Journal of Geophysical Research: Atmospheres

RESEARCH ARTICLE
10.1029/2017JD028002

On the Relationship Between Probabilistic and Deterministic Skills in Dynamical Seasonal Climate Prediction

Dejian Yang[1], Xiu-Qun Yang[1], Dan Ye[1], Xuguang Sun[1], Jiabei Fang[1], Cuijiao Chu[1], Tao Feng[1], Yiquan Jiang[1], Jin Liang[1], Xuejuan Ren[1], Yaocun Zhang[1], and Youmin Tang[2]

[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]Environmental Science and Engineering, University of Northern British Columbia, Prince George, British Columbia, Canada

Key Points:
- A nonlinear, monotonic relationship between probabilistic resolution skill and deterministic correlation skill is analytically derived
- The above theoretical relationship is consistent with what is observed in dynamical seasonal climate prediction

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

Citation:
Yang, D., Yang, X.-Q., Ye, D., Sun, X., Fang, J., Chu, C., et al. (2018). On the relationship between probabilistic and deterministic skills in dynamical seasonal climate prediction. *Journal of Geophysical Research: Atmospheres*, *123*, 5261–5283. https://doi.org/10.1029/2017JD028002

Received 2 NOV 2017
Accepted 7 MAY 2018
Accepted article online 17 MAY 2018
Published online 31 MAY 2018

Abstract Despite the gradually increasing emphasis on assessing the skill of dynamical seasonal climate predictions from the probabilistic perspective, there is a lack of in-depth understanding that an inherent relationship may exist between the probabilistic and deterministic seasonal forecast skills. In this study, we focus on investigating this relationship, through theoretical consideration based on an analytical approach and diagnostic analysis of the historical forecasts produced by multiple dynamical models. The probabilistic forecast skill is gauged in terms of its two different attributes: resolution and reliability, while the deterministic forecast skill is measured in terms of anomaly correlation (AC). Through the theoretical consideration under certain simplified assumptions, a nonlinear, monotonic relationship is analytically derived between the resolution and the AC. Subsequent diagnostic analysis shows that the resolution and AC skills of both the multimodel ensemble and its member single models indeed appear to be approximately monotonically and nonlinearly related, specifically when they are calculated in a zonally aggregated manner by which the impact of finite sample size is reduced. This observed relationship has a specific form that is consistent with what the theory predicts. In short, the theoretical result is well verified by the dynamical model forecasts. Diagnostic analysis also shows that no good relationship exists between the reliability and the AC, signifying the difference of reliability and resolution in nature. A specific application of the proven resolution-AC coherence is also demonstrated. The proved resolution-AC relationship can facilitate comparisons among various assessments of seasonal climate prediction skill from the deterministic or probabilistic perspective alone.

1. Introduction

Seasonal climate prediction is an important element in the climate prediction family, which aims at predicting the climate conditions in the coming one or several seasons and whose accuracy is of vital importance for governments to make rational decisions. In the past two decades, a variety of efforts have been made to develop complex general circulation models (GCMs) to perform seasonal climate prediction (e.g., Jia et al., 2015; Kanamitsu et al., 2002; Liu et al., 2015; Luo et al., 2008; MacLachlan et al., 2015; Merryfield et al., 2013; Molteni et al., 2011; Saha et al., 2006, 2014; Stockdale et al., 1998). Seasonal forecasting using dynamical models 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 address these uncertainties and alleviate their adverse effects on the forecasting accuracy, the strategy of ensembling different predictions has been frequently employed. To tackle the initial condition uncertainties, instead of one single run, an ensemble of predictions starting from slightly different initial conditions is produced with one specific GCM (e.g., Stockdale et al., 1998). This kind of ensemble is often referred to as single-model ensemble (SME). To handle the model uncertainties, rather than one single GCM, multiple GCMs are used to produce forecast trajectories individually, which are then combined to form a multimodel ensemble (MME) of predictions (e.g., Palmer et al., 2004).

Given the multiple runs from an ensemble, the resulting prediction can be made in two formats: the deterministic and probabilistic formats. Deterministic forecast aims to provide a quantitative estimate of the future value of climate elements, usually given by the ensemble mean. However, due to the presence of chaotic, unpredictable components in the observed seasonal atmospheric anomalies, so-called deterministic forecasting cannot be truly deterministic. Instead, a deterministic forecast should be more accurately

considered as a "point forecast" without an a priori estimate of the associated forecast uncertainty. In contrast, probabilistic forecast is not a "point forecast" of a climate variable but provides an estimate of the likelihood the variable may occur in a certain interval. In short, probabilistic forecasts aim at predicting the probability distribution for the future state of the variable. Compared with the deterministic format, the probabilistic format has been argued to be more informative and valuable (e.g., Palmer, 2002; Richardson, 2000, 2006). A forecast of the probability for the event of interest to occur can bring greater economic value for end users than a single deterministic forecast with uncertain accuracy (e.g., Richardson, 2006).

Aside from developing models and forecast systems, assessing and understanding their prediction skills is also an indispensable procedure in the study of dynamical seasonal prediction. The latter practice can provide important feedback for guiding the further activity in the former. Using model hindcast data, extensive works have been done in this respect. In many of these works, the assessments are mainly from the deterministic forecast angle. Several important aspects, such as physical sources of seasonal forecast skill (e.g., Butler et al., 2016; Chowdary et al., 2010; Kumar et al., 2013; Lee, Lee, et al., 2011; Lee, Wang, et al., 2011; Lee et al., 2013; Li et al., 2014; Manzanas et al., 2014; Quan et al., 2006; Rodwell et al., 1999; Scaife et al., 2017; Schlosser & Kirtman, 2005; Stockdale et al., 2015; Yang et al., 1998, 2012), the relation between model fidelity in simulating climatology and skill in predicting seasonal anomaly (e.g., DelSole & Shukla, 2010; Jia et al., 2012; Lee et al., 2010; Sperber & Palmer, 1996), one- versus two-tier modeling strategy (e.g., Beraki et al., 2015; Graham et al., 2005; Guérémy et al., 2005; Kug et al., 2008; Landman et al., 2012; Wang et al., 2005; Zhu & Shukla, 2013), and MME versus SME (e.g., Kang & Shukla, 2006; Krishnamurti, 1999; Krishnamurti et al., 2000; Pavan & Doblas-Reyes, 2000; Peng et al., 2002; Yoo & Kang, 2005) were explored and discussed, in general terms or in the context of predicting specific climate phenomena. In view of the importance of probabilistic forecast, quite a few recent studies have started to pay due attention to or even concentrate on assessing the forecast skill from the probabilistic angle (e.g., Alessandri et al., 2011; Becker & van den Dool, 2015; Hagedorn et al., 2005; Kharin et al., 2009; Kirtman et al., 2014; Min et al., 2017; Palmer et al., 2004; Sohn et al., 2012; Tippett et al., 2017; Wang et al., 2009; Weisheimer et al., 2009; Yan & Tang, 2013; Yang et al., 2016). Many of these studies assessed the probabilistic forecast skill mainly from the aspect of MME versus SME and found that the MME's probabilistic skill is usually significantly better than that of each contributing SME.

While most studies assessed the deterministic and probabilistic skills independently, some studies have tried to compare their differences and similarities and found that probabilistic skill of dynamical seasonal forecasts seems to be related to their deterministic skill (e.g., Alessandri et al., 2011; Cheng et al., 2010; Sooraj et al., 2012; Wang et al., 2009; Yang et al., 2016). Actually, a study of the similar category can be found even in early 1990s in Barnston (1992). This kind of study is quite valuable, since if the differences and similarities of the probabilistic and deterministic skills are made transparent, this would provide potential clues for further understanding of the probabilistic skill. This point is particularly important, given that understanding probabilistic skill per se is, in general, not an easy task due to the explicitly probabilistic nature of the issue. In evaluating the model hindcasts from the Climate Prediction and Its Application to Society project, Wang et al. (2009) found a general nonlinear relationship between grid point probabilistic and deterministic skills for the prediction of tropical precipitation. The probabilistic skill examined by Wang et al. (2009) is the overall skill measured by the Brier skill score (BSS; Wilks, 2011). The overall probabilistic skill can be resolved in terms of two attributes that differ in nature, the reliability and the resolution (e.g., Toth et al., 2006; Wilks, 2011). The reliability refers to how consistent forecast probabilities are with the corresponding observed frequencies, whereas the resolution refers to how different the observed frequencies are from the (unconditional) climatological frequency. With the hindcasts from the Ensemble-Based Predictions of Climate Changes and their Impacts (ENSEMBLES; Weisheimer et al., 2009) project and calculating the skills in an area-aggregated manner, Yang et al. (2016) found little relationship between the reliability and the deterministic skill for the prediction of western North Pacific-East Asian summer monsoon. However, they identified an excellent, approximately monotonic relationship between the resolution and the deterministic skill. Given the meaning of the reliability, there seems to be no big surprise to not see a good relationship between the reliability and the deterministic skill, as argued in Yang et al. (2016). As a measure of the conditional biases of probabilistic forecasts, the reliability is, after all, determined by the biases in the underlying forecast probability density functions (PDFs), which are at the least affected by the biases in both their means and variances. In contrast, the (ensemble-mean) deterministic skill can only be affected by the errors in their means. As such, the nonexistence of a good relationship between the reliability and the deterministic skill seems natural. However,

how a good relationship between the resolution and the deterministic skill is possible remains puzzling. Considering that the resolution is an important source of probabilistic skill and the investigations of its relationship with the deterministic skill in previous studies are only preliminary, a systematic and mechanistic study on this subject is necessary.

In this study, we present an in-depth investigation of the relationship between the probabilistic and deterministic skills of seasonal forecasts through considerations from a theoretical point of view as well as practical diagnostic analysis of GCM forecast data, in order to facilitate comparisons among various assessments of seasonal climate prediction skill from the deterministic or probabilistic perspective alone. In particular, we demonstrate that under certain ideal conditions, a nonlinear, monotonic relationship exists in theory between the resolution skill and the deterministic skill measured by anomaly correlation (AC). This theoretical relationship can be verified by GCM forecast data. The paper is structured as follows. Section 2 describes the used GCM forecast data, verification data, and measures of prediction skill. Section 3 presents an analytical consideration, where a theoretical relationship is derived between the resolution skill and the AC-based deterministic skill. Section 4 investigates the resolution-AC relationship in GCM seasonal forecasts, with a focus on investigating how the theoretical result derived in section 3 is verified by the GCM forecasts. Summary and discussion are provided in section 5.

2. Data and Skill Measures

2.1. Data

The GCM forecast data used in this study are extracted from the archive of the multimodel seasonal-to-annual historical forecast outputs of the ENSEMBLES project. 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. The historical forecasts with the five models were performed for the 46-year period from 1960 to 2005. For each year, 7-month-long seasonal forecasts 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 to 14- month-long annual forecast. For each model, an ensemble of nine integrations starting from different initial conditions of atmosphere and ocean was generated for a forecast. An equally weighted combination of these nine-member SMEs gives rise to a 45-member MME. A detailed description about the ENSEMBLES' models and historical forecasts can be found in Weisheimer et al. (2009).

In this study, the GCM prediction skills are analyzed for 200-hPa geopotential height, 500-hPa geopotential height, surface air temperature (SAT), and precipitation. The former two are typical variables that represent atmospheric circulation, and the latter two are of important socioeconomic impact. The verification data for the geopotential heights and the SAT are from the National Centers for Environmental Prediction/National Center for Atmospheric Research Reanalysis product (Kalnay et al., 1996), while the verification data for precipitation are from the Climate Prediction Center Merged Analysis of Precipitation data set (Xie & Arkin, 1996). While the prediction skills of the geopotential heights and SAT are calculated for the whole period of 1960–2005, the skills of the precipitation are calculated only for the period of 1979–2005, due to the shorter time coverage of the precipitation verification data. We only consider 1-month lead-time seasonal mean forecasts (target months 2–4 averages). As such, the four seasons of DJF (December-January-February), MAM (March-April-May), JJA (June-July-August), and SON (September-October-November) are targeted.

2.2. Measures of Prediction Skill

In this section, the used forecast skill measures are described. For deterministic forecast, the skill is measured by the AC as formulated in Saha et al. (2006). As a measure of the linear association between predicted and observed anomalies, the AC is the most frequently used measure of deterministic skill in the area of seasonal prediction verification. The AC is evaluated for the ensemble-mean prediction, as in previous studies. For probabilistic forecast, the skill is measured in terms of reliability, resolution, and the overall BSS. Three target events are considered, which are the so-called below-, near-, and above-normal events, defined based on the terciles of the observed climatology. The three categorical events so defined have an equal climatological frequency of 1/3. The forecast probability of an event is estimated as the fraction of ensemble members that

forecast the event (i.e., the simple nonparametric counting method). In this study, we calculate prediction skills of historical forecasts based on standardized anomalies; that is, at each grid point, both the single-model hindcasts and observations are normalized with respect to their own local climatologies. To avoid overfitting, the skills are calculated in a cross-validated mode using the so-called leave-one-out method; that is, 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.

