Assessment of the Sea Surface Temperature Predictability Based on Multimodel Hindcasts

SHOUWEN ZHANG AND HUA JIANG

National Marine Environment Forecasting Center, State Oceanic Administration, Beijing, China

HUI WANG

Key Laboratory of Research on Marine Hazards Forecasting, National Marine Environmental Forecasting Center, Beijing, and Institute of Marine Science and Technology, Shandong University, Qingdao, Shandong, China

(Manuscript received 27 February 2019, in final form 9 October 2019)

ABSTRACT

Based on historical forecasts of four individual forecasting systems, we conducted multimodel ensembles (MME) to predict the sea surface temperature anomaly (SSTA) variability and assessed these methods from a deterministic and probabilistic point of view. To investigate the advantages and drawbacks of different deterministic MME methods, we used simple averaged MME with equal weighs (SCM) and the stepwise pattern projection method (SPPM). We measured the probabilistic forecast accuracy by Brier skill score (BSS) combined with its two components: reliability (B_{rel}) and resolution (B_{res}). The results indicated that SCM showed a high predictability in the tropical Pacific Ocean, with a correlation exceeding 0.8 with a 6-month lead time. In general, the SCM outperformed the SPPM in the tropics, while the SPPM tend to show some positive effect on the correction when at long lead times. Corrections occurred for the spring predictability barrier of ENSO, in particular for improvements when the correlation was low or the RMSE was large using the SCM method. These qualitative results are not susceptible to the selection of the hindcast periods, it is as a rule rather by chance of these individual systems. Performance of our probabilistic MME was better than the Climate Forecast System version2 (CFSv2) forecasts in forecasting COLD, NEUTRAL, and WARM SSTA categories for most regions, mainly due to the contribution of B_{rel}, indicating more adequate ensemble construction strategies of the MME system superior to the CFSv2.

1. Introduction

Climate forecasts are subject to many uncertainties and forecast errors because of the stochastic nature of the climate system and the inadequacy of current forecast systems (Palmer 2000). The two main sources of uncertainties are the model initialization and the imperfection of the model itself. The first source of uncertainty often is due to incomplete data coverage, measurement errors, or inappropriate data assimilation procedures and usually is addressed by generating a set of slightly perturbed initial conditions using dynamical models, the so-called ensemble technique (Kalnay 2003; Gneiting and Raftery 2005). This ensemble technique, however, does not take the model imperfections into account; therefore, this technique when performed with a single system usually is overconfident (Slingo and Palmer 2011). The second source of uncertainty often is due to the parameterization of physical processes or to the limited spatial and temporal resolutions of the models, which make it impossible to study some unresolved scales. Three approaches have been pursued to deal with the uncertainties induced by model errors: the introduction of ''stochastic physics'' (Buizza et al. 1999), the ''perturbed parameter'' approach (Pellerin et al. 2003), and the multimodel superensemble technique (Palmer et al. 2004). The multimodel superensemble technique is often referred to as the multimodel ensemble (MME) approach, and it is a relatively recent contribution to climate forecasting as it reduces the systematic and random errors that exist in individual model systems (Fritsch et al. 2000; Peng et al. 2000; Palmer et al. 2004), which is the focus of this study.

The central argument lies in finding the best way to combine the predictions given by different forecast systems from postprocessing procedures, such as MME

Corresponding author: Hui Wang, wangh@nmefc.cn

DOI: 10.1175/WAF-D-19-0040.1

techniques and statistical error correction methods, when applying the multimodel approach. The simple composite method (SCM) is the simplest and most commonly used method that can be constructed by combining the individual ensemble forecasts with equal weights (Hagedorn et al. 2005). In a more sophisticated approach, the individual ensemble forecasts are weighted based on their retrospective performance (hereafter, hindcast). The ability to assess and understand the forecast accuracy of these different methods is indispensable and has been examined by a number of studies dating to the beginning of the multimodel era (Yun et al. 2003; Kug et al. 2008a; Chowdary et al. 2010; Min et al. 2014). In addition to SCM, the two other widely used methods are empirically weighted MMEs with coefficients obtained using multiple linear regression (MRG; Krishnamurti et al. 2000; Yun et al. 2003) and calibrated MMEs, which are estimated as a composite of single-model predictions and corrected based on a stepwise pattern projection method (SPPM; Kug et al. 2008a). Many studies have indicated that SCM method generally outperforms the MRG method, though the MRG method provides optimal weights in the sense of minimization of the mean squared error over the training period (DelSole and Shukla 2009; Rodrigues et al. 2014; Min et al. 2014). This is high possibly resulted from the relatively short training period for estimation of weights and overfitting (Michaelsen 1987; Peng et al. 2002). The model correction using SPPM shows the capability to improve forecast skills of global sea surface temperature (Min et al. 2014; Wang et al. 2017), especially for the tropics. However, these studies are all evaluated using a 1-year-out cross validation, which is completely insufficient for the SPPM and debatable. In this study, to avoid both artificial skill and degeneracy, a 3-year-out cross-validation scheme is used. In addition to deterministic predictions, these forecasts can also be quantified in terms of probabilities as climate forecasts are associated with uncertainties. Probabilistic forecasts can provide an indication of the likelihood for a climate variable to occur at a certain interval, which is of significant value to end-users for decision-making (Richardson 2006; Alessandri et al. 2011). Therefore, to develop a comprehensive assessment of the performance of the MMEs, the forecast quality should be assessed from a deterministic and probabilistic point of view.

Ocean long-term memory allows for forecast models to provide accurate predictions of sea surface temperature anomalies (SSTA) a few months in advance. These forecasts have profound social and economic consequences, in particular, for stakeholders and end-users (Chen et al. 2016). El Niño–Southern Oscillation (ENSO), the most significant coupled ocean–atmosphere phenomenon causing global climate variability on the interannual time scale, is the primary source of predictability at a seasonal time scale. SSTA over the tropical Indian Ocean and North Atlantic Ocean also affects the circulation and precipitation pattern of the Eurasia continent (Guan and Yamagata 2003; Weng et al. 2011). Therefore, to promote the accuracy of SSTA, improved prediction skill could provide valuable references for the prediction of other associated climate variability. Since MME techniques have been seen as an effective way to improve seasonal forecasts, many MME prediction systems are being operational running at different operational centers. For example, the North American Multimodel Ensemble (NMME; Kirtman et al. 2014), the European Seasonal to Interannual Prediction multimodel ensemble system (EUROSIP; https://www.ecmwf.int/en/forecasts/datasets/set-viii) and the Asia–Pacific Economic Cooperation Climate Center (APCC) multimodel ensemble prediction system (Min et al. 2014, 2017). China, by contrast, should do more to set up the multimodel prediction system of its own.

This study operates four different operational dynamical forecast systems maintained and funded by three China institutes, which will be detailed in section 2. We first provided a comprehensive assessment of the deterministic MME forecasts of SSTA by comparing the performance of SCM and SPPM methods. Moreover, for a full assessment and, in particular, to assess the accuracy of the forecasts with respect to the climatological forecast in each category, we used a probabilistic assessment called the Brier skill score (BSS).

The paper is organized as follows: section 2 provides an overview of the MME experiments used to evaluate forecast accuracy of SSTA and describes the approaches to generate deterministic and probabilistic forecasts. Sections 3 and 4 present and discuss the results of the deterministic and probabilistic forecasts, respectively. Section 5 gives the main conclusions.

2. Data and methods

a. Hindcast data

Three China operational and research institutions have contributed four forecasting systems to the MME. These are 1) Integrated Climate Model version 2 (ICMv2) and 2) Flexible Global Ocean–Atmosphere–Land System model Finite Volume version 2 (Fgoals-f) from the Institute of Atmospheric Physics, Chinese Academy of Sciences; 3) First Institute of Oceanography Earth

TABLE 1. Brief description of the forecast systems used in this study.

