*, *19*(5–6), 383–395, doi:10.1007/s00382-002-0245-9.
Kharin, V. V., and F. W. Zwiers (2003), Improved seasonal probability forecasts, *J. Clim.*, *16*(11), 1684–1701.
Kharin, V. V., F. W. Zwiers, Q. Teng, G. J. Boer, J. Derome, and J. S. Fontecilla (2009), Skill assessment of seasonal hindcasts from the Canadian Historical Forecast Project, *Atmos.-Ocean*, *47*, 204–223.
Kirtman, B. P., D. Min, J. M. Infanti, J. L. Kinter III, D. A. Paolino, Q. Zhang, H. van den Dool, S. Saha, M. P. Mendez, and E. Becker (2014), The North American Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction; Phase-2 toward developing intra-seasonal prediction, *Bull. Am. Meteorol. Soc.*, *95*(4), 585–601, doi:10.1175/BAMS-D-12-00050.1.
Kosaka, Y., J. S. Chowdary, S.-P. Xie, Y.-M. Min, and J.-Y. Lee (2012), Limitations of seasonal predictability for summer climate over East Asia and the Northwestern Pacific, *J. Clim.*, *25*(21), 7574–7589, doi:10.1175/jcli-d-12-00009.1.
Kosaka, Y., S.-P. Xie, N.-C. Lau, and G. A. Vecchi (2013), Origin of seasonal predictability for summer climate over the Northwestern Pacific, *Proc. Natl. Acad. Sci. U.S.A.*, *110*(19), 7574–7579, doi:10.1073/pnas.1215582110.
Krishnamurti, T. N. (1999), Improved weather and seasonal climate forecasts from multimodel superensemble, *Science*, *285*, 1548–1550, doi:10.1126/science.285.5433.1548.
Krishnamurti, T. N., C. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran (2000), Multimodel ensemble forecasts for weather and seasonal climate, *J. Clim.*, *13*(23), 4196–4216.
Kug, J.-S., I.-S. Kang, and D.-H. Choi (2008), Seasonal climate predictability with Tier-one and Tier-two prediction systems, *Clim. Dyn.*, *31*(4), 403–416, doi:10.1007/s00382-007-0264-7.
Kumar, A., A. G. Barnston, P. Peng, M. P. Hoerling, and L. Goddard (2000), Changes in the spread of the variability of the seasonal mean atmospheric states associated with ENSO, *J. Clim.*, *13*(17), 3139–3151.
Kumar, A., Q. Zhang, J.-K. E. Schemm, M. L'Heureux, and K.-H. Seo (2008), An assessment of errors in the simulation of atmospheric interannual variability in uncoupled AGCM simulations, *J. Clim.*, *21*, 2204–2217.
Kumar, K. K., M. Hoerling, and B. Rajagopalan (2005), Advancing dynamical prediction of Indian monsoon rainfall, *Geophys. Res. Lett.*, *32*, L08704, doi:10.1029/2004GL021979.
Landman, W. A., D. DeWitt, D.-E. Lee, A. Beraki, and D. Lötter (2012), Seasonal rainfall prediction skill over South Africa: 1- vs. 2-tiered forecasting systems, *Weather Forecast.*, *27*, 489–501, doi:10.1175/WAF-D-11-00078.1.
Lee, D. Y., J. B. Ahn, and K. Ashok (2013), Improvement of multi-model ensemble seasonal prediction skills over East Asian summer monsoon region using a climate filter concept, *J. Appl. Meteorol. Climatol.*, *52*, 1127–1138.
Lee, D. Y., J. B. Ahn, and J. H. Yoo (2014), Enhancement of seasonal prediction of East Asian summer rainfall related to western tropical Pacific convection, *Clim. Dyn.*, *45*, 1025–1042, doi:10.1007/s00382-014-2343-x.
Lee, S.-S., J.-Y. Lee, K.-J. Ha, B. Wang, and J. K. E. Schemm (2011), Deficiencies and possibilities for long-lead coupled climate prediction of the Western North Pacific-East Asian summer monsoon, *Clim. Dyn.*, *36*(5–6), 1173–1188, doi:10.1007/s00382-010-0832-0.
Li, C., R. Lu, and B. Dong (2012), Predictability of the western North Pacific summer climate demonstrated by the coupled models of ENSEMBLES, *Clim. Dyn.*, *39*(1–2), 329–346, doi:10.1007/s00382-011-1274-z.
Li, H., and V. Misra (2014), Global seasonal climate predictability in a two tiered forecast system. Part II: Boreal winter and spring seasons, *Clim. Dyn.*, *42*(5–6), 1449–1468.
Li, S., L. Goddard, and D. G. DeWitt (2008), Predictive skill of AGCM seasonal climate forecasts subject to different SST prediction methodologies, *J. Clim.*, *21*(10), 2169–2186.
Ma, F., X. Yuan, and A. Ye (2015), Seasonal drought predictability and forecast skill over China, *J. Geophys. Res. Atmos.*, *120*, 8264–8275, doi:10.1002/2015JD023185.
McFarlane, N. A., G. Boer, J. Blanchet, and M. Lazare (1992), The Canadian Climate Centre second-generation general circulation model and its equilibrium climate, *J. Clim.*, *5*(10), 1013–1044.
Merryfield, W. J., W.-S. Lee, G. J. Boer, V. V. Kharin, J. F. Scinocca, G. M. Flato, R. Ajayamohan, J. C. Fyfe, Y. Tang, and S. Polavarapu (2013), The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization, *Mon. Weather Rev.*, *141*(8), 2910–2945, doi:10.1175/MWR-D-12-00216.1.