As mentioned in the introduction, the reliability and the resolution are two essentially different attributes that characterize the probabilistic skill. Reliability quantifies how well forecast probabilities are consistent with the corresponding observed frequencies. A good reliability requires that in the cases where our probability forecasts for an event are p, the event indeed occurs on a fraction around p of these cases. This means that the practical evaluation of reliability necessarily needs involving some grouping of prediction cases. The most common way for doing this is the so-called binning method (e.g., Atger, 2004). Let us denote forecast probability as y_j and the corresponding observed outcome as o_j, taking the value 1 if the event actually occurs and 0 otherwise, where the index j denotes a numbering of the considered series of forecast-observation pairs, and partition the whole probability range of [0, 1] into K (taken to 10 in this study, as in previous research) non-overlapping bins of equal width. Then, for the bin k, a binned forecast probability (denoted $\overline{y}_k$) is obtained as the arithmetic mean of all the y_j falling into this bin, and the corresponding observed frequency (denoted $\overline{o}_k$) is obtained as the arithmetic mean of the o_j associated with these y_j. As such, a useful measure of reliability can be defined as (e.g., Wilks, 2011)

$$\mathrm{REL} = \frac{1}{N}\sum_{k=1}^{K} N_k(\overline{y}_k - \overline{o}_k)^2, \tag{1}$$

where N_k is the number of the y_j falling into the bin k and N is the total number of the considered series of forecast-observation pairs. Obviously, REL is negatively oriented; the larger, the worse. In light of the classic statistical interpretation of probability, the concept and formulation of reliability is straightforward. However, reliability alone is not sufficient for a probabilistic forecast system to be skillful. Another important aspect of probabilistic skill is the resolution, which refers to the extent to which the conditional observed frequencies differ from the unconditional observed climatological probability. With the binning method, the resolution can be measured as (e.g., Wilks, 2011)

$$\mathrm{RES} = \frac{1}{N}\sum_{k=1}^{K} N_k(\overline{o}_k - \overline{o})^2, \tag{2}$$

where $\overline{o} = \frac{1}{N}\sum_{j=1}^{N} o_j = \frac{1}{N}\sum_{k=1}^{K} N_k\overline{o}_k$ is the observed climatological probability. The importance of the resolution as a source of probabilistic skill can be shortly elucidated as follows. Besides depending on the reliability, the skill should intuitively also depend on how different the observed frequencies are from the baseline probability of 0.5; the more different, the better the skill. This is because this difference indicates the degree of certainty of the forecasting situation and the skill ought to be positively related to this certainty degree. However, trivial forecasts using historical climatological probability will be also able to display some skill in this sense, as long as it differs from 0.5. Nevertheless, the skill of our real concern should be related to the extra predictive information provided by the forecasts beyond the historical climatological information. Therefore, rather than depending upon the difference between the observed frequencies and 0.5, the skill should eventually depend on the difference between the observed frequencies and the observed climatological probability, which is just measured by the resolution as in equation (2).

The overall probabilistic skill is quantified by the BSS, an integrated measure of both reliability and resolution. The BSS is defined based on the Brier score (BS), which measures the overall accuracy of probability forecasts: $\mathrm{BS} = \frac{1}{N}\sum_{j=1}^{N}\left(y_j - o_j\right)^2$. It turns out that the BS can be decomposed into three terms (e.g., Wilks, 2011):

$$\mathrm{BS} = \mathrm{REL} - \mathrm{RES} + \underbrace{\overline{o}(1-\overline{o})}_{\mathrm{UNC}}, \tag{3}$$

where the first two terms are exactly the reliability and the resolution formulated in equations (1) and (2). The third term is known as the *uncertainty* term, which is independent of forecasts and only a function of event climatological probability. This term just indicates the a priori information provided by the climatology. Actually, the BS of climatological forecasts turns out to be equal to this term, that is, $BS_{clim} = UNC$.

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

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

Unlike the BS, the BSS is positively oriented and contains a builtin 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 the fact that $BS_{clim} = UNC$, the BSS can be further expressed as (Kharin & Zwiers, 2003)

$$BSS = \frac{RES}{UNC} - \frac{REL}{UNC} \equiv BSS_{RES} - BSS_{REL}. \tag{5}$$

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 analyzed in this study. For the tercile-based categorical events considered here, $\overline{o} = 1/3$ by definition, $UNC = \overline{o}(1-\overline{o}) = 2/9$, and therefore BSS_{RES} and RES (also, BSS_{REL} and REL) only differ by a factor of 9/2.

Besides the grid point skill, for the sake of discussion, we also calculate a "zonally aggregated" skill that involves pooling data along a latitude circle. For deterministic forecast, the predicted and observed anomalies that are spatially aggregated are still defined with respect to their own local climatological means. For probabilistic forecast, the very categorical event for which the forecast probabilities and the corresponding observed binary outcomes are spatially aggregated is also still defined for each grid cell based on respective long-term climatologies (Hamill & Juras, 2006).

3. A Theoretical Consideration of the Resolution-AC Relationship

In this section, we study the relationship between the probabilistic resolution skill and the deterministic AC skill from a theoretical point of view. Specifically, we demonstrate that under certain assumptions, a monotonic resolution-AC relationship can be derived in theory. These assumptions include that the underlying forecast PDFs are Gaussian, that the variances of the forecast PDFs are homogeneous (i.e., constant from case to case), and that the means of the forecast PDFs and the corresponding observations are joint Gaussian distributed. The Gaussian assumptions are widely used in the seasonal climate prediction studies and can be justified at least as a first approximation of the truth (e.g., Weigel et al., 2008, 2009; Wilks, 2002, 2011). The homogeneous forecast variance assumption has also been argued to be reasonable for the prediction of seasonally averaged variables (e.g., Kumar et al., 2000; Rowell, 1998; Tang et al., 2008; Van den Dool & Toth, 1991). We note in the end of this paragraph that from a practical prediction viewpoint, since good reliability is generally considered to be achievable through after-the-fact calibrations and the resolution would represent the BSS of the calibrated forecasts, the derived relationship would also represent the relationship of the BSS after calibration with the AC, which may have important practical significance (also see the discussion in the final section).

Let p be a continuous random variable for the forecast probability and O be a discrete random variable representing the corresponding observed outcome for the event, one for occurrence and zero for nonoccurrence. Suppose that we have an infinite number of forecast-observation pairs (N), so that the number of bins (K) can be increased to be sufficiently large, and therefore, each bin width becomes sufficiently small, while the prediction cases pertaining to individual bins yet remain sufficiently many. Then, the *sample* observed frequency $\overline{o}_k$ in equation (2), as a function of $\overline{y}_k$, can be reasonably written in the form of a conditional probability as $P(O = 1|p)$, $\overline{o}$ can be written as an unconditional probability of $P(O = 1)$, and $\frac{N_k}{N}$ can be written as $f_p(p)\mathrm{d}p$, where $f_p(p)$ is a PDF of p such that $f_p(p)\mathrm{d}p$ is the relative frequency

of the forecast probabilities lying in the infinitesimal bin $[p, p + \mathrm{d}p]$. As such, after replacing the summation in equation (2) by an integral, the resolution for the tercile-based events can be formally written as (e.g., Palmer et al., 2000)

$$\mathrm{BSS_{RES}} = \frac{9}{2}\mathrm{RES} = \frac{9}{2}\int_0^1 f_p(p)[P(O=1|p) - P(O=1)]^2 \mathrm{d}p. \tag{6}$$

As argued in Yang et al. (2016), the very expression in the square bracket suggests that the resolution could be understood in terms of a more fundamental concept, the statistical dependence. A similar point of view was also proposed by Bröcker (2015). Specifically, the resolution exists if and only if there is statistical dependence between the probabilistic forecast and the event occurrence, which requires the conditional probability $P(O = 1|p)$ to differ from the unconditional probability $P(O = 1)$.

For theoretical convenience, we further take the ensemble size to infinity. As such, the forecast probability p is free of the sampling error and solely determined by the underlying forecast PDF, specifically as the latter's integral over the range defining the target event. In general cases, describing a forecast PDF requires the knowledge of all its moments. However, if the forecast PDF is Gaussian, it can be fully characterized only by its first two moments, the forecast mean μ and forecast variance σ_e^2, which implies that the forecast probability p is eventually a function of μ and σ_e^2 only. Further, if the forecast variance σ_e^2 is homogeneous, the variation of p can only be caused by the change in μ, which implies that equation (6) can be rewritten as

$$\mathrm{BSS_{RES}} = \frac{9}{2}\int_{-\infty}^{\infty} f_\mu(\mu)[P(O=1|\mu) - P(O=1)]^2 \mathrm{d}\mu, \tag{7}$$

where $f_\mu(\mu)$ represents the PDF of μ. As argued in Yang et al. (2016), the above equation indicates that the resolution could be further understood in terms of the statistical dependence between the ensemble mean (i.e., the forecast mean μ) and the event occurrence. Since $P(O = 1|\mu)$ and $P(O = 1)$ are, after all, determined by the underlying conditional PDF of the original continuous predictand x given μ and the unconditional PDF of x, respectively, the resolution would be ultimately determined by the statistical dependence between μ and x. This understanding has further led Yang et al. (2016) to speculate that a monotonic relationship would be obtainable between the resolution as expressed in equation (7) and the AC skill of the ensemble mean μ predicting the underlying continuous predictand x, if μ and x are joint Gaussian distributed, since the differences and similarities between the general concepts of statistical dependence and linear correlation have been 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 Gaussian distributed variables, they are intrinsically equivalent. In the following, we prove their speculation.

Let us use μ_c (σ_μ^2) to denote the unconditional climatological mean (variance) of the ensemble mean prediction μ and use $x_c(\sigma_x^2)$ to denote the climatological mean (variance) of the continuous predictand x. The correlation between μ and x, that is, the AC skill, is denoted as r. A standard result in statistics states that if μ and x are joint Gaussian distributed, then each of μ and x will have a respective Gaussian marginal PDF and the conditional PDF of x given μ (denoted $f_{x\,|\,\mu}(x|\mu)$) is a Gaussian PDF with mean $x_c + r\sigma_x(\mu - \mu_c)/\sigma_\mu$ and variance $(1 - r^2)\sigma_x^2$. That is, formally we have

$$f_\mu(\mu) = N\left(\mu_c, \sigma_\mu^2\right), \tag{8}$$

$$f_{x\,|\,\mu}(x|\mu) = N\left(\underbrace{x_c + r\frac{\sigma_x}{\sigma_\mu}(\mu - \mu_c)}_{E(x|\mu)}, \underbrace{(1 - r^2)\sigma_x^2}_{\mathrm{var}(x|\mu)}\right), \tag{9}$$

where the conventional notation $N(\cdot,\cdot)$ of Gaussian distribution has been used, with the first argument representing the distributional mean and the second the distributional variance. In equation (9), the notation $E(x|\mu)$ ($\mathrm{var}(x|\mu)$) indicates that the mean (variance) is a conditional mean (variance). Then, the $P(O = 1|\mu)$ in equation (7) can be calculated as the integral of $f_{x\,|\,\mu}(x|\mu)$ over the event range $[x_l, x_r]$,

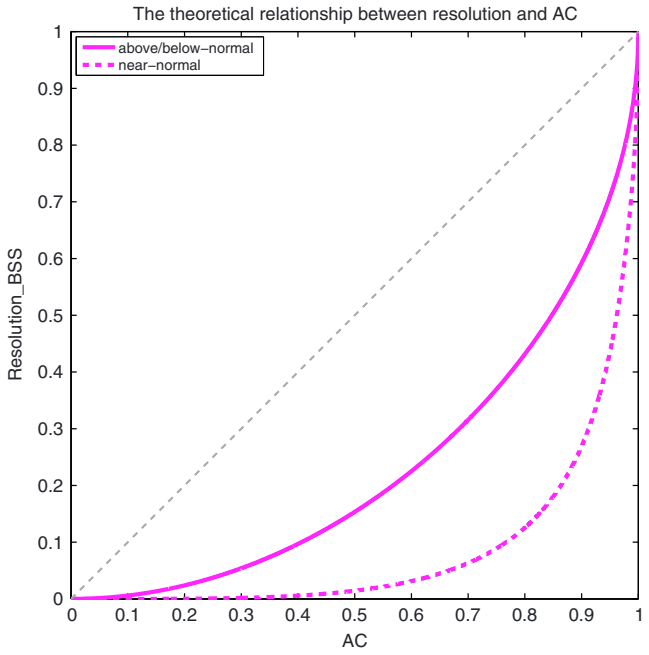

Figure 1. Theoretical relationship between the resolution skill for the probabilistic forecast and the anomaly correlation (AC) skill for the deterministic forecast. The solid curve is for the above- or below-normal event, while the dashed curve is for the near-normal event. BSS = Brier skill score.

which can be further expressed in the form of the cumulative distribution function $\Phi(\cdot)$ of the Gaussian distribution with mean zero and unit variance: $P(O=1|\mu)=\Phi\left((x_r-E(x|\mu))/\sqrt{\mathrm{var}(x|\mu)}\right)-\Phi\left((x_l-E(x|\mu))/\sqrt{\mathrm{var}(x|\mu)}\right)$. For the below-normal event, $x_l=-\infty$ and $x_r=x_c+\sigma_x\Phi^{-1}(1/3)$, where $\Phi^{-1}(\cdot)$ is the inverse function of $\Phi(\cdot)$ and $\Phi^{-1}(1/3)\approx-0.43$; for the above-normal event, $x_l=x_c+\sigma_x\Phi^{-1}(2/3)=x_c-\sigma_x\Phi^{-1}(1/3)$ and $x_r=+\infty$. As a result, we have

$$P(O_{\mathrm{BN}}=1|\mu)=\Phi\left(\frac{x_c+\sigma_x\Phi^{-1}(1/3)-x_c-r\frac{\sigma_x}{\sigma_\mu}(\mu-\mu_c)}{\sqrt{1-r^2}\sigma_x}\right)-0$$
$$=\Phi\left(\frac{\Phi^{-1}(1/3)-r\frac{(\mu-\mu_c)}{\sigma_\mu}}{\sqrt{1-r^2}}\right), \tag{10}$$

$$P(O_{\mathrm{AN}}=1|\mu)=1-\Phi\left(\frac{x_c-\sigma_x\Phi^{-1}(1/3)-x_c-r\frac{\sigma_x}{\sigma_\mu}(\mu-\mu_c)}{\sqrt{1-r^2}\sigma_x}\right)$$
$$=\Phi\left(\frac{\Phi^{-1}(1/3)+r\frac{(\mu-\mu_c)}{\sigma_\mu}}{\sqrt{1-r^2}}\right), \tag{11}$$

$$P(O_{\mathrm{NN}}=1|\mu)=1-P(O_{\mathrm{BN}}=1|\mu)-P(O_{\mathrm{AN}}=1|\mu). \tag{12}$$

Finally, after substituting the explicit expression for Gaussian PDF $f_\mu(\mu)$ into equation (7) and invoking the fact that $P(O=1)=1/3$, we can rewrite the resolution as

$$\mathrm{BSS}_{\mathrm{RES}}=\frac{9}{2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_c}{\sigma_\mu}\right)^2\right]\left[P(O=1|\mu)-\frac{1}{3}\right]^2\mathrm{d}\left(\frac{\mu-\mu_c}{\sigma_\mu}\right), \tag{13}$$

where the integration variable has been changed from μ to $(\mu-\mu_c)/\sigma_\mu$ and $P(O=1|\mu)$ has the explicit expression as in equations (10)–(12). Because the integrand in equation (13) appears as a function of r and $(\mu-\mu_c)/\sigma_\mu$ only, the integration result will be independent of $(\mu-\mu_c)/\sigma_\mu$ and therefore define a deterministic mapping from r to $\mathrm{BSS}_{\mathrm{RES}}$.