Systems	Hindcast period	Members	Lead (month)	Resolution (atmosphere/ocean)
ICMv2	1981–2010	5	12	T63/0.5° at equator–2° at pole
FIO-ESM	1993–2013	10	6	T42/gx1v6
NMEFC-CESM	1981–2017	5	12	f09_gx1v6
Fgoals-f	1981–2017	35	12	f09_gx1v6

System Model (FIO-ESM) (Qiao et al. 2013) from the First Institute of Oceanography, Ministry of Natural Resources; and 4) National Marine Environmental Forecasting Center CESM (NMEFC-CESM) (Li et al. 2015; Zhang et al. 2018) from the National Marine Environmental Forecasting Center, Ministry of Natural Resources. These models are all fully coupled, and initialized at the first day of each month, the lead-1 month is just the initial month. Forecasts are produced monthly with leads up to at least 6 months, whereas individual model ensembles vary from 5 to 35 in hindcast data. Table 1 briefly describes the four forecasting systems used in this study. Initialization, data assimilation schemes, and model physics are left up to the modeling centers. For example, a nudging method is used in ICMv2, NMEFC-CESM, and Fgoals-f, but the setup of the assimilation schemes is different. ICMv2 assimilates sea surface temperature provided by the Hadley Center's Global Sea ice Coverage and Sea Surface Temperature (HadISST; Kennedy et al. 2011), but it assimilates the subsurface temperature (5–400 m) of Global Ocean Data Assimilation System (GODAS) (Behringer and Xue 2004) datasets in the NMEFC-CESM. Fgoals-f assimilates not only the GODAS temperature datasets but also the wind, air temperature, and geopotential height datasets from the Japanese 55-year reanalysis (JRA-55) datasets (Kobayashi et al. 2015; Harada et al. 2016). Before applying MME methods, each member of the individual system is interpolated to a common $1° \times 1°$ grid, which is consistent with the resolution of the observed verification data. Since the member of each system varies, the ensemble mean of each system is taken in the process of making a deterministic forecast.

Considering the periods of the four models and to study the influence of the hindcast period on the SCM and SPPM methods, this study was performed on the basis of two periods. A 30-yr (1981–2010) period for three models and an overlapping period for all models (1993–2010, 18-yr period) were studied, respectively. As shown above, a 3-year-out cross-validation scheme is used. Of the 3 years, one is the test element and the other two are chosen at random without repetition. That is, for the 30 (18) years of hindcast data, the training period is 27 (15) years with 3 years withheld. We obtained the hindcast ensemble mean climatology for each model by using all of the members for the training period (this corrects for systematic bias in the mean), and we then subtracted this climatology from each ensemble member and identified the forecast anomalies of the 3 years withheld.

b. Observations

We verified the SSTA prediction using the optimum interpolation version 2 (OI) monthly mean sea surface temperature (Reynolds et al. 2002), which is constructed by combining observations from different platforms (satellites, ships, and buoys) on a regular global grid. The original resolution of this dataset was 1/4°; however, for the convenience of calculating and comparing with the forecast data, it was linearly interpolated to 1°. The periods to calculate the climatology of SST were consistent with the hindcast data.

c. Methodology of prediction skill

For the deterministic MME prediction methods, we followed two methods operationally implemented in the APCC: SCM and SPPM. For the probabilistic forecast, we measured skills in terms of the overall BSS, reliability, and resolution. A detailed description of each method is given in the following sections.

1) SCM

SCM is the most simple and widely used method for MME. Its prediction equation is defined as follows:

$$S_t = \frac{1}{N}\sum_{i=1}^{N} \mathrm{SSTA}'_{i,t}, \tag{1}$$

where S_t is the ensemble mean results at the forecast time t, $\mathrm{SSTA}'_{i,t}$ is the SSTA forecast results of the ith model at the forecast time t, N is the number of individual models involved, and $1/N$ shows that the same weights are assigned to each participating models.

2) SPPM

The SPPM is a pointwise regression model based on a stepwise pattern projection method (Kug et al. 2008a). It was established based on the assumption that the predictor of the model is a pattern of predicted SSTA in a certain domain and the predictand is SSTA at each

grid point over a global domain. The idea is to predict the predictand at each grid by projecting the spatial pattern of the predictor on to the covariance pattern between the large-scale predictor field and the one-point predictand. After applying the SPPM to individual models, the MME prediction can be obtained by the equal-weighting simple composite of the corrected individual predictions. The model equation is as follows:

$$\mathrm{Cov}(x,y)=\frac{1}{T}\sum_{t=1}^{T}Y(t)\mathrm{SSTA}'(x,y,t). \tag{2}$$

The covariance pattern (COV) is calculated between the observed predictand $Y(t)$ and the predictor field $\mathrm{SSTA}'(x, y, t)$ in a certain domain D. The selection of D plays a crucial role in the predictive skill, and in this study, the correlation coefficients between $Y(t)$ and $\mathrm{SSTA}'(x, y, t)$ are calculated. The results are sorted in descending order, and the first 200 grids are selected as the domain D:

$$P(t)=\sum_{x,y}^{D}\mathrm{Cov}(x,y)\mathrm{SSTA}'(x,y,t), \tag{3}$$

where P shows a projected time series from the covariance pattern and predictor field from the model prediction:

$$Y(t)=\alpha P(t), \tag{4}$$

$$\alpha=\frac{\frac{1}{T}\sum_{t=1}^{T}Y(t)P(t)}{\frac{1}{T}\sum_{t=1}^{T}P^2(t)}, \tag{5}$$

where α is a regression coefficient of the projected time series on the predictand during a training period.

3) Brier score

The Brier score (BS) is used to verify the accuracy of a probability forecast. It is the mean squared difference between the forecast probability and the corresponding observed binary variable defined as one when the event occurs and zero otherwise. Lower values of the BS indicate better forecasts. The BS is defined as follows:

$$\mathrm{BS}=\frac{1}{N}\sum_{i=1}^{N}(P_i-O_i)^2, \tag{6}$$

where N is the number of total verification samples, P_i is the forecast probability, and O_i is the outcome (1 if it happened, 0 if it did not).

The BS can be decomposed into three items: reliability, resolution, and uncertainty as follows:

$$\mathrm{BS}=\frac{1}{N}\sum_{k=1}^{K}N_k(P_k-\overline{O}_k)^2-\frac{1}{N}\sum_{k=1}^{K}N_k(\overline{O}_k-\overline{O})^2+\overline{O}(1-\overline{O}). \tag{7}$$

The probability space can be partitioned into K bins ($K = 10$ in this study), P_k is the averaged forecast probability at bin k, $\overline{O}_k$ is the corresponding observed frequency, and $\overline{O}$ is the climatological probability of the event. Reliability quantifies the consistency of the forecast probabilities compared with the corresponding observed frequencies and resolution quantifies the degree to which these observed frequencies differ from the climatological probability. The Brier score is calculated for each equi-probable category, the so-called COLD, NEUTRAL, and WARM categorical events. The three categorical events so defined have an equal climatological frequency of 1/3 and applied to both observed and forecast results.

Although the BS can determine the accuracy of a forecast, the BSS reveals the relative accuracy of the probabilistic forecast over that of the observed climatology by predicting whether or not an event occurred:

$$\mathrm{BSS}=1-\frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{clim}}}=\mathrm{BSS}_{\mathrm{res}}-\mathrm{BSS}_{\mathrm{rel}}, \tag{8}$$

where $\mathrm{BS}_{\mathrm{clim}}$ is the Brier score of the reference forecast, defined here as the climatological frequency forecast, $\mathrm{BSS}_{\mathrm{res}} = 1$ and $\mathrm{BSS}_{\mathrm{rel}} = 0$ indicate a perfect forecast system.

Hindcast skills of two different systems are considered in the point of probability forecast. As each ensemble member was counted equally in the probabilistic forecast, models with more ensemble members contributed more to the probability forecast, but an excessive emphasis on one individual system usually was overconfident, which resulted in model-specific bias (Slingo and Palmer 2011). Therefore, the MME20 system (20 members) includes 5 members of ICMv2; the 1st, 3rd, 5th, 7th, and 9th member of FIO-ESM; 5 members of NMEFC-CESM; and the 1st, 8th, 15th, 22nd, and 29th member of Fgoals-f. For reference purposes, we used Climate Forecast System version 2. CFSv2 includes 24 members for each initial month covering the period from 1982 to 2010. To free the evaluation from the effects of discontinuities in CFSv2 (Saha et al. 2010), dual climatologies from which to form anomalies are developed (1982–98 and 1999–2009; Barnston and Tippett 2013; Saha et al. 2014). The comparison of results from these two systems enabled us to draw some preliminary