Min, Y.-M., V. N. Kryjov, and S. M. Oh (2014), Assessment of APCC multimodel ensemble prediction in seasonal climate forecasting: Retrospective (1983–2003) and real-time forecasts (2008–2013), *J. Geophys. Res. Atmos.*, *119*, 12,132–12,150, doi:10.1002/2014JD022230.
Misra, V., H. Li, Z. Wu, and S. DiNapoli (2014), Global seasonal climate predictability in a two tiered forecast system: Part I: Boreal summer and fall seasons, *Clim. Dyn.*, *42*(5–6), 1425–1448.
Palmer, T., Č. Branković, and D. Richardson (2000), A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations, *Q. J. R. Meteorol. Soc.*, *126*(567), 2013–2033.
Palmer, T. N., et al. (2004), Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (Demeter), *Bull. Am. Meteorol. Soc.*, *85*(6), 853–872, doi:10.1175/bams-85-6-853.
Rasmusson, E. M., and T. H. Carpenter (1982), Variations in tropical sea surface temperature and surface wind fields associated with the Southern Oscillation/El Niño, *Mon. Weather Rev.*, *110*, 354–384.
Richardson, D. S. (2006), Predictability and economic value, in *Predictability of Weather and Climate*, edited by T. Palmer and R. Hagedorn, pp. 628–644, Cambridge Univ. Press, Cambridge, U. K., doi:10.1017/CBO9780511617652.026.
Ritchie, H. (1991), Application of the semi-Lagrangian method to a multilevel spectral primitive-equations model, *Q. J. R. Meteorol. Soc.*, *117*(497), 91–106.
Rodwell, M. J., and C. K. Folland (2002), Atlantic air-sea interaction and seasonal predictability, *Q. J. R. Meteorol. Soc.*, *128*, 1413–1443.
Saha, S., S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H. Van den Dool, H.-L. Pan, S. Moorthi, and D. Behringer (2006), The NCEP climate forecast system, *J. Clim.*, *19*(15), 3483–3517.
Saha, S., et al. (2014), The NCEP climate forecast system version 2, *J. Clim.*, *27*, 2185–2208.
Shukla, J. (1998), Predictability in the midst of chaos: A scientific basis for climate forecasting, *Science*, *282*(5389), 728–731.
Shukla, J., L. Marx, D. Paolino, D. Straus, J. Anderson, J. Ploshay, D. Baumhefner, J. Tribbia, C. Brankovic, and T. Palmer (2000), Dynamical seasonal prediction, *Bull. Am. Meteorol. Soc.*, *81*(11), 2593–2606.
Smith, T. M., and R. W. Reynolds (2004), Improved extended reconstruction of SST (1854–1997), *J. Clim.*, *17*(12), 2466–2477.
Song, F., and T. Zhou (2014), The climatology and interannual variability of East Asian summer monsoon in CMIP5 coupled models: Does air–sea coupling improve the simulations?, *J. Clim.*, *27*(23), 8761–8777.
Sperber, K. R., H. Annamalai, I. S. Kang, A. Kitoh, A. Moise, A. Turner, B. Wang, and T. Zhou (2012), The Asian summer monsoon: An intercomparison of CMIP5 vs. CMIP3 simulations of the late 20th century, *Clim. Dyn.*, *41*(9–10), 2711–2744, doi:10.1007/s00382-012-1607-6.
Stefanova, L., V. Misra, J. J. O'Brien, E. P. Chassignet, and S. Hameed (2012), Hindcast skill and predictability for precipitation and two-meter air temperature anomalies in global circulation models over the Southeast United States, *Clim. Dyn.*, *38*(1–2), 161–173.
Tang, W., Z. Lin, and L. Luo (2013), Assessing the seasonal predictability of summer precipitation over the Huaihe River basin with multiple APCC models, *Atmos. Oceanic Sci. Lett.*, *6*(4), 185–190.
Tang, Y., H. Lin, and A. M. Moore (2008), Measuring the potential predictability of ensemble climate predictions, *J. Geophys. Res.*, *113*, D04108, doi:10.1029/2007JD008804.
Tang, Y., D. Chen, and X. Yan (2014), Potential predictability of Northern America surface temperature in AGCMs and CGCMs, *Clim. Dyn.*, *45*(1–2), 353–374.