The integral in equation (13) has been numerically evaluated. The calculated resolution as a function of the AC over the interval [0, 1] is depicted in Figure 1. As seen, the resolution appears as a monotonic function of the AC. The theoretical resolution-AC relationships for the above-normal and below-normal events turn out to be the same, which should be the consequence of the symmetric property of Gaussian distribution. Moreover, this theoretical relationship is quite nonlinear: the change of the resolution with the AC is much less sharp when AC is small than when AC is large. The above feature also holds for the near-normal event. However, for the same AC that is not close to zero or one, the corresponding resolution for the near-normal event is much weaker than that for the other two events. This is related to the fact that the dependence of $P(O=1|\mu)$ on the "standardized signal perturbation" $(\mu-\mu_c)/\sigma_\mu$ for the former is usually much weaker than that for the latter (figure not shown). Van den Dool and Toth (1991) also studied the reasons as to why the near-normal categorical event often has a very low skill.

The resolution as a function of the AC over the interval $[-1, 0]$ has also been checked. The resolution-AC relationship for AC ranging from -1 to 0 turns out to appear just as a mirror of the relationship for AC ranging from 0 to 1 (figure not shown). Normally, forecasts with a negative AC should be understood as very bad forecasts, since it indicates that the forecasts tend to vary inversely with the observations. However, these forecasts can have effectively useful information, as once statistically corrected by a regression procedure, they would appear as skillful as the forecasts with a positive AC of the same magnitude. The insensitivity

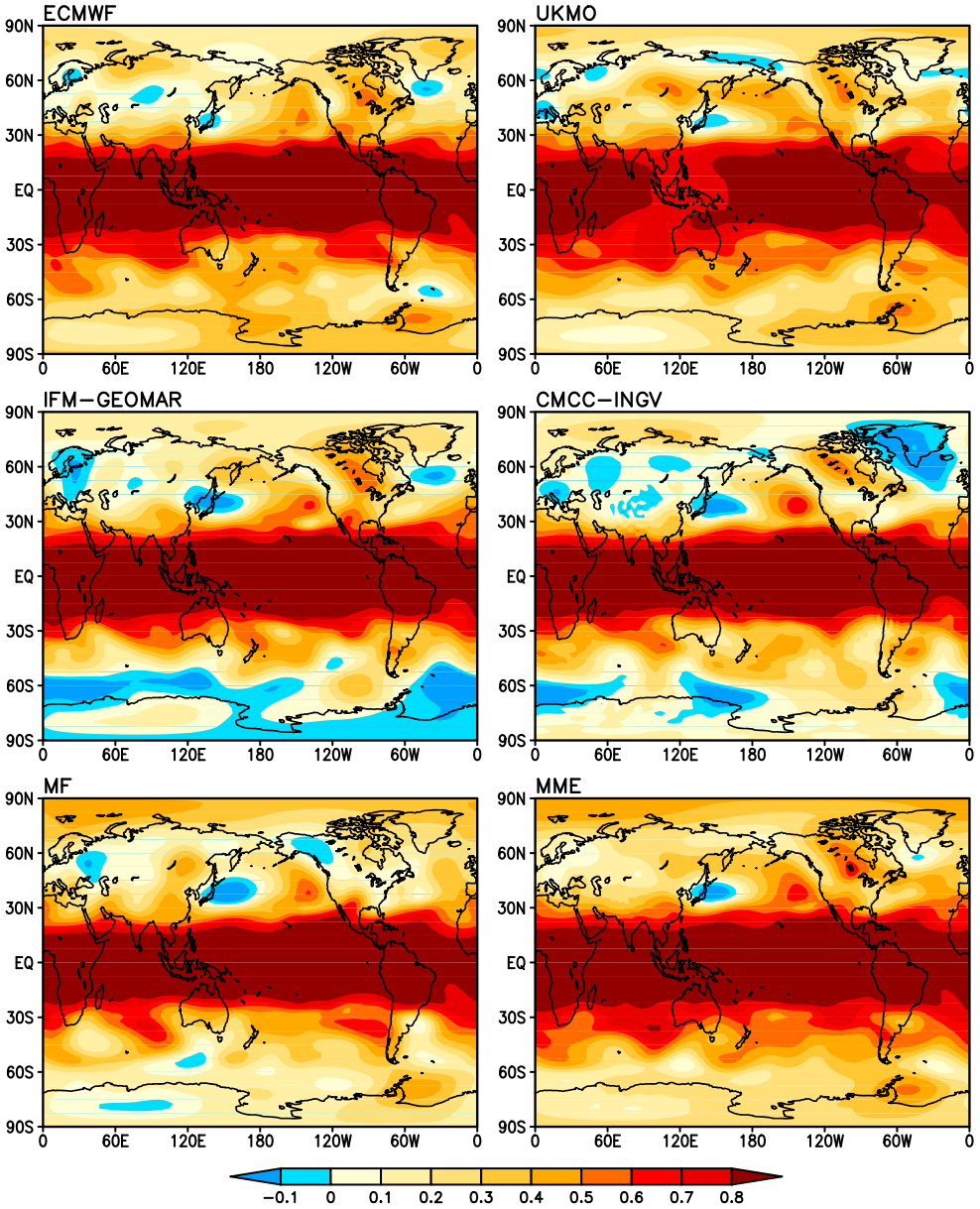

Figure 2. Spatial distributions of the anomaly correlation skill for the 1-month lead deterministic prediction of the December-January-February 200-hPa geopotential height (g200) anomaly by the ENSEMBLES's five models and their multimodel ensemble (MME) over the period of 1960–2005. UKMO = UK Met Office; MF = Météo France; ECMWF = European Centre for Medium-Range Weather Forecasts; IFM-GEOMAR = Leibniz Institute of Marine Sciences at Kiel University; CMCC-INGV = Euro-Mediterranean Centre for Climate Change.

of the resolution to the sign of AC means that it includes an automatic recognition of the effectively useful information provided by the forecasts with negative ACs.

In the following, for simplicity, we restrict our discussion of the resolution-AC relationship only to the case when the AC is nonnegative. As we shall see in the next section, most of the calculated AC skills of the considered GCM forecasts are positive. We also make a supplementary statement here that the monotonicity of the resolution-AC relationship mentioned throughout this paper implicitly refers in particular to when the AC is nonnegative. Apparently, the resolution is not a monotonic function of the AC over the whole range of [−1, 1].

We need to note that the above theoretical derivations have been inspired by Kharin and Zwiers (2003), where a probabilistic interpretation of seasonal potential predictability was proposed based on the BSS metric system. Specifically, they have provided a theoretical expression with a form similar to our equation (13). However, their theoretical expression was derived for describing the relationship between the resolution and the correlation between the true predictable signal in the observed system (denoted as β there) and the observed predictand. This correlation represents the *potential predictability* of the observed system (Kharin & Zwiers, 2003). Under the above relationship, the resolution will be independent of specific forecast models, solely determined by the intrinsic signal-to-noise property of the observed system. Their theoretical expression was derived with explicitly assuming that the observed frequency expressed as $P(O = 1|p)$ is essentially a function of the β only, which, according to our understanding, would be equivalent to impose a *perfect model* assumption that the predicted signal (i.e., the ensemble mean μ) is identical to the true predictable signal β. However, this assumption is generally unrealistic, given that current climate models are still far from perfect in reproducing the reality. Different from Kharin and Zwiers (2003), we have derived a theoretical relationship of the resolution directly with the correlation between the predicted signal and the observed predictand, which represents the *actual skill* we are interested here. This relationship has been argued to exist regardless of whether the predicted signal approaches the true predictable signal or not.

4. The Resolution-AC Relationship in GCM Seasonal Forecasts

In this section, we analyze the prediction skills of the ENSEMBLES's dynamical seasonal forecasts, with a focus on investigating how the theoretical result derived in the preceding section is verified by the GCM forecasts.

We first perform a spatial analysis of the forecast skills for the 1-month lead prediction of the DJF 200-hPa geopotential height (g200). Figure 2 shows the global spatial distributions of the AC-based deterministic skills. The general skill pattern characteristics for every SME and the MME are very similar. The most salient feature is a circumglobal belt of large skill within the tropics (30°S–30°N). Most tropical regions possess a skill level with AC greater than 0.7. This level is certainly quite good, as it means that more than a half of the observed variance can be explained by the prediction. A high deterministic prediction skill in the tropics is a common feature in climate models, primarily determined by more impact of tropical sea surface temperature variability and less impact of internal atmospheric variability (e.g., Charney & Shukla, 1981; Kharin & Zwiers, 2003; Shukla, 1998). Compared to the tropical skill, the prediction skill over the extratropics is generally lower. However, there are some specific geographical regions such as the Pacific-North American region where the prediction skill appears moderately good. A visual comparison between the subplots for the SMEs and the MME in Figure 2 reveals that overall, the MME's deterministic skill seems not impressively superior to those of the SMEs. Yang et al. (2016) also found a similar phenomenon in the prediction of western North Pacific-East Asian summer monsoon and discussed several possible factors that may render the MME improvement in the deterministic skill practically limited.

Figure 3 displays the probabilistic skills for the above- and below-normal events. In view of the fact that the skills for these two events are largely symmetrical, for saving space, the skills shown here are an average of them. Figure 3a shows the resolution skill. As seen, the spatial patterns of the resolution skill bear a strong resemblance to the AC skill patterns shown in Figure 2, characterized by a belt of strong skill within the tropics as well. In sharp contrast to the tropical skill, the extratropical skill is very weak. Nonetheless, signatures of exceptional skill also can be found in the specific regions with significant AC skill, such as the Pacific-North American region. Similar to the AC, the resolution also does not show a very distinguished MME-over-SMEs advantage. Figure 3b shows the reliability skill. For the SMEs, compared to the resolution and AC, the reliability does not show a particularly organized spatial pattern. The reliability contrast between tropics and extratropics is much less dramatic. For the MME, the reliability is largely geographically insensitive and visibly improved compared to the SMEs. Figure 3c shows the overall BSS skill. Dominated by the spatial variation of the resolution, the BSS also exhibits a remarkable meridional decrease from the tropics to the extratropics. In the SMEs, the BSSs in many extratropical areas are negative, signifying that the probabilistic forecasts over there are very bad, even worse than the climatological forecasts. The areas with negative BSS are significantly reduced in the MME.