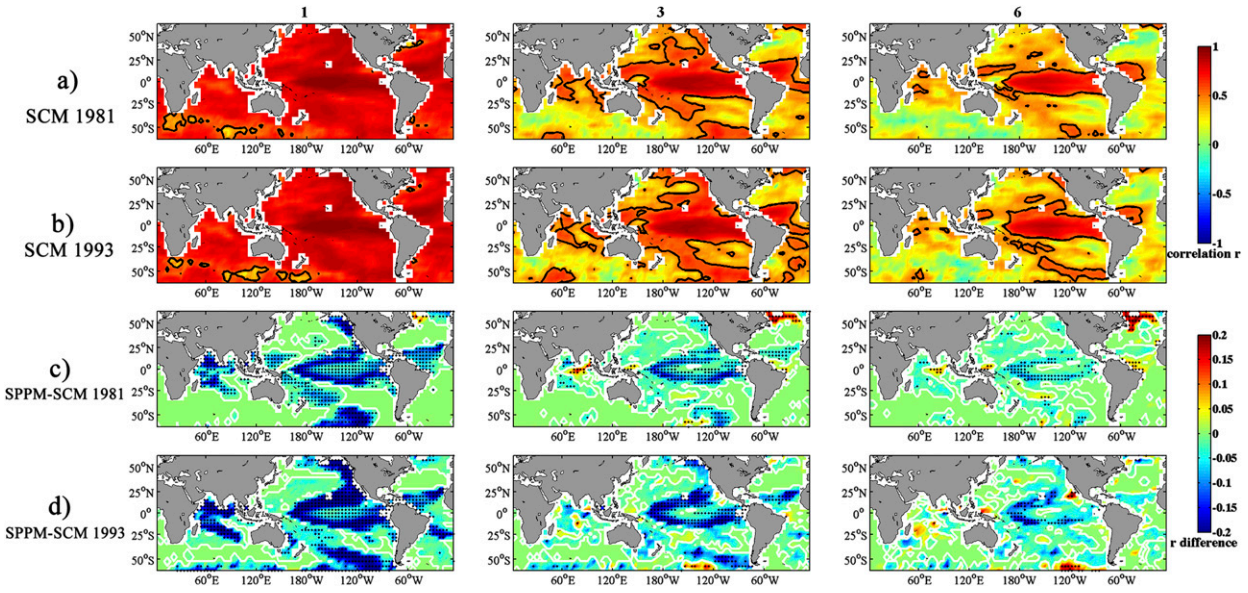

FIG. 1. Spatial distributions of temporal correlation of SCM SSTA for two periods (a) 1981–2010 and (b) 1993–2010 (solid lines are 0.5 isolines, far exceeding the threshold values for the 95% confidence levels). Skill difference of SPPM predictions with respect to the SCM for two periods (c) 1981–2010 and (d) 1993–2010 (black dots represent the skill difference being statistically significant at the 5% level using the Student's *t* statistic).

conclusions about the relative influences of increased model diversity. Note that the overlapping period 1993–2010 of the four models and the CFSv2 is selected to evaluate the probabilistic forecast skill.

3. Performance of the deterministic forecast

We used the temporal correlation to assess the degree of linear association between the MME SSTA and the observed SSTA indices, as shown in Fig. 1. It was obviously that the prediction skills based on the SCM of two periods were identical in the global distribution (Figs. 1a,b). Relatively high levels of prediction skills were notable over the tropical Pacific Ocean, in particular, regardless of the lead time, and the correlation *r* exceeds 0.8 even with a 6-month lead time. Most of this significant skill originated from the influence of ENSO variability (Wang et al. 2009). In contrast, the forecast skills in the Southern Indian Ocean between 30° and 50°S, the Northern Atlantic Ocean, and the regional western oceans off the South American continent were relatively low. The differences among temporal correlations in the SPPM method with respect to the SCM method, which was the reference forecast, were shown in Figs. 1c and 1d. It was obviously that the SPPM method cannot outperform the SCM method for ~1–6-month lead time, especially in the tropics. However, from the comparison of the SCM and SPPM results, skill differences between the SCM and SPPM results tend to be reduced with a 1–6-month lead time. It meant that SPPM method had a positive effect in improving the skill at long lead times. It should also be noted that skill differences between the SPPM and SCM results were positive in the regions where SCM had low correlation coefficients. The influence of hindcast periods on the SCM and SPPM methods were also studied. The most obvious feature showed that the spatial distributions of the skill differences for two periods were identical, but the skill differences with a short hindcast period (1993–2010) were much more significant than those with a long hindcast period (1981–2010), showing forecast skill was more degraded by SPPM when using the shorter hindcast period.

Considering the research on global climate change and the internal features of the MME, we selected the Niño-3.4 index (5°S–5°N, 170°–120°W) as the ENSO indicator to study the forecast accuracy on the basis of two deterministic methods. ENSO is the most important source of predictability at the seasonal time scale, and the assessment of the accuracy of ENSO SSTA predictions is a fundamental requirement for any seasonal forecasting system (Stockdale et al. 2011).

Figure 2 shows temporal correlation and RMSE for SCM, SPPM predictions, and individual models of the Niño-3.4 index for two periods. It was demonstrated that the SCM and SPPM results were better than an individual model results in terms of temporal correlation and RMSE for both periods. The skills of SPPM predictions were

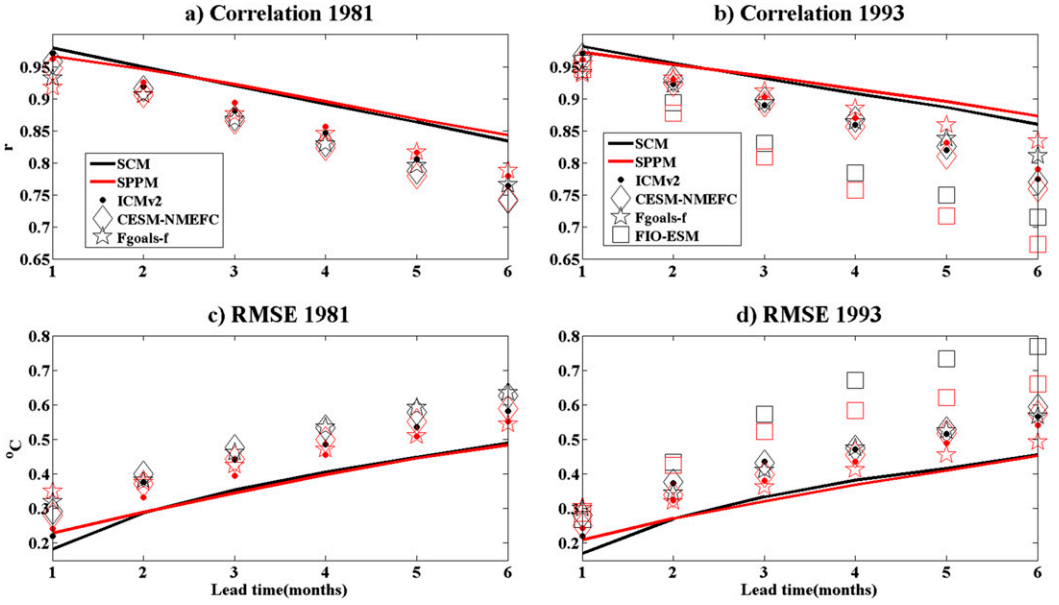

FIG. 2. Temporal correlation and RMSE for SCM, SPPM, and individual models of the Niño-3.4 index for two periods (a),(c) 1981–2010 and (b),(d) 1993–2010. Different symbols represent different individual models as shown in the legend in (a) and (b); black and red solid lines show the results of SCM and SPPM, respectively.

better than that of SCM predictions when the lead time advanced 3 months, though there differences are negligible. There were few cases in which the SPPM method showed no correction (i.e., the results of FIO-ESM for the period of 1993–2010) in terms of temporal correlation, but it tended to be effective when at long lead times. However, compared with the correction of the temporal correlation, the SPPM method was effective for all single models in reducing the RMSE (Figs. 2c,d). It should also be noted that the RMSE of the SCM results had already shown high skill, and the improvement of the SPPM results on the SCM results was negligible. Furthermore, these results were clearly not susceptible to the hindcast periods.