Tippett, M. K., T. DelSole, and A. G. Barnston (2014), Reliability of regression-corrected climate forecasts, *J. Clim.*, *27*(9), 3393–3404.
Toth, Z., O. Talagrand, and Y. Zhu (2006), The attributes of forecast systems: A general framework for the evaluation and calibration of weather forecasts, in *Predictability of Weather and Climate*, edited by T. Palmer and R. Hagedorn, pp. 584–595, Cambridge Univ. Press, Cambridge, U. K., doi:10.1017/CBO9780511617652.026.
Wang, B., and LinHo (2002), Rainy season of the Asian-Pacific summer monsoon, *J. Clim.*, *15*(4), 386–398.
Wang, B., R. Wu, and X. Fu (2000), Pacific-East Asian teleconnection: How does ENSO affect East Asian climate?, *J. Clim.*, *13*(9), 1517–1536.
Wang, B., R. Wu, and K. Lau (2001), Interannual variability of the Asian summer monsoon: contrasts between the Indian and the Western North Pacific-East Asian Monsoons, *J. Clim.*, *14*(20), 4073–4090.
Wang, B., R. Wu, and T. Li (2003), Atmosphere-warm ocean interaction and its impact on Asian-Australian monsoon variation, *J. Clim.*, *16*, 1195–1211.
Wang, B., et al. (2005), Fundamental challenge in simulation and prediction of summer monsoon rainfall, *Geophys. Res. Lett.*, *32*, L15711, doi:10.1029/2005GL022734.
Wang, B., et al. (2008a), How accurately do coupled climate models predict the Asian-Australian monsoon interannual variability?, *Clim. Dyn.*, *30*, 605–619, doi:10.1007/s00382-007-0310-5.
Wang, B., Z. Wu, J. Li, J. Liu, C.-P. Chang, Y. Ding, and G. Wu (2008b), How to measure the strength of the East Asian summer monsoon, *J. Clim.*, *21*(17), 4449–4463, doi:10.1175/2008jcli2183.1.
Wang, B., J.-Y. Lee, I.-S. Kang, J. Shukla, C.-K. Park, A. Kumar, J. Schemm, S. Cocke, J.-S. Kug, and J.-J. Luo (2009), Advance and prospectus of seasonal prediction: Assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004), *Clim. Dyn.*, *33*(1), 93–117.
Wang, B., B. Xiang, and J.-Y. Lee (2013), Subtropical high predictability establishes a promising way for monsoon and tropical storm predictions, *Proc. Natl. Acad. Sci. U.S.A.*, *110*(8), 2718–2722.
Webster, P. J., et al. (1998), Monsoons: Processes, predictability, and prospects for prediction, *J. Geophys. Res.*, *103*, 14,451–14,510.
Weigel, A. P., M. A. Liniger, and C. Appenzeller (2008), Can multimodel combination really enhance the prediction skill of probabilistic ensemble forecasts?, *Q. J. R. Meteorol. Soc.*, *134*, 241–260, doi:10.1002/qj.210.
Weisheimer, A., F. J. Doblas-Reyes, T. N. Palmer, A. Alessandri, A. Arribas, M. Déqué, N. Keenlyside, M. MacVean, A. Navarra, and P. Rogel (2009), ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs, *Geophys. Res. Lett.*, *36*, L21711, doi:10.1029/2009GL040896.
Wilks, D. S. (2011), *Statistical Methods in the Atmospheric Sciences*, *Int. Geophys. Ser.*, vol. 100, 3rd ed., Academic, San Diego, Calif.
Wu, R., and B. P. Kirtman (2005), Roles of Indian and Pacific Ocean air-sea coupling in tropical atmospheric variability, *Clim. Dyn.*, *25*, 155–170, doi:10.1007/s00382-005-0003-x.
Wu, R., B. P. Kirtman, and K. Pegion (2006), Local air-sea relationship in observations and model simulations, *J. Clim.*, *19*, 4914–4932, doi:10.1175/JCLI3904.1.
Xie, P., and P. A. Arkin (1996), Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions, *J. Clim.*, *9*(4), 840–858.
Xie, S.-P., K. Hu, J. Hafner, H. Tokinaga, Y. Du, G. Huang, and T. Sampe (2009), Indian Ocean capacitor effect on Indo-western Pacific climate during the summer following El Niño, *J. Clim.*, *22*(3), 730–747.
Yan, X., and Y. Tang (2013), An analysis of multimodel ensemble for seasonal climate predictions, *Q. J. R. Meteorol. Soc.*, *139*, 1389–1401, doi:10.1002/qj.2019.