Figure 4 displays the probabilistic skills for the near-normal event. For the resolution, while its strength is significantly weaker than that of the resolution for the other two events, its large-scale spatial distribution is still

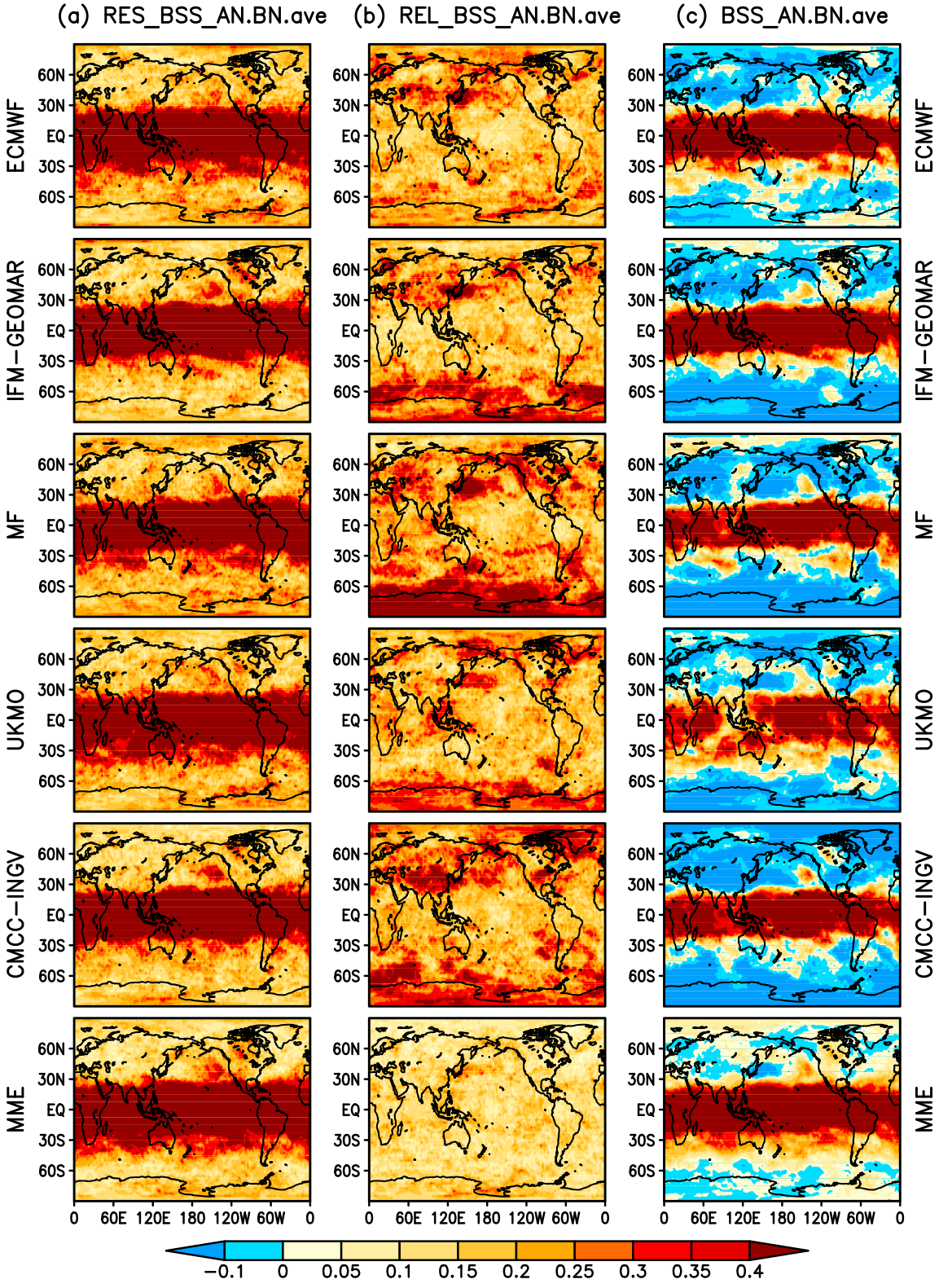

Figure 3. Spatial distributions of (a) resolution, (b) reliability, and (c) Brier skill score (BSS) skills for the 1-month lead probabilistic prediction of the December-January-February 200-hPa geopotential height anomaly by the ENSEMBLES's five models and their multimodel ensemble (MME) over the period of 1960–2005. The results shown here are for an average over the above-normal (AN) and below-normal (BN) events. UKMO = UK Met Office; MF = Météo France; ECMWF = European Centre for Medium-Range Weather Forecasts; IFM-GEOMAR = Leibniz Institute of Marine Sciences at Kiel University; CMCC-INGV = Euro-Mediterranean Centre for Climate Change.

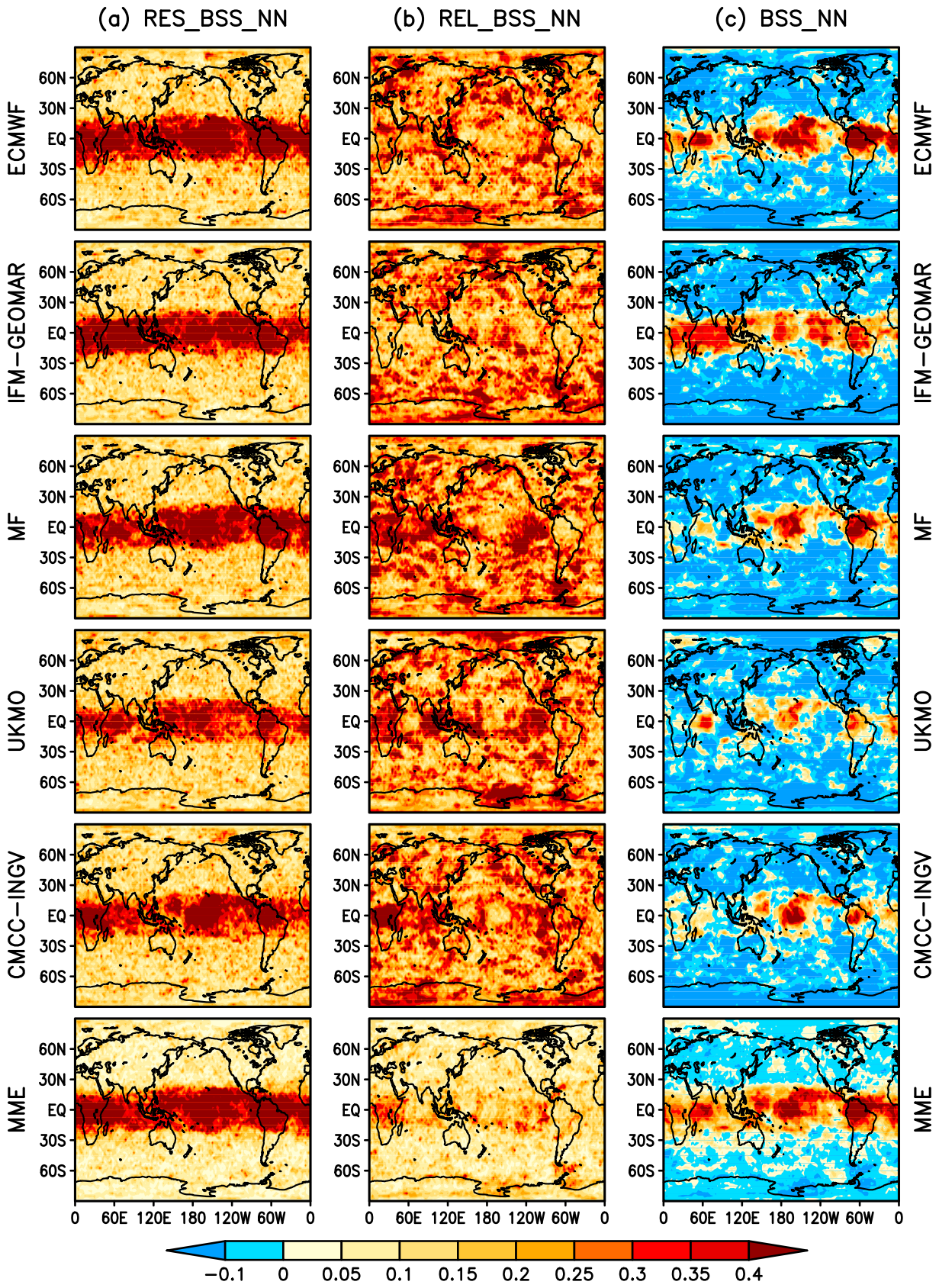

Figure 4. As in Figure 3, but for the near-normal (NN) event.

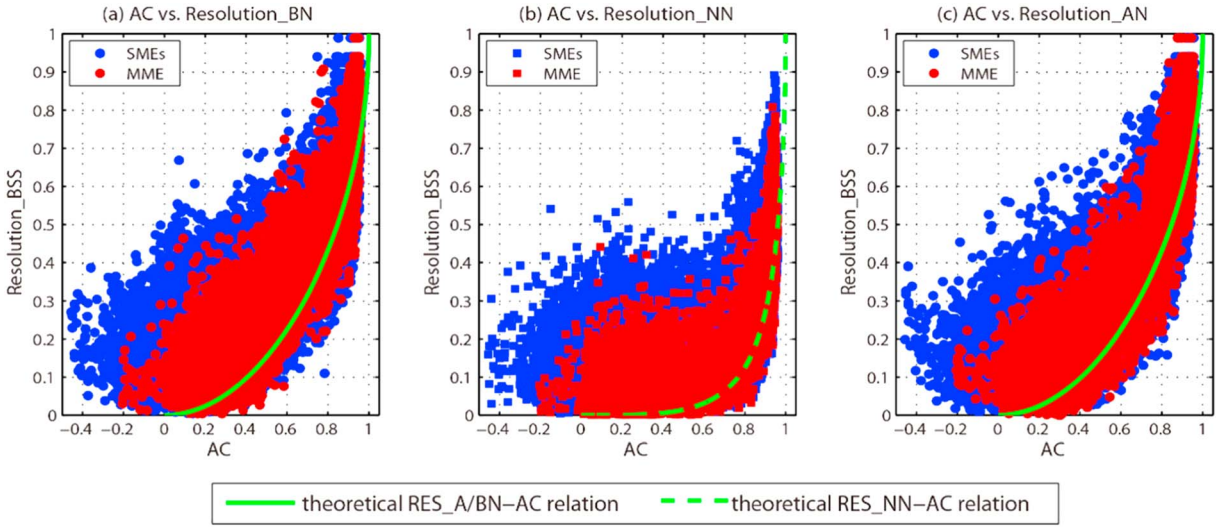

Figure 5. Scatterplots of resolution (*y* axis) versus anomaly correlation (AC; *x* axis) using the data of global grid point skills for the ENSEMBLES's 1-month lead prediction of the December-January-February 200-hPa geopotential height anomaly over the period of 1960–2005 for the (a) below-normal (BN), (b) near-normal (NN), and (c) above-normal (AN) events. The red points are for the multimodel ensemble (MME), while the blue points are for the five single-model ensembles (SMEs). The solid and dashed curves represent the same theoretical resolution-AC relationships as in Figure 1. BSS = Brier skill score.

organized, which fairly resembles those of the latter and the AC. For the reliability, its spatial distribution is very noisy in the SMEs, showing no spatial correspondence with the resolution and AC. The reliability in the MME is significantly improved relative to those in the SMEs. For the BSS, like the resolution, significant skill is mostly confined in the tropics.

All in all, the above skill spatial analysis confirms the qualitative correspondence between the resolution skill and the deterministic skill, which serves as a preliminary support to the theoretical conclusion. On this basis, we further investigate how the theoretically derived resolution-AC relationship is quantitatively verified.

To make a quantitative validation, we choose to depict the scatterplots of resolution versus AC along with the theoretical resolution-AC relationship. Figure 5 shows the scatterplots using the data of grid point skills for the 1-month lead prediction of the DJF g200. As seen, there indeed appears to be a general nonlinear covarying tendency between the resolution and the AC. However, their covarying relationship is far from what could be characterized as a monotonic relationship as suggested by the theory. In addition, since most of the scatters lie above the theoretical relationship, if an empirical resolution-AC relation were fitted, this relation would be biased upward from the theoretical one, especially for the SMEs. In short, the theoretical result is not satisfyingly validated in a quantitative manner by the data of skills calculated on a grid-box basis. However, this does not necessarily disprove the theory. Given the small temporal sample size, the estimated grid point skills of both the AC and the resolution would be usually subject to significant sampling uncertainty. Additionally, besides being susceptible to random sampling error, the grid point resolution skill calculated here would also likely suffer from a systematic overestimation bias. With the small sample size here, the number of forecast cases falling into each of the 10 bins would be not very sufficient for estimating a robust observed frequency. This undersampling within individual bins could lead to an overestimate of the resolution, which has been pointed out by the recent work of Bröcker (2012). Given the above discussions, it is therefore likely that the theoretical resolution-AC relationship being not well verified by the data of grid point skills only reflects the impact of the skill estimation errors, rather than pointing to the failure of the theory.

To better verify the theoretical relationship quantitatively, it seems more appropriate to use the data of skills calculated based on a larger sample size, which can be obtained here through pooling samples from different spatial locations. We select to pool samples on the same latitude circle together. In this way, we would be able to ensure that there are sufficiently many samples being pooled together and the pooled sample as a whole does not badly violate the assumptions required for deriving the theoretical relationship. As mentioned in section 2.2, the skills so calculated are referred to as zonally aggregated skills. A technical caution for