Figures 3 and 4 show the annual cycles of Niño-3.4 SSTA forecast skills based on two deterministic methods in terms of correlation (RMSE) for all 12 targeted months with 1–6 months of lead time. In general, the Niño-3.4 index showed very high forecast skills with the MME predictions, but the spring predictability barrier (SPB; seasonal predictions made during or before the boreal spring have much lower skill than those made at other times of year, which is referred to as the spring predictability barrier; Torrence and Webster 1998) persisted, as shown in Fig. 3a. When the targeted month was June with 4 months of lead time, the minimum correlation had a value of about 0.75. The SPB increased the uncertainty in the ENSO forecasts starting before and during the boreal spring, which mainly resulted from the seasonal cycle in initial condition error growth (Chen and Cane 2008; Jin et al. 2008) and the stochastic forcing (Zheng and Zhu 2010). Note that skills in the SPPM improved with 2–6 months of lead time to forecast the SSTA in June, and the accuracy increased no more than 5% with 4 months of lead time (Figs. 3c,d). Similar characteristics were also evident with the RMSE, whereas SPPM reduced errors when large RMSE happened in the targeted months from June to November with 2–6 months of lead time. Compared with the correction of temporal correlation, the correction using the SPPM method was more evident in the RMSE, with an improvement more than 10%. There are slight differences of the corrected targeted months for two periods (i.e., the concrete corrected targeted months or the lead time), but the correction effect of the SPPM method is the same for two periods.

4. Performance of probabilistic forecast

Due to the inherent uncertainty involved in climate forecasting, we also assessed and described the quality of the probabilistic forecasts. As described previously, the BSS is composed of reliability and resolution. In this part, we first examined these variables globally based on two systems: MME20 and CFSv2, and then we gave a specific analysis to the Niño-3.4 index. The spatial distributions of the BSS scores for three SSTA categories of the MME20 system are shown in Fig. 5. Relatively high accuracy was achieved in the tropical Pacific Ocean, with BSS for the MME20 system at about 0.5 for COLD

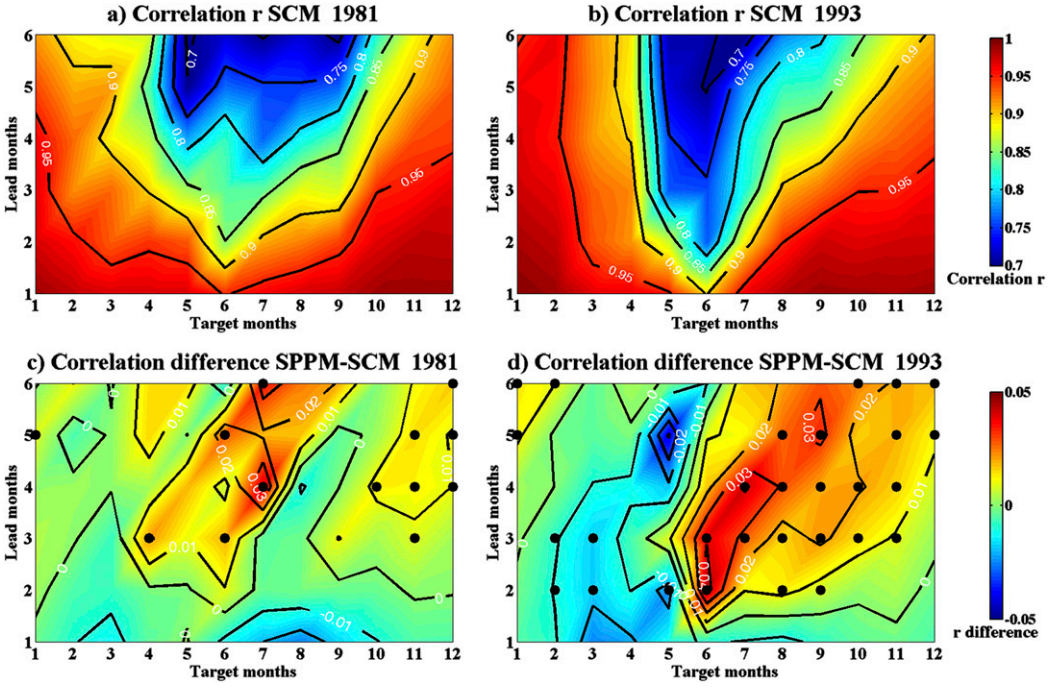

FIG. 3. Anomaly correlation r skill of the Niño-3.4 index for the ensemble's MME predictions and skill difference of SPPM predictions with respect to the SCM with 1–6 months of lead time for the 12 targeted months: (a),(c) 1981–2010 and (b),(d) 1993–2010; black dots in (c) and (d) represent the skill difference being statistically significant at the 5% level using the Student's t statistic.

and WARM categories even with a 6-month lead time. Noted that a relatively high BSS for the NEUTRAL category occurred with only a 1-month lead time, but the BSS was negative with a 3- or 6-month lead time for nearly globally other than the tropical middle Pacific Ocean. As mentioned in Yang et al. (2016), the reliability was very sensitive to the change in the ensemble prediction system itself (uncertainties sampled in initialization and in model errors) and would therefore be a more indicative criterion for testing the ensemble construction strategies. The high skill of the B_{rel} for COLD and WARM categories was shown, as the B_{rel} values were less than 0.2 almost for the whole globe. But for the NEUTRAL category, high B_{rel} values are shown in the latitudinal belts around 40°S and 30°N indicating a relative poor skill there (Fig. 6). Compared with the reliability, the resolution would be more easily impacted by the intrinsic predictability of the real world. The spatial distribution of B_{res} has confirmed the fact that most of the seasonal forecast skill was from the influence of ENSO variability, since B_{res} values were highest for forecasts for the COLD and WARM categories in the tropical mideastern Pacific Ocean, while for the NEUTRAL category, the significant values were only confined to the tropical Pacific Ocean (Fig. 7).

To evaluate the MME20 system with a commonly used forecasting system, we selected the CFSv2 forecasts as a reference. Compared with the CFSv2 forecasts, the MME20 system had higher BSS scores in the tropical Pacific Ocean and the latitudinal belts between 40° and 55°S for all three categorical events especially for a 3- or 6-month lead time (Fig. 8). Seeing the differences of B_{res} and B_{rel} between the MME20 system and CFSv2 forecasts in Figs. 9 and 10, contributions to the different BSS scores between these two systems mainly resulted from B_{rel} differences, indicating the ensemble construction strategies of the MME20 system were obviously superior to the CFSv2.

We compared the prediction skills of the probabilistic MME forecasts in an area-aggregated way. For this purpose, we aggregated skills for the Niño-3.4 index. Figure 11 shows BSS, B_{rel}, and B_{res} of the Niño-3.4 index as a function of lead time for the MME20 system, CFSv2, and four individual systems. BSS was above 0.6 for the COLD and WARM categories of the MME20 system with a 1-month lead time, which was comparable to the results of the North American MME (Becker and van den Dool 2016). Seen from the BSS for the NEUTRAL category, the effective forecast months of the CFSv2 is only within 4 months, but there are still

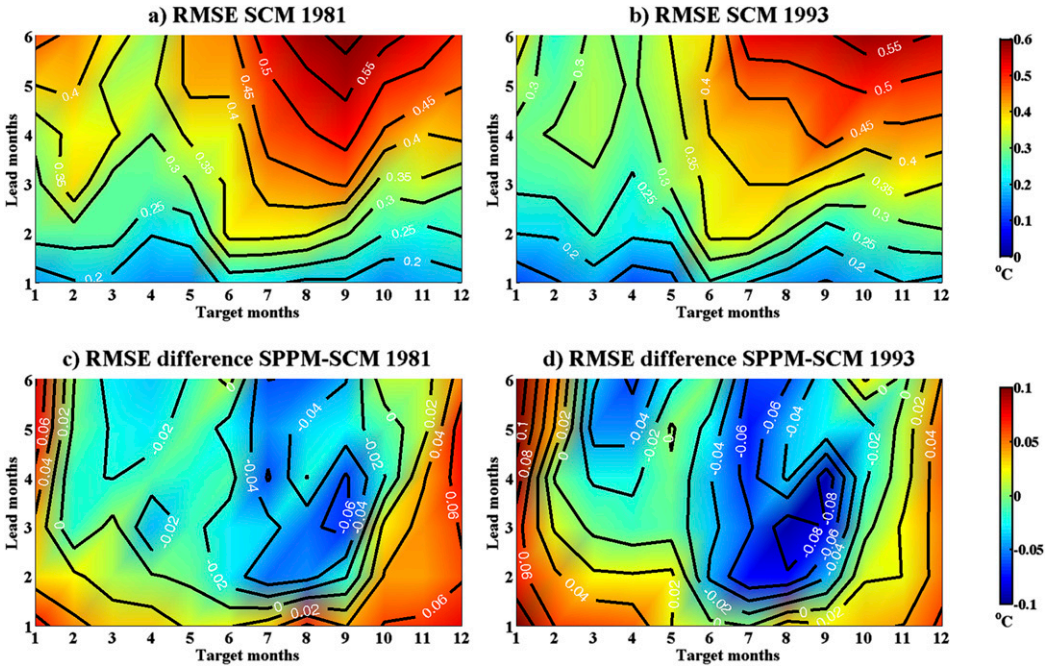

FIG. 4. RMSE of the Niño-3.4 index for the ensemble's MME predictions and skill difference of SPPM predictions with respect to the SCM with 1–6 months of lead time for the 12 targeted months: (a),(c) 1981–2010 and (b),(d) 1993–2010.