Yang, D., Y. Tang, Y. Zhang, and X. Yang (2012), Information-based potential predictability of the Asian summer monsoon in a coupled model, *J. Geophys. Res.*, *117*, D03119, doi:10.1029/2011JD016775.

Yang, X.-Q., J. L. Anderson, and W. F. Stern (1998), Reproducible forced modes in AGCM ensemble integrations and potential predictability of atmospheric seasonal variations in the extratropics, *J. Clim.*, *11*(11), 2942–2959.

Yoo, J. H., and I.-S. Kang (2005), Theoretical examination of a multimodel composite for seasonal prediction, *Geophys. Res. Lett.*, *32*, L18707, doi:10.1029/2005GL023513.

Younas, W., and Y. Tang (2013), PNA predictability at various time scales, *J. Clim.*, *26*, 9090–9114, doi:10.1175/jcli-d-12-00609.1.

Zhu, J., and J. Shukla (2013), The role of air–sea coupling in seasonal prediction of Asia–Pacific summer monsoon rainfall, *J. Clim.*, *26*(15), 5689–5697, doi:10.1175/jcli-d-13-00190.1.

Zou, L., and T. Zhou (2013), Can a regional ocean–atmosphere coupled model improve the simulation of the interannual variability of the western North Pacific summer monsoon?, *J. Clim.*, *26*, 2353–2367.