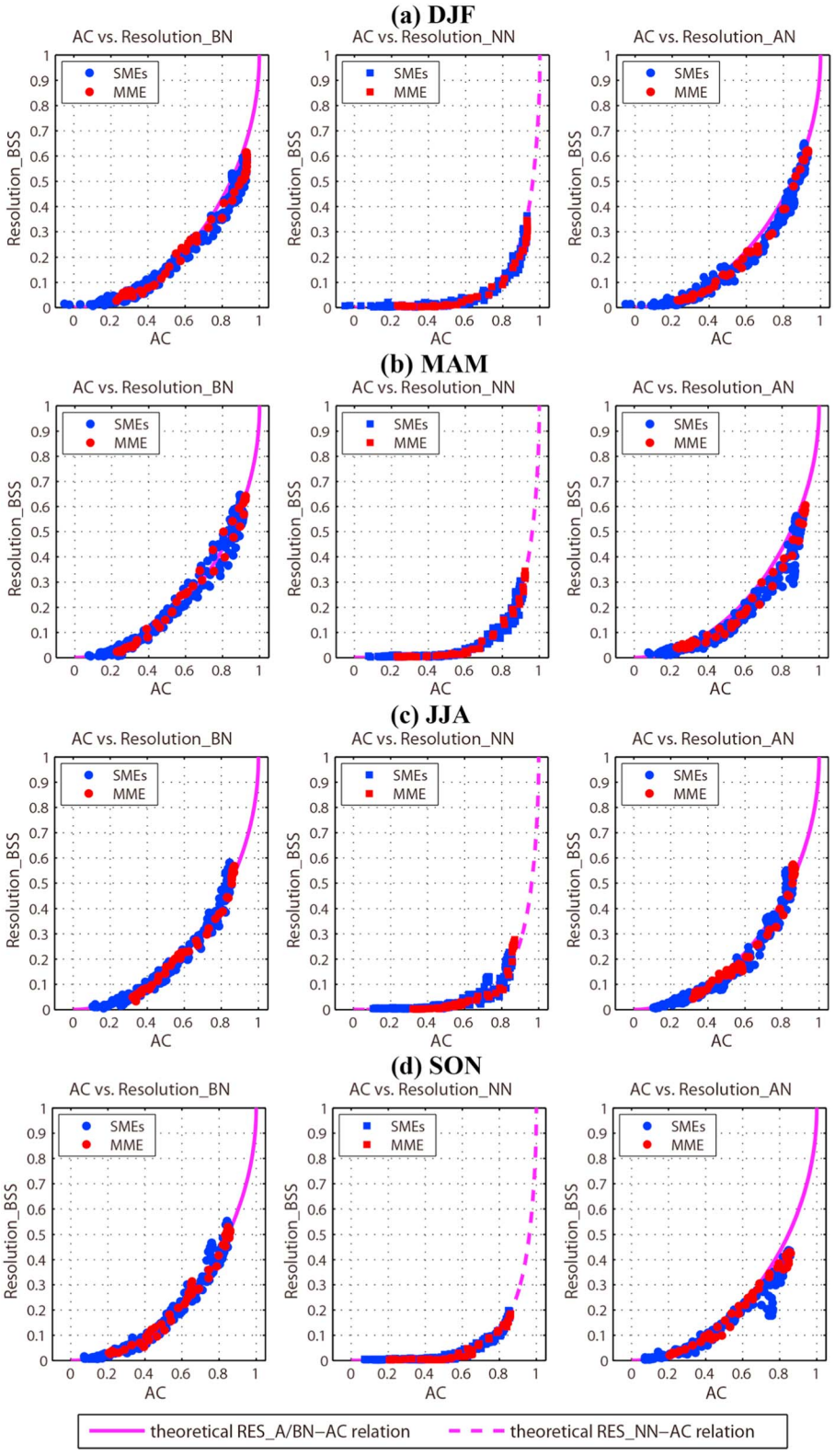

Figure 6. Scatterplots of resolution (*y* axis) versus anomaly correlation (AC; *x* axis) using the data of zonally aggregated skills (60°S–60°N) for the ENSEMBLES's 1-month lead prediction of 200-hPa geopotential height anomaly for the (a) December-January-February (DJF), (b) March-April-May (MAM), (c) June-July-August (JJA), and (d) September-October-November (SON) seasons over the period of 1960–2005. The solid and dashed curves represent the same theoretical resolution-AC relationships as in Figure 1. BSS = Brier skill score; MME = multimodel ensemble; SMEs = single-model ensembles; BN = below normal; NN = near normal; AN = above normal.

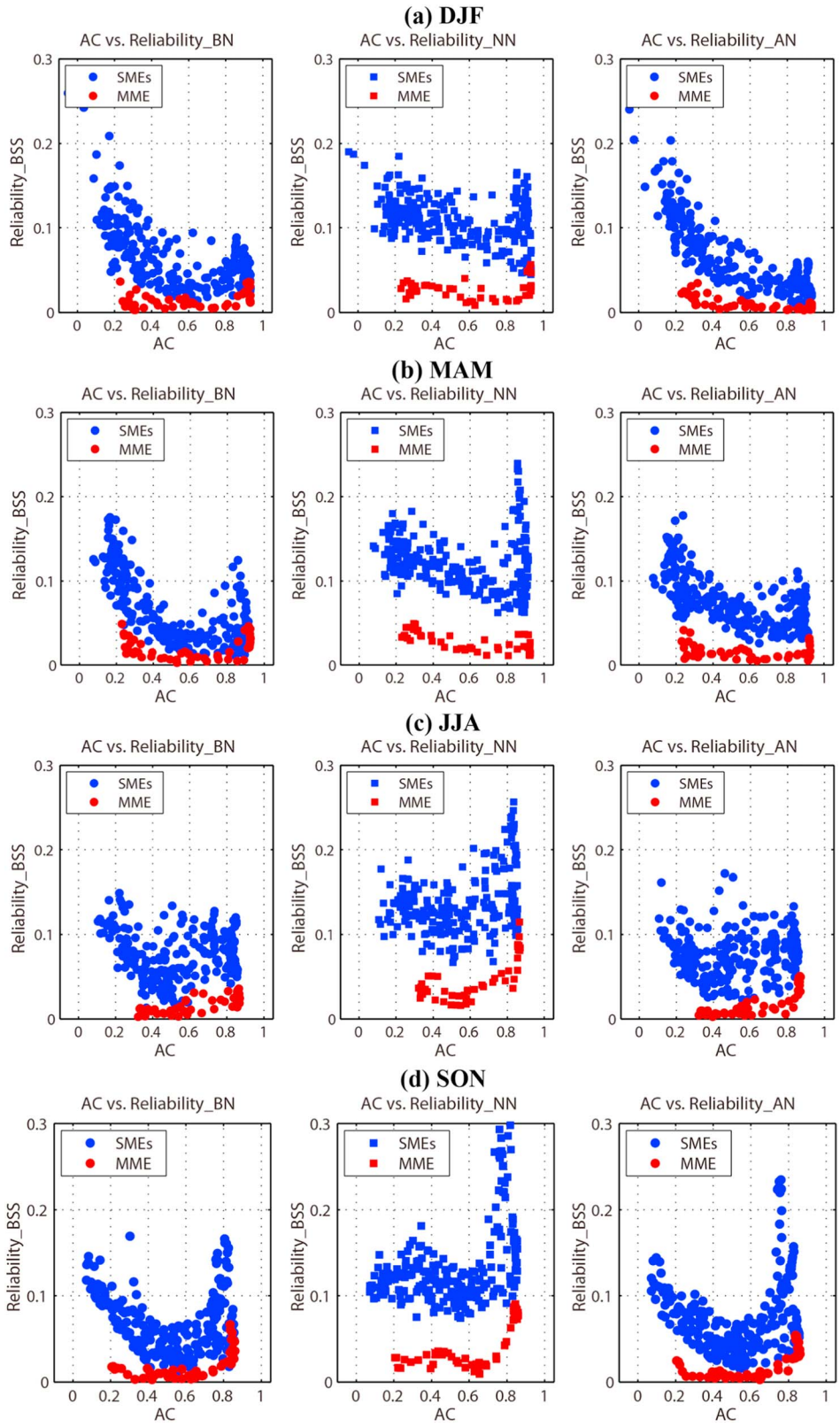

Figure 7. Scatterplots of reliability (*y* axis) versus anomaly correlation (AC; *x* axis) using the data of zonally aggregated skills (60°S–60°N) for the ENSEMBLES's 1-month lead prediction of 200-hPa geopotential height anomaly for the (a) December-January-February (DJF), (b) March-April-May (MAM), (c) June-July-August (JJA), and (d) September-October-November (SON) seasons over the period of 1960–2005. BSS = Brier skill score; MME = multimodel ensemble; SMEs = single-model ensembles; BN = below normal; NN = near normal; AN = above normal.

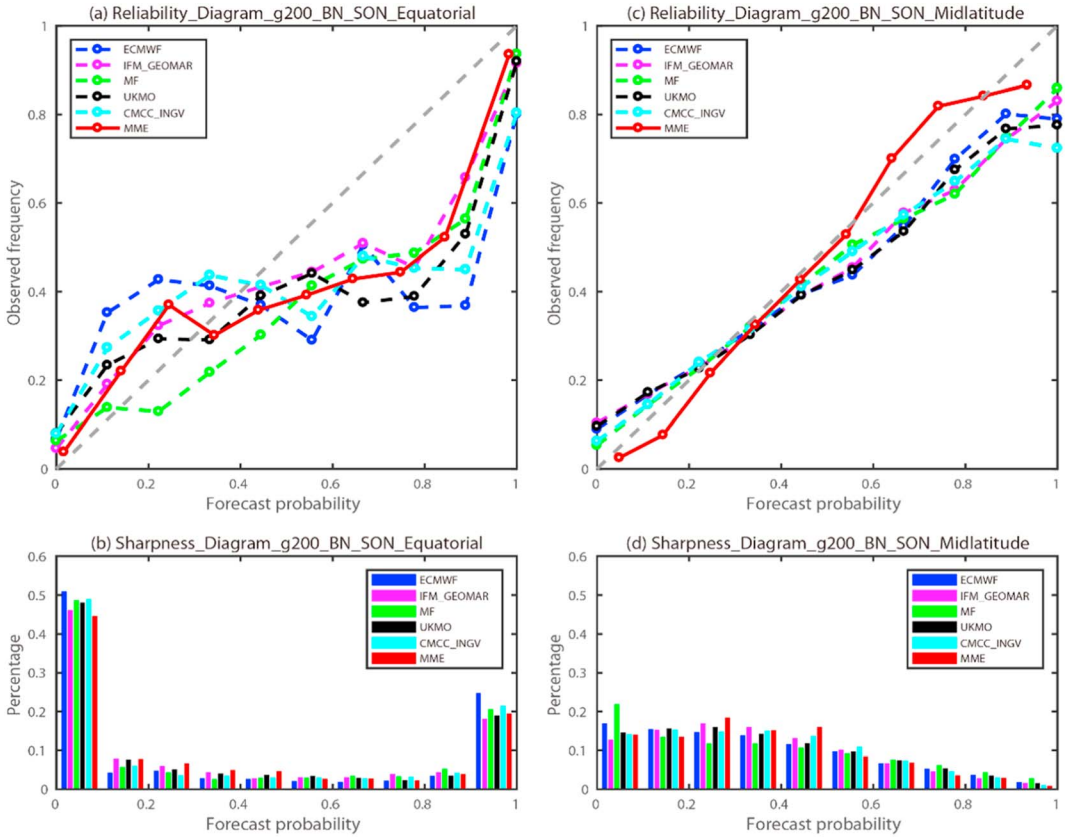

Figure 8. (a and c) Reliability and (b and d) sharpness diagrams for the 1-month lead probabilistic forecasts of the below-normal (BN) event for the September-October-November (SON) season aggregated within the equatorial zone of 10°S–10°N (left column) and for those aggregated within the Southern Hemisphere midlatitude zone of 45°S–25°S (right column), respectively. MME = multimodel ensemble; UKMO = UK Met Office; MF = Météo France; ECMWF = European Centre for Medium-Range Weather Forecasts; IFM-GEOMAR = Leibniz Institute of Marine Sciences at Kiel University; CMCC-INGV = Euro-Mediterranean Centre for Climate Change.

calculating these skills is also made there. Considering that the *effective* sample size at high latitudes might still be not large even after the pooling procedure, we only consider zonally aggregated skills for the latitudes between 60°S and 60°N. Figure 6a shows the scatterplots of resolution versus AC using the data of the zonally aggregated skills for the prediction of the DJF g200. Different from the situation seen in Figure 5, the relationship between the resolution and the AC for both the MME and SMEs and for all the three categorical events becomes strikingly good, able to be approximately characterized as a monotonic relationship, which is highly consistent with the theoretical counterpart. Figures 6b–6d further show the scatterplots for the prediction of g200 for the MAM, JJA, and SON seasons, respectively. As seen, for these seasons, the observed resolution-AC relationships are also well consistent with the theoretical counterpart, although to a slightly lesser extent.

For comparison, Figure 7 shows the scatterplots of zonally aggregated reliability against AC for the prediction of g200 for the four seasons. Overall, the scatters for reliability versus AC are much more complicatedly distributed than those for resolution versus AC shown in Figure 6. On the one hand, when the AC is not high, there tends to be a statistically negative relationship between the values of reliability and AC. Since the reliability calculated here is a negatively oriented quantity as mentioned in section 2.2, the seen negative relationships may just fit many readers' general perception. However, these negative relationships are still far from strong enough to be able to be approximately characterized as a monotonic relationship, which is emphasized throughout this paper. On the other hand, when the AC is high, there seems to be an "unexpected" signature that the values of reliability and AC are positively related. This signature is especially clear for the prediction for the SON season. To further demonstrate this lack of correspondence between the reliability