some skills at 6-month lead time for the MME20 system. In terms of B_{rel}, B_{res}, and BSS, the CFSv2 forecasts were all lower than that for the MME20 system. It was also evident that the BSS differences between the MME20 system and the CFSv2 forecasts resulted mainly from the differences in B_{rel}. It was clearly demonstrated that the MME20 generally performed better than an individual model for B_{res} and BSS for three categories, but for the

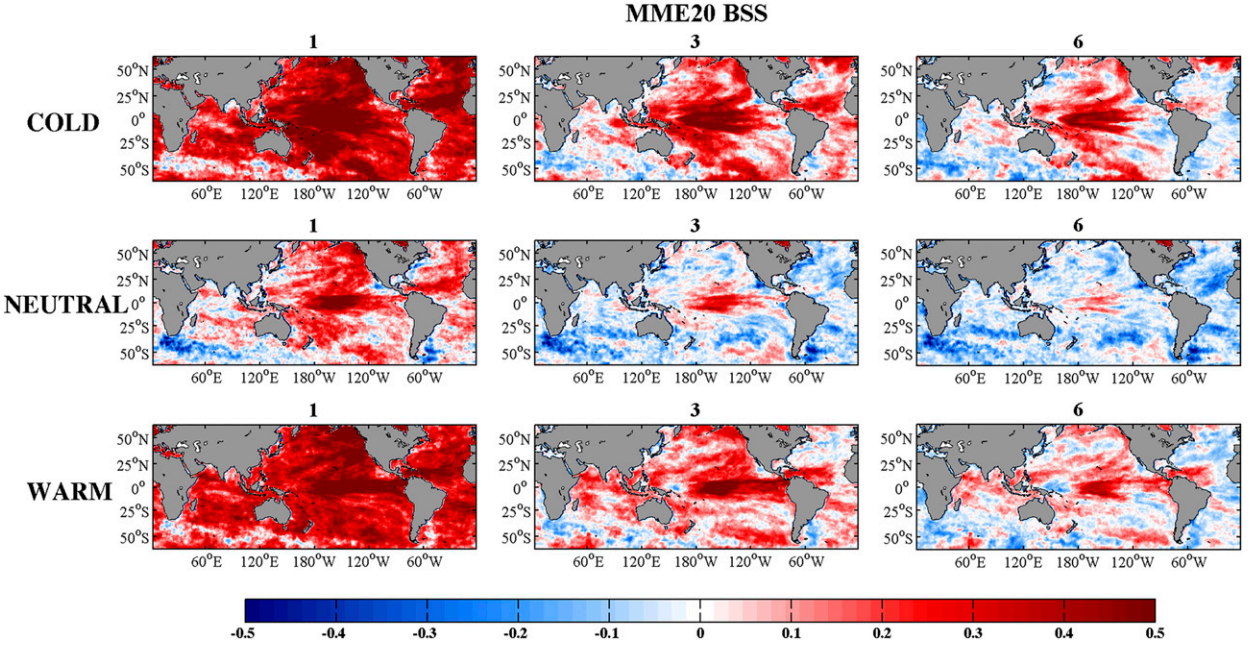

FIG. 5. Spatial distribution of BSS for COLD, NEUTRAL, and WARM SSTA categories with a 1-, 3-, or 6-month lead time of the MME20 system.

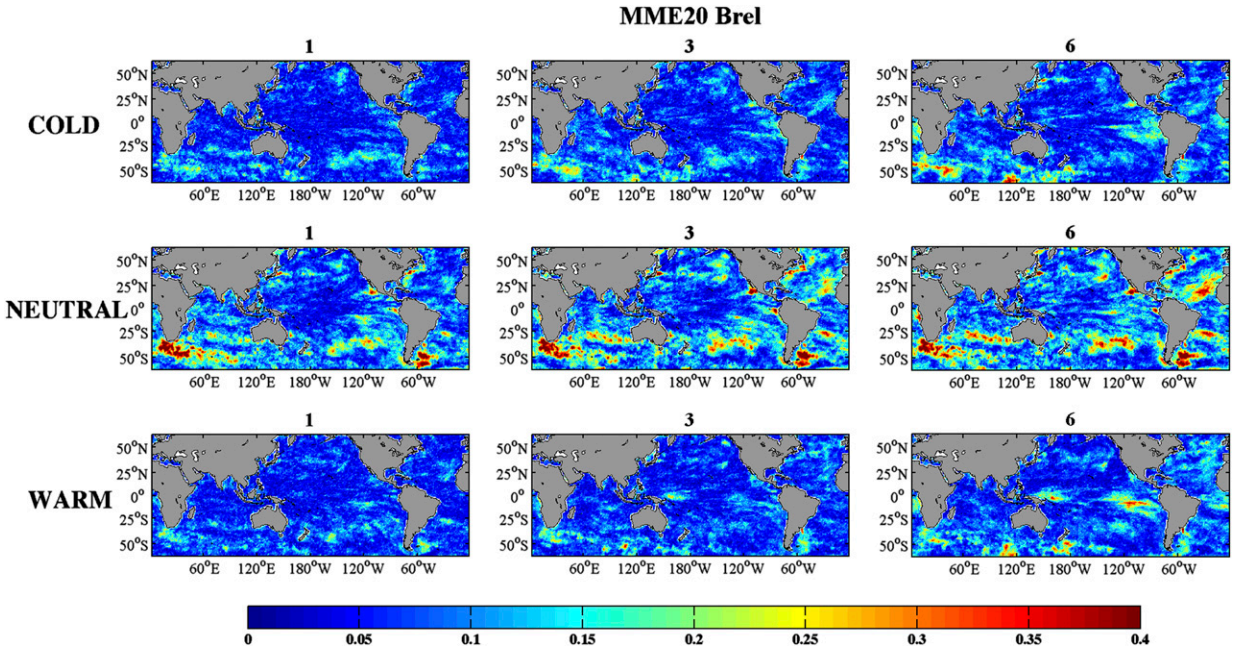

FIG. 6. As in Fig. 5, but for the reliability term B_{rel} of the BSS.

B_{rel}, a single model may perform better than the MME20 results.

5. Conclusions and discussion

In this study, based on the hindcast datasets of the ICMv2, FIO-ESM, NMEFC-CESM, and Fgoals-f ensembles, we investigated the performance of several MME methods to predict the SSTA variability. The metrics of prediction accuracy included two deterministic approaches (SCM and SPPM) and the probabilistic measure of the BSS along with its two components (reliability and resolution). We conducted the study of a deterministic forecast on the basis of two periods,

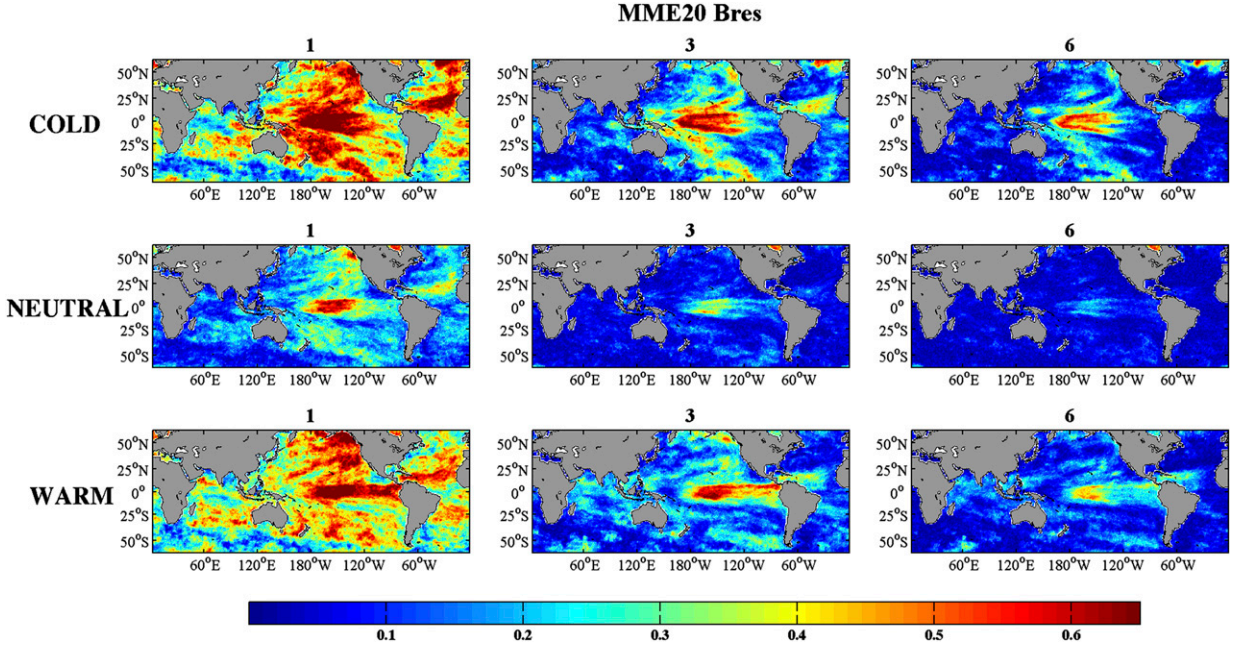

FIG. 7. As in Fig. 5, but for the resolution term B_{res} of the BSS.

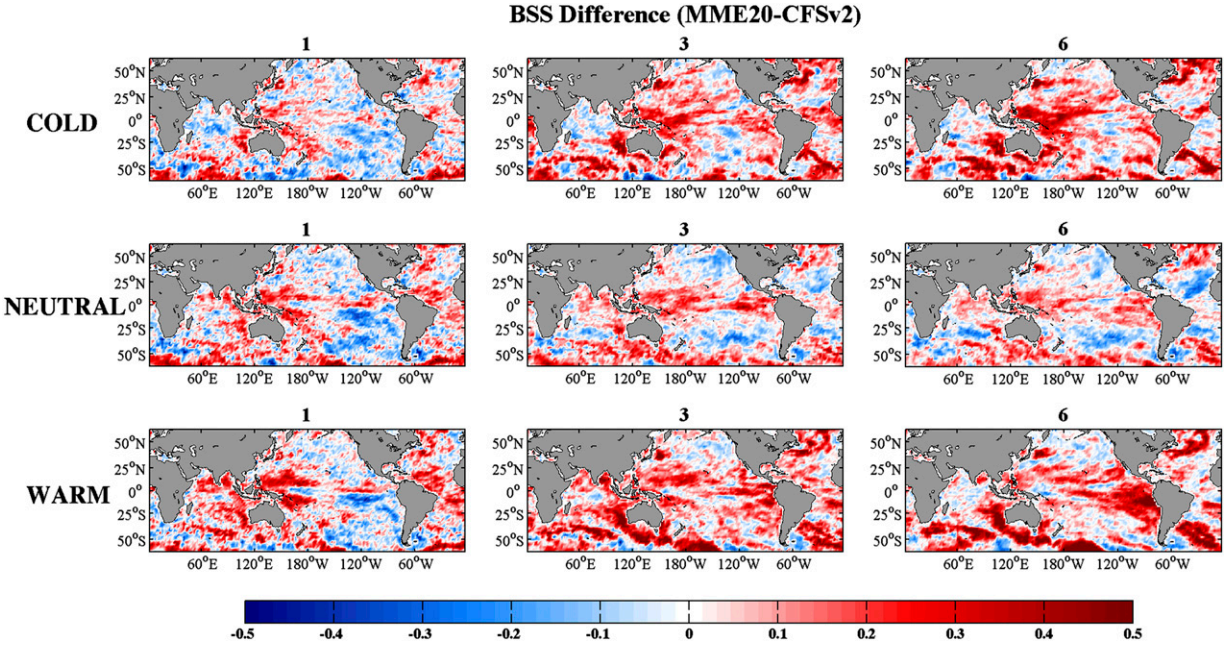

FIG. 8. As in Fig. 5, but for the differences in BSS between MME20 and CFSv2.

1981–2010 for three individual models and 1993–2010 for all four individual models, and we used a 3-year-out cross validation method in the SPPM approach. For the probabilistic forecast, we constructed a 20-member MME20 system, which included 5 members from each model. To study the relative advantages of model diversity, we also used a 24-member CFSv2 forecast as a reference.

For the evaluation of the deterministic forecasts, SCM showed a high predictability in the tropical Pacific Ocean, its temporal correlation skill exceeded 0.8 with a 6-month lead time. In general, the SCM outperformed the SPPM in the tropics where its skills was high, while the SPPM tended to show some positive effect on the correction for the regions where SCM had low skills especially

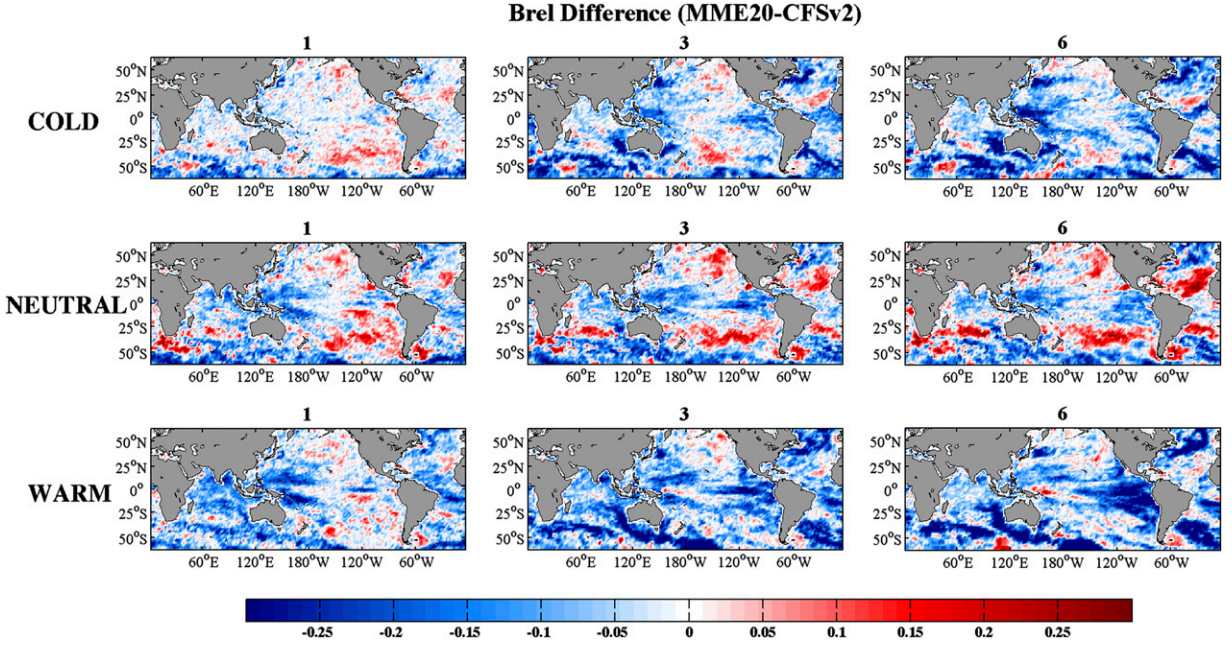

FIG. 9. As in Fig. 5, but for the differences in B_{rel} between MME20 and CFSv2.