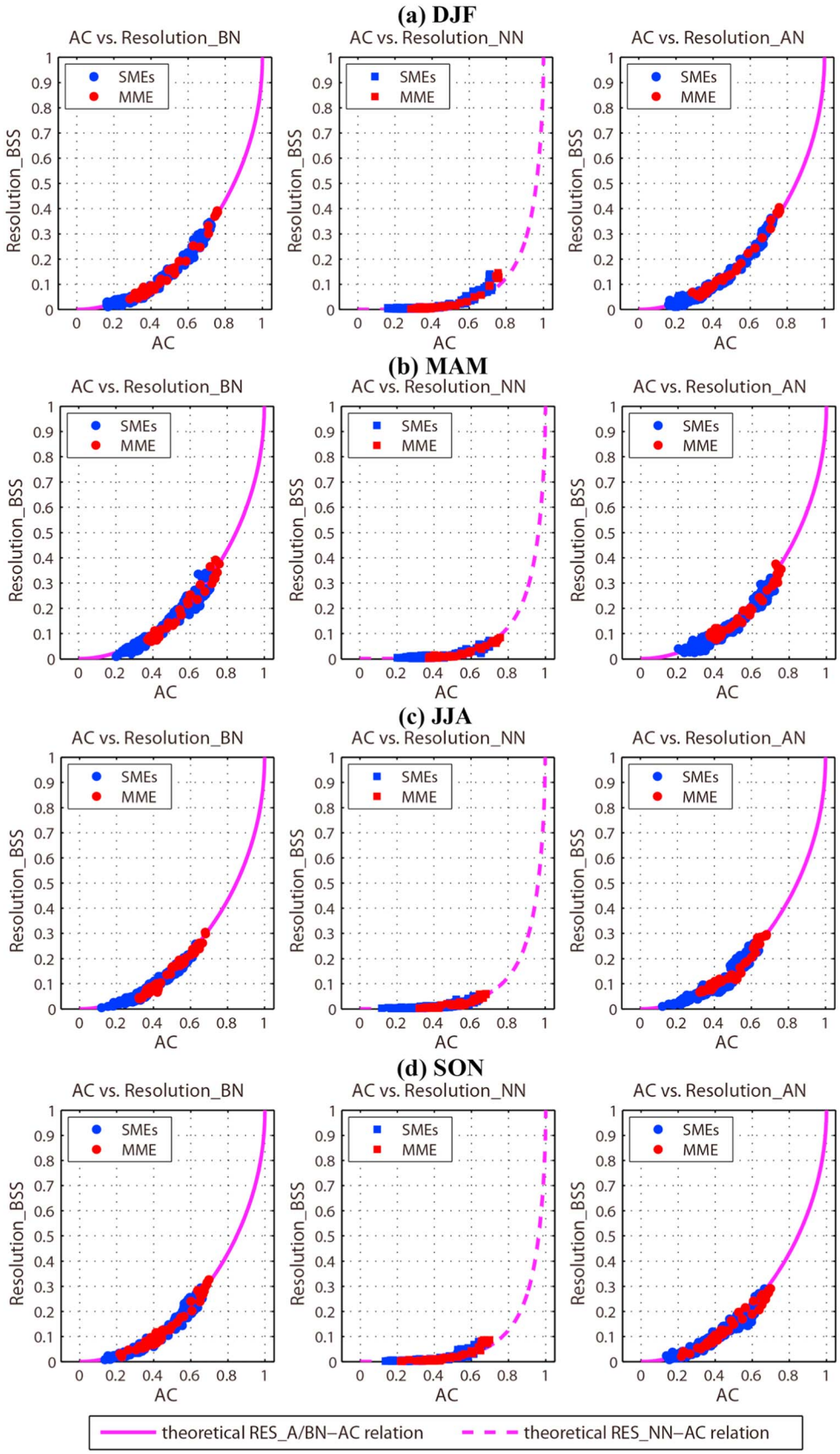

Figure 9. As in Figure 6, but for the prediction of surface air temperature anomaly.

and the AC, we present in Figures 8a and 8c the reliability diagrams, the observed frequency against the forecast probability, for the probabilistic forecasts of the below-normal event for the SON season aggregated within the equatorial zone of 10°S–10°N and for those aggregated within the Southern Hemisphere midlatitude zone of 45°S–25°S, respectively. As checked in the distributions of the SON zonally aggregated skills as a function of latitude (figure not shown), the equatorial zone has the worst reliabilities while the midlatitude zone has the best, despite that the former region has significantly better skills of the AC and resolution than the latter. As shown in Figures 8a and 8c, the tropical prediction's curves are significantly more deviated from the diagonal, the perfect reliability line, than the midlatitude prediction's, clearly illustrating the former's worse reliability characteristics as compared to the latter's. Figures 8b and 8d further show the associated sharpness diagrams, in which the percentages of probability forecasts falling into different probability bins are displayed. As can be seen, the tropical prediction's percentages are significantly higher than the midlatitude prediction's in the two extreme bins, whereas the reverse is usually true in the rest, indicating a larger sharpness for the tropical prediction. However, as seen from here, similar to the resolution, a larger sharpness is not necessarily associated with a better reliability. As argued in Yang et al. (2016), unlike the resolution, the reliability is not slaved to the statistical dependence between forecast and observation. In contrast, it measures the conditional probability bias, that is, the average squared difference between p and $P(O = 1|p)$. Under the assumptions in the preceding section, the reliability would be determined by the bias in the underlying forecast PDF as manifested by the discrepancy between the Gaussian forecast PDF and the Gaussian conditional PDF $f_{x|\mu}(x|\mu)$ and would then be ultimately determined by the biases in its mean and variance as identified by the discrepancy between $E(x|\mu)$ and μ and that between $\mathrm{var}(x|\mu)$ and the forecast variance σ_e^2, respectively. Since there is no general reason to expect a necessary relationship between the AC and the two biases, especially the variance bias, the absence of a good relationship between the reliability and the AC seems not odd. Further studies are needed in the future toward developing nonprobabilistic diagnostics to understand the reliability.

The good agreement between the theoretical relationship of resolution and AC and the relationship that is observed is certainly not only seen in the prediction of g200. An almost equivalently good agreement is also seen in the prediction of the 500-hPa geopotential height (g500; figure not shown). G200 and g500 are the variables characterizing the atmospheric circulations in the upper and middle troposphere, respectively. Figures 9 and 10 further show the scatterplots of resolution versus AC for the predictions of SAT and precipitation, respectively, which are surface climate variables. For the predictions of the SAT and precipitation, the occurrence ranges of AC and resolution are smaller, with the maximum skills, especially for the precipitation, being significantly smaller than those for the predictions of g200 and g500. This is not surprising, given that surface atmospheric variables are usually considered to be less predictable than the overlying atmospheric circulation. As shown in Figure 9, the observed resolution-AC relationship for the prediction of SAT is very consistent with what the theory predicts. In contrast, the observed resolution-AC relationship for the prediction of precipitation is, however, less consistent with the theoretical counterpart (Figure 10). This is probably due to two reasons. First, the Gaussian assumptions may be too ideal for the precipitation, as argued in the literature (e.g., Kharin et al., 2017; Sardeshmukh et al., 2000). Second, the forecast variance for the precipitation may also fail to be homogeneous enough, as compared to the other variables (Yang et al., 2012). However, on the whole, the theoretical relationship can still be regarded as being satisfied fairly well by the zonally aggregated skills of the precipitation, especially in contrast to the situation seen in Figure 5. To test the potential impact of the non-Gaussianity of the precipitation on the result, we further applied a so-called Box-Cox transformation approach detailed in Weigel et al. (2009) to transform the precipitation data to be more Gaussian and then calculated the forecast skills based on the transformed data. As shown in Figure 11, the observed resolution-AC relationship for the transformed data is visibly more consistent with the theoretical counterpart than that for the original data. This implies that the non-Gaussianity of the precipitation is a main reason why the observed resolution-AC relationship for the precipitation prediction fails to be very consistent with the theoretical result.

5. Summary and Discussion

Recently, there has been increasing realization of the importance of assessing the skill of dynamical seasonal predictions from the probabilistic forecast angle. However, there remains a lack of in-depth understanding that the probabilistic skill of seasonal forecasts may be inherently related to its deterministic counterpart.

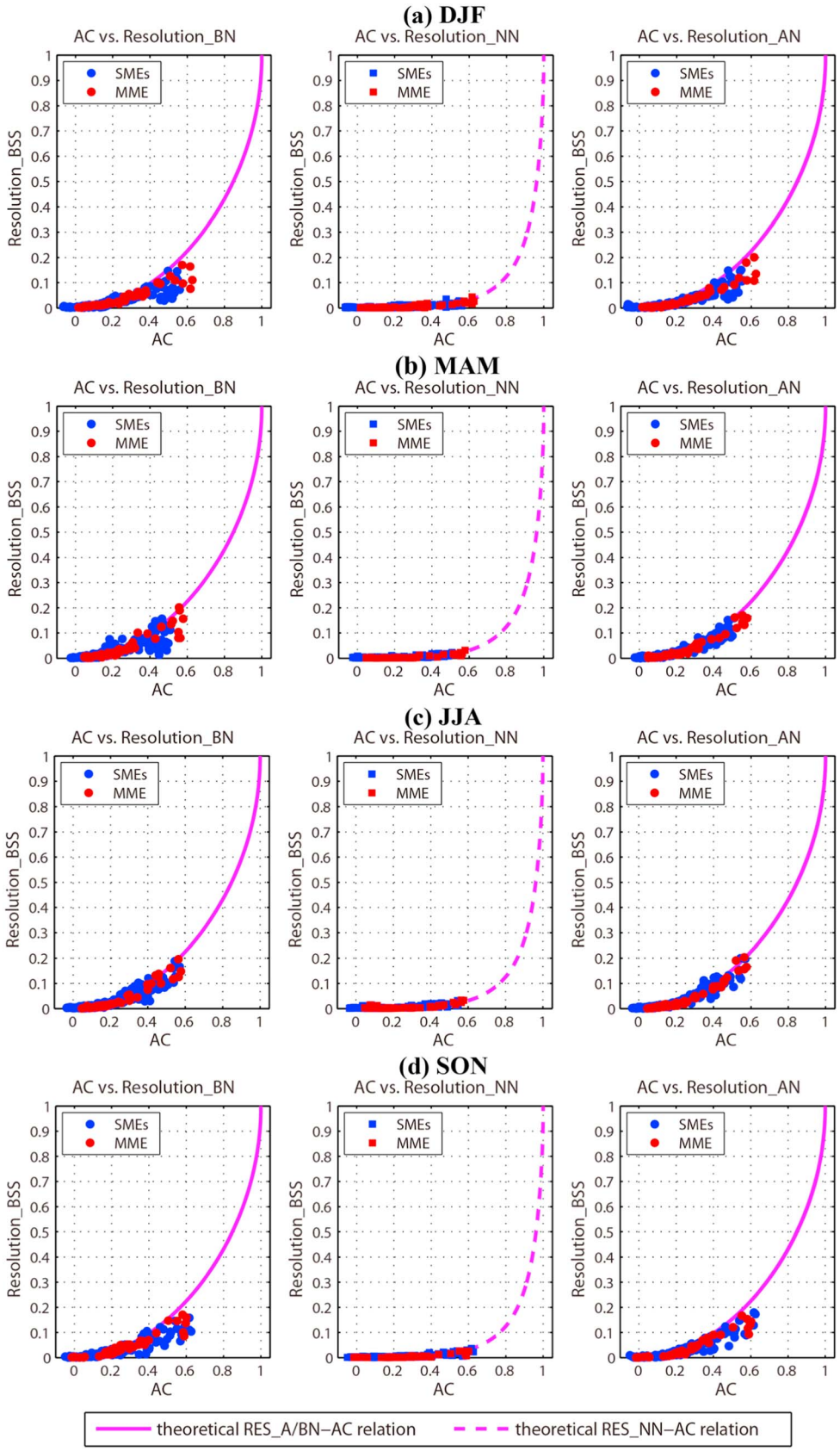

Figure 10. As in Figure 6, but for the prediction of precipitation anomaly over the period of 1979–2005.

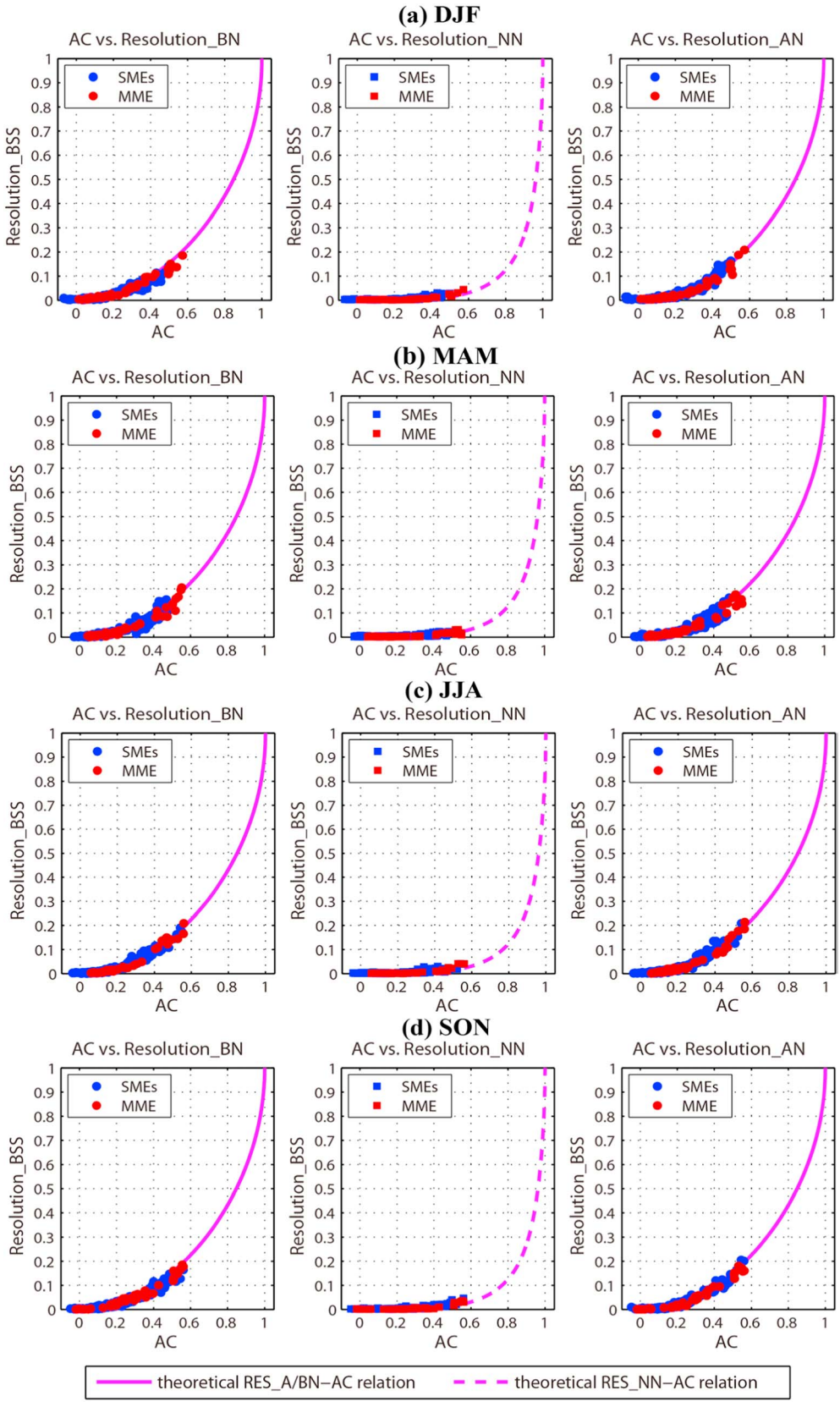

Figure 11. As in Figure 10, but with the forecast skills being calculated based on the transformed precipitation data that are more Gaussian by the Box-Cox transformation approach (see the main text).