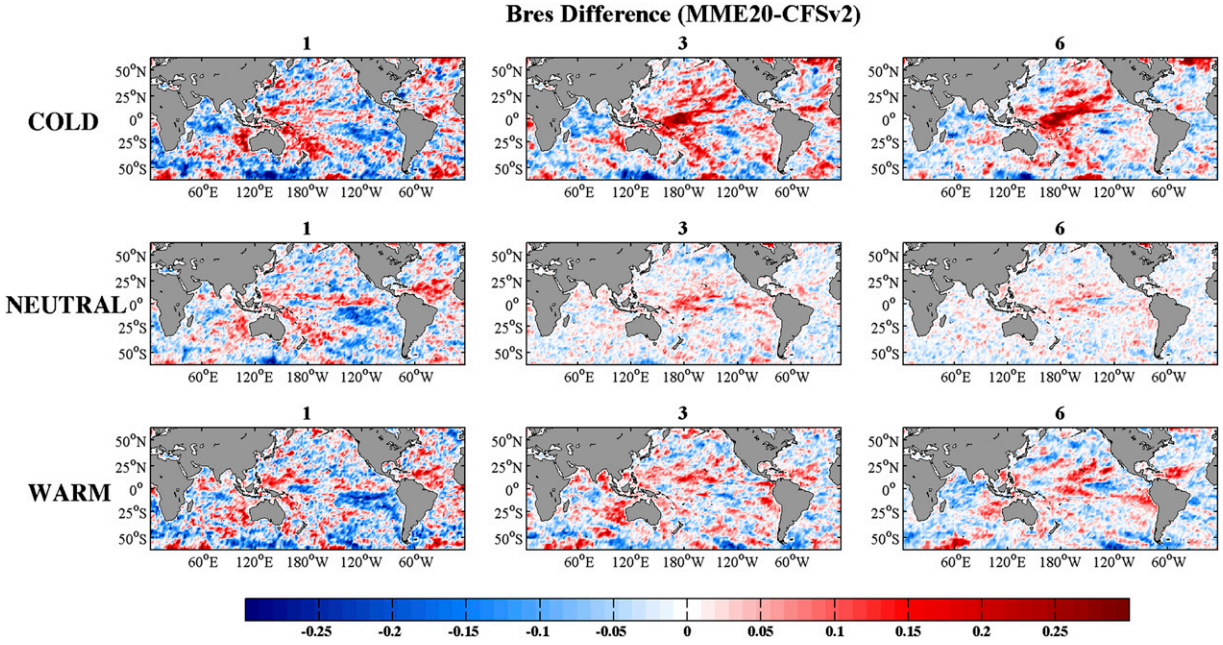

FIG. 10. As in Fig. 5, but for the differences in B_{res} between MME20 and CFSv2.

when at long lead times. The SPPM method contributed to improvements when the correlations were low or the RMSE was large in the Niño-3.4 index when using the SCM method especially for a 2–6-month lead time. Note that the spring predictability barrier phenomenon was reduced by using the SPPM method, especially for the RMSE, which was reduced by more than 10%. These qualitative results were not susceptible to the selection of the hindcast periods.

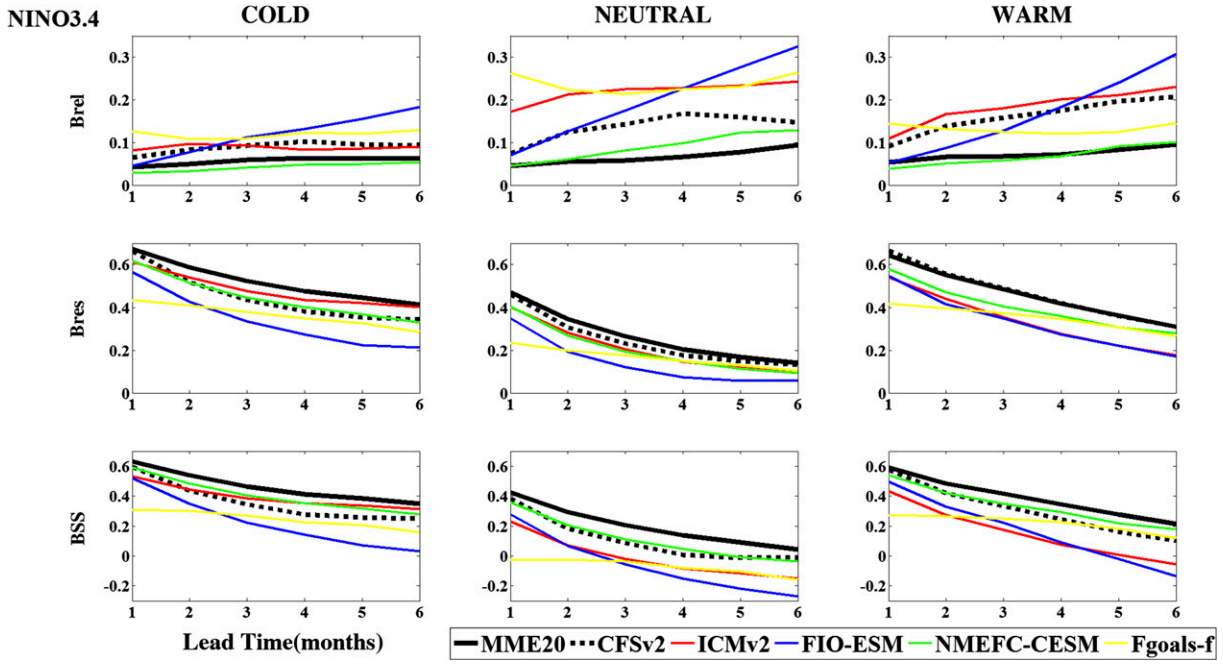

FIG. 11. The reliability term B_{rel}, resolution term B_{res}, and BSS of the Niño-3.4 index as a function of lead time for two ensemble systems (MME20 and CFSv2, thick solid lines and dashed lines) and four individual models (thin solid lines) for COLD, NEUTRAL, and WARM SSTA categories.

The MME20 system had high probabilistic forecasting accuracy compared with the commonly used CFSv2 forecasting system. Its performance was better than the CFSv2 forecasts in most regions in forecasting COLD, NEUTRAL, and WARM categories. The probabilistic skill of the MME was dominated by the characteristic of reliability, whereas the resolution among different systems differed less significantly, showing the high performance of the models.

To summarize, this work was a first attempt to evaluate the performance of MME for all of the individual forecasting systems developed or maintained by Chinese institutes. Currently, this quasi-operational MME system has four individual forecasting systems. We would finally select five to seven of these forecasting systems, which are developed or maintained by China Institutes, to constitute an operational MME system. For now, the study of probabilistic seasonal prediction and verification is still at an early stage, compared with the probabilistic weather forecast. Further studies are required to study the configuration of adequate ensemble construction strategies in developing multimodel prediction systems. Moreover, to further improve seasonal prediction, some empirical/statistical correction methods of model forecasts are still needed. Taking the ENSO forecast as an example, although noticeable progress has been made in the ENSO prediction by using coupled climate models, there are still systematic errors in the tropical Pacific mean states. These models' biases could affect the reliability for ENSO events, to improve the predictive skill, the source and reason of the mean state biases should be understood, and correction methods should be investigated to reduce these biases.

Acknowledgments. This research was supported by the Public science and technology research funds projects of ocean (201505013) and the National Key R&D Program of China (Grant 2017YFA0604203). The authors would also like to thank the three anonymous reviewers and Aifeng Tao, who is affiliated with the Key Laboratory of Coastal Disasters and Defence of Ministry of Education, for their useful comments.

REFERENCES

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. Wea. Rev.*, **139**, 581–607, https://doi.org/10.1175/2010MWR3417.1.

Barnston, A. G., and M. K. Tippett, 2013: Predictions of Nino3.4 SST in CFSv1 and CFSv2: A diagnostic comparison. *Climate Dyn.*, **41**, 1615–1633, https://doi.org/10.1007/s00382-013-1845-2.

Becker, E., and H. van den Dool, 2016: Probabilistic seasonal forecasts in the North American Multimodel Ensemble: A baseline skill assessment. *J. Climate*, **29**, 3015–3026, https://doi.org/10.1175/JCLI-D-14-00862.1.

Behringer, D. W., and Y. Xue, 2004: Evaluation of the global ocean data assimilation system at NCEP: The Pacific Ocean. *Eighth Symp. on Integrated Observing and Assimilation Systems for Atmosphere, Oceans, and Land Surface*, Seattle, WA, Amer. Meteor. Soc., 2.3, https://ams.confex.com/ams/84Annual/techprogram/paper_70720.htm.

Buizza, R., M. Miller, and T. N. Palmer, 1999: Stochastic simulation of model uncertainties in the ECMWF Ensemble Prediction System. *Manage. Decis.*, **125** (560), 494–511.

Chen, C., M. A. Cane, N. Henderson, D. E. Lee, D. Chapman, D. Kondrashov, and M. D. Chekroun, 2016: Diversity, nonlinearity, seasonality, and memory effect in ENSO simulation and prediction using empirical model reduction. *J. Climate*, **29**, 1809–1830, https://doi.org/10.1175/JCLI-D-15-0372.1.

Chen, D., and M. A. Cane, 2008: El Niño prediction and predictability. *J. Comput. Phys.*, **227**, 3625–3640, https://doi.org/10.1016/j.jcp.2007.05.014.

Chowdary, J. S., and Coauthors, 2010: Predictability of summer northwest Pacific climate in 11 coupled model hindcasts: Local and remote forcing. *J. Geophys. Res.*, **115**, D22121, https://doi.org/10.1029/2010JD014595.

DelSole, T., and J. Shukla, 2009: Artificial skill due to predictor screening. *J. Climate*, **22**, 331–345, https://doi.org/10.1175/2008JCLI2414.1.

Fritsch, J. M., J. Hilliker, J. Ross, and R. L. Vislocky, 2000: Model consensus. *Wea. Forecasting*, **15**, 571–582, https://doi.org/10.1175/1520-0434(2000)015<0571:MC>2.0.CO;2.

Gneiting, T., and A. E. Raftery, 2005: Weather forecasting with ensemble methods. *Science*, **310**, 248–249, https://doi.org/10.1126/science.1115255.

Guan, Z., and T. Yamagata, 2003: The unusual summer of 1994 in East Asia: IOD teleconnections. *Geophys. Res. Lett.*, **30**, 1544, https://doi.org/10.1029/2002GL016831.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting–I. Basic concept. *Tellus*, **57A**, 219–233, https://doi.org/10.1111/j.1600-0870.2005.00103.x.

Harada, Y., and Coauthors, 2016: The JRA-55 Reanalysis: Representation of atmospheric circulation and climate variability. *J. Meteor. Sci.*, **94**, 269–302, https://doi.org/10.2151/jmsj.2016-015.

Jin, E. K., and Coauthors, 2008: Current status of ENSO prediction skill in coupled ocean–atmosphere models. *Climate Dyn.*, **31**, 647–664, https://doi.org/10.1007/s00382-008-0397-3.

Kalnay, E., 2003: *Atmospheric Modeling, Data Assimilation and Predictability*. Cambridge University Press, 341 pp.

Kennedy, J. J., N. A. Rayner, R. O. Smith, D. E. Parker, and M. Saunby, 2011: Reassessing biases and other uncertainties in sea surface temperature observations measured in situ since 1850: 1. Measurement and sampling uncertainties. *J. Geophys. Res.*, **116**, D14103, https://doi.org/10.1029/2010JD015218.

Kirtman, B. P., D. H. Min, and J. M. Infanti, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, https://doi.org/10.1175/BAMS-D-12-00050.1.

Kobayashi, S., and Coauthors, 2015: The JRA-55 Reanalysis: General specifications and basic characteristics. *J. Meteor. Sci.*, **93**, 5–48, https://doi.org/10.2151/jmsj.2015-001.

Krishnamurti, T. N., C. M. Kishtawal, D. W. Shin, and C. E. Williford, 2000: Multimodel superensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216, https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Kug, J. S., J. Y. Lee, and I. S. Kang, 2008a: Systematic error correction of dynamical seasonal prediction using a stepwise pattern projection method. *Mon. Wea. Rev.*, **136**, 3501–3512, https://doi.org/10.1175/2008MWR2272.1.

Li, Y., and Coauthors, 2015: An ENSO hindcast experiment using CESM (in Chinese, abstract in English). *HaiyangXuebao*, **37** (9), 39–50.

Michaelsen, J., 1987: Cross-validation in statistical climate forecast model. *J. Climate Appl. Meteor.*, **26**, 1589–1600, https://doi.org/10.1175/1520-0450(1987)026<1589:CVISCF>2.0.CO;2.

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, https://doi.org/10.1002/2014JD022230.

——, ——, ——, and H. J. Lee, 2017: Skill of real-time operational forecasts with the APCC multi-model ensemble prediction system during the period 2008-2015. *Climate Dyn.*, **49**, 4141–4156, https://doi.org/10.1007/s00382-017-3576-2.

Palmer, T. N., 2000: Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.*, **63**, 71–116, https://doi.org/10.1088/0034-4885/63/2/201.

——, and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.*, **85**, 853–872, https://doi.org/10.1175/BAMS-85-6-853.

Pellerin, G., L. Lefaivre1, P. Houtekamer, and C. Girard, 2003: Increasing the horizontal resolution of ensemble forecasts at CMC. *Nonlinear Processes Geophys.*, **10**, 463–468, https://doi.org/10.5194/npg-10-463-2003.

Peng, P., A. Kumar, A. G. Barnston, and L. Goddard, 2000: Simulation skills of the SST-forced global climate variability of the NCEP-MRF9 and Scripps–MPI ECHAM3 models. *J. Climate*, **13**, 3657–3679, https://doi.org/10.1175/1520-0442(2000)013<3657:SSOTSF>2.0.CO;2.

——, ——, H. van den Dool, and A. G. Barnston, 2002: An analysis of multi-model ensemble predictions for seasonal climate anomalies. *J. Geophys. Res.*, **107**, 4710, https://doi.org/10.1029/2002JD002712.

Qiao, F. L., Z. Y. Song, Y. Bao, Y. Song, Q. Shu, C. Huang, and W. Zhao, 2013: Development and evaluation of an Earth System Model with surface gravity waves. *J. Geophys. Res. Oceans*, **118**, 4514–4524, https://doi.org/10.1002/jgrc.20327.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate*, **15**, 1609–1625, https://doi.org/10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

Richardson, D. S., 2006: Predictability and economic value. *Predictability of Weather and Climate*, T. Palmer and R. Hagedorn, Eds., Cambridge University Press, 628–644, https://doi.org/10.1017/CBO9780511617652.026.

Rodrigues, L. R. L., F. J. Doblas-Reyes, and C. A. S. Coelho, 2014: Multi-model calibration and combination of tropical seasonal sea surface temperature forecasts. *Climate Dyn.*, **42**, 597–616, https://doi.org/10.1007/s00382-013-1779-8.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1058, https://doi.org/10.1175/2010BAMS3001.1.

——, and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, https://doi.org/10.1175/JCLI-D-12-00823.1.

Slingo, J., and T. N. Palmer, 2011: Uncertainty in weather and climate prediction. *Philos. Trans. Roy. Soc. London*, **369A**, 4751–4767, https://doi.org/10.1098/rsta.2011.0161.

Stockdale, T. N., and Coauthors, 2011: ECMWF seasonal forecast system 3 and its prediction of sea surface temperature. *Climate Dyn.*, **37**, 455–471, https://doi.org/10.1007/s00382-010-0947-3.

Torrence, C., and P. J. Webster, 1998: The annual cycle of persistence in the El Niño/Southern Oscillation. *Quart. J. Roy. Meteor. Soc.*, **124**, 1985–2004, https://doi.org/10.1002/qj.49712455010.

Wang, B., and Coauthors, 2009: Advance and prospectus of seasonal prediction: Assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). *Climate Dyn.*, **33**, 93–117, https://doi.org/10.1007/s00382-008-0460-0.

Wang, L., and Coauthors, 2017: Statistical correction of ENSO prediction in BCC_CSM1.1m based on stepwise pattern projection method (in Chinese with English abstract). *Meteor. Mon.*, **43** (3), 294–304.

Weng, H. Y., G. X. Wu, Y. M. Liu, S. K. Behera, and T. Yamagata, 2011: Anomalous summer climate in China influenced by the tropical Indo-Pacific Oceans. *Climate Dyn.*, **36**, 769–782, https://doi.org/10.1007/s00382-009-0658-9.

Yang, D. J., and Coauthors, 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, https://doi.org/10.1002/2015JD023781.

Yun, W. T., L. Stefanova, and T. N. Krishnamurti, 2003: Improvement of the multimodel superensemble technique for seasonal forecasts. *J. Climate*, **16**, 3834–3840, https://doi.org/10.1175/1520-0442(2003)016<3834:IOTMST>2.0.CO;2.

Zhang, S. W., and Coauthors, 2018: Evaluation of the hindcasting main SSTA modes of the global key regions based on the CESM forecasting system. *HaiyangXuebao*, **40** (9), 18–30.

Zheng, F., and J. Zhu, 2010: Coupled assimilation for an intermediated coupled ENSO prediction model. *Ocean Dyn.*, **60**, 1061–1073, https://doi.org/10.1007/s10236-010-0307-1.