In this study, through considerations from a theoretical perspective and practical analysis of the ENSEMBLES's GCM historical forecasts, we have targeted on investigating the relationship between the probabilistic and deterministic seasonal forecast skills. The probabilistic skill is measured in terms of resolution, reliability, and the BSS, while the deterministic skill is measured in terms of AC. Forecast skills of four representative variables, including 200-hPa geopotential height, 500-hPa geopotential height, SAT, and precipitation, are examined. For the probabilistic forecast, tercile-based categorical events of the below, near, and above normal are considered.

We have first of all presented a theoretical consideration. This theoretical consideration reveals that under the assumptions that forecast PDFs are Gaussian, that forecast variance does not change from case to case, and that forecast mean and corresponding observation are joint Gaussian distributed, a nonlinear, monotonic relationship can be analytically derived between the resolution and the AC. Specifically, we first demonstrate that under the first two assumptions, the resolution could be understood in terms of the statistical dependence between the forecast mean and the event occurrence. We then demonstrate that under the last assumption, this statistical dependence has a one-to-one coherence with the AC skill, that is, the linear correlation between the forecast mean and the corresponding observation. The specific form of the theoretical resolution-AC relationship depends on the definition of the target event. The theoretical relationships for the above- and below-normal events are the same. For the near-normal event, the theoretical relationship is characterized by a much weaker resolution corresponding to one AC, as compared to the above- and below-normal events. The theoretical resolution-AC relationship has been argued to exist regardless of whether or not the predicted signals in dynamical models approach the true predictable signal in the observed system.

We have subsequently analyzed the prediction skills of the ENSEMBLES's dynamical seasonal forecasts. It is found that the resolution and AC skills of the dynamical forecasts are largely coherent in their large-scale spatial distributions. Further, when calculated in a zonally aggregated manner by which the impact of finite sample size is reduced, the resolution and AC skills for both the MME and SMEs indeed tend to have a very good relationship that can be approximately characterized as a monotonic relationship. The specific form of this relationship, including its nonlinear feature, is greatly consistent with what the theory predicts. In short, the theoretical result is practically well verified by the dynamical model forecasts. In addition, the diagnostic analysis shows that different from the resolution, the reliability does not have a good relationship with the AC. The BSS, as the result of the resolution minus the reliability, also shows certain relationship with the AC. The seen relationship obviously comes from the resolution component.

To sum up, the most important take-home message of this study is that there tends to be a monotonic relationship in dynamical seasonal climate prediction between the probabilistic resolution skill and the deterministic AC skill that can be theoretically established and practically verified. One direct utility of the proved resolution-AC relationship is that it can facilitate comparisons among various assessments of seasonal climate prediction skill from the deterministic or probabilistic perspective alone. In addition, the proven coherence between the resolution and the AC undoubtedly implies that they share the same controlling factors and mechanisms and therefore would facilitate further understanding of the probabilistic forecast skill in many aspects. We provide a specific example for applying this proved coherence as below. As mentioned in the introduction, the MME approach has been documented to be practically effective in improving the probabilistic seasonal forecast skill. However, simple recalibration through appropriately rescaling the SMEs' forecasts a posteriori has been shown to be also able to improve the probabilistic skill (e.g., Doblas-Reyes et al., 2005). What is the fundamental difference between these two kinds of methods and which one is conceptually to be preferred is thereby a crucial question. Using synthetic forecast-observation data generated from an ideal statistical model, Weigel et al. (2009) has numerically explored this question and concluded that while both MME and recalibration can improve the reliability, the former rather than the latter can improve the resolution. We now demonstrate that the latter part of their conclusion actually can be easily inferred with the aid of our proven resolution-AC coherence. On the one hand, the resolution-AC coherence suggests that the question as to whether the MME can improve the resolution is essentially equivalent to the question as to whether the MME can improve the AC, which has been fully discussed in the literature (e.g., Hagedorn et al., 2005; Yoo & Kang, 2005). It is usually considered that through the mechanism of error cancelation, the MME can improve the deterministic skill. Of course, the practical capability of the MME in improving deterministic skill and resolution may at times appear not as pronounced as expected (such as in our case), as a result of the

possible influences from the mutual dependence of the member SMEs' model errors, skill diversity of the member SMEs, and even the intrinsic predictability limit of the nature (e.g., Yang et al., 2016). However, this practical status shall not indicate a failure of the MME concept itself but rather hints to a need for developing sophisticated strategies to make an optimal MME construction. On the other hand, since AC is invariant to rescaling, the resolution-AC coherence would instantly imply that the recalibration would not be able to improve the resolution.

Finally, the success of the theoretical relationship in explaining the observed counterpart in turn certifies that the proposed assumptions that play key roles in deriving the theoretical relationship are indeed approximately true, which may also have implications for understanding the property of seasonal mean atmospheric variability. First, the Gaussianity may indicate that the dynamics underlying the seasonal mean atmospheric variability is, to a large extent, effectively linear, since nonlinear dynamics tends to produce non-Gaussian statistical behavior. Second, because the forecast variance (spread) in seasonal climate prediction mainly reflects the impact of the atmospheric internal dynamics (noise), it being homogeneous may also indicate that the seasonal atmospheric noise characteristics are somewhat insensitive to the interannual variation of external forcings.

Acknowledgments
This work is jointly supported by the National Key Research and Development Program of China (2016YFA0602104), the National Natural Science Foundation of China (41621005, 41305085, and 41330420), and Jiangsu Collaborative Innovation Center for Climate Change. The ENSEMBLES data set is available from http://chfps.cima.fcen.uba.ar/ensemble.html. Authors are grateful to two anonymous reviewers for their constructive comments and suggestions to improve the manuscript.

References

Alessandri, A., Borrelli, A., Navarra, A., Arribas, A., Déqué, M., Rogel, P., & Weisheimer, A. (2011). Evaluation of probabilistic quality and value of the ENSEMBLES multimodel seasonal forecasts: Comparison with DEMETER. *Monthly Weather Review*, *139*(2), 581–607. https://doi.org/10.1175/2010MWR3417.1

Atger, F. (2004). Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, *130*(597), 627–646. https://doi.org/10.1256/qj.03.23

Barnston, A. G. (1992). Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Weather and Forecasting*, *7*(4), 699–709. https://doi.org/10.1175/1520-0434(1992)007,0699:CATCRA.2.0.CO;2

Becker, E., & van den Dool, H. (2015). Probabilistic seasonal forecasts in the North American Multimodel Ensemble: A baseline skill assessment. *Journal of Climate*, *29*, 3015–3026. https://doi.org/10.1175/JCLI-D-14-00862.1

Beraki, A., Landman, W., & DeWitt, D. (2015). On the comparison between seasonal predictive skill of global circulation models: Coupled versus uncoupled. *Journal of Geophysical Research: Atmospheres*, *120*, 11,151–11,172. https://doi.org/10.1002/2015JD023839

Bröcker, J. (2012). Estimating reliability and resolution of probability forecasts through decomposition of the empirical score. *Climate Dynamics*, *39*(3-4), 655–667. https://doi.org/10.1007/s00382-011-1191-1

Bröcker, J. (2015). Resolution and discrimination—Two sides of the same coin. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1277–1282. https://doi.org/10.1002/QJ.2434

Butler, A. H., Arribas, A., Athanassiadou, M., Baehr, J., Calvo, N., Charlton-Perez, A., et al. (2016). The climate-system historical forecast project: Do stratosphere-resolving models make better seasonal climate predictions in boreal winter? *Quarterly Journal of the Royal Meteorological Society*, *142*(696), 1413–1427. https://doi.org/10.1002/qj.2743

Charney, J. G., & Shukla, J. (1981). Predictability of monsoons. In J. Lighthill & R. P. Pearec (Eds.), *Monsoon dynamics* (pp. 99–110). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511897580.009

Cheng, Y., Tang, Y., Jackson, P., Chen, D., & Deng, Z. (2010). Ensemble construction and verification of the probabilistic ENSO prediction in the LDEO5 model. *Journal of Climate*, *23*(20), 5476–5497. https://doi.org/10.1175/2010JCLI3453.1

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

DelSole, T. (2004). Predictability and information theory. Part I: Measures of predictability. *Journal of the Atmospheric Sciences*, *61*(20), 2425–2440. https://doi.org/10.1175/1520-0469(2004)061%3C2425:PAITPI%3E2.0.CO;2

DelSole, T. (2005). Predictability and information theory. Part II: Imperfect forecasts. *Journal of the Atmospheric Sciences*, *62*(9), 3368–3381. https://doi.org/10.1175/JAS3522.1

DelSole, T., & Shukla, J. (2010). Model fidelity versus skill in seasonal forecasting. *Journal of Climate*, *18*, 4794–4806.

Doblas-Reyes, F. J., Hagedorn, R., & Palmer, T. N. (2005). The rationale behind the success of multi-model ensembles in seasonal forecasting. Part II: Calibration and combination. *Tellus*, *57A*, 234–252.

Graham, R., Gordon, M., Mclean, P. J., Ineson, S., Huddleston, M. R., Davey, M. K., et al. (2005). A performance comparison of coupled and uncoupled versions of the Met Office seasonal prediction general circulation model. *Tellus Series A*, *57*(3), 320–339. https://doi.org/10.1111/j.1600-0870.2005.00116.x

Guérémy, J.-F., Déqué, M., Braun, A., & Piedelièvre, J.-P. (2005). Actual and potential skill of seasonal predictions using the CNRM contribution to DEMETER: Coupled versus uncoupled model. *Tellus Series A*, *57*, 308–319.

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

Hamill, T. M., & Juras, J. (2006). Measuring forecast skill: Is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, *132*(621C), 2905–2923. https://doi.org/10.1256/qj.06.25

Jia, L., Yang, X., Vecchi, G. A., Gudgel, R. G., Delworth, T. L., Rosati, A., et al. (2015). Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. *Journal of Climate*, *28*(5), 2044–2062. https://doi.org/10.1175/JCLI-D-14-00112.1

Jia, X., Lin, H., Lee, J. Y., & Wang, B. (2012). Season-dependent forecast skill of the dominant atmospheric circulation patterns over the Pacific North-American region. *Journal of Climate*, *25*(20), 7248–7265. https://doi.org/10.1175/JCLI-D-11-00522.1

Kalnay, E., Kanamitsu, M., Kistler, R., Collins, W., Deaven, D., Gandin, L., et al. (1996). The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, *77*(3), 437–471. https://doi.org/10.1175/1520-0477(1996)077%3C0437:TNYRP%3E2.0.CO;2

Kanamitsu, M., Kumar, A., Juang, H.-M., Schemm, J.-K., Wang, W., Yang, F., et al. (2002). NCEP dynamical seasonal forecast system 2000. *Bulletin of the American Meteorological Society*, *83*(7), 1019–1037. https://doi.org/10.1175/1520-0477(2002)083%3C1019:NDSFS%3E2.3.CO;2

Kang, I.-S., & Shukla, J. (2006). Dynamic seasonal prediction and predictability. In B. Wang (Ed.), *The Asian monsoon* (Chap. 15, pp. 585–612). New York: Springer. https://doi.org/10.1007/3-540-37722-0_15

Kharin, V. V., Merryfield, W. J., Boer, G. J., & Lee, W. S. (2017). A postprocessing method for seasonal forecasts using temporally and spatially smoothed statistics. *Monthly Weather Review*, *145*(9), 3545–3561. https://doi.org/10.1175/MWR-D-16-0337.1

Kharin, V. V., & Zwiers, F. W. (2003). Improved seasonal probability forecasts. *Journal of Climate*, *16*(11), 1684–1701. https://doi.org/10.1175/1520-0442(2003)016%3C1684:ISPF%3E2.0.CO;2

Kharin, V. V., Zwiers, F. W., Teng, Q., Boer, G. J., Derome, J., & Fontecilla, J. S. (2009). Skill assessment of seasonal hindcasts from the Canadian Historical Forecast Project. *Atmosphere-Ocean*, *47*(3), 204–223. https://doi.org/10.3137/AO1101.2009

Kirtman, B. P., Min, D., Infanti, J. M., Kinter, J. L. III, Paolino, D. A., Zhang, Q., et al. (2014). The North American Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction; Phase-2 toward developing intra-seasonal prediction. *Bulletin of the American Meteorological Society*, *95*(4), 585–601. https://doi.org/10.1175/BAMS-D-12-00050.1

Krishnamurti, T. N. (1999). Improved weather and seasonal climate forecasts from multimodel superensemble. *Science*, *285*(5433), 1548–1550. https://doi.org/10.1126/science.285.5433.1548

Krishnamurti, T. N., Kishtawal, C., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., et al. (2000). Multimodel ensemble forecasts for weather and seasonal climate. *Journal of Climate*, *13*(23), 4196–4216. https://doi.org/10.1175/1520-0442(2000)013%3C4196:MEFFWA%3E2.0.CO;2

Kug, J.-S., Kang, I.-S., & Choi, D.-H. (2008). Seasonal climate predictability with Tier-one and Tier-two prediction systems. *Climate Dynamics*, *31*(4), 403–416. https://doi.org/10.1007/s00382-007-0264-7

Kumar, A., Barnston, A. G., Peng, P., Hoerling, M. P., & Goddard, L. (2000). Changes in the spread of the variability of the seasonal mean atmospheric states associated with ENSO. *Journal of Climate*, *13*(17), 3139–3151. https://doi.org/10.1175/1520-0442(2000)013%3C3139:CITSOT%3E2.0.CO;2

Kumar, A., Chen, M., & Wang, W. (2013). Understanding prediction skill of seasonal mean precipitation over the Tropics. *Journal of Climate*, *26*(15), 5674–5681. https://doi.org/10.1175/JCLI-D-12-00731.1

Landman, W. A., DeWitt, D., Lee, D.-E., Beraki, A., & Lötter, D. (2012). Seasonal rainfall prediction skill over South Africa: 1- vs. 2-tiered forecasting systems. *Weather and Forecasting*, *27*(2), 489–501. https://doi.org/10.1175/WAF-D-11-00078.1

Lee, J.-Y., Lee, S.-S., Wang, B., Ha, K.-J., & Jhun, J.-G. (2013). Seasonal prediction and predictability of the Asian winter temperature variability. *Climate Dynamics*, *41*(3-4), 573–587. https://doi.org/10.1007/s00382-012-1588-5

Lee, J.-Y., Wang, B., Ding, Q., Ha, K.-J., Ahn, J.-B., Kumar, A., et al. (2011). How predictable is the Northern Hemisphere summer upper-tropospheric circulation? *Climate Dynamics*, *37*(5-6), 1189–1203. https://doi.org/10.1007/s00382-010-0909-9

Lee, J.-Y., Wang, B., Kang, I. S., Shukla, J., Kumar, A., Kug, J. S., et al. (2010). How are seasonal prediction skills related to models' performance on mean state and annual cycle? *Climate Dynamics*, *35*(2-3), 267–283. https://doi.org/10.1007/s00382-010-0857-4

Lee, S.-S., Lee, J.-Y., Ha, K.-J., Wang, B., & Schemm, J. (2011). Deficiencies and possibilities for long-lead coupled climate prediction of the western North Pacific-East Asian summer monsoon. *Climate Dynamics*, *36*(5-6), 1173–1188. https://doi.org/10.1007/s00382-010-0832-0

Li, C., Lu, R., & Dong, B. (2014). Predictability of the western North Pacific summer climate associated with different ENSO phases by ENSEMBLES multi-model seasonal forecasts. *Climate Dynamics*, *43*(7-8), 1829–1845. https://doi.org/10.1007/s00382-013-2010-7

Liu, X., Wu, T., Yang, S., Jie, W., Nie, S., Li, Q., et al. (2015). Performance of the seasonal forecasting of the Asian summer monsoon by BCC_CSM1.1(m). *Advances in Atmospheric Sciences*, *32*(8), 1156–1172. https://doi.org/10.1007/s00376-015-4194-8

Luo, J.-J., Masson, S., Behera, S., & Yamagata, T. (2008). Extended ENSO predictions using a fully coupled ocean-atmosphere model. *Journal of Climate*, *21*(1), 84–93. https://doi.org/10.1175/2007JCLI1412.1

MacLachlan, C., Arribas, A., Peterson, D., Maidens, A., Fereday, D., Scaife, A., et al. (2015). Global seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1072–1084. https://doi.org/10.1002/qj.2396

Manzanas, R., Frías, M. D., Cofiño, A. S., & Gutiérrez, J. M. (2014). Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. *Journal of Geophysical Research: Atmospheres*, *119*, 1708–1719. https://doi.org/10.1002/2013JD020680

Merryfield, W. J., Lee, W.-S., Boer, G. J., Kharin, V. V., Scinocca, J. F., Flato, G. M., et al. (2013). The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization. *Monthly Weather Review*, *141*(8), 2910–2945. https://doi.org/10.1175/MWR-D-12-00216.1

Min, Y.-M., Kryjov, V. N., & Oh, S. M. (2017). Skill of real-time operational forecasts with the APCC multi-model ensemble prediction system during the period 2008–2015. *Climate Dynamics*, *49*(11-12), 4141–4156. https://doi.org/10.1007/s00382-017-3576-2

Molteni, F., Stockdale, T., Balmaseda, M., Balsamo, G., Buizza, R., Ferranti, L., et al. (2011). The new ECMWF seasonal forecast system (System 4). ECMWF Technical Memorandum 656. Retrieved from http://www.ecmwf.int/publications

Palmer, T., Branković, Č., & Richardson, D. (2000). A probability and decision-model analysis of PROVOST seasonal multi-model ensemble integrations. *Quarterly Journal of the Royal Meteorological Society*, *126*(567), 2013–2033. https://doi.org/10.1256/smsqj.56702

Palmer, T. N. (2002). The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quarterly Journal of the Royal Meteorological Society*, *128*(581), 747–774. https://doi.org/10.1256/0035900021643593

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M., Délécluse, P., et al. (2004). Development of a European multimodel ensemble system for seasonal-to-interannual prediction (Demeter). *Bulletin of the American Meteorological Society*, *85*(6), 853–872. https://doi.org/10.1175/bams-85-6-853

Pavan, V., & Doblas-Reyes, J. (2000). Multimodel seasonal hindcasts over the Euro-Atlantic: Skill scores and dynamic features. *Climate Dynamics*, *16*(8), 611–625. https://doi.org/10.1007/s003820000063

Peng, P., Kumar, A., van den Dool, H., & Barnston, A. G. (2002). An analysis of multi-model ensemble predictions for seasonal climate anomalies. *Journal of Geophysical Research*, *107*(D23), 4710. https://doi.org/10.1029/2002JD002712

Quan, X., Hoerling, M. P., Whitaker, J., & Xu, T. (2006). Diagnosing sources of U.S. seasonal forecast skill. *Journal of Climate*, *19*(13), 3279–3293. https://doi.org/10.1175/JCLI3789.1

Richardson, D. S. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 649–667. https://doi.org/10.1002/qj.49712656313

Richardson, D. S. (2006). Predictability and economic value. In T. Palmer & R. Hagedorn (Eds.), *Predictability of weather and climate* (pp. 628–644). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511617652.026

Rodwell, M. J., Rowell, D. P., & Folland, C. K. (1999). Oceanic forcing of the winter North Atlantic Oscillation and European climate. *Nature*, *398*(6725), 320–323. https://doi.org/10.1038/18648

Rowell, D. P. (1998). Assessing potential seasonal predictability with an ensemble of multidecadal GCM simulations. *Journal of Climate*, *11*(2), 109–120. https://doi.org/10.1175/1520-0442(1998)011%3C0109:APSPWA%3E2.0.CO;2

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., et al. (2014). The NCEP climate forecast system version 2. *Journal of Climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-12-00823.1

Saha, S., Nadiga, S., Thiaw, C., Wang, J., Wang, W., Zhang, Q., et al. (2006). The NCEP climate forecast system. *Journal of Climate*, *19*(15), 3483–3517. https://doi.org/10.1175/JCLI3812.1

Sardeshmukh, P. D., Compo, G. P., & Penland, C. (2000). Changes in probability associated with El Niño. *Journal of Climate*, *13*(24), 4268–4286. https://doi.org/10.1175/1520-0442(2000)013%3C4268:COPAWE%3E2.0.CO;2

Scaife, A. A., Comer, R. E., Dunstone, N. J., Knight, J. R., Smith, D. M., MacLachlan, C., et al. (2017). Tropical rainfall, Rossby waves and regional winter climate predictions. *Quarterly Journal of the Royal Meteorological Society*, *143*(702), 1–11. https://doi.org/10.1002/qj.2910

Schlosser, C. A., & Kirtman, B. P. (2005). Predictable skill and its association to sea surface temperature variations in an ensemble climate simulation. *Journal of Geophysical Research*, *110*, D19107. https://doi.org/10.1029/2005JD005835

Shukla, J. (1998). Predictability in the midst of chaos: A scientific basis for climate forecasting. *Science*, *282*(5389), 728–731. https://doi.org/10.1126/science.282.5389.728

Sohn, S.-J., Min, Y.-M., Lee, J.-Y., Tam, C.-Y., Kang, I.-S., Wang, B., et al. (2012). Assessment of the long-lead probabilistic prediction for the Asian summer monsoon precipitation (1983–2011) based on the APCC multimodel system and a statistical model. *Journal of Geophysical Research*, *117*, D04102. https://doi.org/10.1029/20011JD016308

Sooraj, K. P., Annamalai, H., Kumar, A., & Wang, H. (2012). A comprehensive assessment of CFS seasonal forecast over the tropics. *Weather and Forecasting*, *27*(1), 3–27. https://doi.org/10.1175/WAF-D-11-00014.1

Sperber, K. R., & Palmer, T. N. (1996). Interannual tropical rainfall variability in general circulation model simulations associated with the Atmospheric Model Intercomparison Project. *Journal of Climate*, *9*(11), 2727–2750. https://doi.org/10.1175/1520-0442(1996)009%3C2727:ITRVIG%3E2.0.CO;2

Stockdale, T., Anderson, D., Alves, J., & Balmaseda, M. (1998). Global seasonal rainfall forecasts using a coupled ocean–atmosphere model. *Nature*, *392*(6674), 370–373. https://doi.org/10.1038/32861

Stockdale, T., Molteni, F., & Ferranti, L. (2015). Atmospheric initial conditions and the predictability of the Arctic Oscillation. *Geophysical Research Letters*, *42*, 1173–1179. https://doi.org/10.1002/2014GL062681

Tang, Y., Lin, H., & Moore, A. M. (2008). Measuring the potential predictability of ensemble climate predictions. *Journal of Geophysical Research*, *113*, D04108. https://doi.org/10.1029/2007JD008804

Tippett, M. K., Ranganathan, M., L'Heureux, M., Barnston, A. G., & DelSole, T. (2017). Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. *Climate Dynamics*, 1–22. https://doi.org/10.1007/s00382-017-3721-y

Toth, Z., Talagrand, O., & Zhu, Y. (2006). The attributes of forecast systems: A general framework for the evaluation and calibration of weather forecasts. In T. Palmer & R. Hagedorn (Eds.), *Predictability of weather and climate* (pp. 584–595). Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511617652.026

Van den Dool, H. M., & Toth, Z. (1991). Why do forecasts for "near normal" often fail? *Weather and Forecasting*, *6*(1), 76–85. https://doi.org/10.1175/15200434(1991)006,0076:WDFFNO.2.0.CO;2

Wang, B., Ding, Q., Fu, X., Kang, I.-S., Jin, K., Shukla, J., & Doblas-Reyes, F. (2005). Fundamental challenge in simulation and prediction of summer monsoon rainfall. *Geophysical Research Letters*, *32*, L15711. https://doi.org/10.1029/2005GL022734

Wang, B., Lee, J.-Y., Kang, I.-S., Shukla, J., Park, C.-K., Kumar, A., et al. (2009). Advance and prospectus of seasonal prediction: Assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). *Climate Dynamics*, *33*(1), 93–117. https://doi.org/10.1007/s00382-008-0460-0

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2008). Can multimodel combination really enhance the prediction skill of probabilistic ensemble forecasts? *Quarterly Journal of the Royal Meteorological Society*, *134*(630), 241–260. https://doi.org/10.1002/qj.210

Weigel, A. P., Liniger, M. A., & Appenzeller, C. (2009). Seasonal ensemble forecasts: Are recalibrated single models better than multimodels? *Monthly Weather Review*, *137*(4), 1460–1479. https://doi.org/10.1175/2008MWR2773.1

Weisheimer, A., Doblas-Reyes, F. J., Palmer, T. N., Alessandri, A., Arribas, A., Déqué, M., et al. (2009). ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophysical Research Letters*, *36*, L21711. https://doi.org/10.1029/2009GL040896

Wilks, D. S. (2002). Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, *128*(586), 2821–2836. https://doi.org/10.1256/qj.01.215

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences*, *Int. Geophys. Ser.* (3rd ed., Vol. 100). San Diego, CA: Academic Press.

Xie, P., & Arkin, P. A. (1996). Analyses of global monthly precipitation using gauge observations, satellite estimates, and numerical model predictions. *Journal of Climate*, *9*(4), 840–858. https://doi.org/10.1175/1520-0442(1996)009%3C0840:AOGMPU%3E2.0.CO;2

Yan, X., & Tang, Y. (2013). An analysis of multimodel ensemble for seasonal climate predictions. *Quarterly Journal of the Royal Meteorological Society*, *139*(674), 1389–1401. https://doi.org/10.1002/qj.2019

Yang, D., Tang, Y., Zhang, Y., & Yang, X. (2012). Information-based potential predictability of the Asian summer monsoon in a coupled model. *Journal of Geophysical Research*, *117*, D03119. https://doi.org/10.1029/2011JD016775

Yang, D., Yang, X.-Q., Xie, Q., Zhang, Y., Ren, X., & Tang, Y. (2016). Probabilistic versus deterministic skill in predicting the western North Pacific-East Asian summer monsoon variability with multimodel ensembles. *Journal of Geophysical Research: Atmospheres*, *121*, 1079–1103. https://doi.org/10.1002/2015JD023781

Yang, X.-Q., Anderson, J. L., & Stern, W. F. (1998). Reproducible forced modes in AGCM ensemble integrations and potential predictability of atmospheric seasonal variations in the extratropics. *Journal of Climate*, *11*(11), 2942–2959. https://doi.org/10.1175/1520-0442(1998)011%3C2942:RFMIAE%3E2.0.CO;2

Yoo, J. H., & Kang, I.-S. (2005). Theoretical examination of a multimodel composite for seasonal prediction. *Geophysical Research Letters*, *32*, L18707. https://doi.org/10.1029/2005GL023513

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